Domain adaptive neural networks improve cross-species prediction of transcription factor binding

Kelly Cochran,[1,3] Divyanshi Srivastava,[1,2] Avanti Shrikumar,[3] Akshay Balsubramani,[4] Anshul Kundaje,[3,4,*] and Shaun Mahony[1,2,*]

[1] *Center for Eukaryotic Gene Regulation, Pennsylvania State University, University Park, PA, USA*
[2] *Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA*
[3] *Department of Computer Science, Stanford University, Stanford, CA, USA*
[4] *Department of Genetics, Stanford University, Stanford, CA, USA*
[*] *To whom correspondence should be addressed (akundaje@stanford.edu; mahony@psu.edu).*

The intrinsic DNA sequence preferences and cell-type specific cooperative partners of transcription factors (TFs) are typically highly conserved. Hence, despite the rapid evolutionary turnover of individual TF binding sites, predictive sequence models of cell-type specific genomic occupancy of a TF in one species should generalize to closely matched cell types in a related species. To assess the viability of cross-species TF binding prediction, we train neural networks to discriminate ChIP-seq peak locations from genomic background and evaluate their performance within and across species. Cross-species predictive performance is consistently worse than within-species performance, which we show is caused in part by species-specific repeats. To account for this domain shift, we use an augmented network architecture to automatically discourage learning of training species-specific sequence features. This domain adaptation approach corrects for prediction errors on species-specific repeats and improves overall cross-species model performance. Our results demonstrate that cross-species TF binding prediction is feasible when models account for domain shifts driven by species-specific repeats.

Characterizing where transcription factors (TFs) bind to the genome, and which genes they regulate, is key to understanding the regulatory networks that establish and maintain cell identity. A TF's genomic occupancy depends not only on its intrinsic DNA sequence preferences, but also on several cell-specific factors, including local TF concentration, chromatin state, and cooperative binding schemes with other regulators (Siggers and Gordân 2014; Slattery et al. 2014; Srivastava and Mahony 2020). Experimental assays such as ChIP-seq can profile a TF's genome-wide occupancy within a given cell type, but such experiments remain costly, rely on relatively large numbers of cells, and require either high-quality TF-specific antibodies or epitope tagging strategies (Park 2009; Savic et al. 2015). Accurate predictive models of TF binding could circumvent the need to perform costly experiments across all cell types and all species of interest.

Cross-species TF binding prediction is complicated by the rapid evolutionary turnover of individual TF binding sites across mammalian genomes, even within cell types that have conserved phenotypes. For example, only 12-14% of binding sites for the key liver regulators CEBPα and HNF4α are shared across orthologous genomic locations in mouse and human livers (Schmidt et al. 2010). On the other hand, the general features of tissue-specific regulatory networks appear to be strongly conserved across mammalian species. The amino acid sequences of TF proteins, their DNA binding domains, and intrinsic DNA sequence preferences are typically highly conserved (e.g. both CEBPα and HNF4α have at least 93% whole protein sequence identity between human and mouse). Further, the same cohorts of orthologous TFs appear to drive regulatory activities in homologous tissues. Thus, while genome sequence conservation information is not sufficient to accurately predict TF binding sites across species, it may still be possible to develop predictive models that learn the sequence determinants of cell-type specific TF binding and generalize across species. Indeed, several recent studies have demonstrated the feasibility of cross-species prediction of regulatory profiles using machine learning approaches (Chen et al. 2018; Kelley 2019; Schreiber et al. 2020b; Huh et al. 2018).

Here, we evaluate different training strategies on the generalizability of neural network models of cell-type specific TF occupancy across species. We train our model using genome-wide TF ChIP-seq data in a given

cell type in one species, and then assess its performance in predicting genome-wide binding of the same TF in a closely matched cell type in a different species. Specifically, we focus on predicting binding of four TFs (CTCF, CEBPα, HNF4α, and RXRα) in liver due to the existence of high quality ChIP-seq data in both mouse and human. The models for all TFs showed higher predictive performance for training and test sets from the same species as compared to training and test sets from different species. We show that one source of this cross-species performance gap is a systematic misclassification of transposable elements that are specific to the target species (and which were thus unseen during model training).

We further demonstrate that integrating an unsupervised domain adaptation approach into model training partially addresses the cross-species performance gap. Our domain adaptation strategy involves a neural network architecture with two sub-networks that share an underlying convolutional layer. We train the two sub-networks in parallel on different tasks. One subnetwork is trained with standard backpropagation to optimize classification of TF bound and unbound sequences in one species (the source domain). The other subnetwork attempts to predict species labels from sequences drawn randomly from two species (the source and target domain), but training is subject to a gradient reversal layer (GRL) (Ganin et al. 2016). While backpropagation typically has the effect of giving higher weights to discriminative features, a GRL reverses this effect, and discriminative features are down-weighted. Thus, our network encourages features in the shared convolutional layer that discriminate between bound and unbound sites, while simultaneously discouraging features that are species-specific. Importantly, we neither need nor use TF binding labels from the target species at any stage in training. We show that domain adaptation techniques have the potential to improve cross-species TF binding prediction, particularly by preventing misprediction on species-specific repeats.

Results

Conventionally trained neural network models of TF binding show reduced predictive performance across species

First, we set out to evaluate the ability of neural networks to predict TF binding in a previously unseen species. We chose neural networks due to their ability to learn arbitrarily complex predictive sequence patterns (Avsec et al. 2020). In particular, hybrid convolutional and recurrent network architectures have successfully been applied to accurately predict TF binding

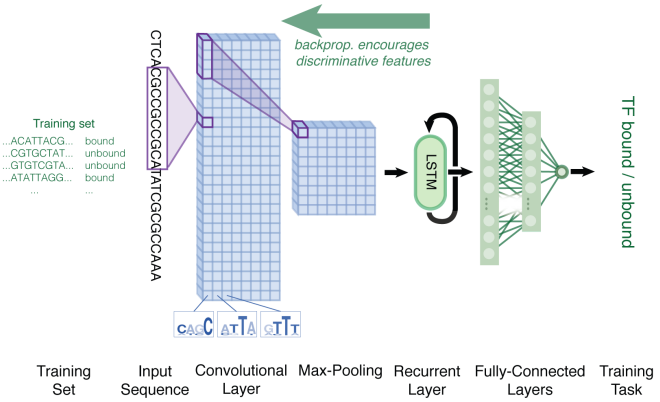

Figure 1: Conventional network architecture. Convolutional filters scan the 500-bp input DNA sequence for TF binding features. The convolutional layer is followed by a recurrent layer (LSTM) and two fully connected layers. A final sigmoid-activated neuron predicts if a ChIP-seq peak falls within the input window.

in diverse applications (Quang and Xie 2019; Srivastava et al. 2020). The motivation behind these architectures is that convolutional filters can encode binding site motifs and other contiguous sequence features, while the recurrent layers can model flexible, higher-order spatial organization of these features. Our baseline neural network is designed in line with these state-of-the-art hybrid architectures (Figure 1).

Using this architecture, named the "conventional model," we trained the network to predict whether a given input sequence contained a ChIP-seq peak or not, using training data from a single source species, and then assessed the model's predictive performance on entire held-out chromosomes in both the source species and a target (previously unseen) species. We chose mouse and human as our species of interest due to the availability of high-quality TF ChIP-seq datasets in liver from both species and the high conservation of key regulator TFs present in both species. For four different TFs, we trained two sets of models: one with mouse as the source species, and the other with human as the source species. To monitor reproducibility, model training was repeated 5 times for each TF and source species.

As models trained for 15 epochs, we monitored source-species and target-species performance on held-out validation sets (Figure 2). Performance was measured using the area under the precision-recall curve (auPRC) which is sensitive to the extreme class imbalance of labels in our TF binding prediction task. We observed that over the course of model training, improvements in source-species auPRC did not always translate to improved auPRC in the target species. Overall,

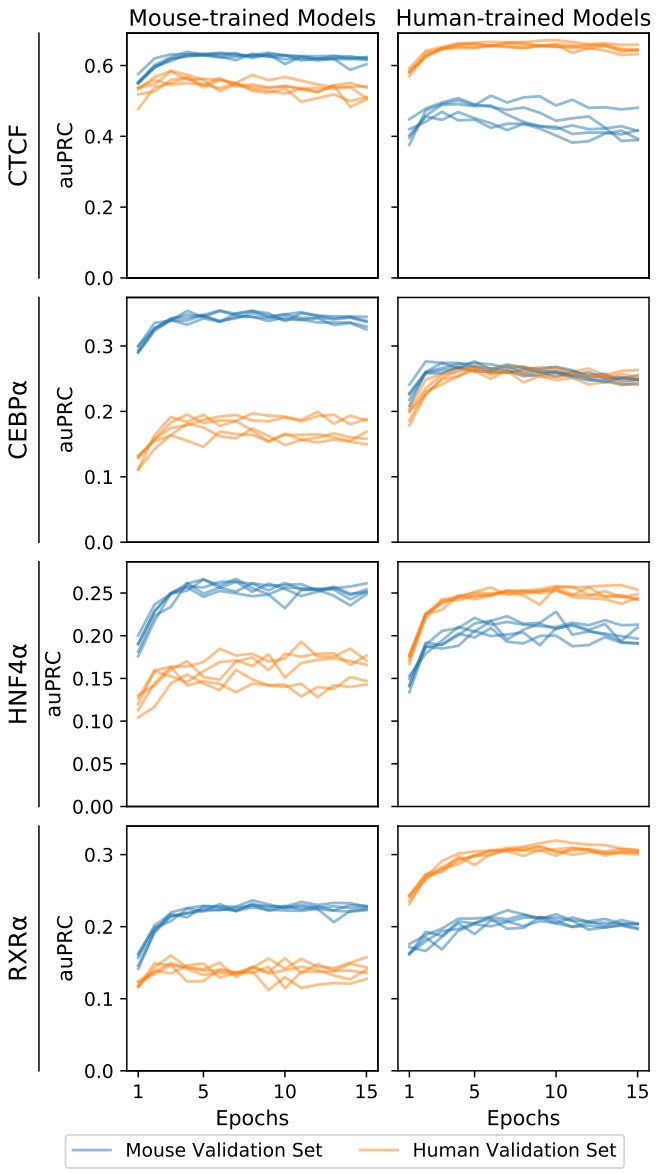

Figure 2: Model performance over the course of training, evaluated on held-out validation data from mouse (blue) and human (orange) chromosome 1. Five models were independently trained for each TF and source species.

cross-species auPRC showed greater variability across epochs and model replicates compared to source-species auPRC. For two TFs, CEBPα and $HNF4\alpha$, the mouse-trained models' performance on the human validation set appeared to split part way through training – based on cross-species auPRC, some model-replicates appeared to become trapped in a suboptimal state relative to other models (see divergence in orange lines in left column of Figure 2); meanwhile, the training-species auPRC did not show a similar trend. Evidently, validation set performance in the source species is not a reliable surrogate for validation set performance in the target species.

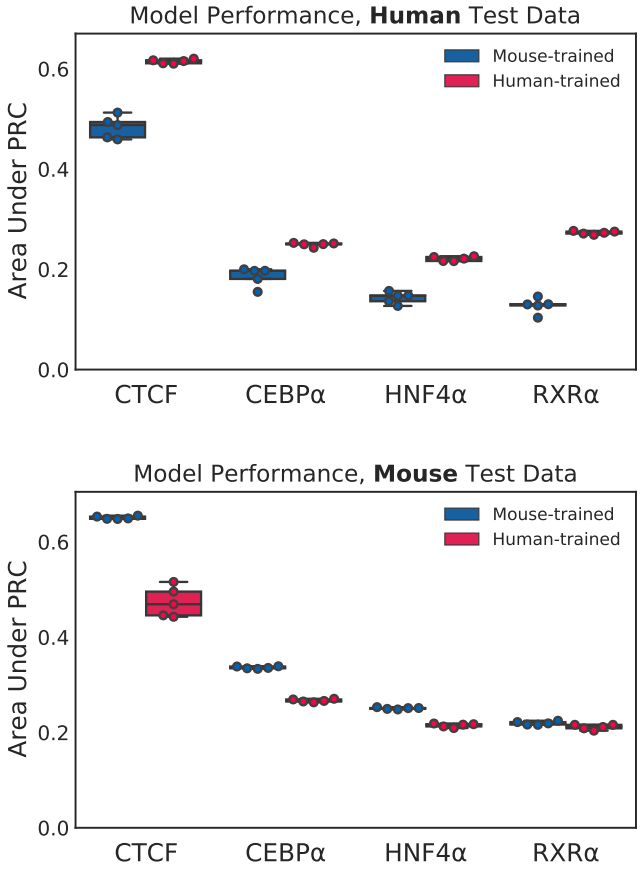

Figure 3: Model performance evaluated on held-out test data: chromosome 2 from human (top) and mouse (bottom). Five models were independently trained for each TF and source species.

Nevertheless, the epochs where models had highest source-species auPRCs were often epochs where models had near-best cross-species auPRC. Thus, we selected models saved at the point in training when source-species auPRC was maximized for downstream analysis. We next evaluated performance on held-out test datasets (distinct from the validation datasets) from each species (Figure 3).

We observe across all TFs that for a given target species, the models trained in that species always outperformed or matched the performance of the models trained in the other species. We refer to this within-species vs. cross-species auPRC difference as a cross-species performance gap, while noting that models trained in either species were still relatively effective at cross-species prediction. Intriguingly, the cross-species

gap was wider for mouse-trained models predicting in human than for human-trained models predicting in mouse. For this reason, subsequent analysis focuses on addressing the mouse-to-human gap.

The mouse-to-human cross-species gap originates from misprediction of both bound and unbound sites

Since the target-species model consistently outperforms the source-species model (on target-species validation), there must be some set of differentially predicted sites that the target-species model predicts correctly, but the source-species model does not. By comparing the distribution of source-model and target-model predictions over all target-species genomic windows, we can potentially identify trends of systematic errors unique to the source-species model. Whether these differentially predicted sites are primarily false positives (unbound sites incorrectly predicted to be bound), false negatives (bound sites incorrectly predicted as unbound), or a combination of both can provide useful insight into the performance gap between the source and target models.

For each TF, we generated predictions over the genomic windows in the human test dataset from both our mouse-trained and human-trained models. Then, we plotted all of the human-genome test sites using the average mouse model prediction (over 5 independent training runs) and the average human model prediction as the x- and y-axis, respectively (Figure 4). Bound and unbound sites are segregated into separate plots for clarity.

For all TFs, the unbound site plots show a large set of windows given low scores by the human model but mid-range to high scores by the mouse model – these are false positives unique to cross-species prediction (Figure 4 right column, bottom/bottom-right region of each plot). These sites are distinct from false positives mistakenly predicted highly by both models, as those common false positives would not contribute significantly to the auPRC gap. Additionally, in the bound site plots of all TFs except CEBPα, we see some bound sites that are scored high by the human model but are given mid-range to low scores by the mouse model – these are cross-species-unique false negatives (Figure 4 left column, top left region of each plot). Hence, our cross-species models are committing prediction errors in both directions on separate sets of sites. The errors for the unbound sites appear more prevelant than the errors for the bound sites.

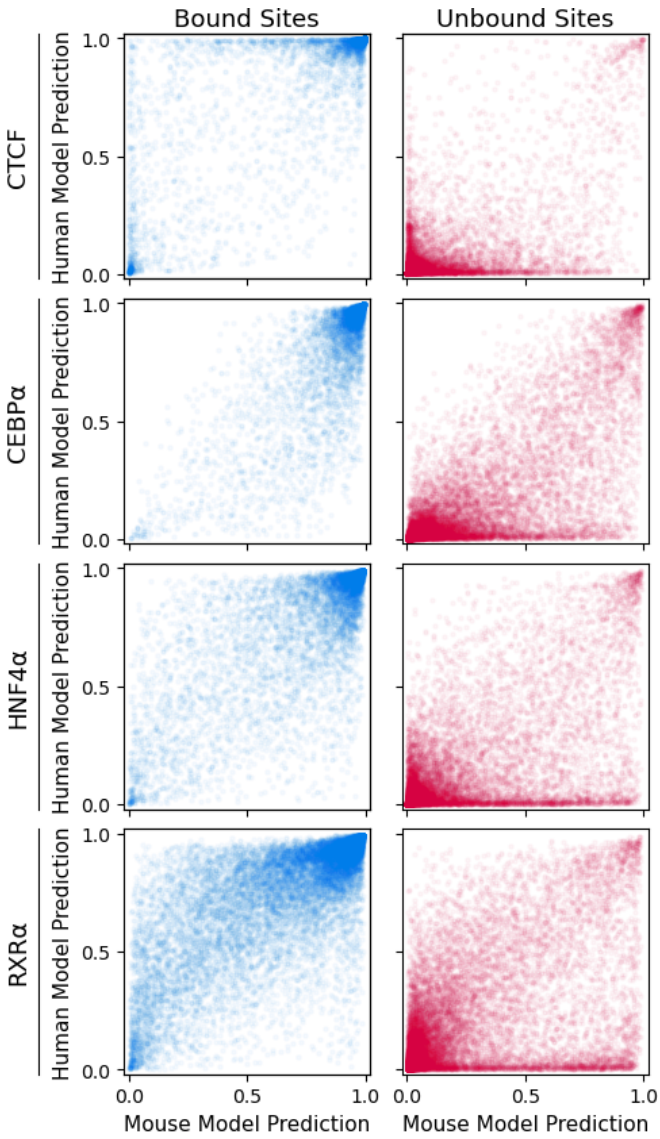

Figure 4: Both bound and unbound sites from human chromosome 2 show evidence of differential binding predictions by human-trained (y-axis) vs. mouse-trained (x-axis) models. For visual clarity, only 25% of bound sites and 5% of unbound sites are shown (sampled systematically).

Motif-like sequence features discriminate between true-positive and false-negative mouse model predictions

Since the only input to our models is DNA sequence, sequence features must be responsible for differential prediction of certain sites across source and target models. Other potential culprits, such as chromatin accessibility changes or co-factor binding, may contribute to TF binding divergence across species without changes to sequence; but without an association between those factors

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
CTCF	12.0%	10.7%	9.6%	21.2%	8.5%	**65.2%**
CEBPα	18.3%	11.5%	0.0%	21.2%	23.2%	**82.2%**
HNF4α	13.6%	9.1%	6.9%	21.3%	16.0%	**87.3%**
RXRα	13.3%	9.7%	4.7%	21.3%	19.2%	**96.9%**

Table 1: Percent of windows overlapping an Alu element, for various categories of genomic windows from the held-out test set. Alu elements dominate the false positives unique to the mouse models. FPs: false positives. FNs: false negatives. See Methods for more details on site categorization.

and sequence, the human-trained model would not be able to gain an advantage over the mouse-trained model by training on sequence input alone. Thus, we focused on genomic sequence to understand differential site prediction.

To begin, we searched for sequence-based determinants of differential prediction of bound sites from the human genome – specifically, we compared bound sequences that both the human-trained and mouse-trained models correctly predicted (true positives) to bound sequences the human-trained model correctly predicted but the mouse-trained model did not (mouse-specific false negatives). We used SeqUnwinder, a tool for deconvolving discriminative sequence features between sets of genomic sequences, to extract motifs that can discriminate between the two groups of sequences and quantitatively assess how distinguishable the sequence groups are (Kakumanu et al. 2017). SeqUnwinder was able to distinguish mouse-specific false negatives from true positives and randomly selected background genomic sequences with area under the ROC curve (auROC) of 0.84, 0.74, 0.83, and 0.88 for CTCF, CEBPα, HNF4α, and RXRα, respectively. Supplemental Figure 1 shows the breakdown of sequence features that are able to distinguish between mouse-specific false negatives and true positives for each TF. Thus, we were able to identify TF-specific motifs that were enriched (or depleted) at mouse-specific false negatives. However, we did not observe systemic sequence features that unanimously contributed to the performance gap across all TFs studied, beyond a poly-A/poly-T motif.

Primate-unique SINEs are a dominant source of the mouse-to-human cross-species gap

One potential source of sequences that could confuse a cross-species model are repeat elements found in the genome of the target species but not the source species. Alu elements, a type of SINE, cover a large portion (10%) of the human genome and are found only in primates (Batzer and Deininger 2002). Several other factors make Alus even more likely candidates for confounding mouse-to-human TF binding predictions: they

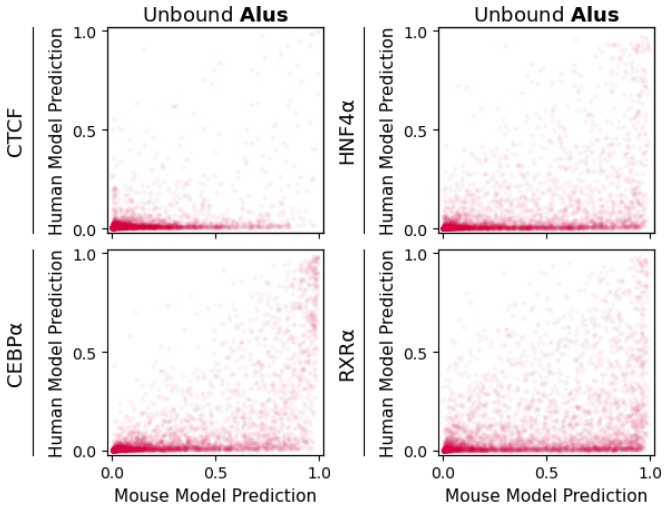

Figure 5: Most unbound sites from the human genome mispredicted by mouse-trained models (x-axis), but not by human-trained (y-axis) models, contain Alu repeats. For visual clarity, only 5% of windows are shown.

are enriched in gene-rich, GC-rich areas of the genome and contain 33% of the genome's CpG dinucleotides (a marker for promoter regions); they may play a role in gene regulation; and in silico studies have previously found putative TF binding sites within Alu sequences (Batzer and Deininger 2002; Schmid 1998; Ferrari et al. 2019; Polak and Domany 2006).

Figure 5 shows only the unbound human-genome windows that overlap annotated Alu elements. Table 1 provides corresponding quantification of Alu enrichment. Note that while Alu elements are typically poorly mappable, and it is thus often difficult to assign them as bound or unbound in ChIP-seq experiments, we focus analyses here only on highly mappable Alu instances (see Methods). Across all four TFs, we see that Alus are substantially enriched in the unbound windows predicted incorrectly only by the mouse model. On average, 83% of these false positives unique to the mouse model overlap with an Alu element, compared to the average overlap rate of 21% for unbound sites overall, or 17% for unbound sites incorrectly predicted by both

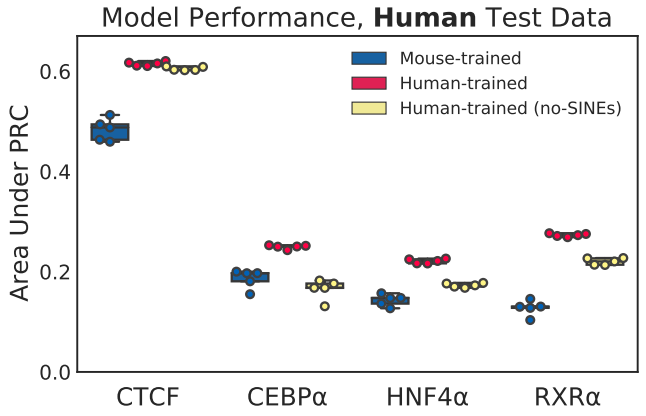

Figure 6: Performance of models that are mouse-trained (blue), human-trained with SINE examples (red), and human-trained without SINE examples (yellow), evaluated on the held-out human chromosome 2. Five models were independently trained for each TF and training species.

models. In contrast, Alus on average only overlap 7% of false negatives unique to the mouse model, which is less than the overlap fraction for bound sites overall (14%) and for false negatives common to both models (10%). We repeated this analysis using other repeat classes, including LINEs and LTRs, and confirmed that no other major repeat family shows an enrichment of comparable strength with either the false positives or false negatives unique to the mouse model (Supplementary Table 1).

Thus, the vast majority of the false positives from the human genome mispredicted only by mouse models can be directly attributed to one type of primate-unique repeat element. We did not observe any similar direct associations between primate-unique elements and the false negatives unique to the mouse model, besides the expected depletion of Alu elements.

Human models trained without SINE examples behave like hybrid mouse-human models

To further characterize how Alu elements are influencing cross-species model performance, we trained additional models on the human dataset after removing all windows from the training dataset that overlap with any SINEs (Figure 6). We filtered out all SINEs, including the primate-specific FLAM and FRAM repeats as well as Alus, to avoid keeping examples that shared any sequence homology with Alus. The no-SINE models were evaluated on the same held-out chromosome test data used previously (which includes SINEs).

Site-distribution plots show that, for unbound sites, no-SINE human-trained models tend to make similar

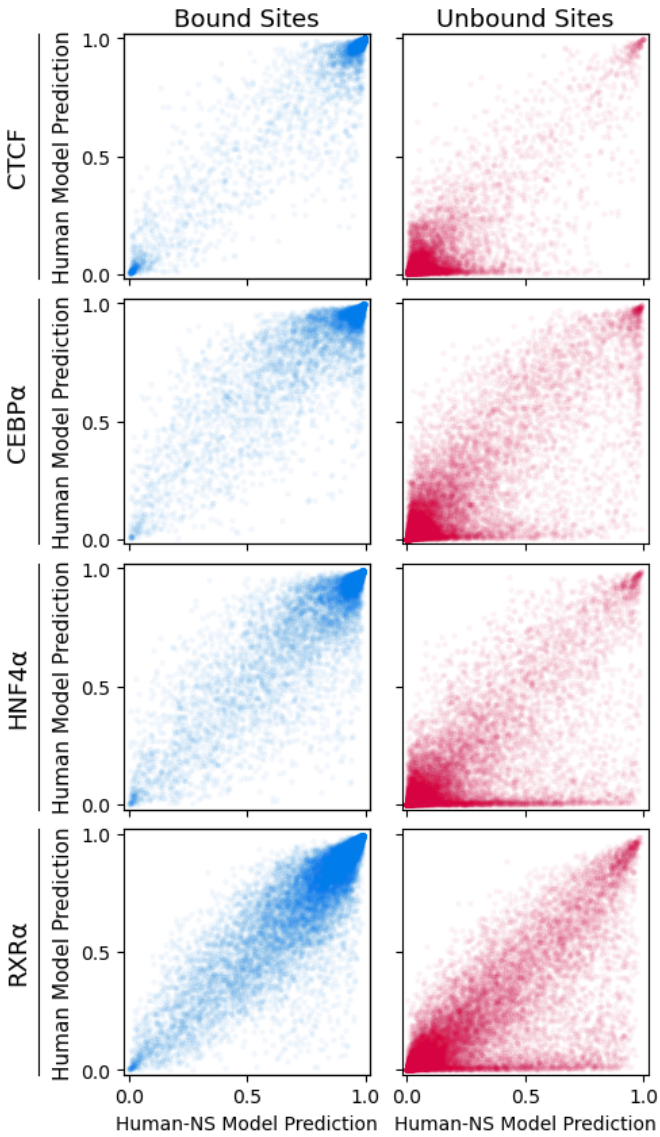

Figure 7: Differential human chromosome 2 site predictions between models trained on human data with or without any examples of SINE windows. Human-NS: models trained on human data with no SINE examples. Similar to mouse-trained models, no-SINE human-trained models systematically mispredict some unbound sites.

prediction mistakes as mouse-trained models (Figure 7). For bound sites, on the other hand, no-SINE human-trained models make predictions that generally agree with predictions from standard human-trained models.

This suggests that the Alu false positives unique to the mouse-trained model may simply be due to the fact that mouse models are not exposed to Alus during training (i.e., Alu elements are "out of distribution"). In addition, the reduction in model-unique false negatives ob-

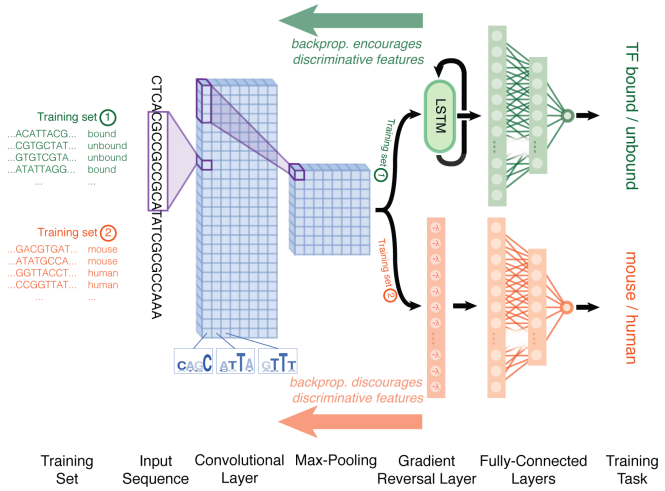

Figure 8: Domain-adaptive network architecture. The top network output predicts TF binding, as before, while the bottom network output predicts the species of origin of the input sequence window. The gradient reversal layer has the effect of discouraging the convolutional filters before it from learning sequence features relevant to the species prediction task.

served when the no-SINE human-trained model is compared to the normal human-trained model suggests that those mispredictions are unrelated to Alus.

Domain-adaptive mouse models can improve cross-species performance

Having observed an apparent "domain shift" across species, partially attributable to species-unique repeats, our next step is to ask how we might bridge this gap and reduce the difference in cross-species model performance. Our problem is analogous to one encountered in some image classification tasks, where the test data is differently distributed from the training data to the extent that the model performs well on training data but much worse on test data (for example, the training images were taken during the day but the test images were taken at sunset). In these situations, various techniques for explicitly forcing the model to adapt across different image "domains" have been shown to improve performance at test time (e.g. Long et al. 2015; Bousmalis et al. 2016; Sun et al. 2016).

One unsupervised domain adaptation method utilizes a gradient reversal layer to encourage the "feature generator" portion of a neural network to be domain-generic (Ganin et al. 2016). The gradient reversal layer's effect is to backpropagate a loss to the feature generator that prevents any domain-unique features from being learned. We chose to test the effectiveness of this version of domain adaptation for our cross-species TF binding prediction problem because we have observed evidence that domain-unique features (species-unique repeat elements) were a major component of the cross-species domain shift.

We modified our existing model architecture to perform training-integrated domain adaptation across species (Figure 8). A gradient reversal layer (GRL) was added in parallel with the LSTM, taking in the result of the max-pooling step (after the convolutional layer) as input. During standard feed-forward prediction, the GRL merely computes the identity of its input, but as the loss gradient backpropagates through the GRL, it is reversed. The output of the GRL then passes through two fully connected layers before reaching a new, secondary output neuron. This secondary output, a "species discriminator," is tasked with predicting whether the model's input genomic window is from the source or target species. The model training process is modified so that the model is exposed to sequences from both species, but only the binding labels of the source species (see Methods). Without the GRL, adding the species discrimination task to the model would encourage the convolutional filters to learn sequence features that best differentiate between the two species – features like species-unique repeats – but with the GRL included, the convolutional filters are instead *discouraged* from learning these features. We hypothesize that this domain-adaptive model will outperform our basic model architecture by reducing mispredictions on species-unique repeats.

We trained domain-adaptive models using the same binding training datasets as before and evaluated performance with the same held-out datasets. We observe that the auPRC for our domain-adaptive models on cross-species test data is moderately higher than the auPRC for the basic mouse models, for all TFs except CTCF, where auPRCs are merely equal (Figure 9, top, blue vs. green boxplots). The domain-adaptive models' auPRCs on mouse test data, meanwhile, is comparable to the auPRCs of basic models (Figure 9, bottom, blue vs. green). While the auPRC improvement is promising, it is also modest in comparison to the full cross-species gap; the domain-adaptive models still do not achieve a level of performance comparable to same-species models (Figure 9, top, green vs. red).

Domain-adaptive mouse models reduce over-prediction on Alu elements

Next, we repeated our site-distribution analysis to determine what constituted the domain-adaptive models' improved performance. The unbound site plots in Fig-

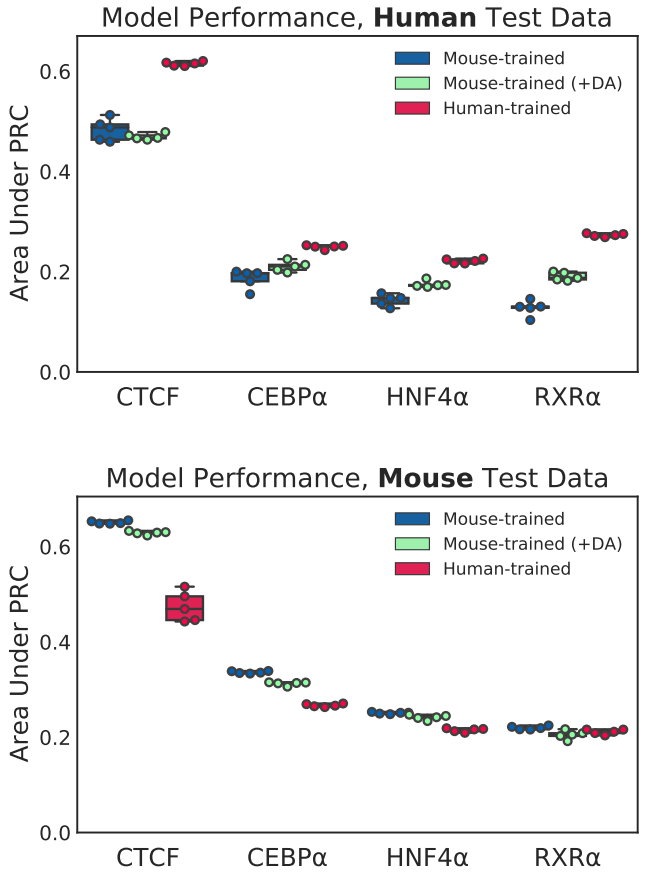

Figure 9: Performance of mouse-trained generic (blue), mouse-trained domain-adaptive (green), and human-trained (red) models, evaluated on human (left) and mouse (right) chromosome 2. Five models were independently trained and evaluated for each TF and training species.

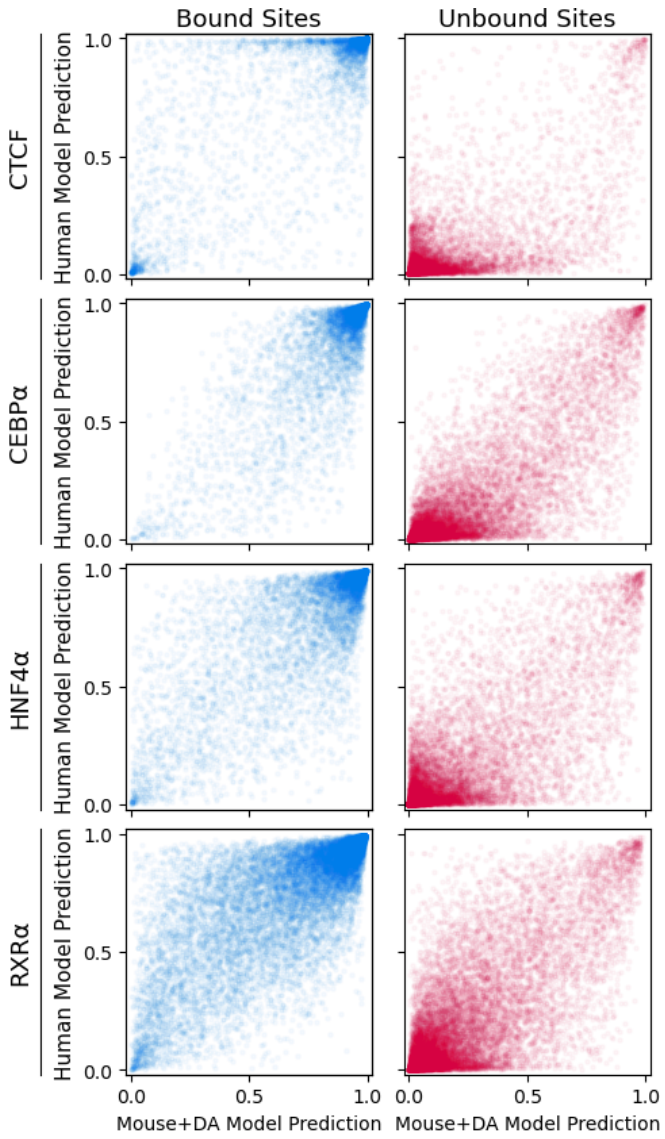

Figure 10: Differential predictions of human genome sites between human-trained and domain-adaptive mouse-trained models. Domain-adaptive mouse models, unlike the original mouse models, do not show species-specific systematic misprediction of unbound sites.

ure 10 compare human genome predictions between domain-adaptive mouse models and the original human models. Alu elements are highlighted in Figure 11.

Compared to Figure 4, the mouse-model-specific false positives have diminished for all TFs. This suggests that the domain-adaptive models are able to correct the problem of false positive predictions from Alus by scoring unbound sites overlapping Alus lower than the basic model did. This effect is even present for CTCF, even though there was no noticeable auPRC difference for CTCF between domain-adaptive and basic mouse models – likely because the initial Alu enrichment in CTCF mouse-model false positives was lower than for other TFs.

In contrast, the site-distribution plots for bound sites demonstrate no noticeable difference from the original plots for the basic model architecture. We applied the same SeqUnwinder analysis to look for sequence features that discriminate between mouse-model false negatives and true positives and discovered similar, but not identical, motif-like short sequence patterns as we did previously (Supplementary Figure 2). Thus, domain adaptation does not appear to have any major influence on bound site predictions.

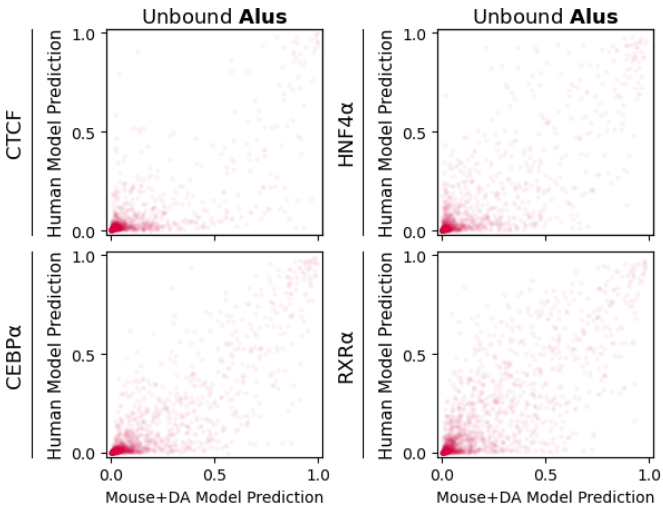

Figure 11: Differential predictions of unbound sites containing Alu elements between domain-adaptive mouse-trained models and human-trained models. Unlike the original mouse models, domain-adaptive mouse models do not show systematic overprediction of Alu repeats.

Discussion

Enabling effective cross-species TF binding imputation strategies would be transformative for studying mammalian regulatory systems. For instance, TF binding information could be transferred from model organisms in cell types and developmental stages that are difficult or unethical to assay in humans. Similarly, one could annotate regulatory sites in non-model species of agricultural or evolutionary interest by leveraging the substantial investment that has been made to profile TF binding sites in human, mouse, and other model organisms (ENCODE Project Consortium 2012; Yue et al. 2014; Roadmap Epigenomics Consortium et al. 2015).

Our results suggest that cross-species TF binding imputation is feasible, but we also find a pervasive performance gap between within-species and cross-species prediction tasks. One set of culprits for this cross-species performance gap are species-specific transposable elements. For example, models trained using mouse TF binding data have never seen an Alu SINE element during training, and often falsely predict that these elements are bound by the relevant TF. Since Alu elements appear at high frequency in the human genome, their misprediction constitutes a large proportion of the cross-species false positive predictions, and thereby substantially affect the genome-wide performance metrics of the model. It should be noted that Alus and other transposable elements can serve as true regulatory elements (Bourque et al. 2008; Sundaram et al. 2014), and thus we don't assume that all transposable elements should be labeled as TF "unbound". Indeed, we minimized the potential mislabeling of truly bound transposable elements as "unbound" by focusing all our analyses on regions of the genome that have a high degree of mappability (and are thereby less likely to be subject to mappability-related false negative labeling issues in the TF ChIP-seq data).

We demonstrated that a simple domain adaptation approach is sufficient to correct the systematic mispredictions of Alu elements as TF bound. Training a parallel task (discriminating between species) but with gradient reversal employed during backpropagation has the effect of discouraging species-specific features being learned by the shared convolutional layers of the network. This approach is straightforward to implement and has the advantage that TF binding labels need only be known in the training species. Our approach accounts for domain shifts in the underlying genome sequence composition, assuming that the general features of TF binding sites are conserved within the same cell types across species.

We note that the underlying assumption of cross-species TF binding prediction - i.e., that the overall features of cell-specific TF binding sites are conserved - may not hold true in all cases. We observe that there are sequence features in bound sites that discriminate between correct and incorrect predictions specific to cross-species models. These discriminative sequence features suggest that cross-species false negative prediction errors could be the result of differential TF activity across the two species. Such differential activities could result from gain or loss of TF expression patterns, non-conserved cooperative binding capabilities, or evolved sequence preferences of the TF itself. We observe that these discriminative features are often preserved after we apply sequence composition domain adaptation, suggesting that our approach does not address the situation where TF binding logic is not fully conserved across species.

Other recent work has also demonstrated the feasibility of cross-species regulatory imputation. For example, Chen, et al. assessed the abilities of support vector machines (SVMs) and CNNs to predict potential enhancers (defined by combinations of histone marks) when trained and tested across species of varying evolutionary distances (Chen et al. 2018). Interestingly, they observed that while CNNs outperform SVMs in within-species enhancer prediction tasks, they are worse at generalizing across species. Our work suggests a possible reason for, and a solution to, this generalization gap. Two other recent manuscripts have applied more complex neural network architectures to impute TF binding and other regulatory signals across species (Kelley 2019;

Schreiber et al. 2020b). Those studies focus on models that are trained jointly across thousands of mouse and human regulatory genomic datasets. They thus assume that substantial amounts of regulatory information has already been characterized in the target species, which may not be true in some desired cross-species imputation settings. In general, however, joint modeling approaches are also likely to benefit from domain adaptation strategies that account for species-specific differences in sequence composition, and our results are thus complementary to these recent reports.

In summary, our work suggests that cross-species TF binding prediction approaches should beware of systematic differences between the compositions of training and test species genomes, including species-specific repetitive elements. Our contribution also suggests that domain adaptation is a promising strategy for addressing such differences and thereby making cross-species predictions more robust. Further work is needed to characterize additional sources of the cross-species performance gap and to generalize domain adaptation approaches to scenarios where training data is available from multiple species.

Methods

Data processing

Datasets were constructed by splitting the mouse (mm10) and human (hg38) genomes into 500 bp windows, offset by 50 bp. Any windows overlapping ENCODE blacklist regions were removed (Amemiya et al. 2019). We then calculated the fraction of each window that was uniquely mappable by 36 bp sequencing reads and retained only the windows that were at least 80% uniquely mappable (Karimzadeh et al. 2018). Mappability filtering was performed to remove potential peak-calling false negatives; otherwise, any genomic window too unmappable for confident peak-calling would be a potential false negative.

Liver ChIP-seq experiments were collected from ENCODE, GEO, and ArrayExpress. Accession IDs are as follows: GSE105829 for human CTCF, ENCSR000CBU for mouse CTCF, E-TABM-722 (PMID 20378774) for CEBPα and HNF4α (in both human and mouse), ENCSR098XMN for human RXRα, and GSM1299600 for mouse RXRα. Corresponding control experiments were utilized during peak calling when available.

ChIP-seq peaks were called using MultiGPS v0.74 with default parameters, excluding ENCODE blacklist regions (Mahony et al. 2014; Amemiya et al. 2019). Peak calls were converted to binary labels for each window in a genome: "bound" (1) if any peak center fell within the window, "unbound" (0) otherwise.

Dataset splits for training and testing

Chromosomes 1 and 2 of both species were held out from all training datasets. For computational efficiency, one million randomly selected windows from chromosome 1 were used as the validation set for each species (for hyperparameter tuning). All windows from chromosome 2 were used as the test sets.

TF binding task training data was constructed identically for all model architectures. Since binary classifier neural networks often perform best when the classes are balanced in the training data, the binding task training dataset consisted of all bound examples and an equal number of randomly sampled (without replacement) unbound examples, excluding examples from chromosomes 1 and 2. To increase the diversity of examples seen by the network across training, in each epoch a distinct random set of unbound examples was used, with no repeated unbound examples across epochs.

Domain-adaptive models also require an additional "species-background" training set from both species for the species discrimination task. Species-background data consisted of randomly selected (without replacement) examples from all chromosomes except 1 and 2. Binding labels were **not** used in the construction of these training sets. In each batch, the species-background examples were balanced, with 50% human and 50% mouse examples, and labeled according to their species of origin (not by binding). The total number of species-background examples in each batch was double the number of binding examples.

Basic model architecture

The network takes in a one-hot encoded 500 bp window of DNA sequence and passes it through a convolutional layer with 240 20-bp filters, followed by a ReLU activation and max-pooling (pool window and stride of 15 bp). After the convolutional layer is an LSTM with 32 internal nodes, followed by a 1024-neuron fully-connected layer with ReLU activation, followed by a 50% Dropout layer, followed by a 512-neuron fully-connected layer with sigmoid activation. The final layer is a single sigmoid-activated neuron.

Domain-adaptive model architecture

The domain-adaptive network builds upon the basic model described above by adding a new "species discriminator" task. The network splits into two output halves following max-pooling after the convolutional

layer. The max-pooling output feeds into a gradient reversal layer (GRL) – the GRL merely outputs the identity of its input during the feed-forward step of model training, but during backpropagation, it multiplies the gradient of the loss by −1. The GRL is followed by a Flatten layer, a ReLU-activated fully connected layer with 1024 neurons, a sigmoid-activated fully connected layer of 512 neurons, and finally a single-neuron layer with sigmoid activation.

Model training

All models were trained with Keras v2.3.1 using the Adam optimizer with default parameters (Chollet 2015; Kingma and Ba 2014). Training ran for 15 epochs, with models saved after each epoch. After training, we selected models for downstream analysis by choosing the saved model with highest auPRC on the training-species validation set.

The basic models were trained by standard procedure with a batch size of 400 (see Section 2.1.2 for training dataset construction). The domain-adaptive models, on the other hand, required a more complex batching setup. Because domain-adaptive models predict two tasks – binding and the species of origin of the input sequence – they require two stages of dataset input per batch. The first stage is identical to a basic model training batch, but with $\lfloor 400/3 \rfloor = 133$ binding examples from the source species. The second stage uses $\lceil 400 * 2/3 \rceil = 267$ examples each from the source species' and target species' "species-background" datasets.

Crucially, the stages differ in how task labels are masked. For each stage, only one of the two output halves of the network trains (the loss backpropagates from one output only). In the first stage, we mask the species discriminator task, so that only the binding task half of the model trains on binding examples from the training species. In the second stage, we mask the binding task, so only the species discriminator task half trains. Thus, the binding task only trains on examples from the source species, while the species discriminator task doesn't see binding labels from either species.

Meanwhile, the weights of the shared convolutional layer are influenced by both tasks. Because these stages occur within a single batch and not in alternating batches, they concurrently influence the weights of the convolutional filters; there is no oscillating "back-and-forth" between the two tasks from batch to batch.

Differentially-predicted site categorization

To quantify site enrichment within discrete categories such as "false positives" and "false negatives", it was necessary to define the boundaries for these labels. In particular, when comparing prediction distributions between models, we needed to define what constitutes, for instance, a "false positive unique to model A." We constructed the following rules for site categorization: 1) unbound sites must have predictions above 0.5 to be labeled false positives, and bound sites must have predictions below 0.5 to be labeled false negatives; 2) a site is considered to be differentially predicted between two source species A and B if $|P_A - P_B| > 0.5$, where P_A and P_B are the predictions from models trained on data from species A and species B, respectively; 3) only sites meeting this differential prediction threshold are labeled as a false positive or negative unique to one model. Thus, if we are comparing models from species A and B, and a site is labeled a false positive unique to model A, then $P_A > 0.5$ and $P_B < 0.5$. To reduce noise in these categorizations, rather than letting P_A and P_B equal the predictions from single models, we trained 5 independent replicate models for each TF and source species, and then let P_A be the average prediction across the 5 replicate models trained on data from species A for a given TF.

Bound site discriminatory motif discovery

SeqUnwinder (v. 0.1.3) (Kakumanu et al. 2017) was used to find motifs that discriminate between true positive predictions and mouse-model-specific false negative predictions using the following command-line settings: "--threads 10 --makerandregs --makerandregs --win 500 --mink 4 --maxk 5 --r 10 --x 3 --a 400 --hillsthresh 0.1 --memesearchwin 16", and using MEME v. 5.1.0 (Machanick and Bailey 2011) internally.

Repeat analysis

All repeat analysis used the RepeatMasker track from the UCSC Genome Browser (Smit et al. 1996). Genome windows were labeled as containing an Alu element if there was any overlap (1 or more bp) with any Alu annotation. For Supplementary Table 1, repeat classes were excluded if fewer than 500 examples of that class were annotated in the test chromosome (before mappability filtering).

Availability

Open source code (MIT license) is available from:
`https://github.com/seqcode/cross-species-domain-adaptation`

Acknowledgements

The authors thank the members of the Center for Eukaryotic Gene Regulation at Penn State for helpful feedback and discussion.

Funding

This work was supported by NIH NIGMS grant R01-GM121613 (to SM), NIH grant DP2GM123485 (to AK) and the Stanford Graduate Fellowship (to KC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**: 9354.

Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A, et al. 2020. Base-resolution models of transcription factor binding reveal soft motif syntax. *BioRxiv* **13**:

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3**: 370–379.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research* **18**: 1752–1762.

Bousmalis K, Silberman N, Dohan D, Erhan D, Krishnan D. 2017. Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 95–104.

Chen L, Fish AE, Capra JA. 2018. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLoS Comput Biol* **14**: e1006484.

Chollet F et al. 2015. Keras. `https://keras.io`.

Ferrari R, Llobet Cucalon LI de, Di Vona C, Le Dilly F, Vidal E, Lioutas A, Oliete JQ, Jochem L, Cutts E, Dieci G, et al. 2019. TFIIIC Binding to Alu Elements Controls Gene Expression via Chromatin Looping and Histone Acetylation. *Molecular Cell* **77**: 475–487.

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research* **17**: 1–35.

Huh I, Mendizabal I, Park T, Yi SV. 2018. Functional conservation of sequence determinants at rapidly evolving regulatory regions across mammals. *PLoS Comput Biol* **14**: e1006451.

Kakumanu A, Velasco S, Mazzoni E, Mahony S. 2017. Deconvolving sequence features that discriminate between overlapping regulatory annotations. *PLoS Comput Biol* **13**: e1005795.

Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. 2018. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Research* **46**: e120.

Keilwagen J, Posch S, Grau J. 2019. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol* **20**: 9.

Kelley DR. 2020. Cross-species regulatory sequence activity prediction. *PLoS Comput Biol* **16**: e1008050.

Kingma DP, Ba J. 2014. Adam: A Method for Stochastic Optimization. arXiv: `1412.6980 [cs.LG]`.

Li H, Quang D, Guan Y. 2019. Anchor: trans-cell type prediction of transcription factor binding sites. *Genome Res.* **29**: 281–292.

Long M, Cao Y, Wang J, Jordan MI. 2015. Learning Transferable Features with Deep Adaptation Networks. arXiv: `1502.02791 [cs]`.

Machanick P, Bailey TL. 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**: 1696–1697.

Mahony S, Edwards MD, Mazzoni EO, Sherwood RI, Kakumanu A, Morrison CA, Wichterle H, Gifford DK. 2014. An Integrated Model of Multiple-Condition ChIP-Seq Data Reveals Predeterminants of Cdx2 Binding. *PLoS Comput Biol* **10**: e1003501.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**: 669–680.

Polak P, Domany E. 2006. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* **7**: 133.

Quang D, Xie X. 2019. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods* **166**: 40–47.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

Savic D, Partridge EC, Newberry KM, Smith SB, Meadows SK, Roberts BS, Mackiewicz M, Mendenhall EM, Myers RM. 2015. CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins. *Genome Res* **25**: 1581–1589.

Schmid CW. 1998. Does SINE evolution preclude Alu function? *Nucleic Acids Research* **26**: 4541–4550.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**: 1036–1040.

Schreiber J, Durham T, Bilmes J, Noble WS. 2020. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biol* **21**: 81.

Schreiber J, Hegde D, Noble W. 2019. Zero-shot imputations across species are enabled through joint modeling of human and mouse epigenomics. *bioRxiv*. DOI: 10.1101/801183.

Siggers T, Gordan R. 2014. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Research* **42**: 2099–2111.

Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordan R, Rohs R. 2014. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences* **39**: 381–399.

Smit A, Hubley R, Green P. 1996-2010. RepeatMasker Open-3.0. http://www.repeatmasker.org.

Srivastava D, Aydin B, Mazzoni EO, Mahony S. 2020. An interpretable bimodal neural network characterizes the sequence and preexisting chromatin predictors of induced TF binding. *bioRxiv*. DOI: 10.1101/672790.

Srivastava D, Mahony S. 2020. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1863**: 194443.

Sun B, Feng J, Saenko K. 2016. Correlation Alignment for Unsupervised Domain Adaptation. arXiv: 1612.01939 [cs].

Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Ressearch* **24**: 1963–1976.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

The Mouse ENCODE Consortium, Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**: 355–364.

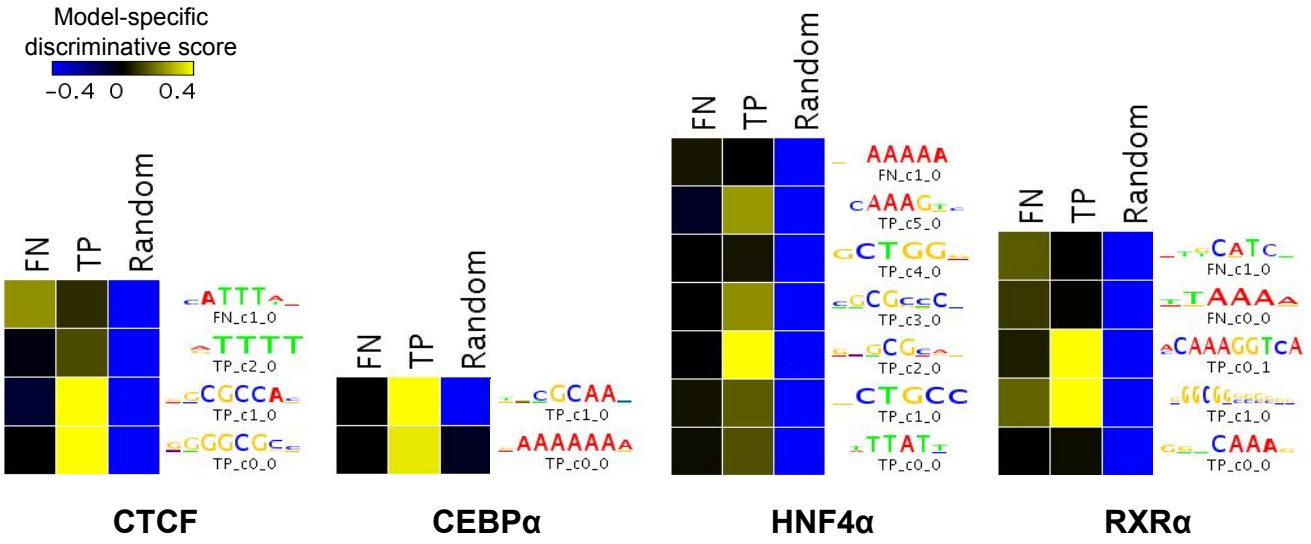

Figure S1: Motif-like sequence features can discriminate between human-genome bound sites correctly predicted by mouse-trained and human-trained models (true positives or TP) and bound sites correctly predicted only by human-trained models (mouse-specific false negatives or FN) for each TF. See Methods for site categorization details.

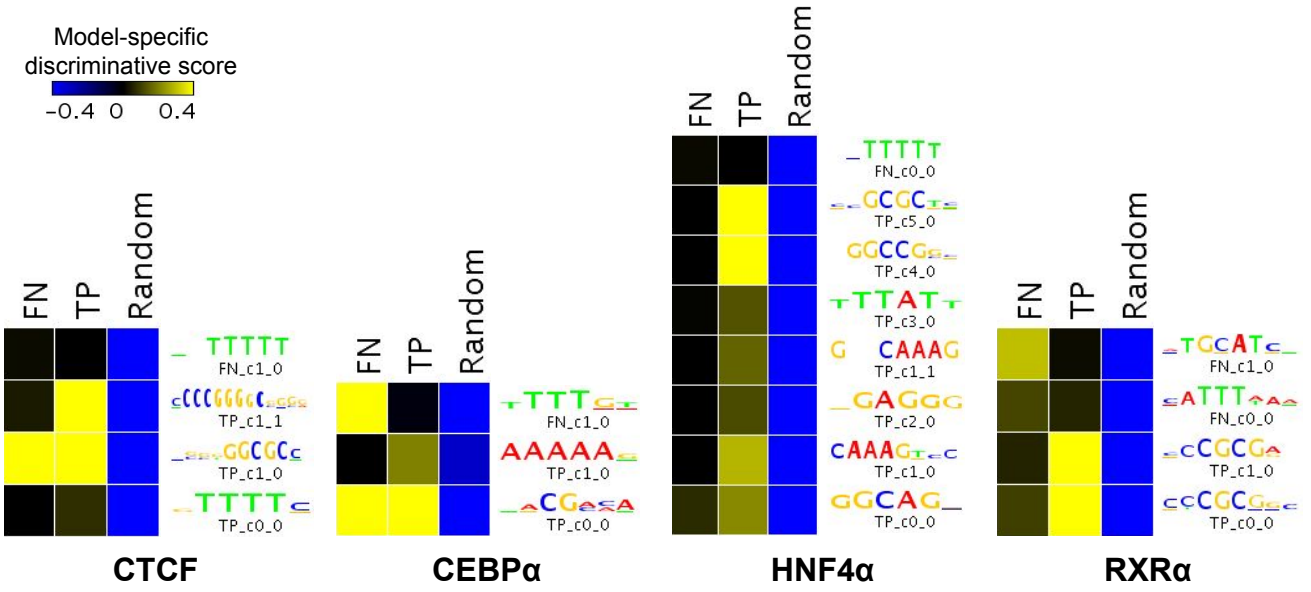

Figure S2: False negative predictions unique to mouse-trained models trained with domain adaptation, compared to human-trained models, can be distinguished from true positive predictions through motif-like sequence features. See Methods for site categorization details.

DNA

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
CTCF	10.1%	11.6%	12.0%	11.4%	7.2%	7.9%
CEBPα	12.3%	11.0%	0.0%	11.4%	13.0%	9.5%
HNF4α	10.8%	11.4%	11.6%	11.4%	9.8%	9.0%
RXRα	10.1%	11.1%	11.0%	11.4%	10.2%	9.4%

LINE

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
CTCF	18.0%	22.2%	24.9%	37.6%	14.6%	24.8%
CEBPα	25.5%	25.7%	0.0%	37.6%	29.0%	32.3%
HNF4α	21.0%	25.4%	30.7%	37.6%	21.8%	29.7%
RXRα	20.8%	28.3%	21.4%	37.8%	21.8%	33.1%

Low-complexity

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
CTCF	2.8%	2.2%	3.0%	1.9%	4.5%	1.5%
CEBPα	1.5%	1.6%	0.0%	1.9%	1.6%	1.4%
HNF4α	2.0%	1.4%	1.5%	1.9%	2.4%	1.5%
RXRα	2.0%	1.4%	1.7%	1.9%	2.2%	1.5%

LTR

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
CTCF	8.7%	12.6%	7.1%	17.6%	13.5%	13.6%
CEBPα	12.8%	11.1%	0.0%	17.6%	19.0%	14.6%
HNF4α	13.3%	14.9%	11.4%	17.6%	19.5%	13.9%
RXRα	11.8%	13.9%	10.9%	17.6%	18.0%	11.7%

Simple Repeat

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
CTCF	13.9%	11.2%	8.1%	11.5%	17.1%	10.6%
CEBPα	9.4%	9.6%	40.0%	11.6%	9.9%	11.7%
HNF4α	12.4%	12.0%	11.2%	11.5%	11.5%	12.1%
RXRα	11.5%	9.7%	14.4%	11.5%	11.1%	13.4%

SINE

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
CTCF	23.5%	25.1%	20.4%	31.2%	16.8%	68.9%
CEBPα	30.8%	22.5%	0.0%	31.1%	35.5%	84.8%
HNF4α	27.1%	21.8%	20.5%	31.2%	28.6%	88.7%
RXRα	26.9%	24.2%	17.5%	31.2%	31.4%	97.1%

Unknown

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
CTCF	0.2%	0.0%	0.2%	0.2%	0.1%	0.0%
CEBPα	0.3%	0.5%	0.0%	0.2%	0.2%	0.1%
HNF4α	0.2%	0.1%	0.2%	0.2%	0.2%	0.0%
RXRα	0.2%	0.2%	0.1%	0.1%	0.2%	0.0%

Table S1: Percent of windows overlapping various RepeatMasker-defined repeat elements, for different categories of genomic windows from the held-out test set. Only RepeatMasker repeat classes with at least 500 distinct annotations within the test set are shown. FPs: false positives. FNs: false negatives. Mouse Only: specific to mouse-trained models. See Methods for more details on site categorization.