# Deep learning of immune cell differentiation

Alexandra Maslova[a,b,1], Ricardo N. Ramirez[c,1], Ke Ma[d], Hugo Schmutz[c], Chendi Wang[a,b], Curtis Fox[d], Bernard Ng[a,b], Christophe Benoist[c,2,3], Sara Mostafavi[a,b,e,f,2,3], and Immunological Genome Project

[a]Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [b]Department of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [c]Department of Immunology, Harvard Medical School, Boston, MA 02115; [d]Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [e]Canadian Institute for Advanced Research, CIFAR AI, Toronto, ON M5G 1M1, Canada; and [f]Vector Institute, Toronto, ON M5G 1M1, Canada

Contributed by Christophe Benoist, August 26, 2020 (sent for review June 24, 2020; reviewed by Anshul Kundaje and Ellen V. Rothenberg)

**Although we know many sequence-specific transcription factors (TFs), how the DNA sequence of cis-regulatory elements is decoded and orchestrated on the genome scale to determine immune cell differentiation is beyond our grasp. Leveraging a granular atlas of chromatin accessibility across 81 immune cell types, we asked if a convolutional neural network (CNN) could learn to infer cell type-specific chromatin accessibility solely from regulatory DNA sequences. With a tailored architecture and an ensemble approach to CNN parameter interpretation, we show that our trained network ("AI-TAC") does so by rediscovering ab initio the binding motifs for known regulators and some unknown ones. Motifs whose importance is learned virtually as functionally important overlap strikingly well with positions determined by chromatin immunoprecipitation for several TFs. AI-TAC establishes a hierarchy of TFs and their interactions that drives lineage specification and also identifies stage-specific interactions, like Pax5/Ebf1 vs. Pax5/Prdm1, or the role of different NF-κB dimers in different cell types. AI-TAC assigns Spi1/Cebp and Pax5/Ebf1 as the drivers necessary for myeloid and B lineage fates, respectively, but no factors seemed as dominantly required for T cell differentiation, which may represent a fall-back pathway. Mouse-trained AI-TAC can parse human DNA, revealing a strikingly similar ranking of influential TFs and providing additional support that AI-TAC is a generalizable regulatory sequence decoder. Thus, deep learning can reveal the regulatory syntax predictive of the full differentiative complexity of the immune system.**

artificial intelligence | gene regulation

The immune system has a wide array of physiological functions, which range from surveillance of the homeostasis of body systems to defenses against a diversity of pathogens. Accordingly, it includes a wide array of cell types, from large polynuclear neutrophils with innate ability to phagocytose bacteria to antibody-producing B cells to spore-like naïve T cells whose effector potential becomes manifest upon antigenic challenge. With the exception of rearranged receptors, all immunocytes share the same genome, and this phenotypic diversity must thus unfold from the genome blueprint, each cell type having its own interpretation of the DNA code. This differential usage is driven by the interplay of constitutive and cell type-specific transcription factors (TFs), regulatory RNA molecules, and possibly yet unknown sequence-parsing molecular entities.

Antigen recognition and effector potential are anchored in the cell's transcriptome, itself a reflection of the conformation of DNA within chromatin that enables the expression of accessible genes, directly or as modulated by triggers from cell receptors and sensors. Recent technical advances reveal chromatin accessibility with high precision and across the entire genome (1), providing reliable charts of chromatin structure through immune cell types (2–5). In these, open chromatin regions (OCRs) reflected quite closely gene expression in the corresponding cells. The question then is to move from these descriptive charts to an understanding of how these chromatin patterns are determined. Analyzing the representation of transcription factors binding motifs (TFBS) in these differentially active OCRs provided some clues as to the TFs potentially responsible for cell specificity, especially by using the cell type-specific expression of the TFs themselves as a correlative prior (2). Although motif enrichment analysis is a mature tool, it relies on imperfect TFBS tables assembled from different sources of data, with unavoidable noise. More importantly, functional and cellular relevance of the sequence patterns is only inferred correlatively from the enrichment.

Artificial neural networks present a powerful approach that can learn complex and nonlinear relationships between large sets of variables and can recognize patterns whose combinations are predictive of multifaceted outcomes. Convolutional neural networks (CNNs) in particular can learn the combinatorial patterns embedded within input examples without the need for alignment to predetermined references. Recent studies have begun to take advantage of CNNs to tackle aspects of gene regulation (6), including models that predict chromatin state (7–9), TF binding (10, 11), polyadenylation (12), or gene expression (7, 13) solely on the basis of DNA (100 bp to 1 Mb) or RNA sequences, with the potential to ferret out relevant motifs.

The ImmGen consortium has recently applied ATAC-seq to generate an exhaustive chart (532,000 OCRs) of chromatin accessibility across the entire immune system of the mouse (81 primary cell types and states directly ex vivo) (2). The data encompass the innate and adaptive immune systems, differentiation

## Significance

**Applying artificial intelligence tools to a highly complex question of immunology, we show that a deep neural network can learn to predict the patterns of chromatin opening across 81 stem and differentiated cells across the immune system, solely from the DNA sequence of regulatory regions. It does so by discovering ab initio the binding motifs for known master regulators, along with some unknown ones, and their combinatorial operation. These predictions validated biochemically, and a mouse-trained neural network predicts human enhancer/promoter activity much better than sequence comparisons would. Beyond serving as a trove of testable functional frameworks, this work is important in showing how massively complex integrated questions of immunology can be handled with such tools.**

Author contributions: A.M., R.N.R., K.M., H.S., C.W., C.F., B.N., C.B., S.M., and I.G.P. designed research; A.M., R.N.R., K.M., H.S., C.W., C.F., B.N., C.B., and S.M. performed research; and A.M., R.N.R., C.B., and S.M. wrote the paper.

Reviewers: A.K., Stanford University; and E.V.R., California Institute of Technology.

The authors declare no competing interest.



A complete list of the Immunological Genome Project can be found in *SI Appendix*.

[1]A.M. and R.N.R. contributed equally to this work.

[2]C.B. and S.M. contributed equally to this work.

[3]To whom correspondence may be addressed. Email: cb@hms.harvard.edu or saram@stat.ubc.ca.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2011795117/-/DCSupplemental.

First published September 25, 2020.

cascades of B and T lymphocyte lineages, and detailed splits of myeloid subsets at baseline or after activation. We reasoned that it might provide the power to push the boundaries of what CNNs can learn, in terms of 1) learning, solely from the DNA sequence of the OCRs, their pattern of activity in primary differentiated cells that span the immune system at high cell-type granularity and 2) robustly interpreting parameters of complex CNNs to identify the sequence motifs and their combination that result in these predictions. The results showed that the CNN model we derived (referred to as "AI-TAC") can learn to accurately predict the fine specificity of cell type-specific OCRs. Further, our model interpretation strategy was able to uncover motifs that are influential in silico and recapitulated the binding sites of their molecular counterparts in "real" chromatin immuno precipitation and sequencing (ChIP-seq) data. Thus, AI-TAC learns the sequence syntax that underlies the globality of immune cell differentiation.

## Results

**AI-TAC Can Predict Enhancer Activity from Sequence Alone.** We developed and trained a deep CNN, hereafter AI-TAC, to predict the chromatin accessibility profiles across 81 immune cell types on the basis of DNA sequences alone. In this way, AI-TAC learns the relationship between the combination of sequence motifs embedded within an OCR and its accessibility profile across varying immune cell types. Fig. 1*A* schematizes the steps of training, interpretation, and biochemical validation. In practice, the model was trained by using as input 90% of 327,927 sequences underlying each of the OCRs defined by our recent ATAC-seq effort (2) to predict as output the profiles of ATAC-seq of each OCR across all measured cell types. The ability of the CNN to learn an accurate mapping between inputs and outputs depends on several hyperparameters (number of hidden layers, filters and their length, loss function), and these were explored systematically (*SI Appendix*, Fig. S1). Bayesian optimization (14) showed that an architecture with three convolutional layers followed by two fully connected layers, with 19-bp sequence detected by the 300 first-layer filters, resulted in lowest achieved error on the validation data (*SI Appendix*, Fig. S1 *A* and *B*). We also found that the form of the loss function resulted in differential ability to predict cell type-specific activity profiles: using Pearson correlation as the loss function metric enhanced the ability of the model to accurately predict sequences whose activity varies across cell types ($P = 10^{-89}$) (*SI Appendix*, Fig. S1 *C* and *D*). On a subset of held-back OCRs, the trained AI-TAC model showed good performance on precisely predicting granularly variable accessibility across all populations, as shown for one example in Fig. 1*B*. Overall, 61% of test OCRs were predicted with a statistically significant correlation coefficient (at false discovery rate [FDR] 0.05) (Fig. 1*C* and *SI Appendix*). We observed a largely monotonic relationship between the predictability of an OCR and the variability of its accessibility across immune cell types, as OCRs with low prediction performance typically had small coefficients of variation (Fig. 1 *D* and *E*). This graph also indicates that the model is not missing out on particular classes of OCRs beyond those that are ubiquitously active (as confirmed in the heat map of Fig. 1*F*).

We assessed the robustness of these predictions by performing several randomization experiments to create 3 different null models (Fig. 1*C* and *SI Appendix*, Fig. S2*A*), as well as performing chromosome leave-out experiments (*SI Appendix*, Fig. S2*B*). In addition, we performed 10 independent trials of 10-fold cross-validation (i.e., 100 trained models) so that each of the 327,927 OCRs was considered as part of 10 different test sets (*SI Appendix*, Fig. S2 *C* and *D*). These data allowed us to confirm that well-predicted OCRs were generally well predicted across different models trained on different subsets of the data, suggesting that regulatory logic captured by the model was generalizable.

**Learned Motifs Are Associated with Known Pioneer Factors and Their Lineage Specificity.** We next sought to interpret AI-TAC's parameters to understand the sequence-based logic it learned that enabled its accurate predictions on variable OCRs. While there is a growing number of approaches for extracting important features from trained neural networks (15), less attention has been given to the uncertainty in feature importance: because of the nonconvexity of the problem, even two runs of the same model on the same training data can result in major differences in the learned model parameters. Here, to robustly identify the regulatory syntax learned by the AI-TAC model, we combined three approaches and concepts: 1) node (filter)-based interpretation coupled with sensitivity analysis, 2) gradient-based methods [e.g., DeepLift (16)] coupled with clustering [TFMoDisco (17)], and 3) reproducibility analysis. As discussed in *Methods*, we concluded that reproducibility analysis is critical to robust feature extraction, and after taken into account, both node (filter)-based and gradient approaches yield similar results in terms of global feature importance (*SI Appendix*, Fig. S3). Thus, we focus below on ensemble filter-based model interpretation.

For each of the 300 first-layer filters, we extracted the short sequence motif that activates it, represented as a position weight matrix (PWM), and defined operational parameters of its robustness: reproducibility (how often its motif recurs in independently trained models), influence (how much it contributes to the prediction accuracy), frequency (how many OCRs in the dataset activate it), and information content (IC) (*Methods*, Dataset S1, and *SI Appendix*, Fig. S4*A*). Combining the characteristic parameters revealed two major groups among these trained first-layer filters (Fig. 2*A*): filters in the first group (e.g., 133, 167, etc.) were rediscovered repeatedly in every or almost every independent training run and had high influence ($>10^{-4}$) and IC, with typically short (8- to 12-bp) consensus motifs reminiscent of typical TF binding sites. The second group (e.g., 259, 37, 249, 241) had far less reproducibility, influence, and IC, with motifs that only included a few scattered bases or were less focused (~15-bp long). Some of these low-influence and nonreproducible filters may represent noise in the neural network (18) or yet unknown regulatory motifs whose similarity structure may escape conventional alignment algorithms. We focused the rest of the analysis on the 99 reproducible filters, as a model retrained using these had only a small drop in performance as compared with the full model (*SI Appendix*, Fig. S4 *B* and *C*) (99 altogether). As illustrated in Fig. 2*B*, reproducible filters partitioned between filters with restricted distribution (activating $10^3$ to $10^4$ OCRs) and generally higher influence and a group of more frequently activated filters with overall lower influence and IC. To identify known motifs associated with the learned PWMs, we searched the Cis-BP database of TF motifs (19) using the TomTom algorithm (20); 101 of the 300 learned PWMs corresponded to at least one known TF motif at q value < 0.05 (Dataset S2), and interestingly, the majority of these annotated PWMs belonged to the set of reproducible filters: 76 of 99 reproducible filters correspond closely to known TF motifs, many with astonishing similarity (as illustrated for Runx, Ets, and Ctcf in Fig. 2*C*). In 10 cases, the model also discovered exact reverse complements of the same motif (e.g., Ctcf in Fig. 2*C*).

The regulatory landscape of chromatin opening throughout immune cell differentiation, as learned de novo by AI-TAC, can thus be summarized by the 99 motifs displayed in Fig. 2*D* (given the known complexities of TF motif assignments, which can reflect promiscuity and variation with cofactors or posttranslational modifications, we opted for caution when several alternative TFs were candidates, annotating several filters at the family level only). We further refined the annotation of the most likely TF to each motif by combining Cis-BP scores with the correlation between activity of the OCR and expression of the TF across cell types

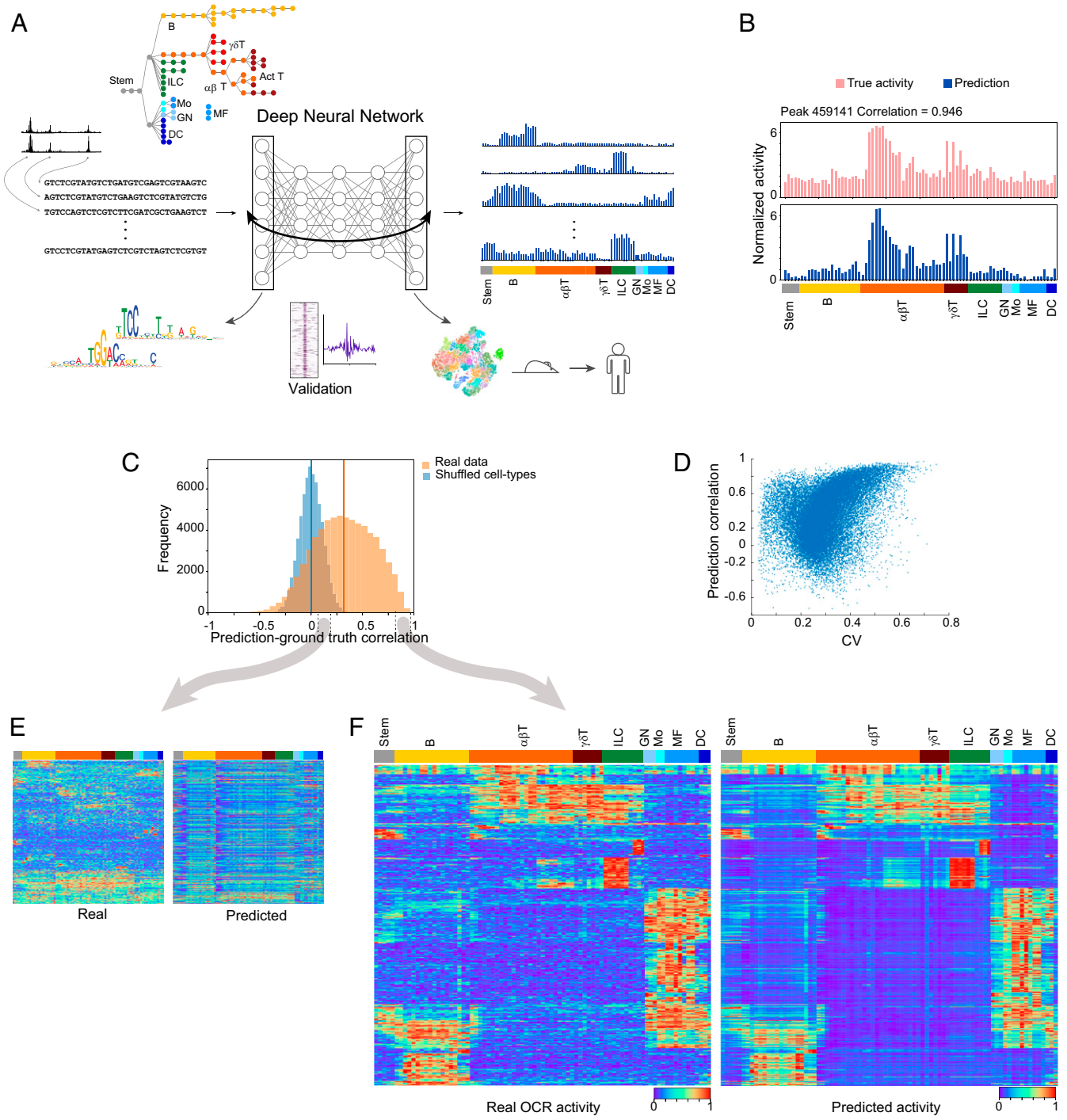


**Fig. 1.** AI-TAC learns to predict cell-specific ATAC-seq activity from sequence composition across the mouse immune system. (*A*) Schematic of the AI-TAC model and its validation. AI-TAC is a deep CNN that takes as input OCR sequences and outputs ATAC-seq accessibility profile for 81 mouse immune cells. The sequence features (motifs) that are predictive of chromatin accessibility are learned during the training process. By analyzing the first- and later-layer filters, we derive important motifs and their combinations that enable the model to make prediction for given OCRs. The predictions and motifs derived by AI-TAC are validated against actual TF binding determined from ChIP-seq experiments. (*B*) Observed (*Upper*) and predicted (*Lower*) chromatin states of 81 immune cell types for a single-test OCR. (*C*) Histogram of AI-TAC test set predictions trained on real data (orange) vs. a model trained and tested on samples with randomly permuted chromatin accessibility profiles (blue). (*D*) The coefficient of variation of the test set OCR chromatin accessibility profile on the *x* axis vs. the AI-TAC prediction correlation for those OCRs on the *y* axis. (*E* and *F*) Observed (*Left*) and predicted (*Right*) chromatin accessibility profiles for real OCRs (rows) with *E* |corr| < 0.1 and with *F* corr > 0.8 across cell types (columns). Color (the legend is shown at the bottom of *F*) indicates normalized value of accessibility ("peak height") for each OCR in each cell type.

(illustrated for Pax5 in *SI Appendix*, Fig. S5*A*); these correlations were comparable for filters annotated to the same TF (Dataset S3 and *SI Appendix*, Fig. S5*B*). The resulting set rediscovered several canonical regulators of lineage differentiation: Pax5, Ebf1, Spi1 (aka PU.1), and Gata3. Other TFs were perhaps less expected in the context of cell-specific expression such as Ctcf, a

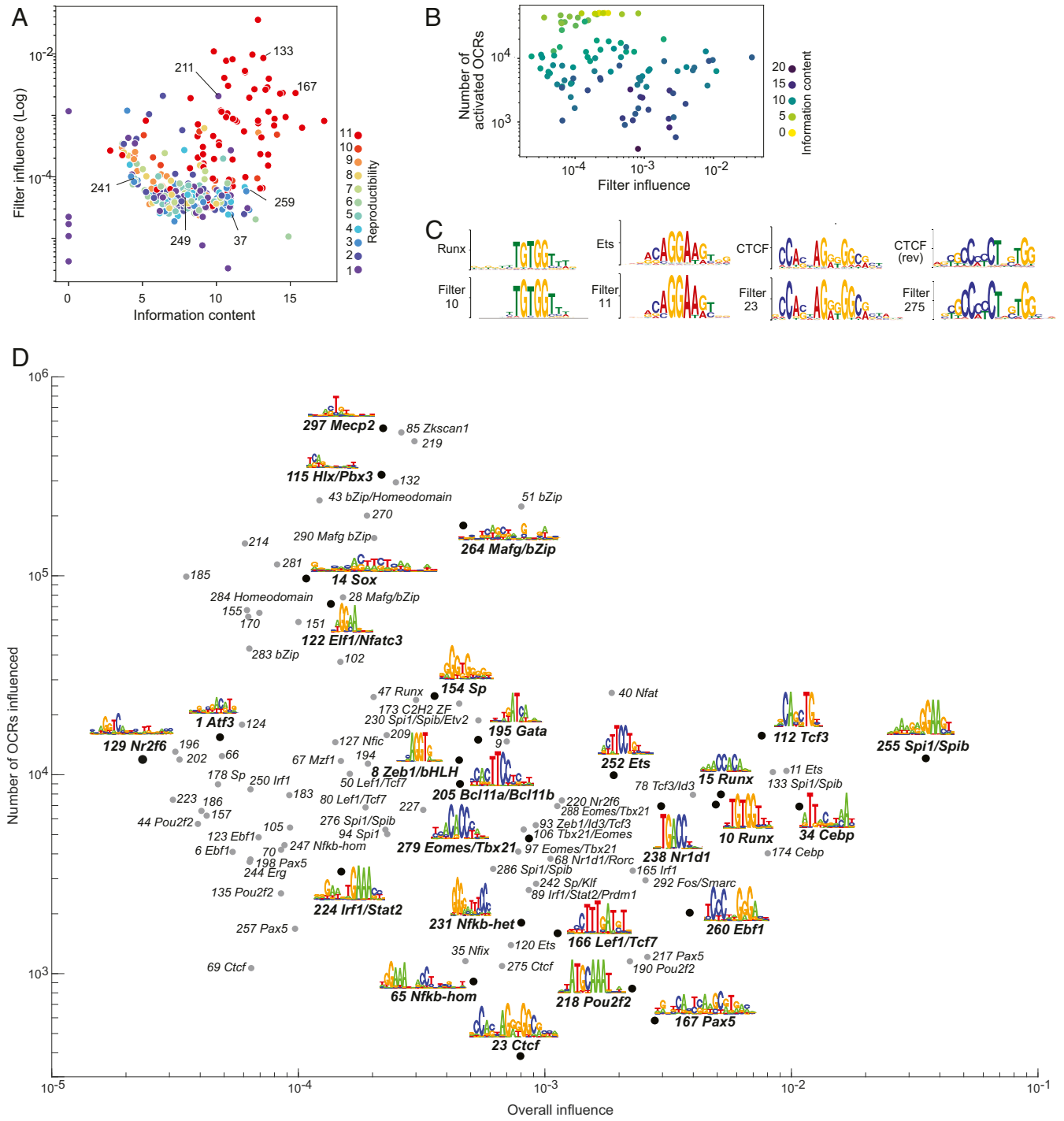


**Fig. 2.** AI-TAC learns a wide range of motifs that together predict immune differentiations. (*A*) Characteristics of each of the 300 AI-TAC first-layer filters: IC (measure of information within the PWM relative to random sequence) plotted vs. influence (contribution to overall correlation) color coded by reproducibility (frequency of recurrence across 10 independent models retrained on a different 90% subset of OCRs; filters defined as "reproduced" in each model if matching at TomTom q < 0.05). (*B*) Filter influence vs. number of OCRs activated (an OCR is considered activated by a filter if filter activation is >1/2 the maximum activation value of that filter across all input sequences) colored by IC. (*C*) Examples of known TFBS (*Upper*) compared with the PWMs of first-layer filters (*Lower*). (*D*) Sequence motifs recognized by 99 top filters, positioned based on influence and number of OCRs activated by each filter.

ubiquitous TF better known for its structural role in nuclear architecture (21), suggesting a cooperative role of Ctcf as an adjunct to lineage-specific factors. A few influential TFs were represented by several filters, usually with slightly different motifs for the same TF (*SI Appendix*, Fig. S5*C*). These nuances may correspond in part to technical noise from model overparameterization (18), but they are also expected from known degeneracy in TF binding specificity, which is further influenced by interactions within dimers, as exemplified by the NF-κB family (22). AI-TAC filters indeed distinguished the canonical NF-κB heterodimer (filter 231) and homodimer (filter 247) motifs. Consistent with distinct biological roles, we also observed specificity in the OCR maximal activation between NF-κB filters, emphasizing unique sets of OCRs based on

heterodimer and homodimer motif preferences (*SI Appendix*, Fig. S5*D*) ($R^2$ = 0.05). For a broader perspective on how AI-TAC understands NF-κB family binding sites, we clustered the PWMs of all filters annotated to NF-κB through 10 independent training runs. Interestingly, the heterodimer motif resurfaced regularly, while other motifs were less frequently discovered or allowed more sequence variation, suggesting gradations in their functional importance (*SI Appendix*, Fig. S5*E*). Finally, several filters corresponded to motifs with no significant matches in Cis-BP or similar databases (*SI Appendix*, Fig. S6*A*). Some were short and plausibly corresponded to half-sites, but others were longer and more complex (e.g., filter 9 in *SI Appendix*, Fig. S6*A*), with relatively high overall influence (*SI Appendix*, Fig. S6*B*), and may correspond to unrecognized TFBS.

**Learned Motifs Associated with Cell-Type Profiles.** To directly assess the relationship between motifs and cell types, we computed a per cell-type influence profile, quantifying the predictive importance

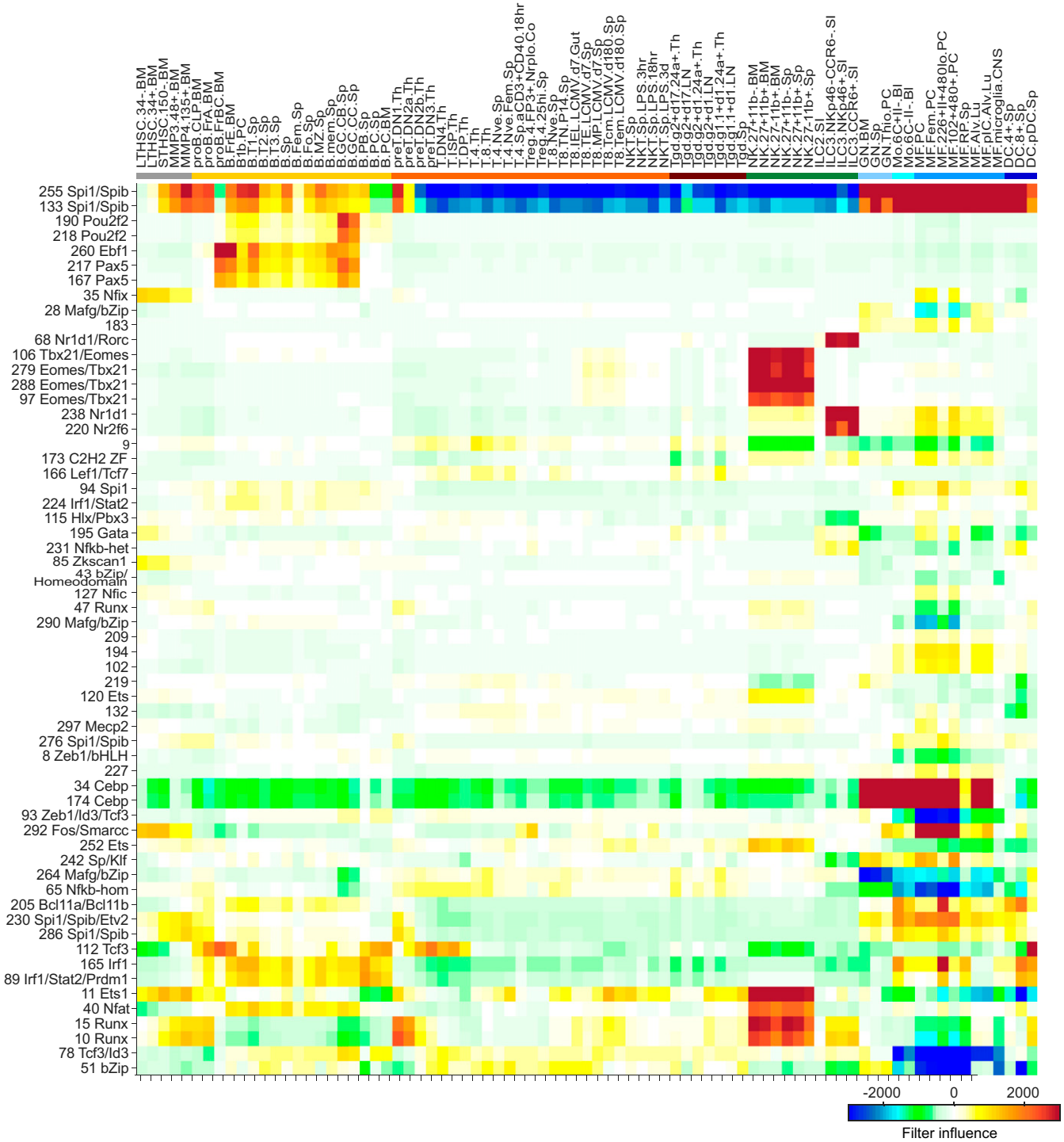


**Fig. 3.** Cell-type profiles of learned motifs. Cell type-specific influence profile for the 99 reproducible filters found in at least 80% of model training iterations.

of each filter in each of the 81 immune populations (as the difference between predicted values with and without each filter) (Fig. 3). This analysis revealed both positive and negative influences (Dataset S1). Several of these positive influences (where the filter is needed for the full activation in the predicted profiles) were consistent with known roles of the corresponding TFs: Pax5 and Ebf1 essential for B cell differentiation, Spi1 and Cebp in myeloid cells, and Tbx21/Eomes in NK cells. AI-TAC identified granular specificity of TFs beyond lineage-level importance: for instance, in the B lineage, Pax5 showed pronounced influence early in pro-B stages and late in germinal center B cells, while Pou2f2 (Oct2) was influential only in the latter. In myeloid cells, CEBP seems particularly influential in neutrophils, monocytes, and tissue macrophages and less so in dendritic cells (consistent with ref. 2) and in central nervous system microglia, an interesting notion given that microglia have a distinct origin from most other macrophage populations. No filter had the same degree of influence for T cells as Cebp/Spi1 or Pax5/Ebf1 had for myeloid and B cells.

More paradoxical were the negative influences (where the predicted activity of OCRs is overestimated in its absence). These occurred most prominently for the myeloid-specifying motifs recognized by Spi1 or Cebp, but not for every strong filter (i.e., not for Pax5 or Tbx21 motifs). Thus, the neural network used the presence of an Spi1 motif in an OCR to enforce its inactivity in T cells, beyond the neutrality that might be expected from a missing factor. Such negative influence may denote a feature of the in silico learning process but may also reflect the known need for Spi1 expression to be turned off for T cell differentiation to occur (23).

**Biochemical Validation of Predicted TF Binding.** While the identities of many motifs learned ab initio were striking and fit known biology, it was important to validate the significance of these observations. We first selected the 500 OCRs most influenced by filter 167 (Pax5) (Fig. 4*A*). In accordance with expectations, these OCRs were active in B cells, but not in thymic DPs (Fig. 4 *A*, *Right*). We then examined the fit between the in silico learned filters and the actual position of the corresponding TFs in the genome, deduced from chromatin immunoprecipitation (ChIP-seq). Overall, we observed a very strong concordance between AI-TAC's predictions and ChIP-seq data. As one example, OCRs predicted to be influenced by filter 255 (Spi1) recapitulated the two main binding sites of Spi1 in the *Il1b* locus (Fig. 4*B*). More generally, the top OCRs influenced by filters 167 (Pax5), 260 (Ebf1), and 166 (Lef1/Tcf7) strikingly overlapped with binding sites defined by ChIP-seq for those factors, relative to control OCRs (0.006 to 0.09; $P < 0.003$) (Fig. 4*C*).

Finally, we analyzed deep ATAC-seq traces in B lymphocytes at nucleotide-level resolution, where one can discern a "footprint" where the binding of a TF prevents or favors accessibility by the Tn5 transposase (24). We superimposed deep (>200 million reads) ATAC-seq traces at positions predicted to activate Pax5 or Ctcf filters in AI-TAC over true binding sites independently determined known from ChIP-seq and motif identification. Here again, AI-TAC–driven predictions accurately coincided with true TF binding, showing the same fine details of accessibility (Fig. 4*D*). Thus, whether in matching the distribution of TF binding or the nucleotide-level traces to biochemically determined ones, AI-TAC in silico predictions are strongly validated by in vivo data.

**Dissecting the Combinatorial Logic of Chromatin Opening.** That enhancer elements tend to occur as repeats has long been a theme, either because those first discovered in viral genomes occurred as tandem repeats or because synthetically engineered enhancers were more effective as strings of the same motif. Thus, we asked whether repeats of the same motif were enriched among active OCRs. While this was not the case for the majority (*SI Appendix*, Fig. S7*D*), two interesting exceptions were GC-rich motifs recognized by Sp (242), consistent with reports that SP1 functions best in the context of repeated GC-rich blocs (25), and filter 231 (NF-κB-het), consistent with the demonstration that NF-κB uses clustered binding sites noncooperatively to incrementally tune transcription (26). On the other hand, activation of the same filter by different OCRs that control the same gene [likelihood determined by regression (2)] showed a significant enrichment for repeats of the same motif compared with chance (*SI Appendix*, Fig. S7*E*). Thus, tandem repeats of controlling motifs within short segments of accessible chromatin are not a regulatory strategy commonly employed to control immune cell differentiation, but motif repetition is provided by independent elements scattered around a gene, likely connected by DNA loops.

Given the size of the vertebrate genomes, combinations of transcriptional regulators are the only practicable solution to encode the complexity of development and cell-type differentiation (27). Pervasive interactions between TFs within multimolecular complexes have been observed in genomic and functional experiments, but an overall perspective on the combinatorial interactions that actually influence transcription remains incomplete. It was thus of interest to ask which combinations of motifs are coinfluential in AI-TAC's predictions. Because the higher-order relationships between first-layer motifs are encoded in the deeper layers of the network, an obvious first attempt at identifying important filter combinations is to look for combinations of motifs assembled by the second-layer convolutional filters (12). We found that in a large number of cases, the second-layer filters recognized similar (or reverse complement) first-layer motifs, indicating that the second layer is perhaps assembling cleaner versions of first-layer motifs rather than learning combinatorial logic (*SI Appendix*, Fig. S8).

As an alternative, we identified for each OCR the set of filters that impact the accuracy of its prediction (i.e., influence) by 5% or more. Of the set of OCRs that were influenced by at least 1 filter at this threshold, many ($n = 23{,}910$, 56%) were influenced by 2 to 6 filters, and a few ($n = 1{,}514$, 4%) were even impacted by 10 or more filters (Fig. 5*A*). This large set of OCRs impacted by multiple filters provided a rich base to identify common coinfluential motifs. To identify influential combinations between different TFs, we computed for each filter pair the number of OCRs that they both impact and compared it with expected coinfluence based on each filter's prevalence. This analysis yielded 493 coinfluential filter pairs (adjusted $P < 0.05$ and number of co-occurrences >100) (Fig. 5*B* and Dataset S5). Interestingly, filters that are broadly influential tended to be significantly coinfluential with each other (e.g., Ebf1 and Pax5, $n = 193$, $P <$ 10e-20; Lef1/Tcf7 and Runx, $n = 471$, $P <$ 10e-50). Among overrepresented pairs, some TFs were highly recurrent, acting as "hubs" of sorts: Tcf3 (filters 78/8/93), Runx (filter 10), Ets (filter 11), and Nfat (filter 40) co-occurred with 40 or more other filters (Dataset S5).

Some of these inferences in terms of motif coinfluence were congruent with existing knowledge (e.g., Tbx21/Runx, Spi1/Cebp, etc.), but to provide proof-of-principle validation, we again turned to ChIP-seq data. Using Pax5 ChIP-seq datasets generated in pro-B and mature B cells (28), we asked what fraction of the OCRs influenced by each AI-TAC filter overlapped with a validated Pax5 binding site. As expected from Fig. 4*C*, OCRs influenced by filters 167, 217, and 257 (all annotated as Pax5) contained a high proportion of true Pax5 binding sites in both pro-B and mature B cells (0.62 to 0.83) (Fig. 5*C*). Interestingly, OCRs influenced by several other filters also contained a high proportion of Pax5 binding sites, in particular filters 260 (Ebf1),

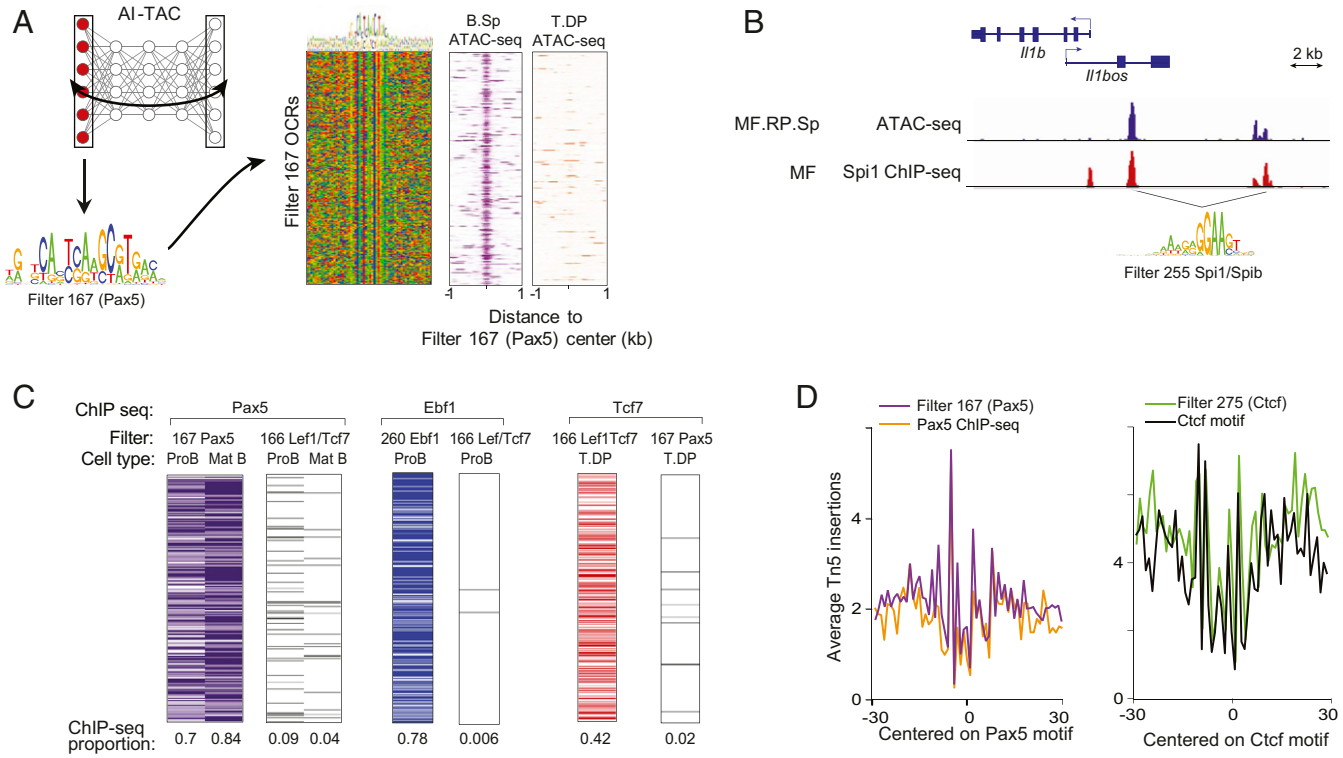


**Fig. 4.** Biochemical validation of AI-TAC learned motifs. (*A*) The top 500 OCRs influenced by filter 167 (Pax5) were selected, their consensus was verified (*Center*), and their ATAC signal in B cells or thymic DPs is displayed in *Right*. (*B*) The two OCRs activated by AI-TAC filter 255 (matches Spi1) around the *Il1b* locus, positioned relative to Spi1 ChIP-seq peaks in macrophages (data from ref. 52). (*C*) Validation of predicted filters against in vivo ChIP-seq data. The top 500 OCRs that activate filters 167 (matches Pax5), 166 (Lef1/Tcf7), and 260 (Ebf1) are shown with their overlap to ChIP-seq peaks in immunoprecipitations with anti-Pax5, -Ebf1, and -Tcf7 from pro- or mature B cells or DP T cells (data from refs. 28 and 32). (*D*) Fine ATAC-seq footprint in OCRs activated by filter 167 (Pax5) or 275 (Ctcf), compared with footprint at Pax5 and Ctcf ChIP-seq footprint, predicted footprint of Ctcf (275) and Ctcf motif.

89 (Irf1/Stat2/Prdm1), or 190 (Pou2f1). Finding Ebf1 associated with Pax5-annotated filters is consistent with the known molecular collaboration between Ebf1 and Pax5 in controlling B cell identity (29–31). This conclusion was borne out by displaying Pax5 and Ebf1 ChIP-seq signals in OCRs active in B cells, showing that some OCRs preferentially bound one of these two TFs and many both (Fig. 5*D*).

As another validation and to further identify combinatorial signals, we compared ChIP-seq signals for Pax5 and Tcf7 (active in T cells) in OCRs predicted to be activated by different filters. OCRs influenced by Lef1/Tcf7 filters (166, 50, 80) were again strongly enriched in Tcf7-bound sites in DPs (32) but had low Pax5 signals in pro-B cells, while OCRs that activate Pax5 filters and its associated Ebf1 and Pou2f filters were enriched in Pax5 ChIP-seq signals in pro-B cells but low in Tcf7 (Fig. 5*E*). OCRs that activated filters annotated to Sp1 (242) or bZip (51) were enriched in Tcf7 ChIP-seq, confirming that these TFs interact with Tcf7 (Dataset S5). Interestingly, AI-TAC predictions recovered regulators Tcf3/E2a (112) and Ets (11, 252, 120) with similar enrichments in Pax5- and Tcf7-bound sites, consistent with known overlapping regulatory function in the specification and maintenance of B and T lineages (33). Thus, combining AI-TAC predictions with in vivo ChIP-seq data parsed TF binding patterns with regulatory coinfluence at different stages of T or B differentiation and resolved regulatory motifs represented in Tcf7-bound sites across disparate T cell states.

**TF cis-Regulatory Syntax Embedded in AI-TAC's Fully Connected Layer.** The last fully connected layer of a neural network represents the final nonlinear embedding of the input examples in the derived feature space. To visualize this space, we represented each well-predicted OCR by its activation values across the 1,000 neurons of the last layer and projected these activation vectors in two dimensions using the t-SNE algorithm (Fig. 6*A*). When OCRs in this space were colored by their accessibility in different lineages, lineage-specific activity mapped to different segments (Fig. 6*B*), indicating that this last layer discriminates well between lineages. Next, we analyzed how the influence of individual first-layer filters (and corresponding TFs) projected in this space. The influence of Pax5 (filter 167) and Ebf1 (filter 260) was highest in closely related poles of the B cell area, overlapping partially (Fig. 6*C*), in accordance with Fig. 5*E*. Similarly, the influence of Spi1 (filter 255) and Cebp (filter 34) in myeloid lineage OCRs was distinguishable, with some OCRs influenced by both (Fig. 6*D*), consistent with the known cooperativity of Spi1 and Cebp across myeloid cell types (34). OCRs influenced by the NF heterodimer (filter 65) and homodimer (filter 231) motifs showed a different cell distribution, revealing a different preference in T and B lineages (*SI Appendix*, Fig. S9 *A* and *B*).

Among the patterns of OCR activity projected in this embedding space, the stratification of OCRs accessible in innate lymphoid cells (ILCs) was intriguing (Fig. 6*B*), as it demarcated a cluster of OCRs distinct from all others. We cannot formally rule out that this unusually strong demarcation of ILC-active OCRs results from a technical artifact, although have no indication in this sense. The dichotomy turned out to reflect a partition between OCRs active in NK cells vs. ILC3 (and to a lesser extent in ILC2, colonic Treg and some Tgd cells) (*SI Appendix*, Fig. S9*C*). OCRs active in NK cells were influenced by Tbx21/Eomes-related filter 106 (*SI Appendix*, Fig. S9*D*), but ILC3-preferential OCRs were mainly influenced by filters annotated to the Nuclear Receptor (NR) family Nr1d1/Rorγ (68) and Nr2f6 (220)

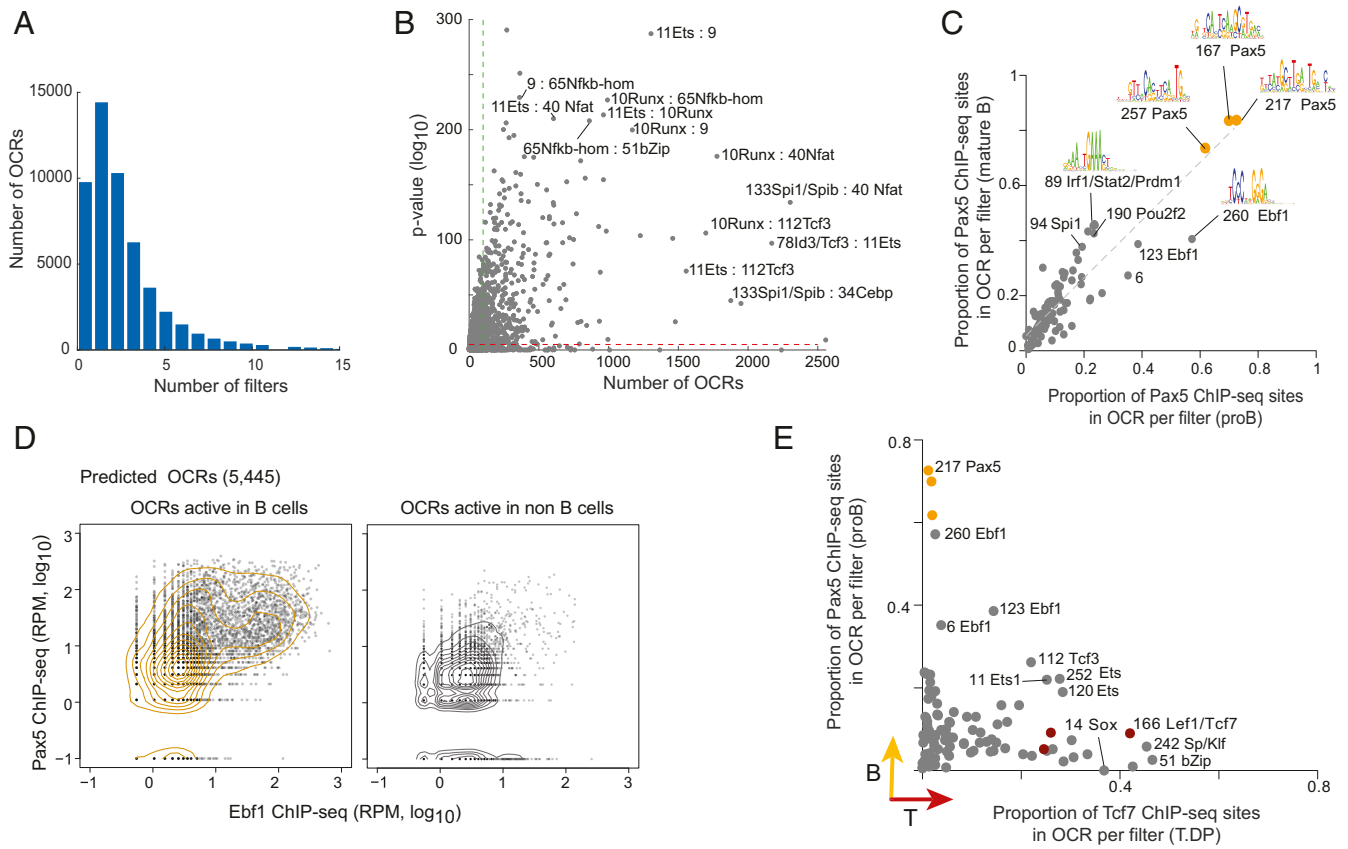


**Fig. 5.** Identifying combinations of motifs that are predictive of immune differentiations. (*A*) Number of filters per OCR that have an influence value of 0.0025 or more, which corresponds to a 5% impact on the correlation of the prediction. (*B*) For each pair of filters, the number of OCRs where both filters were deemed influential is shown on the *x* axis, and the hypergeometric *P* value for the significance of the number of shared OCRs, compared with expectation based on prevalence alone, is shown on the *y* axis. To eliminate technical artifacts, filter pairs whose motifs were similar to each other (PWMEnrich > 0.5) were removed. (*C*) Enrichment of OCRs (top 500 influential OCRs per filter, $n = 49{,}500$ OCRs) bound by Pax5 in AI-TAC reproducible filters ($n = 99$) for pro- and mature B cells. (*D*) In vivo ChIP-seq occupancy for Pax5 and Ebf1 in AI-TAC predicted B cell OCRs ($n = 5{,}443$). Co-occupancy patterns observed in predicted B OCRs and for non-B predicted OCRs ($n = 5{,}443$). (*E*) Enrichment of OCRs (top 500 influential OCRs per filter, $n = 49{,}500$ OCRs) bound by Tcf7 or Pax5 ChIP-seq in AI-TAC reproducible filters ($n = 99$) for T.DP and pro-B cells, respectively.

(*SI Appendix*, Fig. S9*D*; see also *SI Appendix*, Fig. 3*A*). The influence of these NRs is consistent with the demonstration of a role for Nr1d1 in Ilc3 differentiation (35). Thus, chromatin activity and TF control learned in silico appear very different for these groups of ILCs.

Apart from predicting lineage-specific patterns, the last layer also parsed a subset of OCRs with widespread activity across all lineages ("Ubq" in Fig. 6*A*). These small clusters were characterized by the influence of the ubiquitous TFs Sp/Klf (filter 242) and Ctcf (filter 23/275) (*SI Appendix*, Fig. S9*E*), suggesting common structural motifs. Interestingly, the influence by Ctcf filters was also observed in clusters of more cell type-specific OCRs, a disposition consistent with the notion that Ctcf partakes in the generic organization of DNA topologies in the nucleus but also cooperates with cell type-specific TFs to form specific loop and domain structures (21). Thus, AI-TAC's final-layer embedding of OCRs had the ability to refine lineage and cell specificity through TF influence patterns, suggesting that marginal influence estimates can serve as a proxy for the biological regulatory impact.

**Cross-Species Generalization of AI-TAC Predictions.** The ultimate test of generalizability of a trained machine learning model requires assessing its performance on independent/external datasets. Because the human and mouse immune systems share many regulatory nodes (36–38), and TFs and their motifs are conserved across far wider evolutionary distance, we used cross-species testing to assess AI-TAC predictions on unseen human OCRs defined by a prior ATAC-seq analysis in 25 hematopoietic cell types (5). We first used the ImmGen pipeline to preprocess the human dataset, identifying 539,611 OCRs of 251-bp length. We then directly applied the mouse-trained AI-TAC model on these human sequences and predicted their accessibility across the eight cell types from the mouse model that had a counterpart in the human dataset (Dataset S7). AI-TAC thus applies the cis-regulatory logic it has learned on mouse OCRs (including the composition of regulatory motifs and their distance preferences) to human OCR sequences to make predictions about cell-type activity. The correlation between predicted and observed accessibility profiles was significant for a large number of these OCRs (Fig. 7*A*). Note that cross-species identification of orthologous cis-regulatory sequences is very inefficient with standard sequence alignment tools (39): with the standard sequence alignment using LiftOver (39), only 13% of human OCRs could be aligned to any mouse OCR (at >95% bp remap). In contrast, by reasoning about hierarchical motif composition, the CNN is better able to ferret out orthologous OCRs.

We then explored the degree of conservation of the important AI-TAC TF motifs. After fine tuning the AI-TAC model on human data, we obtained influence scores for each filter based on its prediction performance on the set of well-predicted

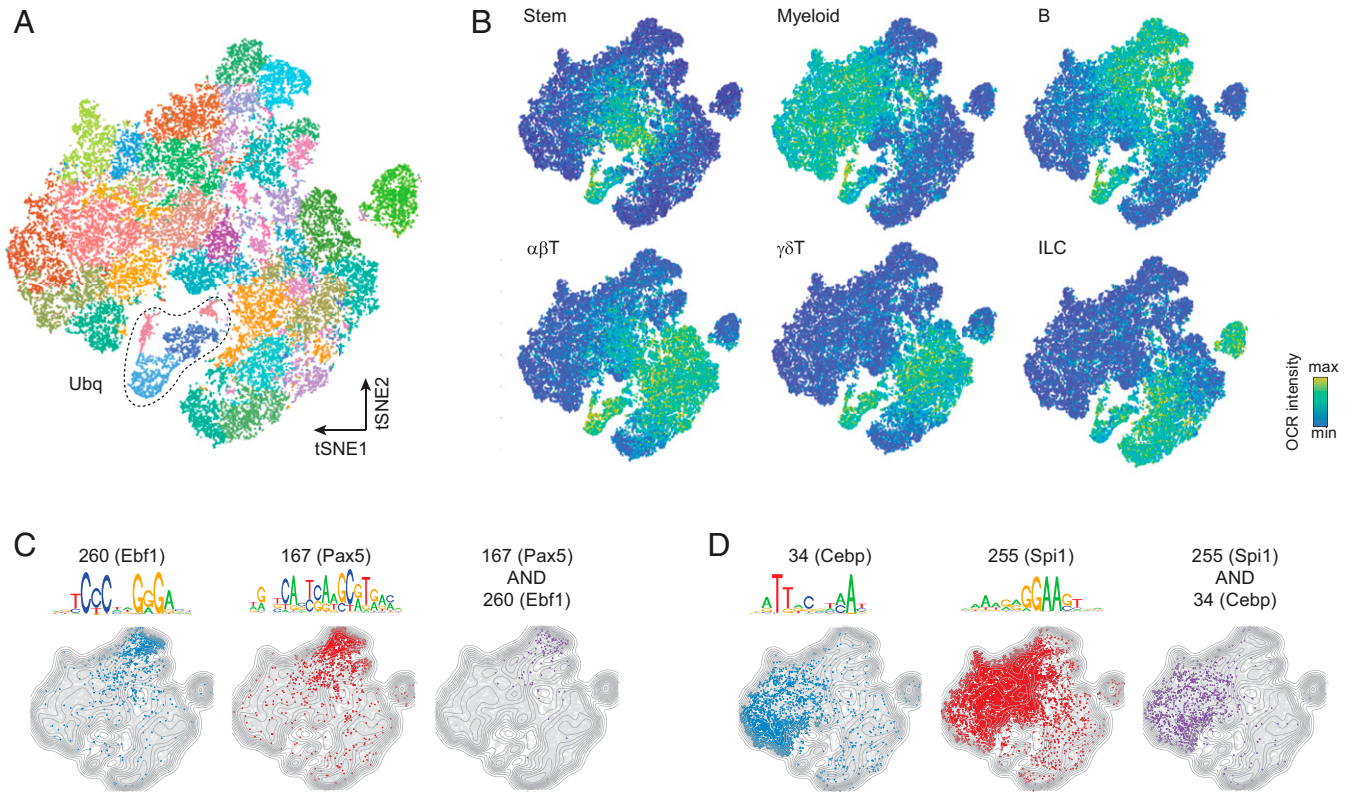


**Fig. 6.** Identifying combinatorial regulatory syntax embedded in AI-TAC's fully connected layer. (*A*) t-SNE representation and clustering of well-predicted OCRs ($n = 30{,}875$) based on their scores across the last layer (695 nodes) of the trained AI-TAC model. (*B*) ATAC-seq intensity of OCRs across immune lineages. (*C*) OCRs influenced by filters 167 (Pax5) and 260 (Ebf1) and co-occurring. (*D*) OCRs influenced by filters 34 (Cebp) and 255 (Spi1) and co-occurring.

human OCRs. We observed a striking correlation in terms of predictive influence of a filter in mouse and human datasets, indicating preservation of overall regulatory impact on the immune cells profiled here (Fig. 7*B*). Only a few outliers were noted (for example, Gata), which in this case, may be explained by the addition of erythroid cells in the human dataset.

Finally, to assess whether there exist broad classes of human OCRs that are not predictable by the mouse model, we trained a CNN directly on the human dataset and compared its prediction performance over the test set of OCRs with the mouse AI-TAC model. We observed a strong correlation between prediction performance of the mouse and human models on these human OCRs (Fig. 7*C*), with only a minor shoulder of OCRs that were better predicted by the human-trained model. This indicates that the regulatory code that is predictive of immune cell chromatin activity in regulatory regions is strongly conserved between human and mouse species.

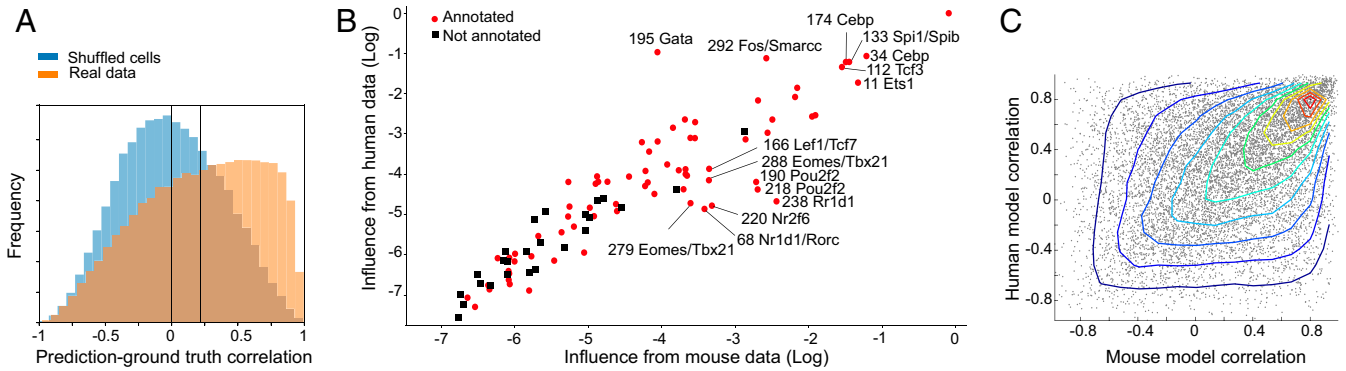


**Fig. 7.** AI-TAC model is predictive of human OCR accessibility profiles. (*A*) The trained AI-TAC model was directly applied to predict accessibility profile of human sequences underlying OCRs across eight cell types that overlapped between mouse and human datasets. The figure shows a histogram of AI-TAC predictions (measured by Pearson correlation between observed and the model's predictions) on real human 251-bp sequences underlying 539,611 OCRs (orange) vs. randomly permuted human OCR sequences. (*B*) Influence of AI-TAC's filters in mouse (*x* axis) and human (*y* axis) on the basis of nullification of each filter at a time. (*C*) Prediction performance (Pearson correlation) for test-set human OCRs based on AI-TAC trained on mouse data (*x* axis) and a model directly trained on the human ATAC-seq training set (*y* axis).

## Discussion

Differentiated cell states and functions, deeply encoded in the DNA sequence, unfold through the coordinated action of TFs and of the transcriptional modifiers they co-opt. We show here that an artificial neural network can emulate these biological decoders and predict, based on sequence alone, cell-specific patterns of chromatin accessibility across the entire immune system. It does so with high accuracy for most cell type-specific OCRs, matching biochemical validation data and in the process, rediscovering ab initio the binding sites for known TFs. By probing the sequence cues that the CNN detects and integrates, we infer the sequence information that is necessary for the biological decoders to unfold an entire immune system, yielding a broad portrait of the sequence motifs and TFs that govern immune cell differentiation, strikingly conserved in human and mouse systems.

There is growing interest in applying deep learning computation to predict chromatin state and more broadly, transcriptional activity from nucleotide sequences (7, 13). A major breakthrough in using CNNs to accurately predict "activity profile" from sequence, which AI-TAC also benefits from, has been the utilization of multitasking frameworks that model multiple prediction tasks at once (e.g., predictions of accessibility across multiple tissues or cell types). The multitask models can learn generalizable features whose combinations are predictive of different but related outcomes; this attribute is especially powerful in regulatory biology, where combinations of a finite set of sequence motifs underlie cellular differentiation. However, the representation of training data and the criteria for providing feedback to the model during the learning phase are of key importance, on which AI-TAC differs from previous work, allowing it to establish accurate sequence-based prediction of chromatin state at cell-type resolution and across the entire immune system. By modeling continuous accessibility values across 81 cell populations that represent fine-scale differences in immunocyte differentiation and then measuring the model's prediction error based on Pearson correlation, AI-TAC parameters were optimized to identify sequence features that are predictive of differences in profiles rather than ubiquitous activity levels, a feature that proved essential to its performance. Our study also differs from previous work by its emphasis on robust extraction of learned motifs and its validation with epigenomic data. To go beyond prediction capabilities of NNs and to understand the underlying regulatory logic learned by the model, we combined three strategies and showed that accounting for reproducibility is an important factor in robust extraction of sequence motifs.

The tight fit between AI-TAC's "interpretation" and biochemical data gave high confidence that they were valid projections of the true regulation of chromatin accessibility across immunocytes. Furthermore, the cell-specific influence of these filters recapitulated prior knowledge about cell-type specificity for several TFs (e.g., Pax5 and Ebf1 in B cells, Eomes/Tbx21 in NK cells, Spi1 and Cebp in myeloid cells), also reproducing and broadening the results obtained by enrichment and regression analysis of motifs in deoxyribonuclease hypersensitivity or ATAC-seq data (2, 5, 40). Several observations are worth highlighting.

1) A high-resolution ranked landscape of chromatin regulation across the entire immune system is provided by AI-TAC. Even if many players were recognized, in particular by the studies mentioned above, their dominance (Fig. 3) was not necessarily appreciated: knockouts only identify the stage at which a TF becomes essential for further differentiation, potentially distinct from those involved in overall specification of cell-specific chromatin architecture. Less expected was the dominant influence of Eomes/Tbx21 filters for NK cells or of NRs for ILC2/3, which proved quite different from T cells (even the most differentiated NKT or effector CD8s), contradicting the oversimplification that ILCs are basically TCR-less T cells. This unique influence of NRs in ILC2/3, partially shared with RORg+ Tregs and some γδT, might prompt the speculation that these TFs and the OCRs they control are primarily active in cells at the microbial interface; it is also possible that these ILCs are further differentiated than any other cells in the dataset, a stage at which the NR family becomes more prominent.
2) T cells are different? Dominantly influential controllers were identified for B, myeloid, and ILCs, but no strong equivalent emerged for T cells (influenced more weakly by Lef1/Tcf7, Tcf3, Ets, Runx, and Gata). There are some technical caveats that might underlie this observation (e.g., redundancy or wobble of motifs recognized by a controlling TF might reduce the apparent influence of individual motifs, or its motif might only be reconstructed by the model in deeper layers). These caveats notwithstanding, one may speculate that the lack of dominant factors is that T cell differentiation follows a different strategy from other lineages: that T cells are a lineage adopted when other avenues are no longer possible (i.e., by having terminally extinguished Spi1 and Pax5) or that the functional and phenotypic diversity of T cells involves several different controllers, not a single dominant master regulator.
3) Twenty-one motifs were identified by AI-TAC (*SI Appendix*, Fig. S6). Some of the unannotated filters may represent "half-sites," perhaps mere building blocks used by the CNN (41) or perhaps biologically relevant half-sites as reported for NF-κB or NRSF (22, 42). Others appeared like typical TF binding sites (short and continuous blocks of preferred bases), and they may represent unrecognized TFs or alternative sites for known TFs and require further investigation [e.g., by directly optimizing the ability of the model to learn complete motifs (43)]. Also intriguing were the poorly reproducible filters, which typically recognized scattered conserved bases; their low individual influence and nonreproducibility in different training runs would suggest that they only represent noise, but we cannot rule out that they correspond to a different regulatory syntax, perhaps read by noncoding RNAs.
4) The repeat structure (few tandem repeats but pervasive motif repeats in different enhancers connected to the same gene) suggests that eukaryotic genes do exploit cooperative multimeric interactions by repeats of the same factor but do so by recruiting several spaced OCRs rather than by locally dense tandems, a solution that may provide both transcriptional and evolutionary flexibility.
5) TF combinations. Deeper insights of co-occurring TF motifs were gained from combinatorial predictions (Fig. 5), again strongly validated by biochemical data. Some associations were expected (e.g., Pax5 and Ebf1), and the combination of AI-TAC and ChIP-seq validation data revealed patterns of differential association in B cell stages, as well as factors with broadly distributed coinfluence (Tcf3 and Ets). However, AI-TAC also identified 493 significant interactions, many previously unreported and some encompassing unannotated filters (e.g., filter 9, associated with NF-κB, Runx, and Ets).

The underlying logic in this work is that, by analyzing how a deep neural network can decipher the cis-regulatory code of immune cell differentiation, we can infer how the biological network in live cells actually does. Some caveats need to be stated, however. Choosing correlation to determine the loss function improved predictions for variably active loci but penalized predictions of ubiquitously active OCRs. Another caveat is that CNNs leverage repeated effects and will fail to identify very specific TF combinations that act only on one or two genes that may nevertheless be functionally critical [e.g., the λ5 enhancer (29) or the fine interplay between Tbx21 and Eomes during effector T cell

differentiation (44)]. TFs have varying degrees of dependence on sequence-specific DNA binding: none for TFs such as Aire (45) and variable for others such as the estrogen receptor [strictly dependent on a canonical motif at some loci, co-opted in a looser manner at others (46)]. AI-TAC would clearly miss TFs that do not rely on specific binding. Similarly, some TFs are "opportunistic," only binding to chromatin already made accessible by other factors; FoxP3 is in this category (47), and it is interesting that no TF of the Forkhead family was discovered by AI-TAC, suggesting that Forkhead family factors may not be pioneers in hematopoietic lineages cells as they are in mesenchymal cells (48). Recent approaches that consider TF gene expression along with sequence features may be able to better parse the contribution of such opportunistic factors (49). TFs whose binding specificity is very dependent on dimer formation or on cofactors might be difficult for AI-TAC to recognize, although it is interesting to note that it is able to ferret out motifs for NF-κB, a TF family notorious for its combinatorial specificity and tolerance to variation (22). Relatedly, two factors competing for the same motifs may be poorly resolved by AI-TAC: for instance, the motif bound by Bcl11a and Bcl11b, essential for differentiation of many lymphoid, myeloid, and even erythroid lineages (50, 51), scores in AI-TAC as mainly influential in myeloid and B cells. Finally, AI-TAC cannot read the influence of other means of regulation like specific DNA methylation, and there should be potential in integrating multiple data modalities into CNNs to further improve performance.

In conclusion, a deep learning approach to genome-wide chromatin accessibility revealed modalities and complex patterns of immune transcriptional regulators that arise directly from the DNA sequence. Although some blind spots remain, this draft regulatory road map should provide a foundation to graft additional layers of human- or machine-generated results and a springboard for experimental exploration.

## Methods

The AI-TAC CNN model (*SI Appendix*, Fig. S1*A*) (https://github.com/smaslova/AI-TAC) trained on the ImmGen ATAC-seq dataset (2) (input is DNA sequence of 251-bp OCRs, predicts as output chromatin activity across 81 immune cell types) consists of three convolutional and two fully connected layers, trained using one correlation as a loss function. For parameter interpretation, 1) a node-based strategy (7) was applied to derive 300 PWMs corresponding to each of the first-layer filters, and 2) a gradient back-propagation strategy (DeepLift and TFMoDisco) was applied (16). Reproducible filters (based on PWM representation) were identified using "occurrence count" across 11 separately trained model. PWMs were annotated using TomTom (20) to search the Cis-BP database of TFBS (19) (FDR 0.05). Filter influence values were computed using an ablation strategy: each filter was removed in turn, and the average of squared delta in model's error was computed across all examples. For biochemical validation, raw ChIP-seq datasets for Pax5 (28), Ebf1 (28), Spi1 (52), and Tcf1 (32) downloaded from Gene Expression Omnibus (GEO) were peak called (53) and intersected with AI-TAC predictions. To visualize high-order sequence logic, AI-TAC's embedding captured by node activation in the last shared layer was obtained ($n$ = 1,000) and projected in two dimensions using t-SNE.

**Data Availability.** All ATAC-seq datasets from the ImmGen project are available from http://www.immgen.org/ and GEO (accession no. GSE100738). All software and scripts for generating the AI-TAC model are available from https://github.com/smaslova/AI-TAC/. All study data are included in the article and *SI Appendix*.

**ACKNOWLEDGMENTS.** We thank Drs. A. Stark, B. Kee, and S. Ghosh for insightful discussions. This research was enabled in part by computing support provided by WestGrid and Compute Canada. This work was supported by NIH Grant AI072073 (to ImmGen) and Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Research Program Grant DGN (to S.M.). R.N.R. was partially supported by NIH Supplement Grant 3R01AI116834-03S1, K.M. was partially supported by NSERC Undergraduate Student Research Awards, and S.M. was partially supported by an NSERC CREATE scholarship.

1. J. D. Buenrostro *et al.*, Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
2. H. Yoshida *et al.*; Immunological Genome Project, The cis-regulatory atlas of the mouse immune system. *Cell* **176**, 897–912.e20 (2019).
3. D. Lara-Astiaso *et al.*, Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
4. D. Calderon *et al.*, Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).
5. M. R. Corces *et al.*, Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
6. G. Eraslan, Ž. Avsec, J. Gagneur, F. J. Theis, Deep learning: New computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389–403 (2019).
7. D. R. Kelley, J. Snoek, J. L. Rinn, Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
8. J. Zhou, O. G. Troyanskaya, Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
9. D. A. Cusanovich *et al.*, A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324.e18 (2018).
10. B. Alipanahi, A. Delong, M. T. Weirauch, B. J. Frey, Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
11. H. Zeng, M. D. Edwards, G. Liu, D. K. Gifford, Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics* **32**, i121–i127 (2016).
12. N. Bogard, J. Linder, A. B. Rosenberg, G. Seelig, A deep neural network for predicting and engineering alternative polyadenylation. *Cell* **178**, 91–106.e23 (2019).
13. J. Zhou *et al.*, Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
14. J. Snoek, H. Larochelle, R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms" in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS), 2012*, F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger, Eds. (Curran Associates Inc., 2012), pp. 2951–2959.
15. S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions. arXiv: 1705.07874v2 (25 November 2017).
16. A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences. arXiv:1704.02685v2 (12 October 2019).
17. A. Shrikumar *et al.*, Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. arXiv:1811.00416v5 (30 April 2020).
18. Z. Allen-Zhu, Y. Li, Y. Liang, Learning and generalization in overparameterized neural networks, going beyond two layers. arXiv:1811.04918v5 (28 May 2019).
19. M. T. Weirauch *et al.*, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
20. S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, W. S. Noble, Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
21. C. T. Ong, V. G. Corces, CTCF: An architectural protein bridging genome topology and function. *Nat. Rev. Genet.* **15**, 234–246 (2014).
22. M. C. Mulero, V. Y. Wang, T. Huxford, G. Ghosh, Genome reading by the NF-κB transcription factors. *Nucleic Acids Res.* **47**, 9967–9989 (2019).
23. E. V. Rothenberg, H. Hosokawa, J. Ungerbäck, Mechanisms of action of hematopoietic transcription factor PU.1 in initiation of T-cell development. *Front. Immunol.* **10**, 228 (2019).
24. Z. Li *et al.*, Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 45 (2019).
25. D. Gidoni, W. S. Dynan, R. Tjian, Multiple specific contacts between a mammalian transcription factor and its cognate promoters. *Nature* **312**, 409–413 (1984).
26. L. Giorgetti *et al.*, Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs. *Mol. Cell* **37**, 418–428 (2010).
27. P. J. Mitchell, R. Tjian, Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371–378 (1989).
28. R. Revilla-I-Domingo *et al.*, The B-cell identity factor Pax5 regulates distinct transcriptional programmes in early and late B lymphopoiesis. *EMBO J.* **31**, 3130–3146 (2012).
29. S. L. Nutt, B. L. Kee, The transcriptional regulation of B cell lineage commitment. *Immunity* **26**, 715–725 (2007).
30. R. Li *et al.*, Dynamic EBF1 occupancy directs sequential epigenetic and transcriptional events in B-cell programming. *Genes Dev.* **32**, 96–111 (2018).
31. H. Singh, K. L. Medina, J. M. Pongubala, Contingent gene regulatory networks and B cell fate specification. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 4949–4953 (2005).
32. A. O. Emmanuel *et al.*, TCF-1 and HEB cooperate to establish the epigenetic and transcription profiles of $CD4^+CD8^+$ thymocytes. *Nat. Immunol.* **19**, 1366–1378 (2018).
33. E. V. Rothenberg, T. Taghon, Molecular genetics of T cell development. *Annu. Rev. Immunol.* **23**, 601–649 (2005).
34. S. Pundhir *et al.*, Enhancer and transcription factor dynamics during myeloid differentiation reveal an early differentiation block in Cebpa null progenitors. *Cell Rep.* **23**, 2744–2757 (2018).
35. Q. Wang *et al.*, Circadian rhythm-dependent and circadian rhythm-independent impacts of the molecular clock on type 3 innate lymphoid cells. *Sci. Immunol.* **4**, eaay7501 (2019).
36. T. Shay *et al.*; ImmGen Consortium, Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2946–2951 (2013).

37. A. B. Stergachis *et al*., Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* **515**, 365–370 (2014).
38. C. A. de Graaf *et al*., Haemopedia: An expression atlas of murine hematopoietic cells. *Stem Cell Rep.* **7**, 571–582 (2016).
39. A. S. Hinrichs *et al*., The UCSC genome browser database: Update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
40. A. J. González, M. Setty, C. S. Leslie, Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nat. Genet.* **47**, 1249–1259 (2015).
41. P. K. Koo, M. Ploenzke, Improving convolutional network interpretability with exponential activations. bioRxiv:10.1101/650804 (27 May 2019).
42. D. S. Johnson, A. Mortazavi, R. M. Myers, B. Wold, Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
43. P. K. Koo, M. Ploenzke, Improving representations of genomic sequence motifs in convolutional networks with exponential activations. bioRxiv:10.1101/2020.06.14.150706v1 (15 June 2020).
44. A. M. Intlekofer *et al*., Effector and memory CD8+ T cell fate coupled by T-bet and eomesodermin. *Nat. Immunol.* **6**, 1236–1244 (2005).
45. D. Mathis, C. Benoist, Aire. *Annu. Rev. Immunol.* **27**, 287–312 (2009).
46. J. Gertz *et al*., Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol. Cell* **52**, 25–36 (2013).
47. R. M. Samstein *et al*., Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* **151**, 153–166 (2012).
48. K. S. Zaret, J. S. Carroll, Pioneer transcription factors: Establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).
49. S. Nair, D. S. Kim, J. Perricone, A. Kundaje, Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics* **35**, i108–i116 (2019).
50. S. K. Durum, Bcl11: Sibling rivalry in lymphoid development. *Nat. Immunol.* **4**, 512–514 (2003).
51. W. J. R. Longabaugh *et al*., Bcl11b and combinatorial resolution of cell fate in the T-cell gene regulatory network. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 5800–5807 (2017).
52. D. Gosselin *et al*., Environment drives selection and function of enhancers controlling tissue-specific macrophage identities. *Cell* **159**, 1327–1340 (2014).
53. S. Heinz *et al*., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).