Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

An intrinsically interpretable neural network architecture for sequence to function learning

Ali Tuğrul Balcı [1,2], Mark Maher Ebeid [1,2], Panayiotis V Benos[4], Dennis Kostka[1,2]*, Maria Chikina[1,2]*

[1]Joint Carnegie Mellon University-University of Pittsburgh Program in Computational Biology, Institution, Pittsburgh, 15213, United States and
[2]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, 15213, Unites States and
[3]Department of Developmental Biology, University of Pittsburgh, Pittsburgh, 15213, Unites States and
[4]Department of Epidemiology, University of Florida, Gainesville, 32610, Unites States.

*kostka@pitt.edu (D.K.) and mchikina@pitt.edu (M.C.).

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Sequence-based deep learning approaches have been shown to predict a multitude of functional genomic readouts, including regions of open chromatin and RNA expression of genes. However, a major limitation of current methods is that model interpretation relies on computationally demanding post-hoc analyses, and even then, we often cannot explain the internal mechanics of highly parameterized models. Here, we introduce a deep learning architecture called tiSFM (totally interpretable sequence to function model). tiSFM improves upon the performance of standard multi-layer convolutional models while using fewer parameters. Additionally, while tiSFM is itself technically a multi-layer neural network, internal model parameters are intrinsically interpretable in terms of relevant sequence motifs.
Results: tiSFM's model architecture makes use of convolutions with a fixed set of kernel weights representing known transcription factor (TF) binding site motifs. We analyze published open chromatin measurements across hematopoietic lineage cell-types and demonstrate that tiSFM outperforms a state-of-the-art convolutional neural network model custom-tailored to this dataset. We also show that it correctly identifies context specific activities of transcription factors with known roles in hematopoietic differentiation, including Pax5 and Ebf1 for B-cells, and Rorc for innate lymphoid cells. tiSFM's model parameters have biologically meaningful interpretations, and we show the utility of our approach on a complex task of predicting the change in epigenetic state as a function of developmental transition.
Availability and implementation : The source code, including scripts for the analysis of key findings, can be found at https://github.com/boooooogey/ATAConv, implemented in Python.
Contact: atb44@pitt.edu

1 Introduction

Functional genomics assays have accelerated our understanding of how non-coding regions of genomes contribute to cellular and organismal function by enabling their identification and characterization. Large consortium projects such as ENCODE Project Consortium, 2012; Gal-Oz *et al.*, 2019 have generated extensive datasets profiling diverse functional properties of the non-coding genome across many tissues and cell-types, including transcription factor (TF) binding, regions of open chromatin, and biochemical modification of N-terminal histone tails. While such genomic readouts are believed to be largely determined by DNA sequence, the precise sequence-to-function (S2F) relation is complex and remains poorly understood. Nevertheless, recent developments have shown that by using deep learning models with millions of parameters it is indeed possible to learn S2F mappings, predict epigenetic readouts, or even characterize gene expression (Maslova *et al.*, 2020; Avsec *et al.*, 2021b; Kelley *et al.*, 2016; Zhou and Troyanskaya, 2015; Quang and Xie, 2016). Given these

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

successes, it is worthwhile to ask what scientific insights that go beyond advancing our understanding of applying machine learning engineering principles to genomics data such models can provide.

Primarily, a model can be considered useful (in the canonical supervised machine learning sense) if it is able to make reliable out-of-sample predictions. This is a major promise of complex deep learning models, given they can potentially predict functional output of arbitrary sequences (e.g., individual genomes, haplotypes with disease-associated genetic variants, or synthetic constructs). However, while existing models have made progress to this end and significantly outperform baseline approaches on these tasks, performance can still be variable; with existing models not yet being capable of substituting experiments. For example, the current SOTA in the field of accessible chromatin (as measured by ATAC-seq) prediction from sequence, `AI-TAC`, suffers from these issues. Specifically, although Maslova *et al.*, 2020 is able to train models that perform well (in terms of established performance metrics, and by production of certain biological insights), this is only after considerable effort is put forward to extract meaningful information from essentially a black box model featuring a large parameter space that obfuscates straightforward interpretations. Additionally, `AI-TAC`'s reported competitive performance is not consistent since the model fails to capture relevant biological information throughout all training iterations.

Moving past out-of-sample predictions, a successful model can also provide information about underlying biochemical principles or mechanisms. For example, if increasing the receptive field of a S2F model from 40KB to 200KB increases model performance (as was recently demonstrated for gene expression predictions (Avsec *et al.*, 2021a)), then we can conclude that there is indeed evidence for relevant biochemical interactions that occur in the scope of 200KB. However, questions about the actual biochemical entities involved, or how they may interact, or be organized into higher order structures, cannot be answered from observing model performance. Instead, model parameters need to be subjected to transparent post-hoc interpretation. For example, many deep learning S2F models' architectures include an initial convolution layer on a locus' DNA sequence, allowing for the possibility to use a priori information by interpreting this layer's learned kernel weights as position weight matrices (PWM) and matching them to databases of known transcription factor motifs (Alipanahi *et al.*, 2015; Maslova *et al.*, 2020).

Successful incorporation of such strategies into current work is highly dependent on the effective tuning of hyperparameters; for example, kernel weights may only represent partial TF motifs that are aggregated in subsequent layers of the network (Koo and Eddy, 2019). Moreover, even when successful, since neither sign nor the magnitude of a TF's contribution is known, such approaches can only reveal which TFs are involved, but not how they contribute. As an alternative approach one can perform attribution analysis, which refers to a class of algorithms that can attribute a prediction to a specific subset of the input. In many cases, this type approach will highlight regions of input DNA that can subsequently be matched to motifs of known TFs. While such attribution approaches overcome the problem with partial motifs, they do not provide a universally interpretable solution. Most attribution techniques rely on propagating "blame" for a prediction though layers of the network using gradients, and existing methods differ in the exact calculation performed.

Importantly, these approaches require considerable computation and expertise, and the results will vary depending on the method chosen. We note that the two types of approaches can also be combined, revealing both: TFs involved, and their specific contributions (sign and magnitude) to predicting functional genomic readouts, as was done in `AI-TAC` (Maslova *et al.*, 2020). Specifically, `AI-TAC` focused on a hematopoeitic development dataset, combining attribution and kernel analysis to create a TF by cell-type map. However, while effective in highlighting many known drivers of hematopoeitic differentiation, it required complex post-processing steps. Moreover, the final interpretation product, the TF by cell-type contribution, only partially reveals the underlying mechanism of prediction performance. For example, the `AI-TAC` method has 3 convolutions layers and 3 linear layers – an architecture that was extensively optimized for the dataset. However, the kernel-motif matching combined with attribution tells us nothing about how the internal parameters of the model contribute to the predictions. Do subsequent convolution layers encode interactions or simply fine-tune motifs?

In this work we introduce a "totally interpretable sequence-to-function model", `tiSFM`. While the model is technically a multi-layer CNN, each layer is interpretable as either DNA sequence or TF binding. Thus, the final linear layer directly maps TFs to outputs and can produce a meaningful TF by output matrix with no post processing. Moreover, the internal parameters of the model, such as learnable pooling or interaction attention, are also directly interpretable and can reveal additional information about mechanisms by which highly parameterized models produce high accuracy predictions.

2 Methods

2.1 Data Processing

We trained `tiSFM` on the ImmGen ATAC-seq dataset (Gal-Oz *et al.*, 2019) that consists of assayed open chromatin region (OCR) activity of 81 immune cell-types in the mouse hemtopoeitic lineage. This dataset was extensively evaluated in a similar deep learning prediction and interpretation context in a seminal paper by Maslova *et al.*, 2020. Similarly to this previous work, we focus on predicting correlations across cell-types rather than raw chromatin accessibility. However, rather than changing the form of the loss function as was done in the original study, we achieve the same effect by z-scoring the activity of each OCR across cell-type and using the standard mean squared error (MSE) loss on the z-scores. Additionally, we focus most of our benchmarking and evaluation on predicting data aggregated over the each of the eight main lineage types, rather than the individual cell-types. This is done for two reasons: **(1)** Difficulty — as was demonstrated in the original paper, the within-lineage performance is considerable worse than across-lineages, **(2)** Interpretability — in many cases, we have substantial knowledge about the TFs that play a role in regulating major lineages, but information about further developmental decisions within a lineage is comparably sparser.

We also train our `tiSFM` on the full dataset; however, rather than using raw data or the z-scores of OCR activities, we focus on OCR activity differences along the cell lineage tree of hemtopoeitic differentiation. Specifically, if a cell-type x has a parent $P(x)$ we predict the difference between $\texttt{OCR_activity}(x)$ and $\texttt{OCR_activity}(P(x))$. This transformation prevents large differences between lineages from dominating the training objective, instead shifting the focus on developmental transitions. However, this also has the effect of making the problem more difficult and the (R^2) values achieved for all methods are considerably lower than previously reported. We refer to this dataset as "tree-diff".

The input dataset is summit-centered, and for our model we restricted the sequence to 300 bp centered on the peak summit. We use 862 position weight matrices (PWMs) of mouse TFs from the cisBP database (Weirauch *et al.*, 2014a) as weights for our initial convolution kernels.

2.2 Model training

All models are implemented in PyTorch (Paszke *et al.*, 2019). Because of the differences in the objective and train/test splitting we also re-trained the original `AI-TAC` model architecture using our version of the data, objective, train/validation/test splits, and stopping criteria for comparison.

All 512,596 reported genomic regions were divided into 10 approximately similar-sized folds. There are no coordinate overlaps between different peaks and fold assignment was performed at random. In turn, we used 8 folds to form a training set, 1 fold for validation, and 1 fold for testing. We trained models starting with a learning rate of 0.01, decaying by a factor of 0.1 if the validation loss does not improve for 10 epochs; once the learning rate is below 5×10^{-7} the training stops. We repeat this procedure 10 times, with each fold singled out as the test set once; the performance was averaged across test set folds. `tiSFM` is trained first by freezing the first layer kernel weights to be cisBP PWMs, the kernels are then unfrozen and trained further (for approximately 10 epochs, while monitoring validation improvement as defined above). A path algorithm for L_1 or the minimax concave penalty (MCP) regularization is applied only to the final layer of the model by a modified version of the ADAM optimizer that uses the proximal operators for L_1 and MCP penalties as described in Yun *et al.*, 2021.

2.3 Model Architecture

All variations of `tiSFM` take one-hot-encoded DNA sequence of length L as input, which is then is augmented by the integration of positional embedding and attention. Mathematically, this step can be represented as follows:

$$\begin{aligned} y &= x + p \\ a &= \mathrm{MultiHeadAttention}\left(y\right) \\ x_{\text{aug}} &= x + a \end{aligned}$$

where x is the one-hot embedded input ($\in \mathbb{R}^{4\times L}$) and $p \in \mathbb{R}^L$ is a vector of positional embedding parameters. The first sum is carried out by broadcasting p over the rows of x, while the dimensions of a and x match for the second sum. These transformations enable the model to account for positional effects; for instance, TFs are known to preferentially bind in the center of ATAC-seq peaks, and to incorporate nucleotide context information. In the subsequent layer, we compute a convolution of the augmented input with a n fixed PWM matrices for motif binding sites taken from the CIS-BP database (Weirauch *et al.*, 2014b). Although the PWMs weights are fixed, every kernel in the convolutional layer (PWM) has two trainable parameters: a scale and a bias, with a sigmoid function as activation. This layer maps raw PWM match scores to $[0, 1]$ in a PWM specific manner. Next, each channel is globally pooled (across the sequence) using a programmable pooling layer to extract a single value for each PWM/kernel. We use programmable global pooling which works similarly to the attention pooling popularized by the Enformer model (Avsec *et al.*, 2021b). Every channel is assigned a learnable parameter that scales the channel before applying softmax. If the scale is close to zero, the pooling layer functions essentially as average pooling; on the other hand, for high parameter values the operation is essentially maximum pooling. The programmable pooling layer computes the following for each channel:

$$o = \frac{\sum_i^L \exp\left(ax_i\right) x_i}{\sum_j^L \exp\left(ax_j\right)}$$

where x is a vector of length L, corresponding to a specific channel in the output of convolutional layer, a is the trainable weight that is specific to this channel, and o is the pooled value, which is a channel-specific scalar. The sigmoid activation combined with the ability to learn a trade-off between average and maximum pooling enables this architecture to approximate motif match counting, which has been shown to be useful in modeling TF-DNA binding (Roider *et al.*, 2006).

After global pooling, every input sequence is reduced to a vector $y \in \mathbb{R}^n$ of n TF/PWM activities. These values go into a fully connected layer with sigmoid activation to generate a "mask" for global protein activity that takes protein-protein interactions into account. Mathematically this step can be formulated as follows:

$$\begin{aligned} m &= \sigma\left(Ay\right) \\ z &= m \odot y \end{aligned} \tag{1}$$

where A is a $\mathbb{R}^{n\times n}$ matrix of parameters, y is the input vector for this step, σ is the sigmoid function, and $\odot$ represents element-wise multiplication. The output, z, goes into a final fully-connected layer to predict functional output for each input region.

2.4 Proximal Operators for MCP/L_1 regularization

To make our model sparse and therefore more interpretable, we impose sparsity regularization such as L_1 and MCP on the coefficients of the final layer of our model. Yun *et al.*, 2021 formulated the proximal operator for the L_1 regularized ADAM as

$$\begin{aligned} \widehat{\theta}_{t,i} &= \theta_{t,i} - \alpha_t \frac{m_{t,i}}{C_{t,i} + \delta} \\ \theta_{t+1,i} &= \mathrm{sign}\left(\widehat{\theta}_{t,i}\right)\left(\left|\widehat{\theta}_{t,i}\right| - \frac{\alpha_t \lambda}{C_{t,i} + \delta}\right) \end{aligned}$$

where $C_{t,i}$ is the preconditioner matrix, $m_{t,i}$ is the momentum estimate, α_t is the learning rate, δ is a small constant added to avoid division by 0, θ is the coefficient of the model, and finally λ is the hyperparameter for L_1. In a similar vein, they formulated the closed form solution of projecting through MCP regularization as

$$\begin{aligned} \widehat{\theta}_t &= \theta_t - \alpha_t \left(C_t + \delta I\right)^{-1} m_t \\ \theta_{t+1,i} &= \mathrm{sign}\left(\widehat{\theta}_{t,i}\right) \min\left\{\frac{b \max\left\{\left|\widehat{\theta}_{t,i}\right| - \frac{\alpha_t \lambda}{C_{t,i}+\delta}, 0\right\}}{b-1}, \left|\widehat{\theta}_{t,i}\right|\right\} \end{aligned}$$

where the parameters are the same with L_1 case, with the addition of b which is the second hyperparameter for MCP.

3 Results

3.1 The tiSFM model outperforms a state of the art CNN at OCR activity prediction

An overview of the `tiSFM` architecture can be seen in **Fig. 1A**. Briefly, the `tiSFM` is multi-layer CNN that consists of several elements that have been successfully used across a variety of deep learning models, tasks, and datasets, such as: self-attention mechanisms, positional embedding, convolutional layers, and programmable pooling. In order to quantify the performance contribution of different parts of the model we implement several variants of this architecture that selectively leave out some of the layers from the full model for comparison (ablation analysis). We also trained all models with fixed kernels (cisBP PWM), and then subsequently fine-tuned kernels.

Fig. 1B and **Fig. 1C** show the performance of different model variants quantified on the test data of all 10 cross validation folds and reported as mean and standard error. We choose to compare the change in R^2, defined as the difference between `tiSFM`'s and `AI-TAC`'s.

As a baseline we consider the `AI-TAC` CNN architecture proposed by Maslova *et al.*, 2020, retrained using the train/validation/test splits and training procedure (see Methods for details). As this architecture was extensively optimized on this dataset we consider it to be SOTA for this learning task. Overall, we find that the full `tiSFM`'s model architecture

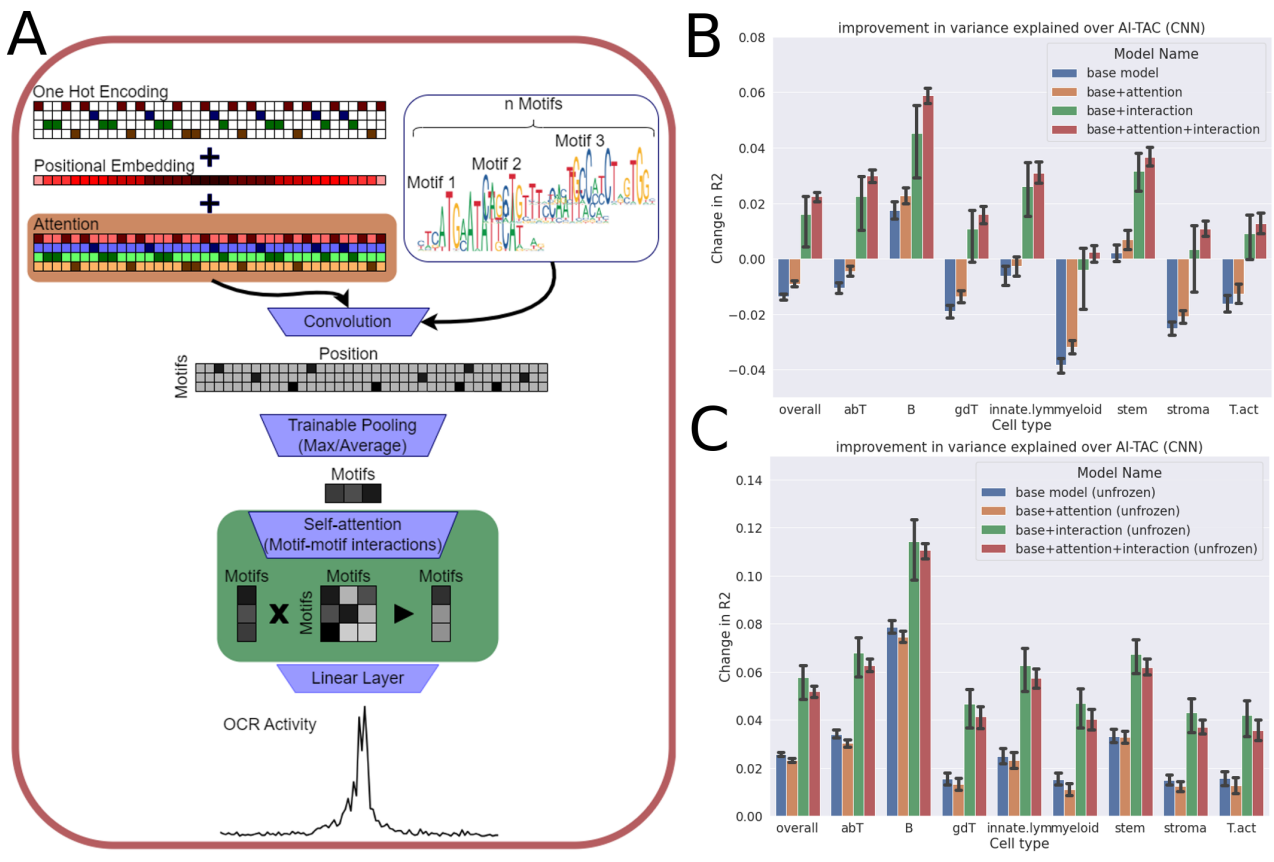

Fig. 1. tiSFM improves on prediction accuracy when compared to the current SOTA. (A) A graphical display of the architecture of tiSFM. (B) The improvement of tiSFM, measured via the change in R^2 on the full model; additionally, different iterations of tiSFM were tested with the inclusion/exclusion of certain model components in order to weigh their contribution to the overall performance. (C) similar to (B), but after the kernels of the convolutional layer were fine-tuned during training.

(red bars), using either frozen or un-frozen kernels outperforms the `AI-TAC` CNN model across all cell-types (**Fig. 1B and C**).

This observation is particularity striking as `tiSFM` has 966,182 parameters (including the kernels) while `AI-ATAC` has 2,974,908 — a nearly 3-fold increase. This result highlights the importance of adding biological prior knowledge in the form of PWMs and suggests that learning the appropriate kernel weights (or PWMs) is the main source of prediction accuracy. This interpretation is consistent with a recent results demonstrating that combining a max-pooled convolution layer with a fully connected linear layer is competitive with multi-layer CNN networks on a variety of sequence-to-function prediction models (Novakovsky *et al.*, 2022).

Studying the unfreezing of PWM/kernel weights further supports this line of thought (Figure 1C), as we find that even the base model that lacks interaction and attention layers (but still has PWM specific learnable logistic and pooling parameters) can outperform `AI-TAC` after kernel fine-tuning.

While our results strongly indicate that the first layer convolution kernels, which approximate the binding preferences of transcription factors, are the most critical part of the model, other aspects of the model architecture yield additional performance gains. Investigating the impact of attention and/or interaction we find that both contribute an appreciable amount to the performance, with the biggest contribution coming from inclusion of the modeling TF interactions (**Fig. 1C**). This trend is the same regardless of whether kernels are frozen or not. Attention, on the other hand, improves the performance when the kernels are kept fixed and interaction is included; however, the performance is slightly reduced when the kernels are fine-tuned.

One concern with the fine-tuning of the kernels is that they will no longer resemble the initial PWM which would hamper interpretability. However, on this data, this problem does not arise. PWMs before and after weight unfreezing are very similar, with an average MSE of 0.04.

Additionally, we examined the pairwise distance between the initial kernels and the final fine-tuned kernels. For 92% of the fine-tuned kernels, the most similar motifs were their pre-fine-tuning counterparts; if a change was detected between the two, the fine-tuned motif is often a different motif for the same TF, as is the case with Tcfcp2l1, or a motif in the same family — for example, Dlx1 in the fine-tuned kernel is most similar to Dlx3 in the initial kernel. We also note that our observation that a relatively minor parameter change (i.e., the difference between fine-tuned and initial kernel weights) can dramatically increase performance further highlighting the fact that getting the first layer kernels are the most important parameters of S2F CNNs.

3.2 tiSFM parameters are intrinsically interpretable in terms of lineage specific transcription factors

In the original `AI-TAC` manuscript, the authors combined kernel-PWM matching (also using the mouse CIS-BP database) with attribution analysis using DeepLift (Shrikumar *et al.*, 2017) to create a cell-type by TF contribution map. In our approach both of these steps are unnecessary, as the final linear layer already contains this information. We visualize the `tiSFM` TF-by-cell-type contribution from a single run of the model in **Fig.**

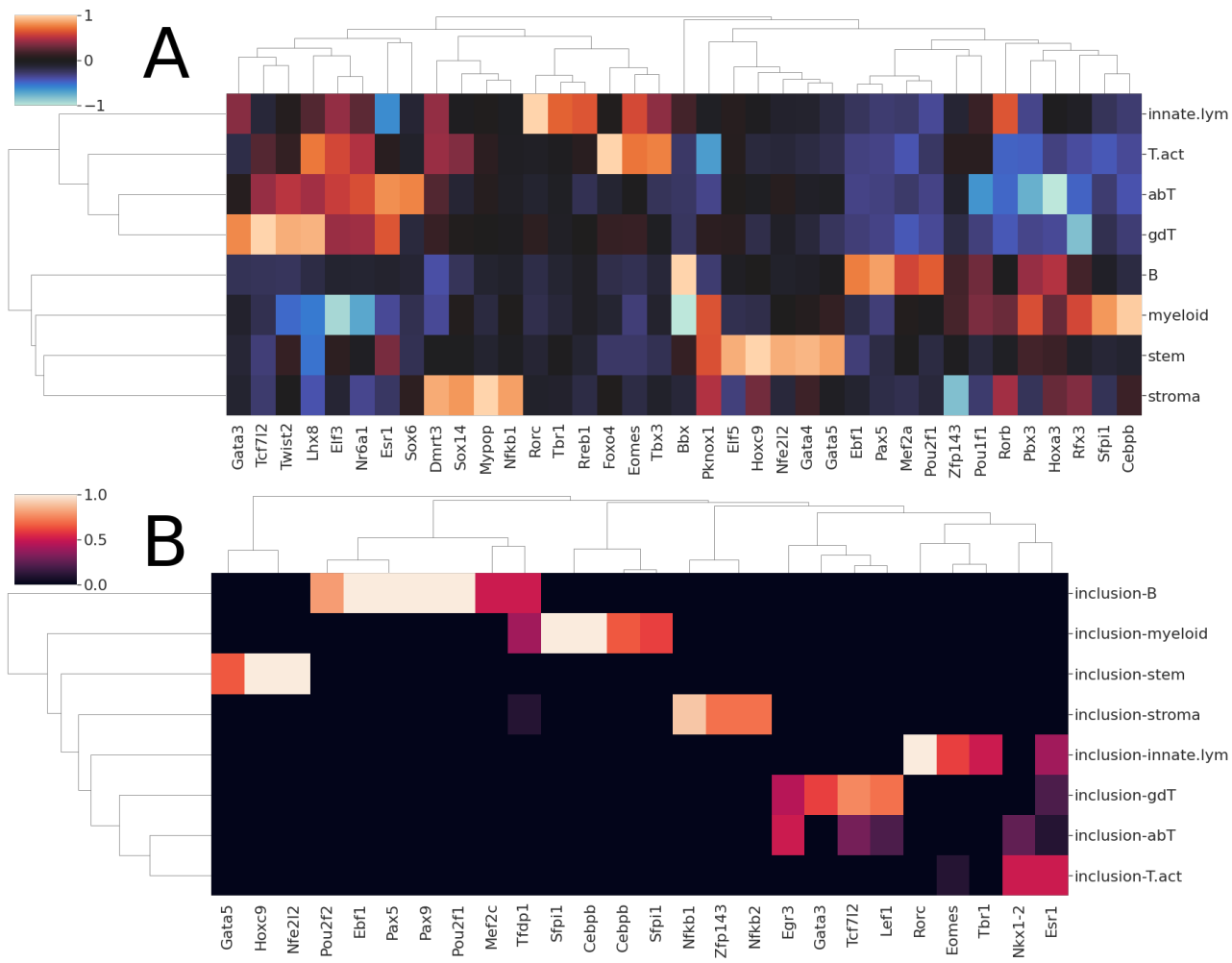

Fig. 2. tiSFM consistently finds motifs that are relevant to immune cell differentiation. (A) The final layer of a single fully trained model, after the rows were normalized to [-1, 1], while preserving the sparsity. The 5 motifs with the highest absolute weight for each cell type are shown. (B) Inclusion ratio was defined as the number of times a motif appeared among the top 10 motifs across the 10 folds in our k-fold cross validation procedure, with the highest weights in the final layer for each cell type. The motifs included in more than 50% of the models are shown here.

2A. Coefficients from the final layer of tiSFM are normalized to $[-1, 1]$ without breaking sparsity.

We find that `tiSFM` effectively highlights known regulators of immune cell differentiation. For B cells the model selects Pax5 (with Pax9, an additional TF in the family) and Ebf1, which are known as *the* key early regulators of B-cell lineage commitment (Somasundaram *et al.*, 2021; Hagman *et al.*, 2011). The model also highlights Cebpb and Sfpi1 (a.k.a. Spi1/PU.1) for myeloid cells and Rorc and Eomes for innate lymphoid cells. These TFs are also well-characterized master regulators of their respective lineages (Suñer *et al.*, 2022; Hoorweg *et al.*, 2012; Kiekens *et al.*, 2021). Importantly, these TFs were also highlighted in the original `AI-TAC` analysis and overall the `AI-TAC` and `tiSFM` TF by cell-type matrices are highly similar; an important difference, though, is that for the `tiSFM` model these are simply model parameters no additional post-hoc calculations are required.

Additionally, we use cross-validation folds from model training to assess the consistency of our approach by calculating a per-motif inclusion ratio, defined as the number of times a motif appears in the top 10 motifs with the highest absolute weights from each cell type. Several of the motifs reported consistently appeared the top 10. Indeed, all the well known lineage regulators discussed above (Pax5, Ebf1, Cebp, Sfpi1, and Rorc) have an inclusion ratio of 1 (**Fig 2B**), suggesting that TFs with similarly high inclusion ratio are likely true positives. One such TF is Hoxc9, which is highlighted by `tiSFM` as being important for hematopoietic stem cells (HSCs). While no such function for Hoxc9 has been established, Hoxa9 is a highly similar motif (cisBP database the Hoxa9 consensus is a subset of the Hoxc9 consensus) that is known to play a critical role in hematopoiesis. Defects in Hoxa9 lead to an inability of HSCs to repopulate an irradiated host (Lawrence *et al.*, 2005). On the other hand, over-expression of Hoxa9 increases the efficiency of hematopoiesis (Ramos-Mejía *et al.*, 2014).

We also note that the consistency of top transcription factors for T cells is notably lower (**Fig 2B**); nevertheless, among the ones shown we find several that are canonically associated with maintaining T cell function, including Lef1, Tcf7l2 (Tcf7 family), Egr3, and Gata3 (Xing *et al.*, 2016; Shan *et al.*, 2021; Ho *et al.*, 2009; Li *et al.*, 2012). Notably, while Gata3 has a known role in early T cell development, other of these TFs are associated with maintaining the function of specific T cell sub-types and are not necessarily establishing T cell identity. This result mirrors the original `AI-TAC` finding and conclusion that the observed weak T-cell attribution implied that the T-cell lineage was a fall back plan that did not require specific regulators. Rothenberg, 2011 supports this notion with experimental evidence .

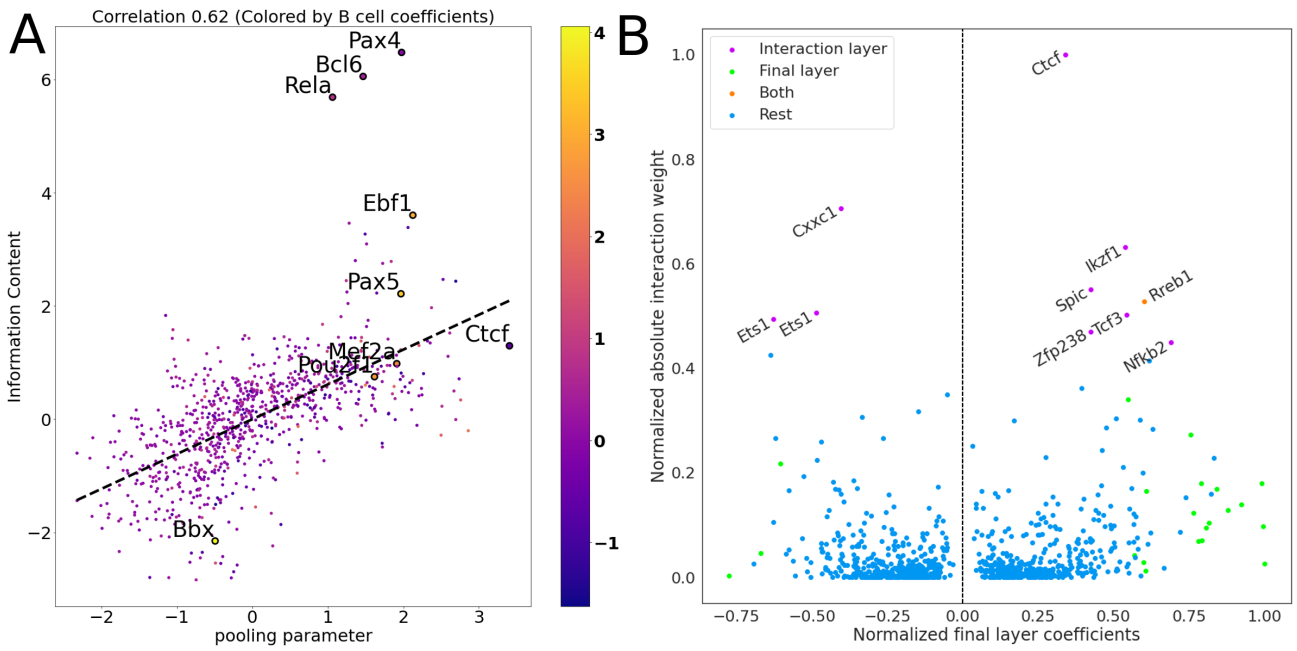

Fig. 3. (A) A scatter plot of pooling parameters vs. the information content (IC) calculated from motif PWMs. The color annotation of the points correspond to the final layer coefficients for B-cell, OCR activity prediction. The 5 motifs with the highest absolute weight for B cells are annotated (Ebf1, Pax5, Mef2a, Pou2f1, Bbx) along with some outliers to the pooling/IC trend (Pax4, Bcl6, Rela, Ctcf). (B) A scatter plot contrasting the coefficients in the final linear layer (pooled from absolute values of the cell-type prediction coefficients corresponding to the same TF, then the coefficients are normalized to $[-1, 1]$, range preserving the sparsity) with total contribution to the interaction layer. The two are notably different, only one TF influences the other TFs while contributing to the cell-type prediction significantly, as some TFs contribute far more to the interaction matrix but have near 0 linear coefficients.

3.3 Internal tiSFM parameters have intuitive biological interpretations

We have demonstrated that the final linear layer of our `tiSFM` model is interpretable and highly consistent with known biology. However, other internal parameters of the model also have biological and biochemical interpretations.

For example, our model includes a learnable pooling parameter that can interpolate between max and average pooling (see Methods). As the input to the pooling is passed through a sigmoid activation with TF specific scaling and offset, the learnable pooling model can in theory count binding events, rather than simply retain the maximal motif match. From a biochemical perspective we expect that TFs with weak protein DNA interactions to be more dependent on multiple instances of a binding motif. As such, we expect that motifs with relatively lower information content (as proxy for low sequence specificity) would prefer average pooling. This is indeed what we observe by plotting motif information content against pooling parameter in **Fig. 3A**. There we see that the information content and pooling parameter are strongly correlated. We note that Ctcf has the largest pooling parameter, indicating that the strength of the strongest Ctcf site is the most useful featurization of local Ctcf activity. In the figure weights of the final model layer for B-cells are color coded; no trend is observed, indicating that the association of motif information content and pooling parameter is independent of final layer weights. We highlighted 5 motifs with the highest absolute weight for B cell prediction: Ebf1, Pax5, Mef2a, Pou2f1, Bbx, along with some others that were outliers relative to the overall trend between information content and pooling parameter (Rela, Bcl6, Pax4).

Fig. 3B displays the total interaction influence as computed as the column sum of the absolute values of the matrix A in Equation (1). The corresponding interaction layer enables the final pooled score of one motif to influence the score of another motif, while maintaining directionality (i.e., A is not necessarily symmetric). Plotting interaction influence against the final layer coefficients (we pooled the maximum from the absolute value of coefficients, one for each cell-type, for a TF) we find that the top interaction TFs are not the same as the ones with the most influence on the final output. Indeed, Ctcf has the largest overall interaction influence, but does not appear as a top TF for any cell type. This is consistent with known biology as Ctcf is an important regulator of 3D genome organization is expressed in every cell-type and thus should not be predictive of cell-type identity.

The other two motifs that score highly for interaction, Cxxc1 and Ikzf1, also have low linear layer coefficients. While these TFs do have roles in specifying cell-fate, they have short and thus low-affinity recognition sequences and partake in complex interaction networks with other TFs that increase their binding affinity (Kiuchi *et al.*, 2021; Marke *et al.*, 2018).

3.4 Sparsity constraints improve performance for the most cell-types and increase interpretability of `tiSFM`

We emulate the path algorithm, first put forward by Park and Hastie, 2007 in their seminal work. Specifically, we utilized the proximal operators for L_1 and MCP to enforce sparsity on the final layer of `tiSFM` and to select a few motifs that are crucial for prediction of celltype-specific OCR activity. The path-finding algorithm starts with the fully trained model and progressively increases the intensity of the penalty (by increasing MCP's hyperparameter, lambda) on the final layer of the model, starting at each step from the model that was converged on at the previous step. Results are summarized in **Fig. 4**.

While we observe that the sparsity does not improve model performance significantly (**Fig. 4A**), it enables us to obtain sparser, more interpretable models that can perform similarly to the full model; sparser representations allow us to focus on consistently reappearing motifs, indicating their importance in contributing to the `tiSFM` model. **Fig. 4B** shows the regularization path for innate lymphocytes. We observe that choosing the best-performing model induces sparsity (many PWMs'

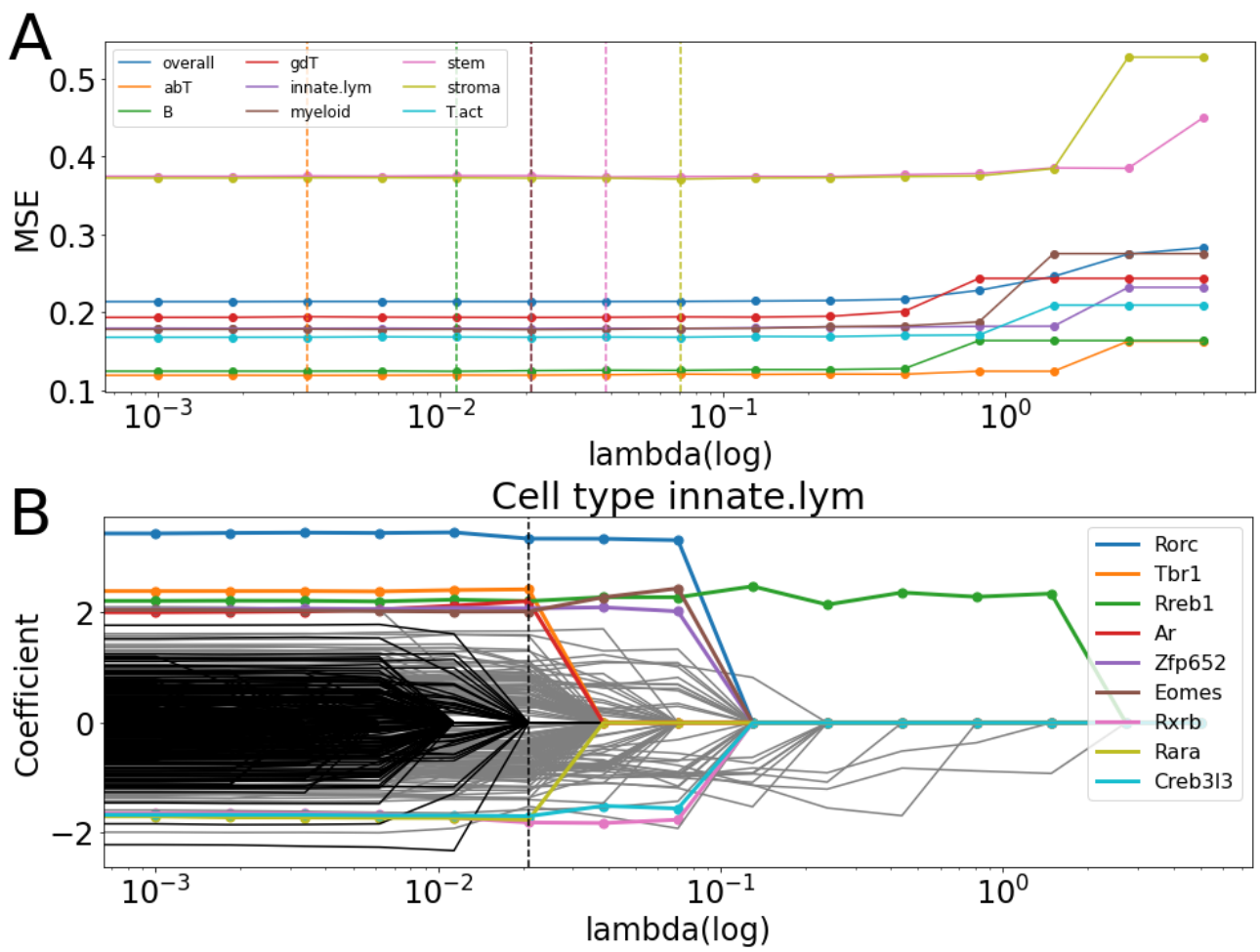

Fig. 4. Sparsity constraints improve the performance for the most cell-types and increase interpretability of the tiSFM. (A) MSE vs the MCP hyperparameter (the second hyperparameter is fixed at 3). The lines for every cell type and the overall performance are assigned a color. The vertical line indicates the hyperparameter that resulted in the best performance for the corresponding cell type. (B) Coefficients vs the MCP hyperparameter path plot. The best model in terms of MSE is marked by a dashed vertical line. Colored, are the top 9 motifs with the highest absolute coefficients from the best model, and the black lines are the motifs with 0 coefficient in the best model.

contribution is zero); PWMs with the largest coefficients are colored, and we note that Rorc is consistently selected by the path algorithm as one of the proteins that contributes the most to the epigenetic characteristics of the innate lymphoid cells, as seen in **Fig. 4B**.

3.5 tiSFM identifies key regulatory events along differentiation trajectories

For our analyses, we use `tiSFM` to analyze OCR activity changes between parents and children in the homeopoietic lineage tree, which we term "tree-diff" (see Methods). This problem is considerably more challenging than predicting open chromatin in cell-types per se. For instance, using lineage-aggregated data (i.e., pooling all cell-types within a lineage), all models are able to achieve a mean (across cell-types) R^2 of about 0.2 or a Pearson correlation of about 0.4 with `AI-TAC`'s R^2 range being 0.1 (stem cell) to 0.31 (myeloid). On the "tree-diff" data the corresponding values are around 0.05 (R^2) and $\sqrt{0.05} \approx 0.22$ (correlation). There are large differences in performance, with some changes having R^2 values near 0. We focus our analysis on those lineage differences that had $R^2 > 0.05$ (equivalently, correlation > 0.22). For visualization in **Fig. 5** these edges are scaled according to R^2 and color-coded according to the transition identity. The same colors are used to indicate TF importance (as inferred from final model layer) for each well-predicted transition (heatmap inset). The first column in the heatmap's row annotations indicate the lineage of the parent node, the second the lineage of the child node, and the third the specific edge.

Despite relatively few edges with high quality predictions, we find striking consistency with known biology among the transitions that we predict well. The lineage transitions that stand out and are modeled most clearly are those leading to NK, ILC, and myeloid subtypes from their respective common progenitors. The TF influence map for these transitions are highly correlated indicating that we capture largely general development and only to a lesser extent differentiation into specific subtypes. For these transitions TF contributions mirror closely the ones we observed when analyzing lineage aggregated data, for example B and innate lymphoid cells.

There is comparatively few well predicted transitions in the lymphocyte portion of the tree. Strikingly though, one of the transitions that is predicted well is the CD4/CD8 split. In fact, for this transition, `tiSFM` highlights Zbtb7b as a key regulator of CD4 cells — a result not observed when the T cell lineage is predicted in aggregate. Zbtb7b (often referred to in the literature as ThPOK) is a repressor gene that is well known to be essential for CD4$^+$ T cell differentiation (Basu *et al.*, 2021). As this protein is a

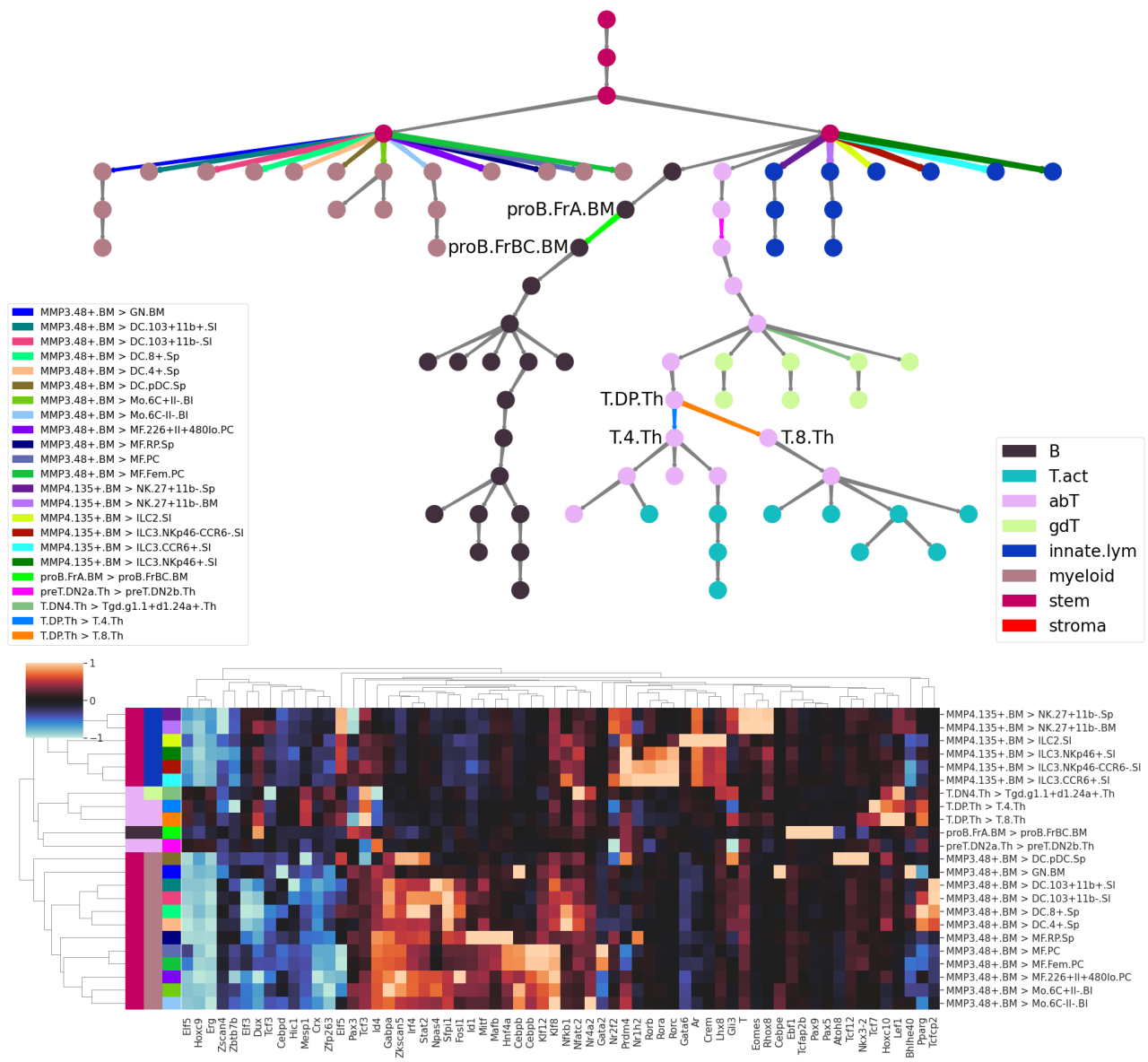

Fig. 5. tiSFM predictions and TF contributions applied to the problem of predicting differentiation transitions. In this setting the model predicts the output corresponding to each edge along the differentiation tree computed as the difference between child and parent. Edges are scaled according to R^2 and those with a value of > 0.05 are selected for TF contribution analysis depicted in the color matched heatmap. The included motifs are among top 5 with the highest absolute coefficients for at least one target.

repressor, and in fact represses the alternative CD8 lineage, the observed negative coefficient is consistent with known biology. Our "tree-diff" analysis also recovers Pax5 and Ebf1 as master B-cell regulators; however, it places their influence at a specific developmental transition denoted as "proB.FrBC.BM-proB.FrA.BM" thus pinpointing the time point at which these regulators induce lineage commitment.

4 Discussion

Non-coding regulatory grammar, encoded into the nucleotide sequence of genomes, mediates genome function through complex interactions between transcription factors and their DNA binding motifs. Here we propose `tiSFM`, a novel CNN-based architecture for predicting functional genomic readouts directly from sequence. Our approach is capable of matching and exceeding the performance of SOTA architectures, while at the same time providing immediately interpretable model parameters.

`tiSFM` has less parameters (as compared to current SOTA models like `AI-TAC`), instead relying on prior knowledge regarding sequence preferences of TFs. The fact that it outperforms SOTA standard CNN architecture highlights that learning the right PWMs (i.e., convolution kernels) is key to performance in sequence to function modeling. This result strongly indicates that subsequent CNN layers in deep CNNS are more likely to perform kernel refinement rather than compute complex regulatory grammar.

Using a dataset of hematopoietic development we show that our model re-discovers essential regulators previously highlighted by a much more complex post-hoc interpretation approach (Maslova *et al.*, 2020), without the need of any additional analysis. We also demonstrate that other internal

model parameters, such as the learnable pooling parameters and TF-TF interactions are interpretable in terms of biochemical principles and known biology.

Further, applying our model to above-mentioned data, and considering accessibility changes across the hematopoietic lineage tree, we identify specific developmental transitions that are highly predictable from sequence; this indicates that they correspond to large chromatin remodeling events that are driven by local TF activity.

Overall, our work demonstrates progress towards a model that has high predictive capacity but is also interpretable in the context of current knowledge. Significant challenges remain, however. First, it is not always possible to identify a true regulator from the PWM relevance, because many TFs come in families with highly similar PWMs. For example, our model highlighted Hoxc9 as important for hematopoietic stem cells, while the correct protein involved is most likely Hoxa9. This problem may be partially resolved by cross-referencing the motif databases and results genes that are actually expressed in the cell-type or sample of interest. Second, while we show that explicitly accounting for TF-TF interactions increases model performance and, that the interaction coefficients we recover are consistent with known biology, the interaction model could be made more biochemically relevant. For example, our current interaction layer is applied after global pooling and thus does not consider distance between TF motifs in the input sequence. The interaction layer is also not cell-type specific, which is an important limitation as biochemical interactions can only occur if both species are present, but not all TFs are expressed in all cell-types. Expanding on the TF-TF interaction model will be the subject of future efforts. Nevertheless, `tiSFM` presents a step forward in interpretable sequence-to-function modeling and can readily be applied to contexts other than modeling open chromatin during blood cell differentiation.

Acknowledgments

We acknowledge Sara Mostafavi for help with the Immgen data and helpful technical discussions.

Funding

This work has been supported by R01HG009299-5, DARPA N6600119C4022, R01AI04360321

References

Alipanahi, B. *et al.* (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.

Avsec, Ž. *et al.* (2021a). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.

Avsec, *et al.* (2021b). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, **18**(10), 1196–1203. Number: 10 Publisher: Nature Publishing Group.

Basu, J. *et al.* (2021). Essential role of a ThPOK autoregulatory loop in the maintenance of mature CD4+ T cell identity and function. *Nat. Immunol.*, **22**, 969–982.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

Gal-Oz, S. T. *et al.* (2019). ImmGen report: sexual dimorphism in the immune system transcriptome. *Nature Communications*, **10**(1), 4295.

Hagman, J. *et al.* (2011). B lymphocyte lineage specification, commitment and epigenetic control of transcription by early b cell factor 1. In *Current Topics in Microbiology and Immunology*, pages 17–38. Springer Berlin Heidelberg.

Ho, I.-C. *et al.* (2009). GATA3 and the T-cell lineage: essential functions before and after T-helper-2-cell differentiation. *Nat. Rev. Immunol.*, **9**(2), 125.

Hoorweg, K. *et al.* (2012). Functional differences between human NKp44- and NKp44 RORC innate lymphoid cells. *Frontiers in Immunology*, **3**.

Kelley, D. R. *et al.* (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**(7), 990–999.

Kiekens, L. *et al.* (2021). T-BET and EOMES accelerate and enhance functional differentiation of human natural killer cells. *Frontiers in Immunology*, **12**.

Kiuchi, M. *et al.* (2021). The Cxxc1 subunit of the Trithorax complex directs epigenetic licensing of CD4+ T cell differentiation. *J. Exp. Med.*, **218**(4).

Koo, P. K. and Eddy, S. R. (2019). Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Comput. Biol.*, **15**(12), e1007560.

Lawrence, H. J. *et al.* (2005). Loss of expression of the hoxa-9 homeobox gene impairs the proliferation and repopulating ability of hematopoietic stem cells. *Blood*, **106**(12), 3988–3994.

Li, S. *et al.* (2012). The Transcription Factors Egr2 and Egr3 Are Essential for the Control of Inflammation and Antigen-Induced Proliferation of B and T Cells. *Immunity*, **37**(4), 685–696.

Marke, R. *et al.* (2018). The many faces of IKZF1 in B-cell precursor acute lymphoblastic leukemia. *Haematologica*, **103**(4), 565.

Maslova, A. *et al.* (2020). Deep learning of immune cell differentiation. *Proc. Natl. Acad. Sci. U.S.A.*, **117**(41), 25655–25666.

Novakovsky, G. *et al.* (2022). ExplaiNN: interpretable and transparent neural networks for genomics.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(4), 659–677. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2007.00607.x.

Paszke, A. *et al.* (2019). Pytorch: An imperative style, high-performance deep learning library.

Quang, D. and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, **44**(11), e107.

Ramos-Mejía, V. *et al.* (2014). HOXA9 promotes hematopoietic commitment of human embryonic stem cells. *Blood*, **124**(20), 3065–3075.

Roider, H. G. *et al.* (2006). Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**(2), 134–141.

Rothenberg, E. V. (2011). T cell lineage commitment: identity and renunciation. *Journal of immunology (Baltimore, Md. : 1950)*, **186**(12), 6649.

Shan, Q. *et al.* (2021). Tcf1 and Lef1 provide constant supervision to mature CD8+ T cell identity and function by organizing genomic architecture. *Nat. Commun.*, **12**(5863), 1–20.

Shrikumar, A. *et al.* (2017). Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Somasundaram, R. *et al.* (2021). EBF1 and PAX5 control pro-b cell expansion via opposing regulation of the imyc/i gene. *Blood*, **137**(22), 3037–3049.

Suñer, C. *et al.* (2022). Macrophage inflammation resolution requires CPEB4-directed offsetting of mRNA degradation. *eLife*, **11**.

Weirauch, M. T. *et al.* (2014a). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**(6), 1431–1443.

Weirauch, M. T. *et al.* (2014b). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**(6), 1431–1443.

Xing, S. *et al.* (2016). Tcf1 and Lef1 transcription factors establish CD8+ T cell identity through intrinsic HDAC activity. *Nat. Immunol.*, **17**(6), 695.

Yun, J. *et al.* (2021). Adaptive Proximal Gradient Methods for Structured Neural Networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24365–24378. Curran Associates, Inc.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**(10), 931–934.