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

Kelly Cochran, Divyanshi Srivastava, Avanti Shrikumar, et al.

Genome Res. published online January 18, 2022
Access the most recent version at doi:10.1101/gr.275394.121

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] Ross C. Hardison,[1,2] 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).*

Abstract

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.

Introduction

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 CEBPA and HNF4A 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 CEBPA and HNF4A 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 2020; Schreiber et al. 2020; 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, CEBPA, HNF4A, and RXRA) in liver due to the existence of high quality ChIP-seq data in

both mouse and human. We proceed to investigate gaps in performance between within-species and cross-species models, with the aim of identifying specific genomic patterns that are associated with systematic misprediction specifically across species.

We further evaluate the model performance improvement gained from integrating an unsupervised domain adaptation approach into model training. This 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 aims to encourage features in the shared convolutional layer that discriminate between bound and unbound sites, while simultaneously discouraging features that are species-specific. Importantly, this approach does not use TF binding labels from the target species at any stage in training. We conclude by assessing the effectiveness of domain adaptation in terms of reducing systematic mispredictions.

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. 2021a; Avsec et al. 2021b; Fudenberg et al. 2020; Kelley 2018; Koo et al. 2021). In particular, hybrid convolutional and recurrent network architectures have successfully been applied to accurately predict TF binding in diverse applications (Quang and Xie 2016; 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).

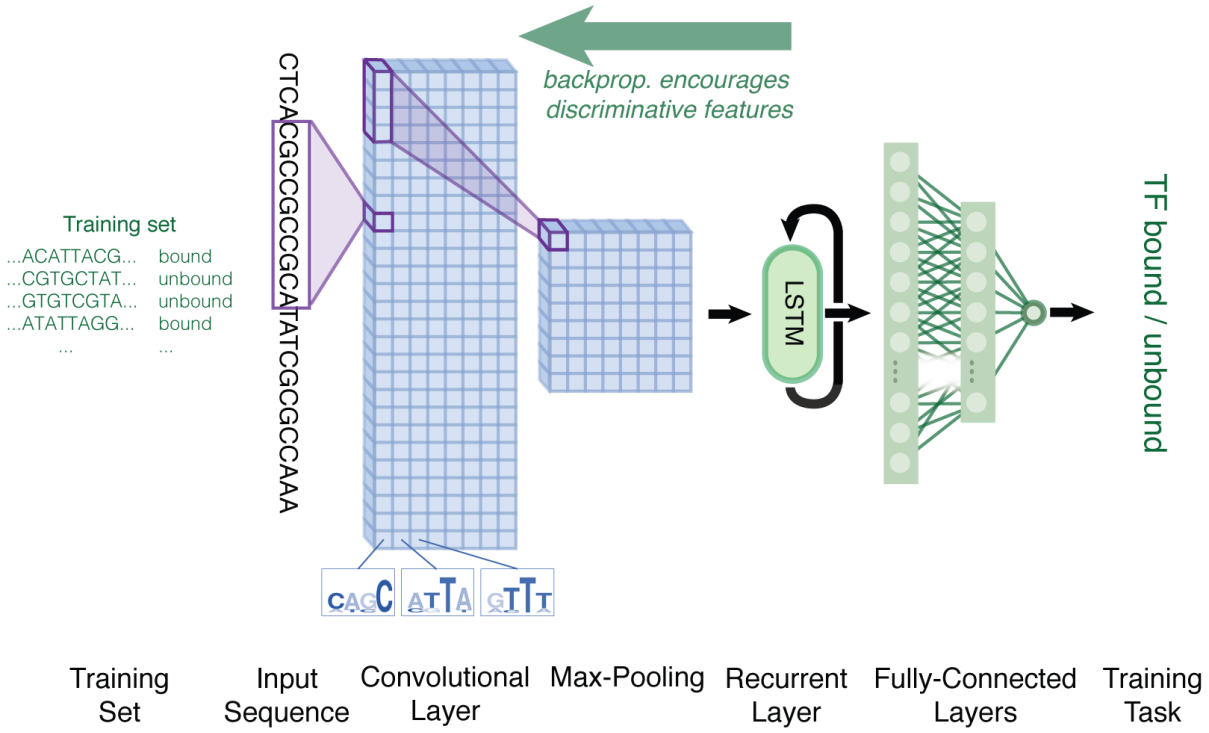

Figure 1: Conventional network architecture. Convolutional filters scan the 500-bp input DNA sequence for TF binding features. The convolutional layer is followed by a recurrent layer (LSTM) and two fully connected layers. A final sigmoid-activated neuron predicts if a ChIP-seq peak falls within the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

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 from epoch to epoch did not always translate to improved auPRC in the target species. Generally, cross-species auPRCs showed greater variability across epochs and model replicates compared to source-species auPRCs. For HNF4A in particular, 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 red 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 an ideal surrogate for validation set performance in the target 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. Because we observe a wider cross-species gap for mouse-trained models predicting in human than for human-trained models predicting in mouse, subsequent analysis focuses on addressing the mouse-to-human gap.

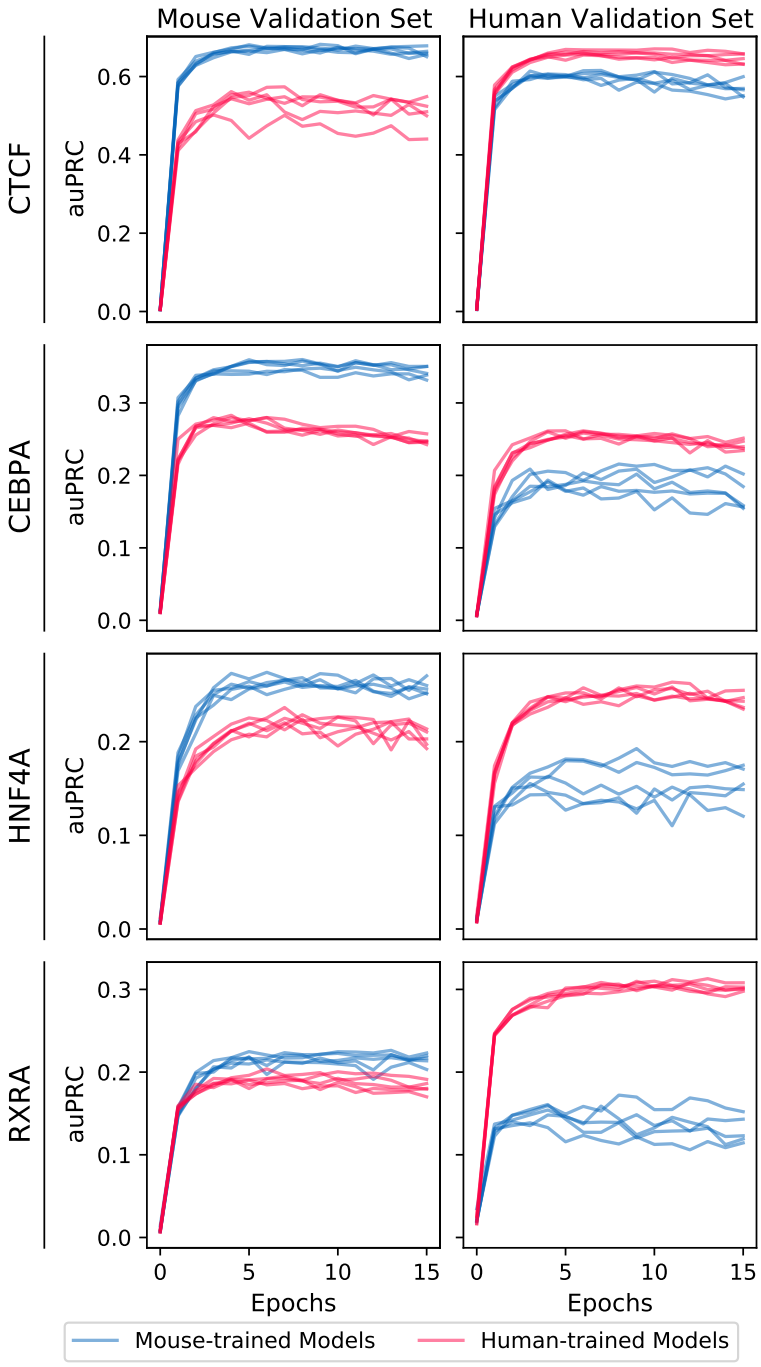

Figure 2: Model performance over the course of training, evaluated on held-out validation data from mouse (left) and human (right) Chromosome 1. Five models were independently trained for each TF and source species (mouse-trained models in blue, human-trained models in red). Values at epoch 0 are evaluations of models after weight initialization but before training (akin to a random baseline). Note that auPRCs are not directly comparable between different validation sets because ground truth labels are derived from a different experiment for each dataset; the auPRC will depend on the fraction of sites labeled bound as well as model prediction correctness.

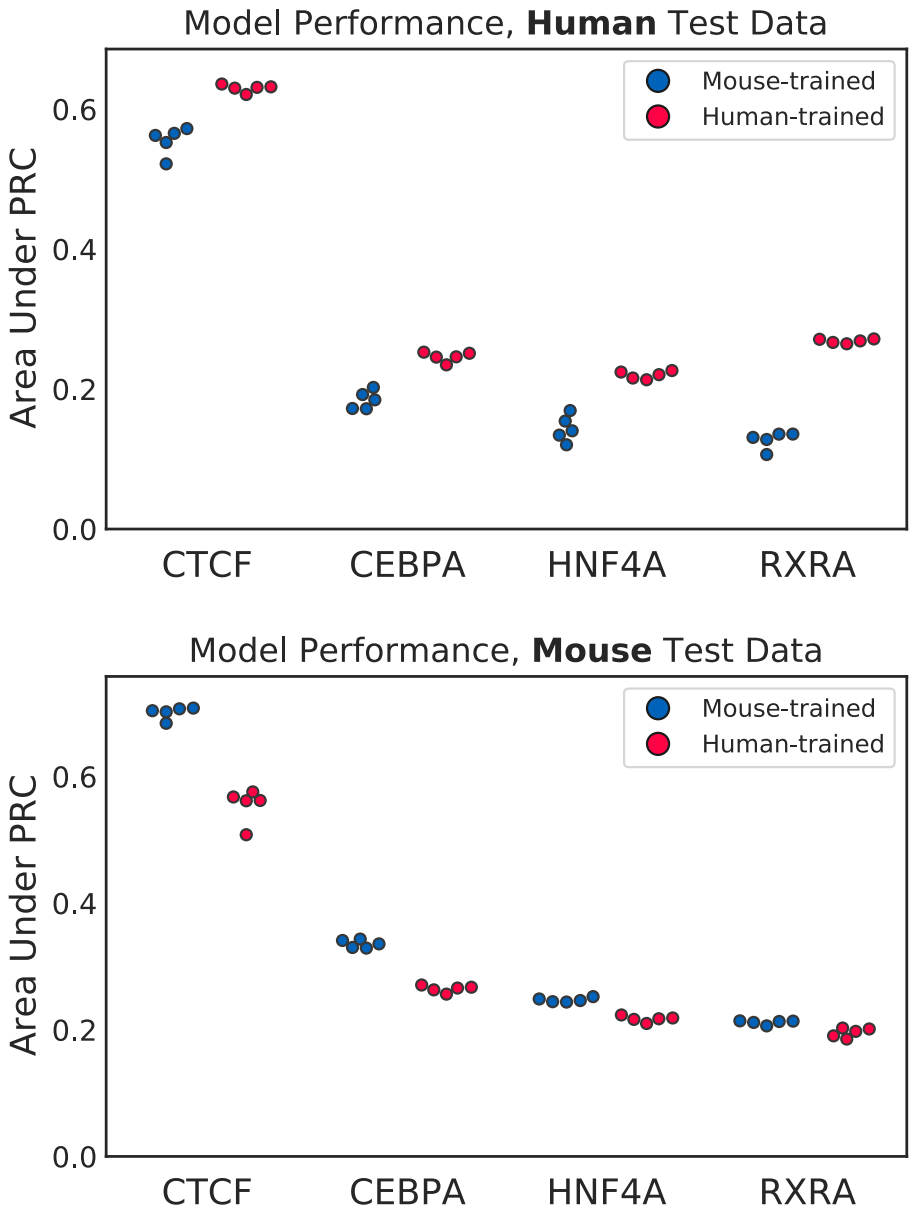

Figure 3: Model performance evaluated on held-out test data: Chromosome 2 from human (top) and mouse (bottom). Five models were independently trained for each TF and source species.

To get a sense of how specific to our model design or training strategy this cross-species gap might be, we applied multiple sufficiently different machine learning approaches to the same problem and datasets and assessed whether the cross-species gap persists. First, we trained gapped *k*-mer support vector machines, or gkSVMs, to classify a balanced sample of bound vs. unbound windows for each TF and species (Ghandi et al. 2014; Lee 2016). We then evaluated those models on the set of non-overlapping windows in each test dataset (Supplemental Fig. S1). We observe that the cross-species gap persists, although it shrinks in absolute magnitude, presumably due to the drastically lower auPRC values across the board. These auPRCs also demonstrate that our neural network approach can indeed outperform related methods on this task.

Next, we sought to assess the cross-species performance of another state-of-the-art deep learning model trained on a related TF binding prediction task, distinct from our binary classification setup. We applied a BPNet-like profile model, which predicts the distribution of the raw, base-resolution ChIP-seq read profile at a given genomic window rather than a 0-1 binary label, to both our mouse and human datasets across our four TFs (Avsec et al. 2021b). The profile models were trained using a peak-enriched subset of the training data used by the binary models, and performance was evaluated on the same test datasets (see Methods).

First, we investigated how well individual profile predictions transfer across species (Supplemental Fig. S2, bottom). We observe that overall, within-species profile models are usually able to predict both the location and the shape of peaks accurately. Cross-species profile models tend to predict the peak location nearly as well as within-species models, but for some TFs, there is a clear discrepancy between the predicted and true profile shape. Specifically, there are apparent non-biological differences in experimental protocol or quality between our matched datasets across species; this can cause profile models that learned how reads typically distribute around binding sites from one experiment to appear to generalize imperfectly to other datasets with different read distributions about binding sites.

Next, we quantified the performance of the profile models, using the predicted total number of reads across a genomic window as a proxy for binary label prediction (Supplemental Fig. S2, top). We again observe cross-species performance gaps for most datasets. We also note that the auPRC values attained by the profile models are comparable to those attained by our conventional model in most cases, so we decided to focus on understanding the cross-species gap in the context of the conventional model in the remainder of the study.

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 three of the four 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. Even for CTCF, the exception to the pattern, there is an enrichment of unbound sites that can be characterized as mispredictions specific to mouse models. Additionally, in the bound site plots of all TFs except CEBPA, 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 site, although the errors in the unbound sites appear more prevalent than the errors in the bound sites.

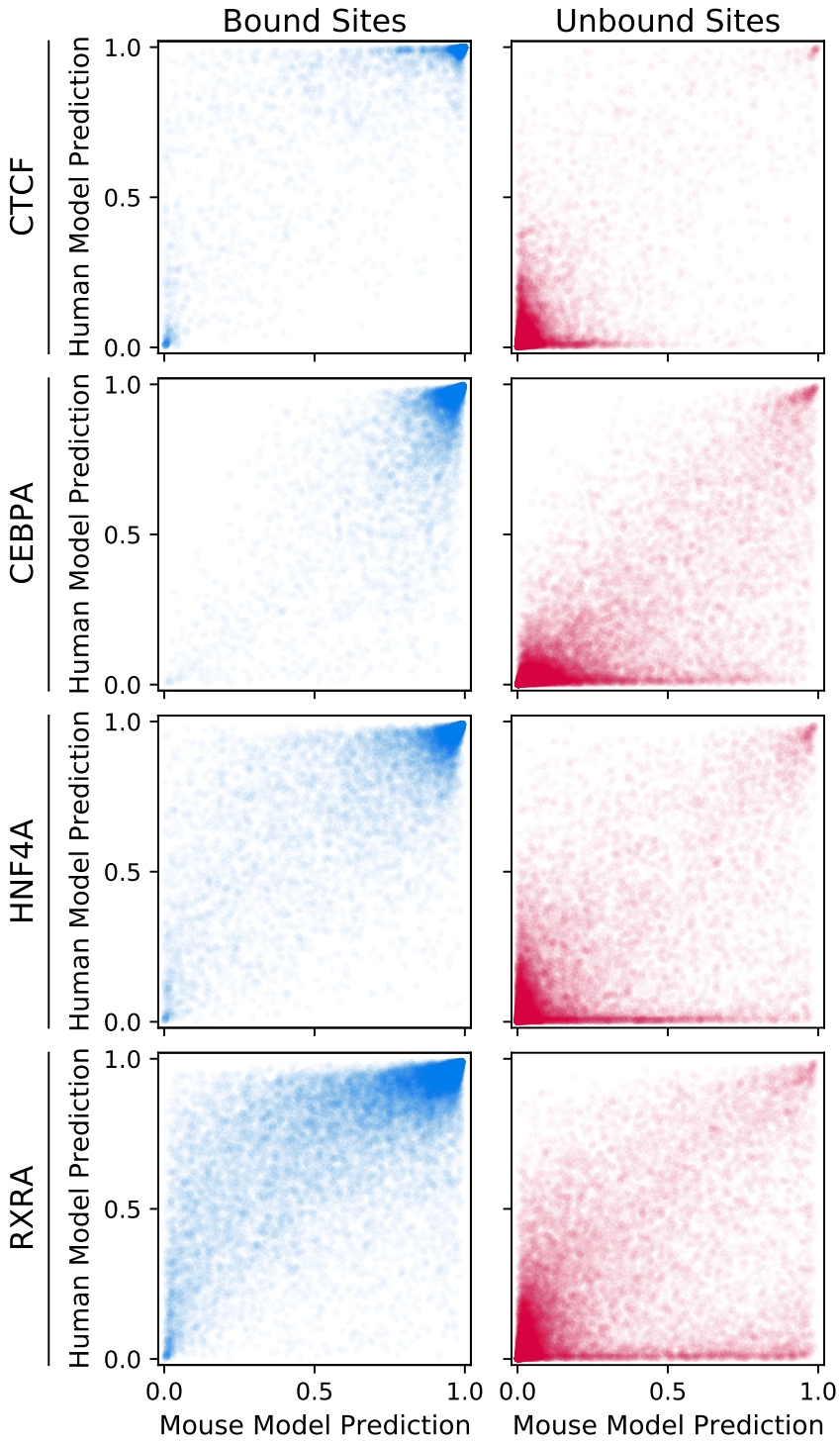

Figure 4: Both bound and unbound sites from human Chromosome 2 show evidence of differential binding predictions by human-trained (y-axis) vs. mouse-trained (x-axis) models. For visual clarity, only 25% of bound sites and 5% of unbound sites are shown (sampled systematically).

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 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 features associated with 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 0.78, 0.79, 0.80, and 0.87 for CTCF, CEBPA, HNF4A, and RXRA, respectively. Supplemental Fig. S3 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 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 mispredicted only by the mouse model. On average, 89% 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 18% for unbound sites incorrectly predicted by both models. In contrast, *Alus* on average only overlap 6% of false negatives unique to the mouse model, which is less than the overlap fraction for bound sites overall (15%) and for false negatives common to both models (11%). 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 (Supplemental Table S1). Investigating the enrichment of individual *Alu* subfamilies in mouse-model-unique false positives showed that this phenomenon is not restricted to a single subtype of *Alu*, but that subfamilies are enriched at different levels in a manner that is TF-specific and varies particularly between the *AluJ*, *AluS*, and *AluY* subfamily groupings (Supplemental Fig. S4).

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.

Model interpretation reveals sequence features driving divergent mouse and human model predictions

To understand why mouse and human models make divergent predictions at some sites, we compared base-pair resolution importance scores from both models at selected example sites. Specifically, we implemented a strategy similar to in silico mutagenesis (ISM) where a base's score was determined by the differential model output between the original sequence and the sequence with 5bp centered on that base replaced with bases from a dinucleotide-shuffled reference (Alipanahi et al. 2015). We observed that this strategy

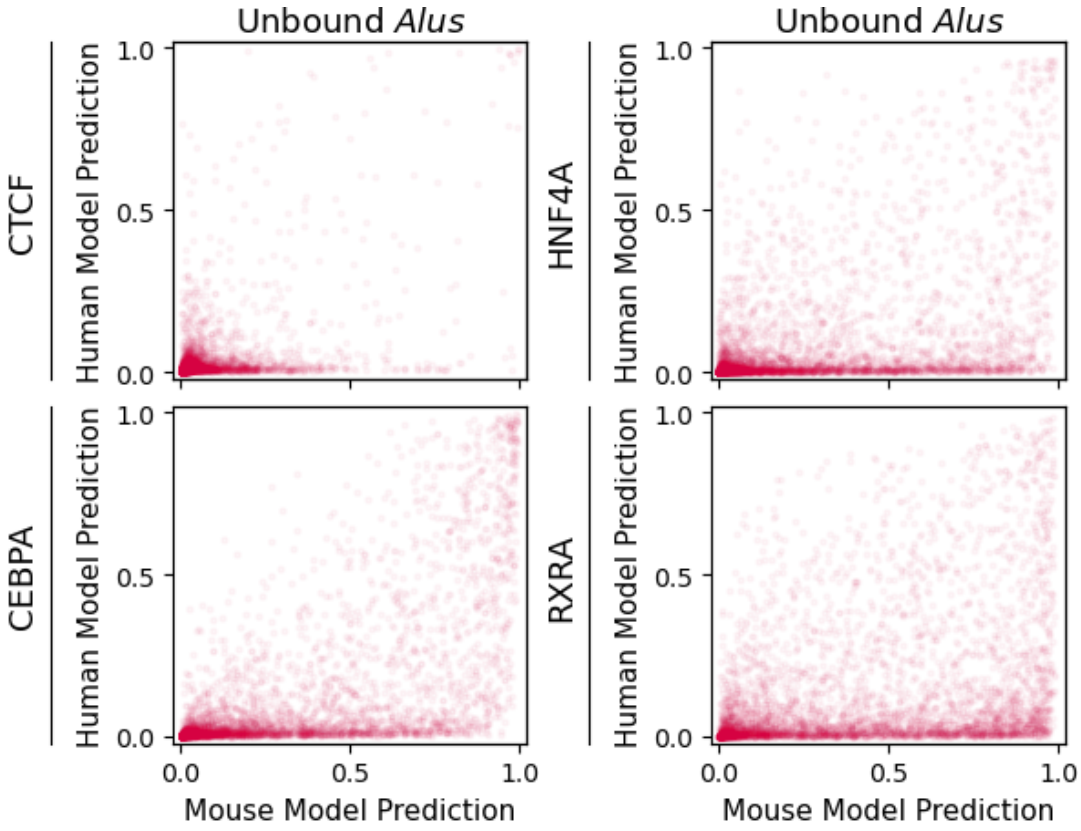

Figure 5: Most unbound sites from the human genome mispredicted by mouse-trained models (x-axis), but not by human-trained (y-axis) models, contain *Alu* repeats. For visual clarity, only 5% of windows are shown.

TF	Bound	FN (Both Models)	FN (Mouse Only)	Unbound	FP (Both Models)	FP (Mouse Only)
CTCF	12.6%	12.8%	9.9%	21.3%	10.0%	**78.6%**
CEBPA	18.3%	11.1%	0.0%	21.3%	22.9%	**84.8%**
HNF4A	13.6%	10.4%	8.0%	21.3%	16.9%	**95.1%**
RXRA	13.7%	10.6%	5.5%	21.4%	20.3%	**97.4%**

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.

outperformed backpropagation-based scoring methods, potentially by avoiding gradient instability.

First, we compared importance scores between the mouse and human models at example bound sites that both models predicted correctly (Supplemental Fig. S5). If the two models learned to use similar logic to make binding predictions, we would expect to see similar sequence features highlighted in the importance scores. Overall, we observe that the scores generated by the mouse and human models are reasonably concordant, although the extent of agreement varies noticeably across TFs. CTCF and CEBPA show the greatest tendency for agreement in importance scores across models. HNF4A showed a slightly weaker trend of score agreement, while RXRA importance scores were the most likely to disagree across models, including instances where motifs are highlighted by high scores from one model but given near-zero scores by the other model. However, across all TFs, instances of the primary cognate motif for the appropriate TF are common in the sequences marked by higher importance scores from either model.

Next, we repeated the analysis on example unbound windows classified as mouse-model-unique false positives (Supplemental Fig. S6). At these sites, the mouse model's prediction scores overshoot those of the human model by at least 0.5. Importance scores in this set of sites show much greater disagreement between the two models. Commonly across all four TFs, we observed two trends: first, the mouse models often assigned high importance to motif-sized contiguous stretches of bases which were not similarly recognized by the human models. These pseudo-motifs can superficially resemble approximate matches to the TF's cognate motif. Second, the human models commonly showed apparent sensitivity to specific, often sparse features which received negative scores of moderate to high magnitude. These observations imply that the human model has learned to ignore certain sequence features that the mouse model's scores suggest are favorable for binding. Furthermore, the human model may be adopting that strategy based on whether or not there are nearby sequence contexts that indicate that the sequence is not a binding site.

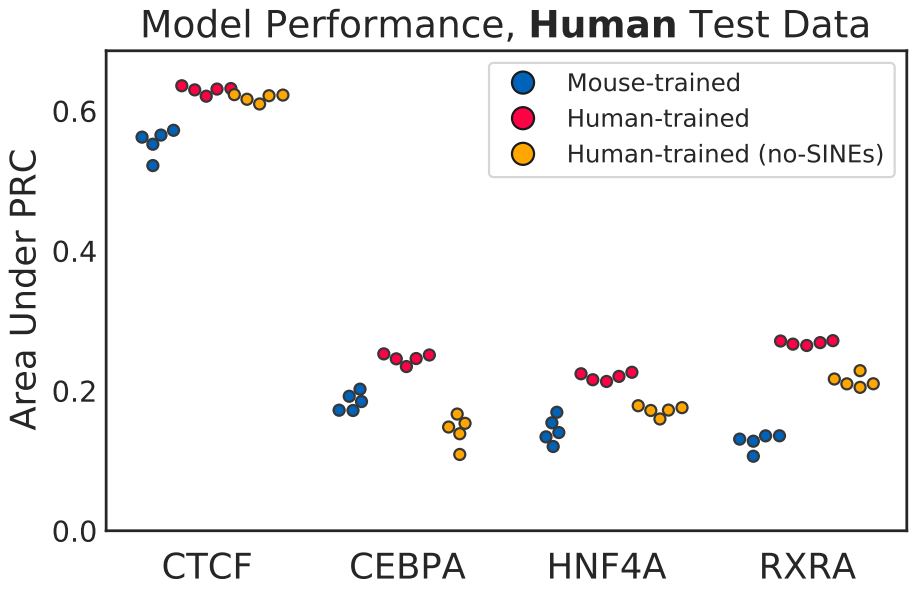

Figure 6: Performance of models that are mouse-trained (blue), human-trained with SINE examples (red), and human-trained without SINE examples (yellow), evaluated on the held-out human Chromosome 2. Five models were independently trained for each TF and training species.

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). For all TFs except CTCF, the no-SINE models perform substantially worse than models trained using the complete human training sets.

Site-distribution plots show that, for unbound sites, no-SINE human-trained models make mispredictions in a pattern similar to mouse-trained models; there is a similarly-sized subset of unbound sites mispredicted by the no-SINE human-trained models but not by the standard human-trained models (Figure 7). Plotting only the sites that overlap with *Alus* confirms that the false positives unique to the no-SINEs model are predominantly *Alu* elements (Supplemental Fig. S7). For bound sites, on the other hand, no-SINE human-trained models make predictions that generally agree with predictions from standard human-

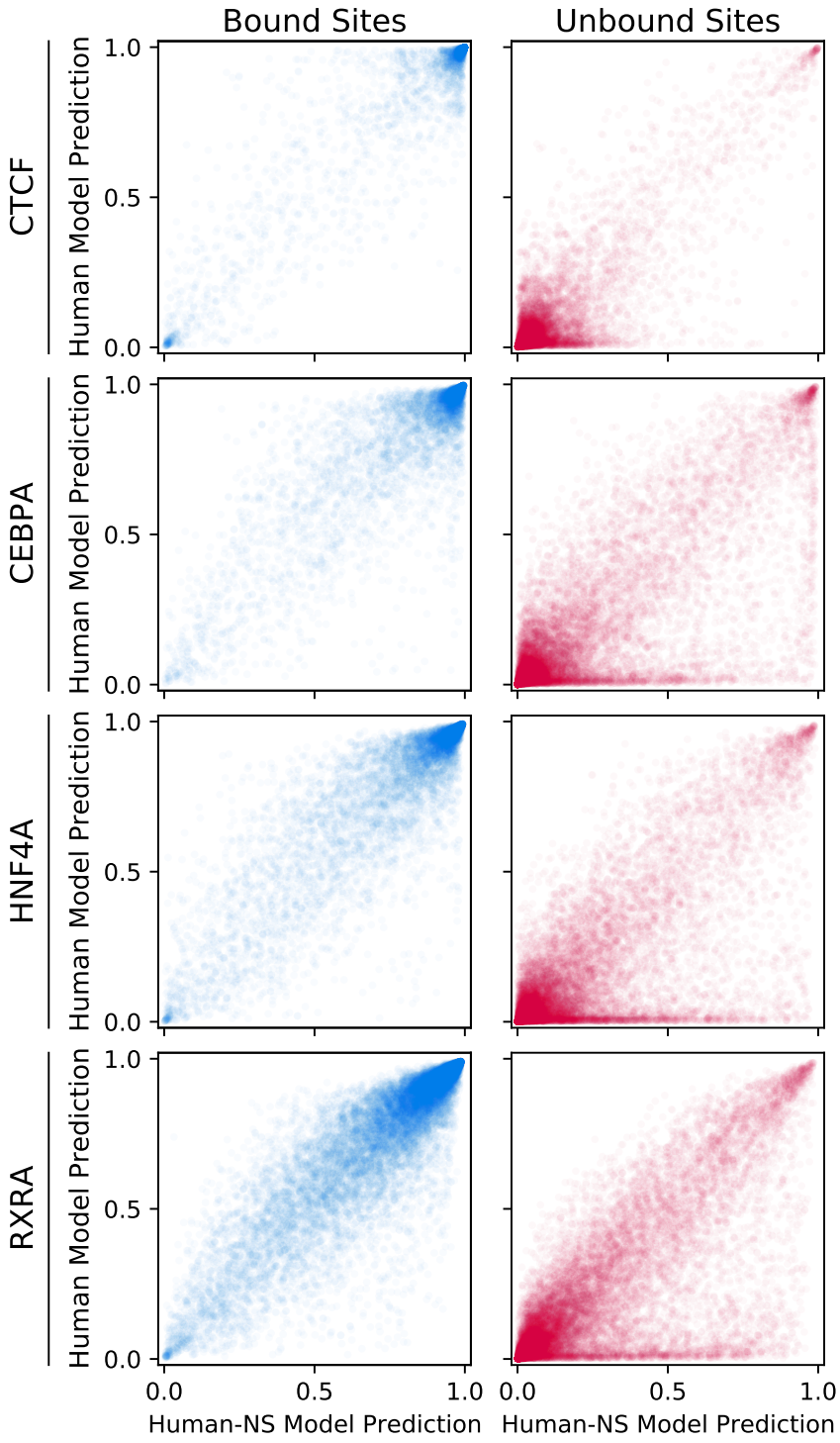

Figure 7: Differential human Chromosome 2 site predictions between models trained on human data with or without any examples of SINE windows. Human-NS: models trained on human data with no SINE examples. Similar to mouse-trained models, no-SINE human-trained models systematically mispredict some unbound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

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 observed 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

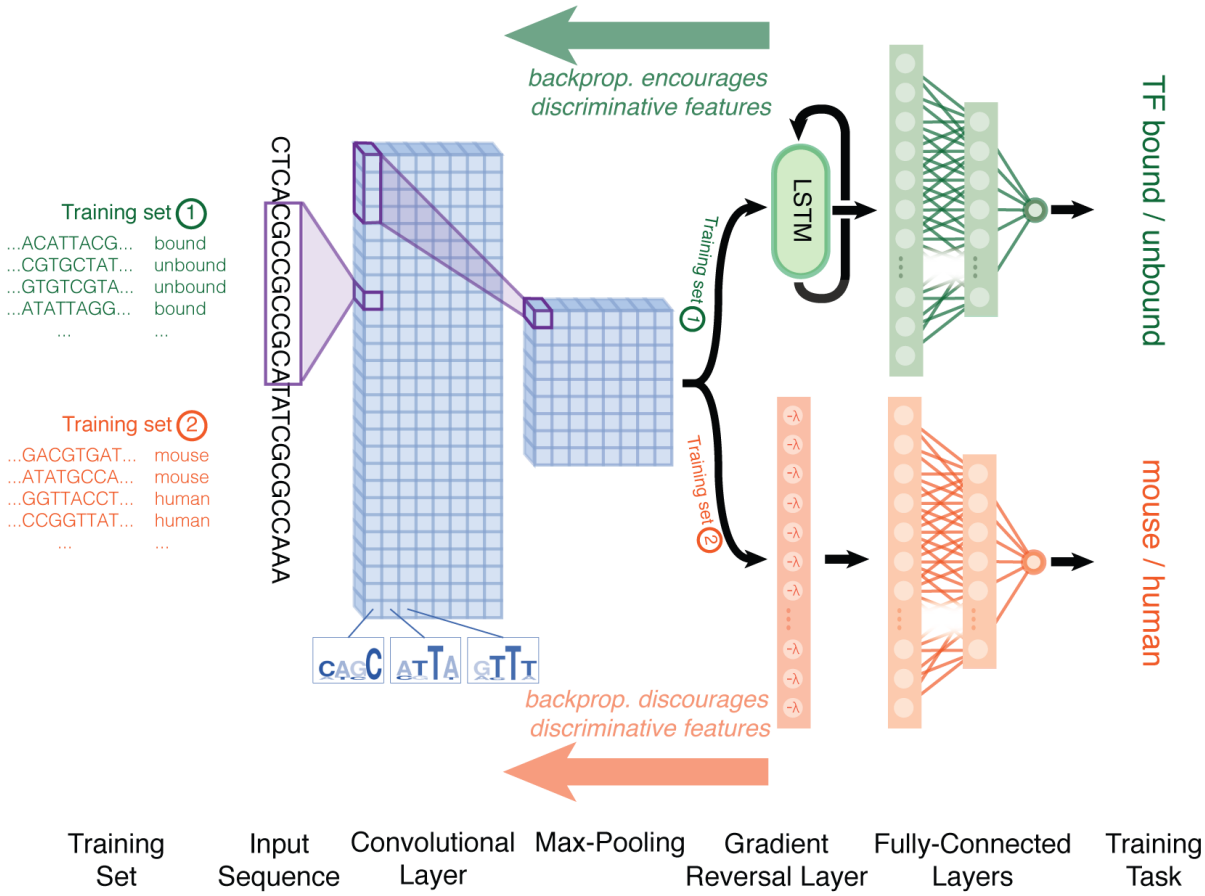

Figure 8: Domain-adaptive network architecture. The top network output predicts TF binding, as before, while the bottom network output predicts the species of origin of the input sequence window. The gradient reversal layer has the effect of discouraging the underlying convolutional filters from learning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

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/left vs. green/middle dots). The domain-adaptive models' auPRCs on mouse test data, meanwhile, is comparable to the auPRCs of basic models (Figure 9, bottom, blue/left vs. green/middle). 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/middle vs. red/right).

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 Figure 10 compare human genome predictions between domain-adaptive mouse models and the original human models. *Alu* elements are highlighted in Figure 11, with quantification in Supplemental Table S2.

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 discov-

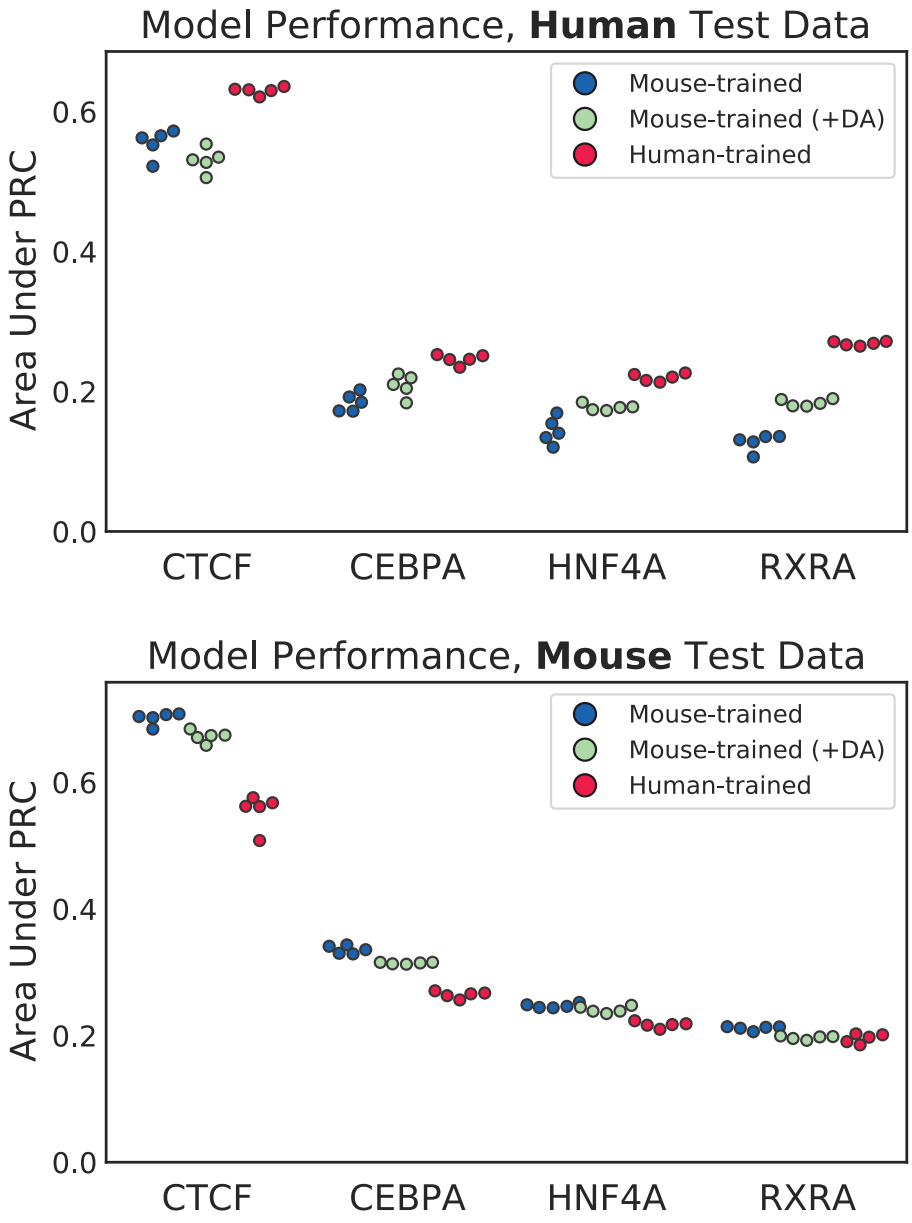

Figure 9: Performance of mouse-trained generic (blue), mouse-trained domain-adaptive (green), and human-trained (red) models, evaluated on human (top) and mouse (bottom) Chromosome 2. Five models were independently trained and evaluated for each TF and training species.

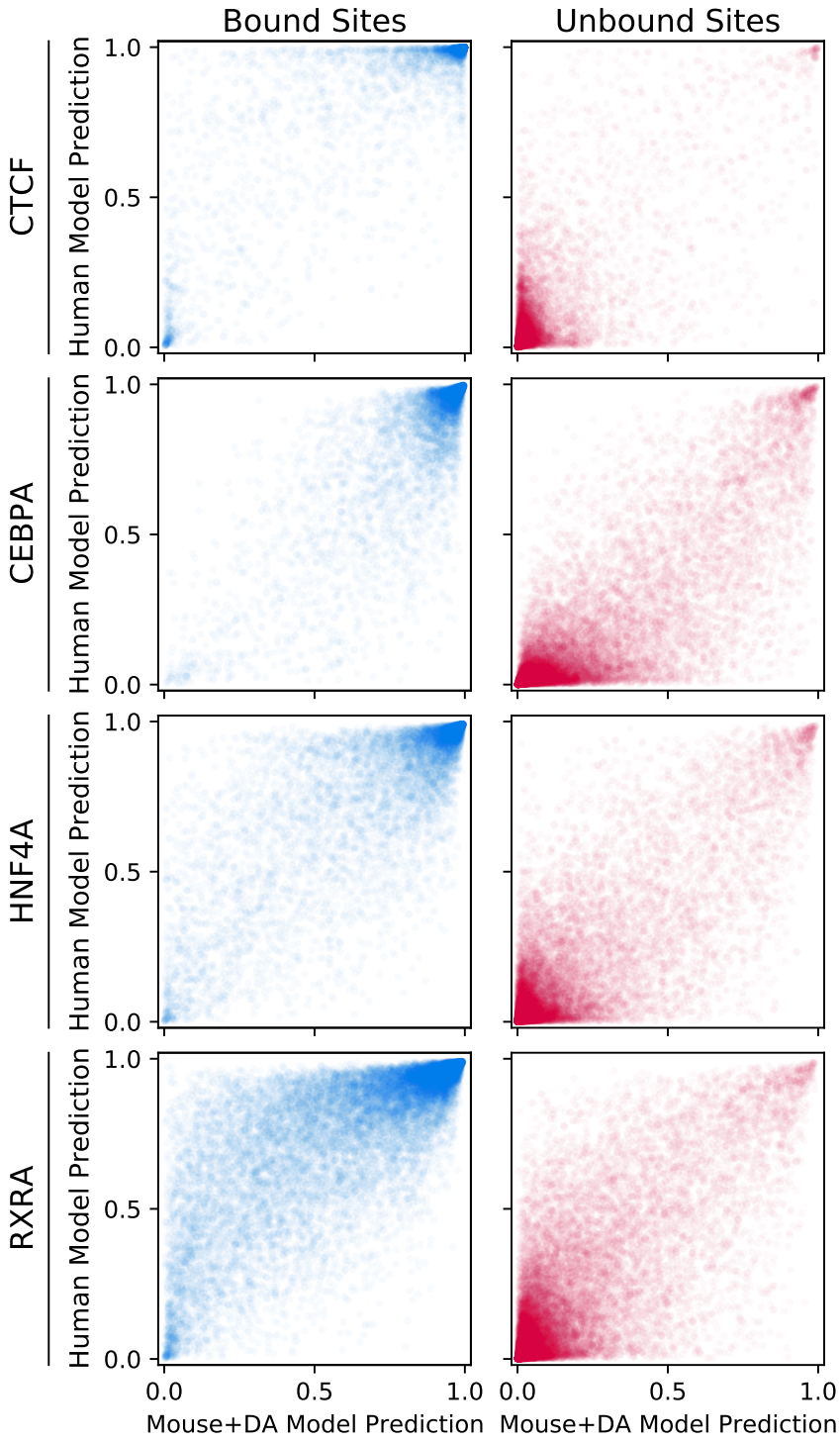

Figure 10: Differential predictions of human genome sites between human-trained and domain-adaptive mouse-trained models. Domain-adaptive mouse models, unlike the original mouse models, do not show species-specific systematic misprediction of unbound sites.

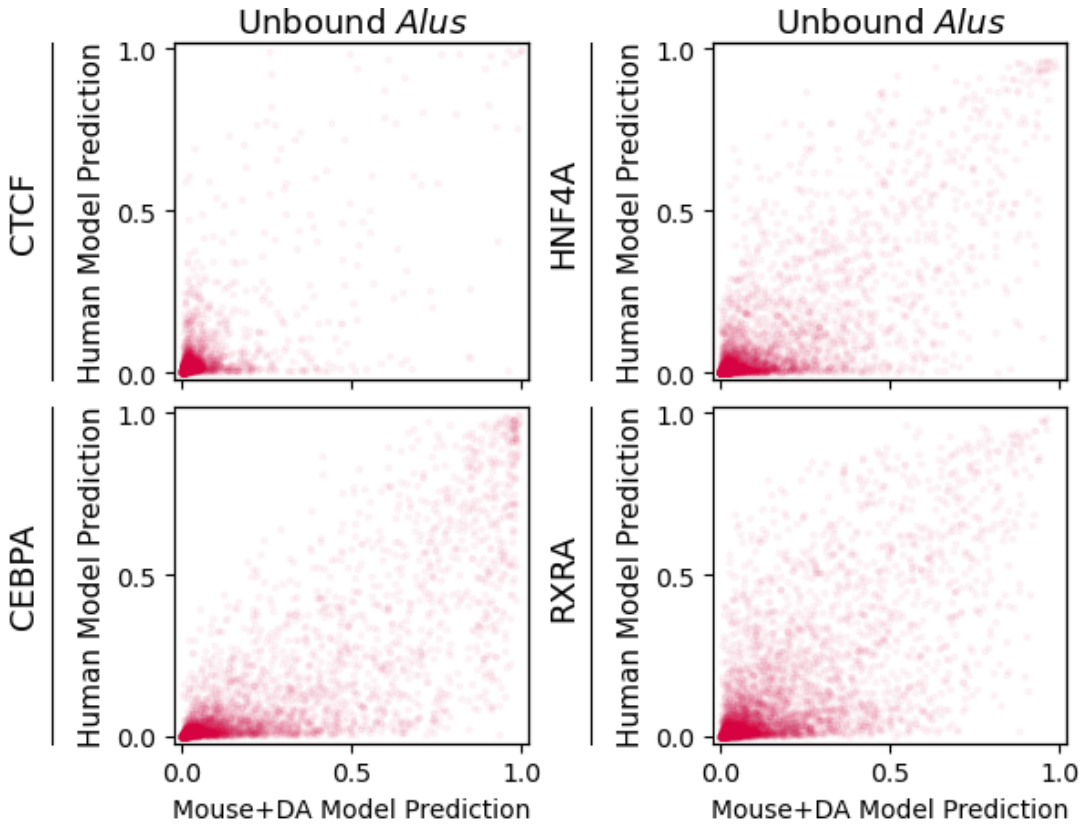

Figure 11: Differential predictions of unbound sites containing *Alu* elements between domain-adaptive mouse-trained models and human-trained models. Unlike the original mouse models, domain-adaptive mouse models do not show systematic overprediction of *Alu* repeats.

ered similar, but not identical, motif-like short sequence patterns as we did previously (Supplemental Fig. S8). Thus, our domain adaptation approach does not appear to have any major influence on bound site predictions.

Alus commonly drive mouse-model false positives across diverse cell types

Finally, we asked whether the observed over-prediction of species-specific repeats is a general issue of concern in cross-species TF binding prediction, or whether it is particular to the examined liver TFs. We thus widened our analyses to 53 additional pairs of ChIP-seq datasets targeting orthologous TFs across 8 additional equivalent human and mouse cell types (see Methods). One caveat is that the expanded set of paired datasets typically focus on cell lines and cell types that are more difficult to closely match across species than liver samples. Thus, the additional experiments examined here may not be as comparable across species as the previously examined liver datasets.

Our expanded analyses confirm that the cross-species performance gap is present in most tested TFs and cell types (Supplemental Table S3). A large portion of mouse-to-human false positive predictions is attributable to *Alu* elements. In 43 of the 53 additional examined datasets, *Alu* elements overlap a third or more of the mouse-model-unique false positive predictions (Supplemental Table S4). Our domain adaptation procedure is successful in reducing *Alu*-related false positive predictions in 46 of the 53 additional examined datasets (Figure 12; Supplemental Table S4). However, in megakaryocyte and hematopoietic progenitor datasets, we generally see a smaller percentage of mouse-model-unique false positives being attributable to *Alus*. The false positive predictions that do overlap *Alus* are also generally less likely to be corrected by our domain adaptation approach in these cell types (Figure 12). Therefore, our observations may not apply uniformly to all cell types.

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

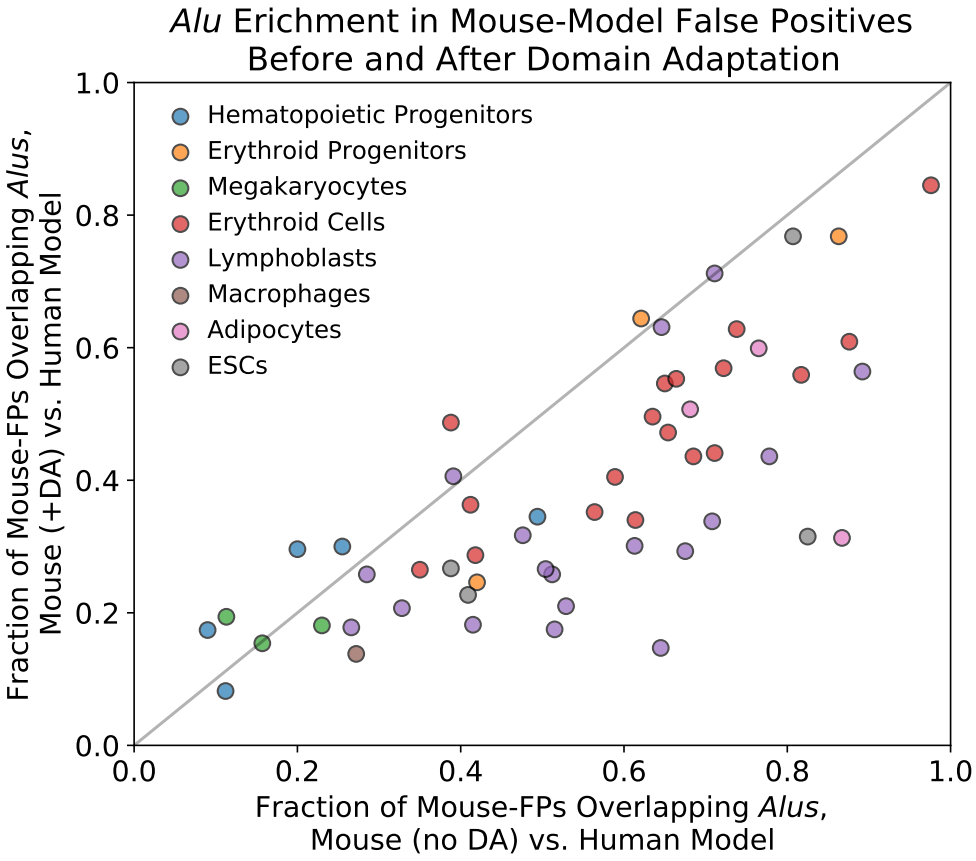

Figure 12: The fraction of mouse-model-unique false positives that overlap *Alus* when either the basic mouse model (x-axis) or the domain-adaptive mouse model (y-axis) are compared against the human model, across our additional paired datasets. The black diagonal line shows $y = x$; points below the line represent TFs where the fraction of *Alus* in mouse-model-unique false positives decreased with our domain adaptation strategy.

other model organisms (The 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. For some TFs, concordant importance scores between mouse and human models across true-positive bound sites suggests that both models learned similar representations of the TF's cognate motif. However, for other TFs, the same analysis suggests that the models' representations of the sequences important for binding may not completely agree. We also observe, particularly for those TFs with less concordant importance scores across species, that there are sequence features in bound sites that discriminate between correct and incorrect predictions specific to cross-species models. Therefore, 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. Our sequence composition domain adaptation approach is unlikely to address situations 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). 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 2020; Schreiber et al. 2020). 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, excluding sex chromosomes, 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.

ChIP-seq experiments and corresponding controls (where available) were collected from ENCODE, GEO, and ArrayExpress. Database accession IDs for all data used in this study are listed in Supplemental Tables S5, S6, and S7. We chose to focus our initial analyses on liver, as several previous studies have provided matched ChIP-seq experiments characterizing orthologous TF binding across mammalian liver samples (Schmidt et al. 2010; Odom et al. 2007). Our expanded analyses use erythroid, lymphoblast, and ES cell line experiments that were previously compared across species by Denas, et al. (Denas et al. 2015). We also analyzed matched adipocyte datasets that were performed on adipocyte cell lines within the same labs (Schmidt et al. 2011; Mikkelsen et al. 2010). Additional datasets were sourced by searching the literature for ChIP-seq data targeting orthologous TFs in erythroid progenitor, megakaryocyte, macrophage, and hematopoietic progenitor cell types (Tijssen et al. 2011; Hu et al. 2011; Pham et al. 2012; Pencovich et al. 2013; Kaikkonen et al. 2013; Beck et al. 2013; Yue et al. 2014; Huang et al. 2016; Goode et al. 2016).

For cell types where all data was sourced from the mouse and human ENCODE projects (i.e., erythroid, lymphoblast, and ES cell lines), we downloaded ChIP-seq narrow peak calls from the ENCODE portal. For liver and all other cell types, we first aligned the FASTQ files to the mm10 and hg38 reference genomes using Bowtie (version 1.3.0) (Langmead and Salzberg 2012). We then called ChIP-seq peaks using MultiGPS v0.74 with default parameters, excluding ENCODE blacklist regions (Mahony et al. 2014; Amemiya et al. 2019). Corresponding control experiments were utilized during peak calling when available. 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. Supplemental Table S5 shows the numbers of peaks called for liver datasets, as well as the number of bound windows retained after filtering and the fraction of all retained windows that are bound; Supplemental Tables S6 and S7 show the same information for all other datasets. Candidate datasets were discarded from the analysis if the numbers of called peaks was less than 1000 in mouse or human.

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. Chromosomes X and Y were not used to avoid confounding because our matched datasets across species did not always match in sex.

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, 2, X, and Y. 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.

Model performance evaluations were computed with the sci-kit-learn v0.23 implementation of the average_precision_score function, which closely approximates the area under the precision-recall curve (auPRC).

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 discriminative 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 Supplemental Table S1, repeat classes were excluded if fewer than 500 examples of that class were annotated in the test chromosome (before mappability filtering).

Gapped *k*-mer SVMs

The gkmtrain and gkmpredict utilities from the lsgkm package were used for gkmSVMs gkm training and prediction generation, respectively (Lee 2016). For training, 50000 examples each were selected randomly from the set of all bound windows and unbound windows in the original neural network model training sets. Every 10th example from the original test set (in other words, sampling windows such that all selected windows were non-overlapping) was considered in evaluation for computational efficiency. All default parameters were used in running lsgkm (center-weighted + truncated *l*-mer kernel, word length 11, maximum 3 mismatches).

Profile models

Our profile model consists of a dilated convolutional residual model architecture that closely resembles the BPNet architecture (Avsec et al. 2021b), with the following modifications: 1) 21bp-long filters in the first convolutional layer, rather than 25bp; 2) 8 dilated convolutional layers, rather than 9; 3) a learning rate of 0.001; 4) 2114 bases of sequence input. The first three hyperparameters were selected by tuning on the source-species validation set loss; the sequence input length was chosen based on what would produce a 1000bp-long profile prediction given the 8-layer architecture's receptive field.

The profile models were trained using the same task and loss scheme as in Avsec et al. 2021b, with the loss function value of λ set to 10. Training lasted 30 epochs, with early stopping used to select the best model according to the source-species validation set profile (multinomial) loss. The training data used was sampled from regions in the training set used by the binary models: specifically, each epoch the profile model saw a 3:1 ratio of windows centered on peaks from training set chromosomes, with up to 200bp jitter, and windows not overlapping peaks with a GC-content distribution that matched the set of peak-centered windows. Hyperparameter tuning was performed using a combination of the BPNet multinomial

loss for the profile task, calculated on peaks from Chromosome 1, and auPRCs calculated using the same validation set of 1 million random windows from Chromosome 1 that the binary models used. Final model evaluation was performed on the full original test sets from Chromosome 2 used by the binary models.

Importance scoring

For a given 500bp window and model, importance scores were generated using a method similar to in silico mutagenesis, which measures the change in model prediction when a given base and the region immediately around it are ablated. First, ten independent dinucleotide-shuffled versions of the original sequence were generated to serve as reference sequences unlikely to contain motifs. Next, the 5bp region centered at a particular base was replaced with the corresponding 5bp region from one of the ten shuffled sequences, and the post-sigmoid difference in model output for this ablated sequence was recorded. This was repeated for all ten shuffled sequences, with the average model prediction differential reported as the score for the base that the ablated region centered on. This process was repeated for all bases in the sequence being scored.

Software availability

Open source code (MIT license) is provided in a Supplemental Code file and is also 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 and Jacob Schreiber for helpful feedback and discussion. The authors also thank Daniela Uribe, Edgar Roman, and Yishu Chen for their work replicating the findings pertaining to profile models in Supplemental Fig. S2.

Funding

This work was supported by NIH NIGMS grant R01GM121613 and NSF CAREER 2045500 (both to SM), NIH NIGMS grant DP2GM123485 (to AK), and the Stanford Graduate Fellowship (to KC). RCH is supported by NIH NIDDK grant R24DK106766. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interest statement

A.K. is a scientific co-founder of Ravel Biotechnology Inc, consultant with Illumina Inc., and on SAB of OpenTargets, SerImmune and PatchBio. The other authors have no potential conflicts of interest to declare.

References

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**: 831–838.

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

Avsec Ž, Agarwal V, Visentin D, Ledsam J, Barwinska AG, Taylor K, Assael Y, Jumper J, Kohli P, Kelley D. 2021. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*. DOI: https://doi.org/10.1038/s41592-021-01252-x.

Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A, et al. 2021. Base-resolution models of transcription factor binding reveal soft motif syntax. *Nat Genet* **53**: 354–366.

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

Beck D, Thoms JAI, Perera D, Schütte J, Unnikrishnan A, Knezevic K, Kinston SJ, Wilson NK, O'Brien TA, Göttgens B, et al. 2013. Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. *Blood* **122**: e12–22.

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 Res* **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`.

Denas O, Sandstrom R, Cheng Y, Beal K, Herrero J, Hardison RC, Taylor J. 2015. Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. *BMC Genomics* **16**: 87.

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. *Mol Cell* **77**: 475–487.

Fudenberg G, Kelley DR, Pollard KS. 2020. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* **17**: 1111–1117.

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V. 2016. Domain-Adversarial Training of Neural Networks. *J Mach Learn Res* **17**: 1–35.

Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS Comput Biol* **10**: e1003711.

Goode DK, Obier N, Vijayabaskar MS, Lie-A-Ling M, Lilly AJ, Hannah R, Lichtinger M, Batta K, Florkowska M, Patel R, et al. 2016. Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. *Dev Cell* **36**: 572–587.

Hu G, Schones DE, Cui K, Ybarra R, Northrup D, Tang Q, Gattinoni L, Restifo NP, Huang S, Zhao K. 2011. Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res* **21**: 1650–1658.

Huang J, Liu X, Li D, Shao Z, Cao H, Zhang Y, Trompouki E, Bowman T, Zon LI, Yuan GC, et al. 2016. Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. *Dev Cell* **36**: 9–23.

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.

Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, Chun HB, Tough DF, Prinjha R, Benner CK, et al. 2013. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell* **51**: 310–325.

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 Res* **46**: e120.

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

Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. 2018. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* **28**: 739–750.

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

Koo PK, Majdandzic A, Ploenzke M, Anand P, Paul SB. 2021. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLoS Comput Biol* **17**: e1008925.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lee D. 2016. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**: 2196–2198.

Long M, Cao Y, Wang J, Jordan MI. 2015. Learning Transferable Features with Deep Adaptation Networks. *2015 International Conference of Machine Learning (ICML)*, 97–105.

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.

Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, Lander ES, Rosen ED. 2010. Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**: 156–69.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**: 730–732.

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

Pencovich N, Jaschek R, Dicken J, Amit A, Lotem J, Tanay A, Groner Y. 2013. Cell-autonomous function of Runx1 transcriptionally regulates mouse megakaryocytic maturation. *PLoS One* **8**: e64248.

Pham TH, Benner C, Lichtinger M, Schwarzfischer L, Hu Y, Andreesen R, Chen W, Rehli M. 2012. Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. *Blood* **24**: e161–171.

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. 2016. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* **44**: e107.

— 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 Res* **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.

Schmidt SF, Jørgensen M, Chen Y, Nielsen R, Sandelin A, Mandrup S. 2011. Cross species comparison of C/EBPα and PPARγ profiles in mouse and human adipocytes reveals interdependent retention of binding sites. *BMC Genomics* **12**: 152.

Schreiber J, Hegde D, Noble W. 2020. Zero-shot imputations across species are enabled through joint modeling of human and mouse epigenomics. *ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. DOI: 10.1145/3388440.3412412.

Siggers T, Gordan R. 2014. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res* **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 Biochem Sci* **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. *Genome Biol* **22**: 20.

Srivastava D, Mahony S. 2020. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *BBA Gene Regul Mech* **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 Res* **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.

Tijssen M, Cvejic A, Joshi A, Hannah R, Ferreira R, Forrai A, Bellissimo D, Oram SH, Smethurst P, Wilson N, et al. 2011. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev Cell* **20**: 597–609.