

“Frustratingly Easy” Domain Adaptation for Cross-Species Transcription Factor Binding Prediction

Mark Maher Ebeid¹[0000–0002–9113–7082], Ali Tuğrul Balci¹[0000–0002–1461–6733],
Maria Chikina¹[0000–0003–2550–5403], Panayiotis V.
Benos^{1,2}[0000–0003–3172–3132], and Dennis Kostka¹[0000–0002–1460–5487]

¹ Department of Computational & Systems Biology, University of Pittsburgh School of Medicine and Joint Carnegie Mellon—University of Pittsburgh Ph.D. Program in Computational Biology, University of Pittsburgh, Pittsburgh, PA, USA

² Department of Epidemiology, University of Florida, Gainesville, FL, USA
kostka@pitt.edu

Abstract. How DNA sequence encodes gene regulation remains a central challenge in regulatory genomics. Transcription factors (TFs) are key mediators of this process, binding to specific sequence motifs to control gene expression. Yet, predicting where they bind from sequence alone remains a challenging problem. A cross-species angle offers two complementary benefits: it tests whether trained models have learned conserved, biochemically grounded rules that generalize across species, and it enables binding prediction in species where experimental data is scarce. Key challenges in this context are that TF binding sites undergo rapid evolutionary turnover, and that there are systematic distributional differences between species’ genomes. We present MORALE, a domain adaptation framework for cross-species TF binding prediction. By aligning the first and second moments of sequence embeddings across species during training, MORALE learns species-invariant representations without adversarial training, additional parameters, or architectural changes. Applied to liver ChIP-seq data from two species (human, mouse) and five species (adding rhesus macaque, rat, and dog), MORALE consistently matches or outperforms gradient reversal (the adversarial baseline) across all TFs, and avoids the performance degradation below the no-adaptation baseline that gradient reversal can exhibit. In the five-species setting, MORALE surpasses a human-only model, demonstrating that moment alignment can unlock cross-species generalization that neither multi-species training nor adversarial adaptation achieves alone. MORALE also recovers TF binding motifs more faithfully than the adversarial approach, suggesting its representations capture biologically meaningful sequence features. As a closed-form, parameter-free regularizer, MORALE integrates into any embedding-based sequence model.

Keywords: Evolutionary Genomics · Gene Regulation · Transcription Factor Binding · Sequence To Function Modeling · Cross-Species Comparison · Domain Adaptation ·

1 Introduction

Genomic regulatory activity is largely governed by transcription factors (TFs) that bind to DNA and influence gene expression. Sequence-to-function models, typically deep neural networks, have become a cornerstone for predicting TF binding and regulatory activity from sequence, enabling downstream interpretation, functional element discovery, and *in silico* sequence design [14,24,7,26,8,4,6]. A key open question is the extent to which these models learn universal biochemical principles that generalize across species.

Training on multi-species data offers a promising path toward such generalization. However, TF binding sites undergo rapid evolutionary turnover (even between closely related species), making cross-species transfer non-trivial [29]. At the same time, TF DNA-binding domains are highly conserved across species, suggesting a shared regulatory vocabulary that multi-species models could exploit [16,3,11,9]. Realizing this potential requires methods that can bridge systematic differences between species’ genomes’ nucleotide distributions.

Cochran et al. [10] addressed this with a gradient reversal layer (GRL) approach: an adversarial discriminator penalizes the model for learning species-specific features, encouraging a species-invariant representation. While effective in some settings, GRL requires a separate discriminator branch, substantially increasing model complexity, and, as we show, can degrade performance below a no-adaptation baseline.

We propose MORALE, which instead aligns the first and second moments of sequence embeddings across species [31] using a closed-form operation that requires no additional parameters and integrates into any embedding-based model. We compare MORALE against GRL on human–mouse TF binding prediction across four TFs, and extend the evaluation to a five-species setting. MORALE consistently matches or outperforms GRL and learns a robust species-invariant feature set.

2 Materials & Methods

2.1 Data

Two-species. Following Cochran et al. [10], we processed ChIP-seq data for CTCF, HNF4 α , RXRA, and CEBPA in human and mouse liver (ENCODE: ENCSR000CBU, ENCSR911GFJ, ENCSR098XMN; ArrayExpress: E-TABM-722; GEO: GSM1299600). Sequences were tiled into 500-bp windows (50-bp overlap), ENCODE blacklist regions removed [2], and aligned to GRCh38/GRCm38 with BowTie2 [19]. Peaks were called with multiGPS v0.75 [22]. Windows covering a peak center were labeled “bound.” Chr1/Chr2 were held out for validation/testing; sex chromosomes excluded.

Multi-species. We used published liver ChIP-seq for CEBPA, FOXA1, ONECUT1, and HNF4 α across five mammals (human, rhesus macaque, mouse, rat, dog) [5] (ArrayExpress E-MTAB-1509), tiled into 1000-bp windows. Validation/test chromosomes were chosen via a linear program to approximate a target

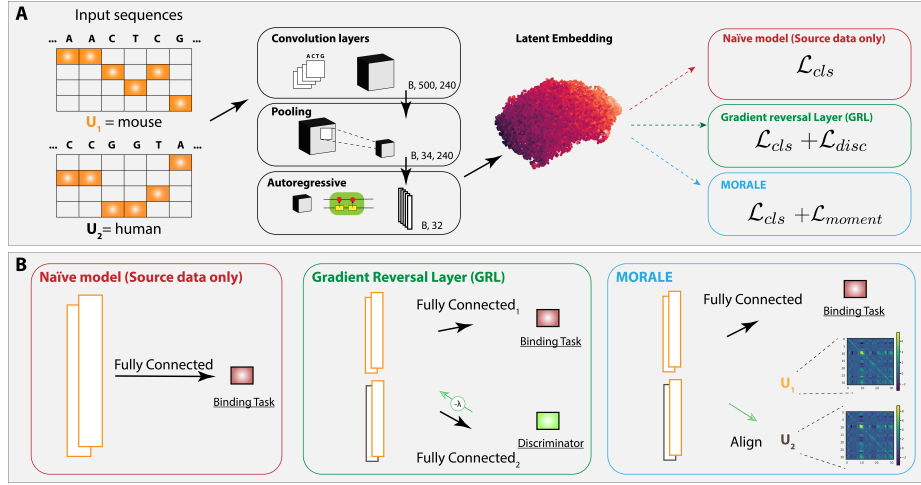


Fig. 1: **Overview of MORALE and comparison with alternative domain adaptation approaches.** (A) Input sequences from source and target species are one-hot encoded and passed through convolution, pooling, and autoregressive layers to produce a latent embedding, which feeds all downstream tasks. Three training procedures are compared: (1) source-only (naive baseline), (2) gradient reversal (GRL), and (3) moment alignment (MORALE). (B) Architectural detail of each approach. In the source-only model, only source embeddings are passed to the classification head. GRL adds an adversarial discriminator branch that predicts species of origin; gradients from this branch are reversed to discourage species-specific representations. MORALE forgoes the discriminator entirely, instead adding a moment alignment loss between source and target embeddings directly to the training objective.

fraction of positive windows while maintaining class balance (R package lpSolve; see **STabs.** 7, 8).

2.2 Model Architecture

Two-species. We use the architecture of Cochran et al. [10]: Conv(240 × 20bp) → MaxPool(15) → LSTM(32) → dense head (613K params). GRL adds an adversarial branch (9.5M params total). MORALE adds no parameters; moments are computed on the 32-d LSTM embedding via ADAPT [23].

Multi-species. We adapted a bidirectional GRU from gReLU [18]: Conv(240 × 20bp, ReLU) → MaxPool(16) → BiGRU(240) → FC × 2 → Conv1 × 1 (63ch) → GlobalAvgPool (63-d embedding, 960K params).

2.3 Hyperparameter Tuning

For GRL, to determine λ , we search a grid of values from 0.0 to 10.0 with a step size of 0.50 and select the maximal auPRC value for each transcription factor.

For mouse-to-human we select 1.5, 6.0, 0.5, 7.5 for CTCF, CEBPA, HNF4 α , and RXRA, respectively. For human-to-mouse direction, we select 6.5, 8.5, 10.0, 1.0.

For MORALE we use a range from 0 to 10 with a step size of 1. In the mouse-to-human direction we select 4, 7, 8, and 8 for CTCF, CEBPA, HNF4 α , and RXRA, respectively. In the human-to-mouse direction, we select 4, 8, 6, and 7. For the multi-species analysis we use a grid between 1 and 8 with a step-size of 1.

2.4 Training & Evaluation

Two-species models used TensorFlow 2.18 [1], Adam (lr=1e-3), 15 epochs, batch size 400 (200 bound/unbound from source) plus 400 background per species for domain adaptation. Multi-species used PyTorch 2.5.1 [25], same optimizer, mini-batches of 200 bound/unbound (balanced across sources) plus 250 background (50/species). Performance was assessed by auPRC (scikit-learn [27]).

2.5 Attribution & Downstream Analyses

Per-nucleotide attribution scores were computed via expected integrated gradients (CREsted [17]). Motifs were discovered with TF-MoDISco lite [30] on up to 2,000 high-confidence bound sites (sigmoid ≥ 0.98) and annotated with Tomtom [13]. Repeat analyses used RepeatMasker tracks from UCSC [28]; phylogenetic trees were built with phyloT v2 [20].

3 Results

3.1 Approach

Our approach, MORALE, is a straightforward generalization of the Deep CORAL method [31], applied to TF binding data. We assume labeled (sequence) data from n different source domains and unlabeled data from a target domain; in our application, the domains are species. We also assume an encoder model (see **Section 2.2**), mapping input sequences $x^{i|j}$ to vector embeddings $z^{i|j} \in \mathbb{R}^d$, where $i = 1, \dots, m_j$ and $j = 1, \dots, n + 1$ (including the target domain), and m_j is the number of sequences in the j -th domain. Vector representations are input to a classification model predicting the sequence label (e.g., whether the input sequence binds a specific TF), with an associated loss (e.g., cross entropy) denoted $\mathcal{L}_{\text{label-classification}}$.

Given a mini-batch of $b < m_j$ sequence representations $\{z^{i|j}\}_{i=1}^b$, we denote the sample mean by $\hat{\mu}^j$ and the sample covariance matrix by \hat{C}^j . For a mini-batch containing sequences from each domain, the MORALE loss is calculated as

$$\mathcal{L}_{\text{MORALE}} = \frac{2}{n(n+1)} \sum_{l \neq k} \left(\|\hat{C}^l - \hat{C}^k\|_F^2 + \|\hat{\mu}^l - \hat{\mu}^k\|_2^2 \right),$$

quantifying the difference in first and second moments between all domain pairs. Here $\|\cdot\|_F$ denotes the Frobenius norm. We estimate $\hat{\mu}^j$ and \hat{C}^j for $\mathcal{L}_{\text{MORALE}}$ using the background portion of the mini-batch (described in **Section 2.1**), which contains an equal number of unlabeled sequences sampled from each domain (source(s) and target). For source domains, the vector representations $\{z^{i|j}\}_{i=1}^b$ are additionally used for label classification.

The MORALE loss is added to the label classification loss, encouraging sequence representations that are moment-aligned across domains:

$$\mathcal{L} = \mathcal{L}_{\text{label-classification}} + \lambda \mathcal{L}_{\text{MORALE}}. \quad (1)$$

We note that gradient reversal layers (GRL) are a successful adversarial approach to encouraging domain-invariant representations [12,21,15]. Briefly, a domain classifier and associated loss for predicting embeddings’ domain of origin (e.g., cross entropy) is added to the model for an overall loss of

$$\mathcal{L} = \mathcal{L}_{\text{label-classification}} + \mathcal{L}_{\text{domain-classification}}. \quad (2)$$

Backpropagated gradients from $\mathcal{L}_{\text{domain-classification}}$ are penalized by a factor of $-\lambda$ (for $\lambda > 0$, hence “gradient reversal”), encouraging representations that are uninformative about the domain of origin.

In our analyses we compare MORALE (**Equation 1**) with the GRL approach (**Equation 2**) and find it generally outperforms GRL. We also note that MORALE does not require additional parameters beyond the encoder and label classification model, whereas GRL requires additional design and parameterization of the domain classifier. **Figure 1** summarizes all three approaches.

3.2 Cross-Species TF-Binding Prediction Between Human and Mouse

MORALE Improves Cross-Species TF Binding Prediction Performance.

First, we applied our framework to re-analyze a dataset introduced by Cochran et al. [10], in which the binding of four TFs (CTCF, HNF4 α , RXRA, and CEBPA) was assayed in liver tissue from humans and mice, giving two domains.

We assess TF binding site prediction for each species as target, with the other as source. Test set performances are summarized in **Figure 2**. We compare four models: source-only (no domain adaptation), target-trained (upper bound), MORALE, and GRL. Domain adaptation methods have access to target embeddings but not target labels. As expected, source-only performs worst and target-trained performs best, with domain adaptation methods in between. Interestingly, for CTCF with human as target, GRL performs worse than the no-adaptation baseline, while MORALE does not suffer this drop. Likewise, with mouse as target, GRL falls below the source baseline for three of four TFs (CEBPA, HNF4 α , RXRA); MORALE does not. For TFs where GRL does outperform the source baseline, MORALE either matches or exceeds GRL. Numerical values are in **Table 1**.

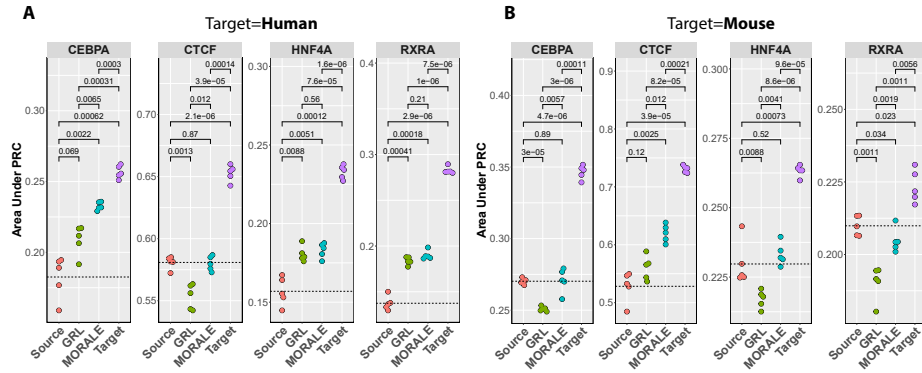


Fig. 2: Moment alignment improves cross-species TF binding site predictions. Prediction performance is shown for four models: (1) source-only (red), (2) gradient reversal (green), (3) MORALE (blue), and (4) target-trained (purple), across four TFs. **(A)** Results when adapting a mouse-trained model to human data. MORALE outperforms or matches GRL in each case without suffering degradation. **(B)** The same analysis in the reverse direction, human-to-mouse. Performance degradation is persistent under GRL, while MORALE meets or exceeds the source baseline and outperforms GRL.

Upon examining results (**Table 1**), we observe that in the mouse-to-human direction the GRL model makes more positive predictions overall, but with disproportionately more false positives compared with MORALE. This leads to worse performance, especially for CTCF. MORALE has lower false positive rates (e.g., 1.38% versus GRL’s 1.85% for CTCF) while suffering only slightly increased false negative rates (0.079% vs. 0.068%, again for CTCF). In the human-to-mouse direction this pattern holds for CTCF. For CEBPA and HNF4 α , MORALE has more true positives and fewer false positives than GRL, and consequently shows better performance. For RXRA, MORALE’s performance is close to the source model, while GRL performs somewhat worse.

From these analyses, we conclude that MORALE improves over GRL for cross-species TF binding prediction on this dataset. Next, we compare the quality of predictions in more detail.

MORALE Improves Cross-Species TF Binding Prediction Quality. To quantify and assess the quality of TF binding predictions, we compare how well MORALE and GRL adapt the source-trained model toward the target-trained model, using the latter as a reference. We use two complementary metrics, both based on per-nucleotide importance scores from post-hoc attribution analysis (see **Section 2.5**).

First, we compare the correlation of GRL/MORALE importance scores with those of the target model, focusing on sequences where the source model disagrees with the target: differential false positives (dFPs), where the source model

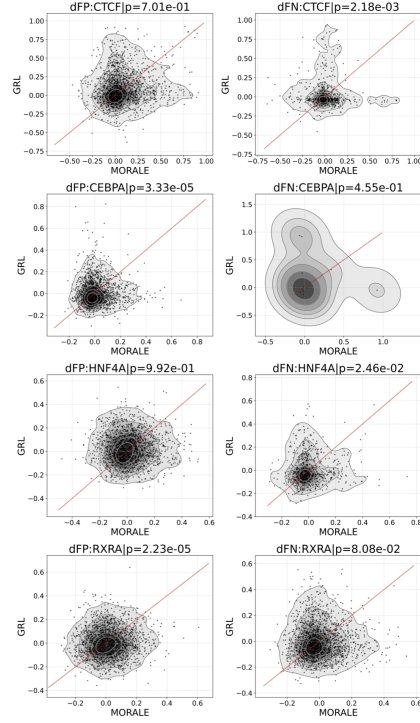
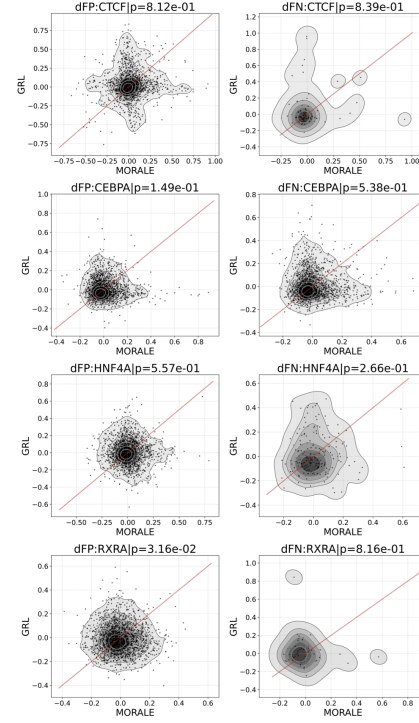
A Target = human**B Target = mouse**

Fig. 3: MORALE attribution scores align more closely with the target model than GRL. Pearson correlation coefficients between MORALE's importance scores and the target model (x-axes) and GRL's importance scores and the target model (y-axes), for differentially false positive sites (dFPs, where the source model incorrectly predicts binding but the target model does not) and differentially false negative sites (dFNs, where the source model misses a binding event correctly predicted by the target model). Panel A shows mouse-to-human and panel B human-to-mouse. P-values for a one-sided Wilcoxon Rank Sum Test are indicated for each plot.

Table 1: **We display the true positives (TPs), false positives (FPs), false negatives (FNs) with the auPRC across four TFs and three methods in this study.** In order to generate this confusion matrix across five-folds of data, we average the sigmoid value over the folds and compute the relevant prediction types w.r.t. the ground truth label, percentages are with respect to the complete dataset. We observe that the GRL more often predicts positives than either the source model, or MORALE, however this comes at much higher occurrence of FPs, leading to performance degradation as in CTCF when the target is human. MORALE, on the other hand, takes a much more conservative step towards the target model, finding a middle ground that strictly results in improvement or meeting baselines.

Target=Human												
TF	TPs (%)			FPs (%)			FNs (%)			auPRC		
	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source
CTCF	0.413	0.402	0.401	1.847	1.384	1.397	0.068	0.079	0.08	0.553	0.58	0.581
CEBPA	0.542	0.545	0.547	10.436	9.77	11.938	0.048	0.045	0.044	0.209	0.233	0.183
HNF4A	0.605	0.578	0.574	9.137	8.496	10.806	0.136	0.164	0.168	0.181	0.183	0.157
RXRA	1.133	1.035	1.076	9.981	8.023	13.181	0.468	0.566	0.526	0.184	0.19	0.137
Average	0.673	0.64	0.649	7.85	6.918	9.33	0.18	0.213	0.205	0.282	0.296	0.264

Target=Mouse												
TF	TPs (%)			FPs (%)			FNs (%)			auPRC		
	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source
CTCF	0.7	0.699	0.697	3.937	3.237	3.591	0.042	0.044	0.046	0.56	0.62	0.528
CEBPA	0.988	0.996	0.952	7.158	6.66	5.732	0.335	0.328	0.372	0.251	0.271	0.27
HNF4A	0.824	0.828	0.813	11.026	10.388	9.802	0.122	0.118	0.133	0.217	0.233	0.23
RXRA	0.831	0.827	0.843	17.891	16.039	19.208	0.087	0.091	0.075	0.19	0.205	0.21
Average	0.836	0.837	0.826	10.003	9.081	9.583	0.146	0.145	0.156	0.304	0.332	0.31

wrongly predicts TF binding but the target model does not; and differential false negatives (dFNs), where the source model misses a binding event that the target model correctly predicts. **Figure 3** shows contour plots of these correlation coefficients, stratified by TF and sequence type (dFP/dFN). Higher correlation indicates that the domain adaptation method better reproduces the importance scores of the target-trained model.

For mouse-to-human and dFP sequences, MORALE importance scores are significantly more correlated with the target model than GRL scores for CEBPA and RXRA, while results are comparable for CTCF and HNF4 α . For dFN sequences, MORALE’s correlation is significantly closer to the target for RXRA and HNF4 α . For human-to-mouse and dFPs, MORALE scores are more correlated with the target for RXRA, while other comparisons are not significant. Overall, MORALE improves upon GRL in highlighting meaningful sequence positions after domain adaptation, and we have not observed the reverse.

Second, we identified sequence motifs from attribution scores using TF-MoDISco [30] (see **Section 2.5**). **Figure 4** summarizes our findings. For CTCF binding sites, the target model’s most frequently found motif corresponds to CTCF, which is recapitulated by MORALE. GRL, like the source-only model, finds a larger fraction of motifs corresponding to other TFs. In panels B to E, the target-trained, MORALE, and source-only models all find CTCF as their most frequent motif match. Surprisingly, GRL shifts the motif distribution such that poorly matching motifs become more frequent.

Table 2: **We display the performance (auPRC) for windows that overlap both SINE and LINE repeats, and their average.** We observe that MORALE performs competitively in the performance across windows that contain species-specific repeats in the mouse-to-human direction, and outperforms the GRL in the human-to-mouse direction. Notably, for HNF-4 α and RXR α , the GRL does a convincingly better job windows that overlap SINE elements in human, which would be Alus in the majority.

Target= Human									
TF	LINE			SINE			Average		
	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source
CTCF	0.149	0.173	0.17	0.181	0.191	0.177	0.165	0.182	0.173
CEBPA	0.039	0.042	0.034	0.04	0.04	0.027	0.039	0.041	0.031
HNF4A	0.043	0.04	0.03	0.048	0.034	0.023	0.046	0.037	0.026
RXRA	0.058	0.056	0.036	0.062	0.052	0.029	0.06	0.054	0.033
Average	0.072	0.078	0.068	0.083	0.079	0.064	0.077	0.079	0.066

Target= Mouse									
TF	LINE			SINE			Average		
	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source
CTCF	0.103	0.124	0.111	0.102	0.129	0.104	0.102	0.126	0.107
CEBPA	0.077	0.08	0.082	0.097	0.103	0.108	0.087	0.092	0.095
HNF4A	0.04	0.043	0.045	0.061	0.063	0.062	0.051	0.053	0.053
RXRA	0.027	0.029	0.026	0.041	0.042	0.035	0.034	0.036	0.03
Average	0.062	0.069	0.066	0.075	0.084	0.077	0.068	0.077	0.071

Finally, we explored model performance on test windows overlapping LINE and SINE repeat elements, following Cochran et al. [10], who reported that GRL reduces over-prediction at such sites in the mouse-to-human direction. Results are in **Table 2**. Indeed, GRL outperforms both MORALE and the source model for HNF4 α and especially RXR α in this direction. Nevertheless, MORALE outperforms GRL for CTCF and CEBPA. Furthermore, in the human-to-mouse direction, MORALE outperforms GRL for all four TFs and achieves the best average performance for three of the four.

Overall, MORALE outperforms GRL in most comparisons and rarely performs worse, demonstrating robust domain adaptation across multiple evaluation criteria.

3.3 Learning TF-Binding in Human by Leveraging Data Across Five Mammals

Next, we explored whether multi-species training combined with domain adaptation can improve TF binding prediction in human. Using liver ChIP-seq data from rhesus macaque, mouse, rat, and dog, we evaluate generalization to human test data across four TFs (FOXA1, HNF4 α , HNF6, and CEBPA), comparing three approaches: human-only training, multi-species training without adaptation, and multi-species training with MORALE. Results are in **Figure 5**. We make two observations. First, as in the two-species case, a source-only model trained on a single non-human species cannot approach human-only performance. However, a plain multi-species model trained on all five species surpasses the human-only model for all four TFs — a result not seen in the two-species

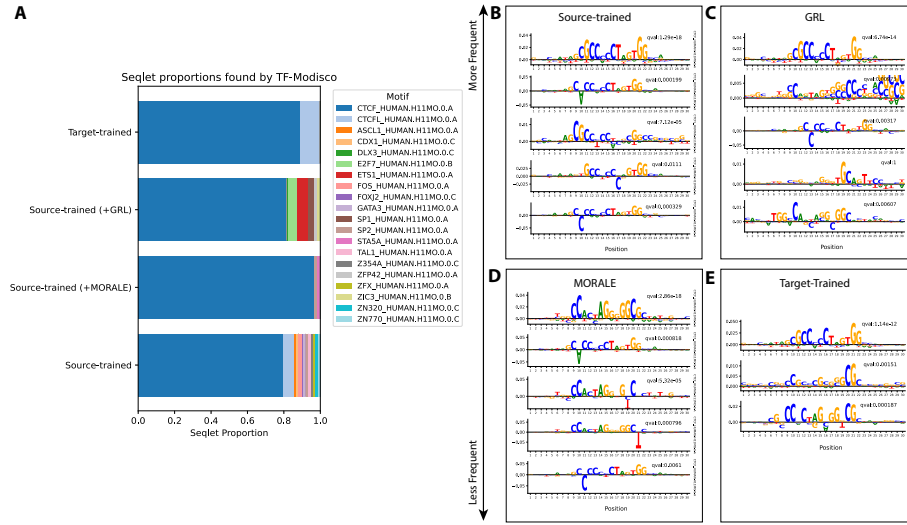


Fig. 4: MORALE discovers de-novo motifs more similar to CTCF. De-novo motifs were identified from attribution scores across source, source-adapted, and target models, using 2,000 randomly sampled bound sites for CTCF. **(A)** Proportions of all motifs found across the four models. The target model finds only CTCF (and its paralog CTCFL). Source-trained and GRL-adapted models find CTCF as their primary motif but at a lower proportion than MORALE, which reports nearly exclusively CTCF. **(B–E)** Top 5 de-novo motifs for the source-trained, GRL-adapted, MORALE-adapted, and target-trained models respectively, with TomTom match annotations and p-values. The top match of MORALE strongly resembles the established CTCF motif with a significant q-value.

setting. Second, applying MORALE to the multi-species model further increases performance, making MORALE the top performer for every TF. This demonstrates MORALE’s broad applicability to settings with more than two domains.

We then examined the contribution of individual species. The phylogenetic tree in **Figure 6** shows the evolutionary relationships between species. Based on evolutionary distance, we would expect rhesus macaque holdout to have the largest effect on human prediction, followed by mouse, rat, and dog. This expectation holds for three of four TFs (CEBPA, FOXA1, HNF4 α), where rhesus macaque holdout indeed has the largest impact. However, for HNF6 the expected trend does not hold: rat (rn7) holdout has an impact comparable to rhesus macaque, likely because rn7 is enriched for bound sites relative to the other species for this TF (see **STab. 8**).

The group holdout analysis in panel B confirms this picture. Successively holding out rhesus macaque, then mouse, then rat leads to a monotonic drop in

Fig. 5: **MORALE** attains higher performance when training on multiple source species and predicting in human as the target. Across all four TFs, MORALE consistently increases performance compared to multi-species training without domain adaptation. Unlike the two-species case, training on multiple source species allows the multi-species model to surpass the human-only model — previously unattained.

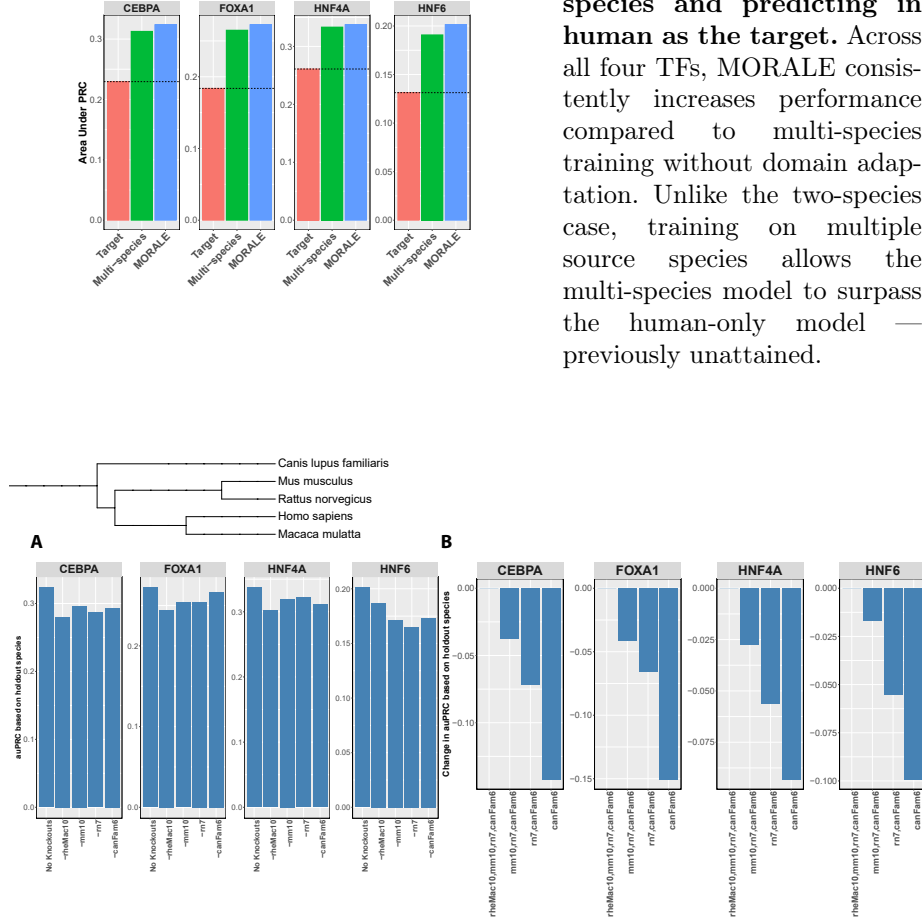


Fig. 6: **Species contributions to multi-species performance gains are species-dependent but consistently beneficial.** We quantify the effect of holding out species individually (**A**) and in groups (**B**) on model performance under MORALE, with human as the target. In (**A**), ‘No knockout’ is the model trained on all source species; each subsequent bar removes one species. In (**B**), species are removed successively from left to right, from all included to a single source species, illustrating that more source species consistently aids performance.

performance, emphasizing that each additional source species contributes to the gain over the human-only model.

4 Discussion

We develop and apply MORALE, a domain adaptation framework for cross-species TF binding prediction based on moment alignment of sequence embeddings. The core idea is to encourage a species-invariant latent representation by aligning the first and second moments of embeddings across all pairs of domains (species) during training — without requiring an adversarial component, additional parameters, or extra architectural decisions. This contrasts with gradient reversal (GRL), the prevailing approach, which requires a separate discriminator branch that substantially increases model complexity.

We evaluate MORALE in two settings. In the two-species case, we train on liver ChIP-seq data from one species (human or mouse) and predict TF binding in the other, for four TFs (CTCF, HNF4 α , RXRA, CEBPA). Here, MORALE consistently matches or outperforms GRL, and critically avoids the performance degradation below the no-adaptation baseline that GRL exhibits for several TF/direction combinations. In the five-species case, using liver data from rhesus macaque, mouse, rat, and dog to predict human TF binding, MORALE improves over both the no-adaptation multi-species model and — notably — over a human-only model, a result not observed in the two-species setting. This suggests that the combination of diverse training signal and moment alignment unlocks cross-species generalization that neither approach achieves alone.

Beyond predictive performance, MORALE recovers sequence motifs that more closely match known TF binding motifs compared to GRL, indicating that its invariant representations capture biologically meaningful features rather than superficial domain differences. MORALE’s simplicity — it integrates into any embedding-based model without modifying gradient computation — makes it broadly applicable to cross-species regulatory genomics and, more generally, to multi-domain sequence modeling tasks where domain-invariant representations are desirable.

Several limitations deserve mention. First, MORALE introduces a regularization weight λ that requires tuning; though the same is true of GRL, and optimal values varied across TFs and adaptation directions for both methods. This means performance gains may depend on access to held-out target data for validation, which may not always be available. Second, despite consistent improvements over GRL, a substantial gap between domain-adapted and target-trained models remains in the two-species setting, indicating that moment alignment alone cannot fully bridge species-specific differences when only a single source species is available — and that further methodological advances are needed. Third, all experiments use liver tissue and TFs with relatively strong, conserved binding motifs. Whether MORALE generalizes to tissues with more divergent regulatory landscapes, or to TFs with weaker sequence preferences, remains an open question.

In summary, MORALE is a simple, parameter-efficient domain adaptation method that improves cross-species TF binding prediction by aligning sequence embedding distributions across species. It is robust to the performance degradation seen with adversarial approaches, scales naturally to multiple source species,

and recovers more accurate sequence motifs — making it a practical and interpretable tool for cross-species regulatory genomics.

Acknowledgments. We thank Shaun Mahony for the helpful discussions and the processed data used in this work. This work was partly supported by the University of Florida and the University of Pittsburgh School of Medicine.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), <https://www.tensorflow.org/>, software available from tensorflow.org
2. Amemiya, H.M., Kundaje, A., Boyle, A.P.: The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**(9354), 1–5 (Jun 2019). <https://doi.org/10.1038/s41598-019-45839-z>
3. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., Kelley, D.R.: Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (Oct 2021). <https://doi.org/10.1038/s41592-021-01252-x>
4. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., Zeitlinger, J.: Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (Mar 2021). <https://doi.org/10.1038/s41588-021-00782-6>
5. Ballester, B., Medina-Rivera, A., Schmidt, D., González-Porta, M., Carlucci, M., Chen, X., Chessman, K., Faure, A.J., Funnell, A.P.W., Goncalves, A., Kutter, C., Lukk, M., Menon, S., McLaren, W.M., Stefflova, K., Watt, S., Weirauch, M.T., Crossley, M., Marioni, J.C., Odom, D.T., Flicek, P., Wilson, M.D.: Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *eLife* (Oct 2014). <https://doi.org/10.7554/eLife.02626>
6. Brennan, K.J., Weilert, M., Krueger, S., Pampari, A., Liu, H.y., Yang, A.W.H., Morrison, J.A., Hughes, T.R., Rushlow, C.A., Kundaje, A., Zeitlinger, J.: Chromatin accessibility in the *Drosophila* embryo is determined by transcription factor pioneering and enhancer activation. *Dev. Cell* **58**(19), 1898–1916.e9 (Oct 2023). <https://doi.org/10.1016/j.devcel.2023.07.007>
7. Brix, G., Durrant, M.G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G.A., King, S.H., Li, D.B., Merchant, A.T., Naghipourfar, M., Nguyen, E., Ricci-Tam, C., Romero, D.W., Sun, G., Taghibakshi, A., Vorontsov, A., Yang, B., Deng, M., Gorton, L., Nguyen, N., Wang, N.K., Adams, E., Baccus, S.A., Dillmann, S., Ermon, S., Guo, D., Ilango, R., Janik, K., Lu, A.X., Mehta, R.,

- Mofrad, M.R., Ng, M.Y., Pannu, J., Re, C., Schmok, J.C., St. John, J., Sullivan, J., Zhu, K., Zynda, G., Balsam, D., Collison, P., Costa, A.B., Hernandez-Boussard, T., Ho, E., Liu, M.Y., McGrath, T., Powell, K., Burke, D.P., Goodarzi, H., Hsu, P.D., Hie, B.: Genome modeling and design across all domains of life with evo 2. *bioRxiv* (2025). <https://doi.org/10.1101/2025.02.18.638918>, <https://www.biorxiv.org/content/early/2025/02/21/2025.02.18.638918>
8. Chen, K.M., Wong, A.K., Troyanskaya, O.G., Zhou, J.: A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.* **54**, 940–949 (Jul 2022). <https://doi.org/10.1038/s41588-022-01102-2>
 9. Chen, L., Fish, A.E., Capra, J.A.: Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLoS Comput. Biol.* **14**(10), e1006484 (Oct 2018). <https://doi.org/10.1371/journal.pcbi.1006484>
 10. Cochran, K., Srivastava, D., Shrikumar, A., Balasubramani, A., Hardison, R.C., Kundaje, A., Mahony, S.: Domain-adaptive neural networks improve cross-species prediction of transcription factor binding. *Genome Res.* **32**(3), 512–523 (Jan 2022). <https://doi.org/10.1101/gr.275394.121>
 11. Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A.H., Oteri, F., Dallago, C., Trop, E., de Almeida, B.P., Sirelkhatim, H., Richard, G., Skwark, M., Beguir, K., Lopez, M., Pierrot, T.: Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* pp. 1–11 (Nov 2024). <https://doi.org/10.1038/s41592-024-02523-z>
 12. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. *Journal of Machine Learning Research* **17**(59), 1–35 (2016), <http://jmlr.org/papers/v17/15-239.html>
 13. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., Noble, W.S.: Quantifying similarity between motifs. *Genome Biol.* **8**(2), 1–9 (Feb 2007). <https://doi.org/10.1186/gb-2007-8-2-r24>
 14. Hu, Y., Horlbeck, M.A., Zhang, R., Ma, S., Shrestha, R., Kartha, V.K., Duarte, F.M., Hock, C., Savage, R.E., Labade, A., Kletzien, H., Meliki, A., Castillo, A., Durand, N.C., Mattei, E., Anderson, L.J., Tay, T., Earl, A.S., Shores, N., Epstein, C.B., Wagers, A.J., Buenrostro, J.D.: Multiscale footprints reveal the organization of cis-regulatory elements. *Nature* **638**, 779–786 (Feb 2025). <https://doi.org/10.1038/s41586-024-08443-4>
 15. Javanmardi, M., Tasdizen, T.: Domain adaptation for biomedical image segmentation using adversarial training. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 04–07. IEEE. <https://doi.org/10.1109/ISBI.2018.8363637>
 16. Kelley, D.R.: Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.* **16**(7), e1008050 (Jul 2020). <https://doi.org/10.1371/journal.pcbi.1008050>
 17. Kempynck, N., Mahieu, L., Ekşi, E.C., Konstantakos, V., Blaauw, C., De Winter, S., Hulselmans, G., Taskiran, I., Aerts, S.: CREsted: Cis regulatory element sequence training, explanation, and design (2024)
 18. Lal, A., Gunsalus, L., Nair, S., Biancalani, T., Eraslan, G.: gReLU: A comprehensive framework for DNA sequence modeling and design. *bioRxiv* p. 2024.09.18.613778 (Sep 2024), <https://doi.org/10.1101/2024.09.18.613778>
 19. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (Apr 2012). <https://doi.org/10.1038/nmeth.1923>

20. Letunic, I.: phyloT : a phylogenetic tree generator (Mar 2025), <https://phylo.t.biobyte.de>, [Online; accessed 19. Mar. 2025]
21. Mahmood, F., Chen, R., Durr, N.J.: Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training. *IEEE Trans. Med. Imaging* **37**(12), 2572–2581 (Jun 2018). <https://doi.org/10.1109/TMI.2018.2842767>
22. Mahony, S., Edwards, M.D., Mazzoni, E.O., Sherwood, R.I., Kakumanu, A., Morrison, C.A., Wichterle, H., Gifford, D.K.: An Integrated Model of Multiple-Condition ChIP-Seq Data Reveals Predeterminants of Cdx2 Binding. *PLoS Comput. Biol.* **10**(3), e1003501 (Mar 2014). <https://doi.org/10.1371/journal.pcbi.1003501>
23. de Mathelin, A., Deheeger, F., Richard, G., Mougeot, M., Vayatis, N.: Adapt: Awesome domain adaptation python toolbox. arXiv preprint arXiv:2107.03049 (2021)
24. Pampari, A., Shcherbina, A., Kvon, E.Z., Kosicki, M., Nair, S., Kundu, S., Kathiria, A.S., Risca, V.I., Kuningas, K., Alasoo, K., Greenleaf, W.J., Pennacchio, L.A., Kundaje, A.: ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. *bioRxiv* p. 2024.12.25.630221 (Jan 2025), <https://doi.org/10.1101/2024.12.25.630221>
25. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv (Dec 2019). <https://doi.org/10.48550/arXiv.1912.01703>
26. Patel, A., Singhal, A., Wang, A., Pampari, A., Kasowski, M., Kundaje, A.: DART-Eval: A Comprehensive DNA Language Model Evaluation Benchmark on Regulatory DNA. arXiv (Dec 2024). <https://doi.org/10.48550/arXiv.2412.05430>
27. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
28. Perez, G., Barber, G.P., Benet-Pages, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, C.M., Nassar, L.R., Raney, B.J., Speir, M.L., van Baren, M.J., Vaske, C.J., Haussler, D., Kent, W.J., Haeussler, M.: The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* **53**(D1), 1243–1249 (Jan 2025). <https://doi.org/10.1093/nar/gkae974>
29. Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., Talianidis, I., Flicek, P., Odom, D.T.: Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science* **328**(5981), 1036–1040 (Apr 2010). <https://doi.org/10.1126/science.1186176>
30. Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S., Kundaje, A.: Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. arXiv (Oct 2018). <https://doi.org/10.48550/arXiv.1811.00416>
31. Sun, B., Saenko, K.: Deep CORAL: Correlation Alignment for Deep Domain Adaptation. arXiv (Jul 2016). <https://doi.org/10.48550/arXiv.1607.01719>

A Appendix

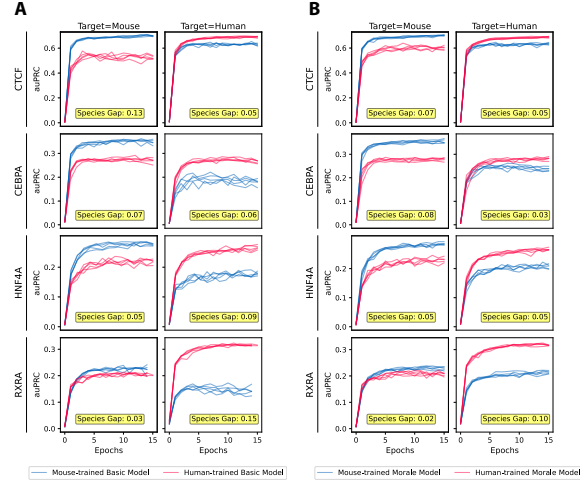


Fig. 7: The species gap is closed during training when using MORALE. We display the training, in the two-species case, of the five-fold cross validation performance over 15 epochs between the basic model, and MORALE. The gap between the peak performance between the source-on-target and the target-on-target models are annotated at the bottom of each plot. In (A) we have the basic model across each TF-target pair, and in (B) we display the same information, but with MORALE.

Table 3: The confusion matrix for mouse-adapted models when evaluating on the test set (Chr2) from human. The table displays the percentage of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs) for each TF. They were calculated as a ratio over all the windows in the test set and attach the auPRC value based on model type.

TF	TPs (%)			FPs (%)			TNs (%)			FNs (%)			auPRC		
	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE
CTCF	0.401	0.413	0.435	1.397	1.847	1.384	98.122	97.672	98.135	97.625	0.08	0.068	0.079	0.046	0.038
CEBPB	0.547	0.542	0.545	0.501	11.938	9.77	87.472	88.974	89.64	93.753	0.04	0.048	0.045	0.080	0.183
ENHFA	0.574	0.605	0.616	0.578	10.806	9.137	8.406	6.716	88.452	90.702	0.068	0.136	0.164	0.126	0.157
HNF4A	0.635	0.635	0.635	10.806	9.137	8.406	88.452	90.702	90.376	86.734	0.068	0.468	0.566	0.237	0.184
HMGB1	0.776	1.163	1.055	1.364	13.818	8.023	11.665	85.218	88.413	90.376	0.068	0.468	0.566	0.237	0.184

Table 4: The confusion matrix for human-adapted models when evaluated on the test set (Chr2) from mouse. The table displays the percentage of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs) for each TF. We include them as a ratio over all the windows in the test set and attach the auPRC value based on model type.

TF	TPs (%)			FPs (%)			TNs (%)			FNs (%)			auROC		
	Source	GRL	MoRALE Target	Source	GRL	MoRALE Target	Source	GRL	MoRALE Target	Source	GRL	MoRALE Target	Source	GRL	MoRALE Target
CITFC	0.697	0.699	3.591	3.457	1.514	95.666	96.32	97.743	0.946	0.942	0.044	0.528	0.56	0.271	
CECBA	0.788	0.786	1.126	7.156	6.387	92.914	91.518	92.532	0.935	0.928	0.196	0.757	0.721	0.482	
CECBA	0.952	0.946	5.732	11.601	10.388	89.252	88.027	88.666	0.9635	0.965	0.188	0.733	0.735	0.263	
CECBA	0.833	0.828	0.824	17.801	16.659	79.873	83.043	80.469	0.937	0.932	0.037	0.217	0.235	0.235	
CECBA	0.831	0.827	0.763		8.612				0.885	0.891	0.155	0.263	0.263	0.223	

Table 5: We compare across all repeat types (with at least 500 instances in our test (human) chromosome, Chr2) in the source (mouse) genome between the two mouse-adapted models. The last row is the average auPRC across all repeat types

TF	DNA		LINE		Low complexity		LTR		Simple repeat		SINE		Unknown		Average	
	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL
CITE	0.047	0.053	0.044	0.039	0.044	0.041	0.031	0.033	0.045	0.047	0.04	0.04	0.052	0.055	0.043	0.052
CEPFA	0.056	0.056	0.044	0.043	0.044	0.049	0.033	0.036	0.062	0.059	0.064	0.023	0.087	0.109	0.105	0.095
INFA	0.073	0.078	0.055	0.058	0.057	0.07	0.047	0.053	0.086	0.085	0.082	0.029	0.11	0.139	0.109	0.074
Average	0.067	0.068	0.06	0.072	0.077	0.066	0.048	0.055	0.087	0.086	0.079	0.064	0.122	0.177	0.171	0.083
															0.096	0.096

Table 6: We compare across all repeat types (with at least 500 instances in our test (mouse) chromosome, Chr2) in the source (human) genome between the two human-adapted models. The last row is the average auPRC across all repeat types

TP	DNA			LINE			Low complexity			LTR			Simple repeat			SINE			Unknown			Average		
	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source	GRL	MORALE	Source
CTCF	0.151	0.18	0.168	0.103	0.124	0.111	0.128	0.148	0.133	0.086	0.108	0.097	0.121	0.147	0.127	0.102	0.129	0.104	0.122	0.156	0.138	0.116	0.142	0.125
EBVNA	0.101	0.115	0.121	0.077	0.08	0.082	0.084	0.08	0.092	0.073	0.078	0.083	0.086	0.091	0.094	0.097	0.103	0.108	0.096	0.106	0.101	0.089	0.1	0.106
EBVNA	0.086	0.093	0.095	0.063	0.065	0.065	0.065	0.065	0.065	0.063	0.063	0.063	0.063	0.063	0.063	0.061	0.063	0.063	0.063	0.063	0.063	0.061	0.061	0.063
EBVNA	0.084	0.090	0.093	0.077	0.079	0.079	0.077	0.085	0.083	0.078	0.081	0.077	0.088	0.088	0.083	0.081	0.082	0.085	0.086	0.087	0.088	0.084	0.084	0.088
Average	0.09	0.102	0.1	0.062	0.069	0.066	0.08	0.087	0.082	0.057	0.066	0.063	0.076	0.084	0.078	0.075	0.084	0.077	0.091	0.111	0.105	0.076	0.086	0.092

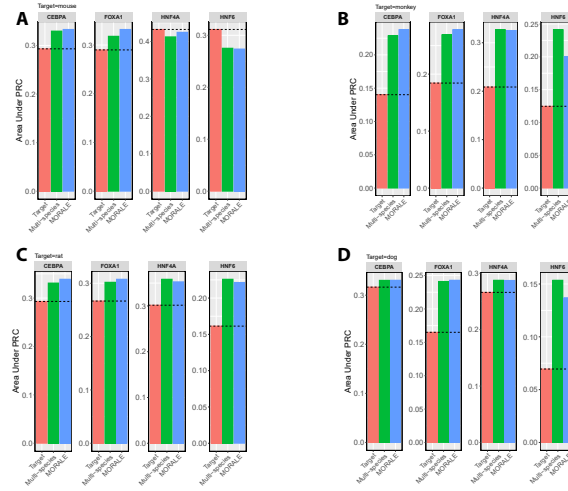


Fig. 8: We display performance across all targets in the multi-species case. In (A) we show the performance across all 4 transcription factors when the target is mouse, (B) displays monkey, (C) displays rat, and (D) display dog.

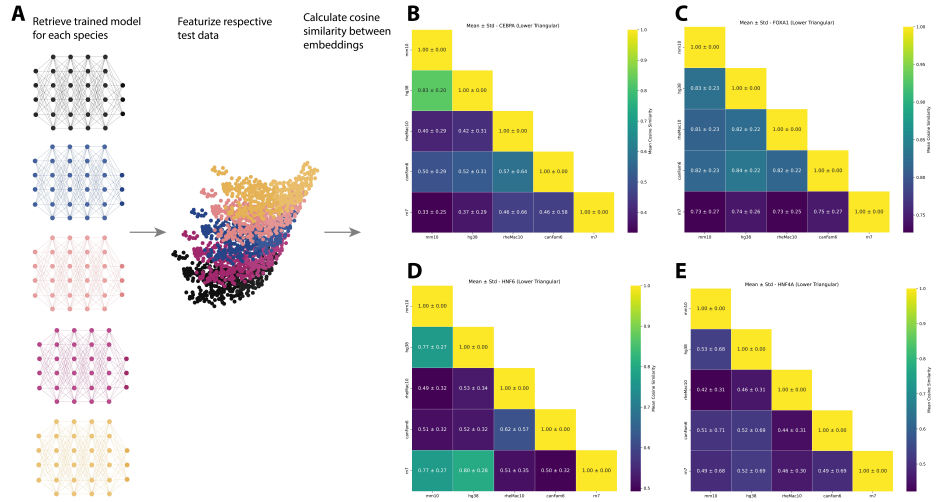


Fig. 9: We construct heatmaps based to understand relatedness of learned embeddings through the different species models. We do so for each TF under study in the multi-species case. For each TF we use the models trained to predict in each species and run the test data through the feature extractor in our models to capture the embeddings. In (B) we show the lower triangular for CEBPA, (C) FOXA1, (D) HNF6, and (E) HNF4A.

Table 7: **The binding site information for the four transcription factors used in the two-species case.** The following quantities are listed: the number of peaks called across the entire genome; the number of called peaks within the filtered window set, merged if within 500 bp of each other; the number of windows in the filtered window set labeled bound due to peak overlap; the fraction of the filtered window set labeled bound; and the database accession ID (ENCODE, GEO, or ArrayExpress). The size of the filtered window sets for the mouse and human genomes were 41883806 and 48742577, respectively.

TF	Species	Raw Peaks	Filtered Peaks	Bound Windows	Frac. Bound	Accession ID
CTCF	Mouse	32006	28943	296117	0.71%	ENCSR000CBU
	Human	29067	26477	270100	0.55%	ENCSR911GFJ
CEBPA	Mouse	62636	48812	566945	1.35%	E-TABM-722
	Human	32243	28545	298066	0.61%	E-TABM-722
HNF4A	Mouse	44800	36540	415846	0.99%	E-TABM-722
	Human	42766	34714	387077	0.79%	E-TABM-722
RXRA	Mouse	46443	33751	404284	0.97%	GSM1299600
	Human	95085	71032	854289	1.75%	ENCSR098XMN

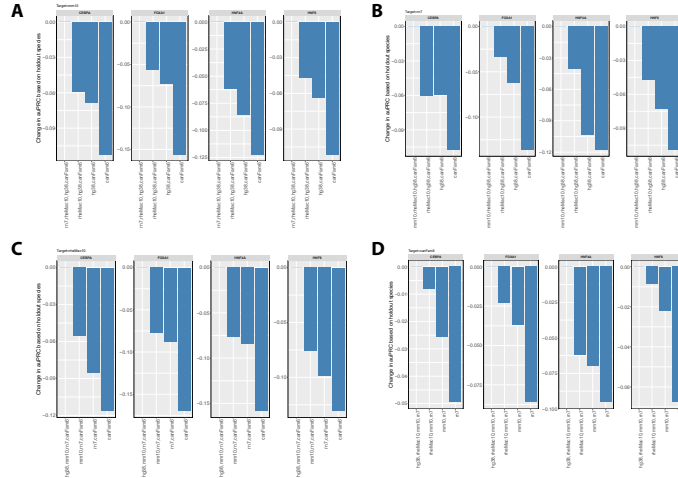


Fig. 10: We include group holdouts under each species (other than human) as the target species.

Table 8: **The binding site information for the four transcription factors used in the multi-species case.** The following quantities are listed: the number of peaks called across the entire genome; the number of called peaks within the filtered window set, merged if within 1000 bp of each other; the number of windows in the filtered window set labeled bound due to peak overlap; the fraction of the filtered window set labeled bound; and the database accession ID (ArrayExpress).

TF	Species	Raw Peaks	Filtered Peaks	Bound Windows	Frac. Bound	Accession ID
CEBPA	Mouse	50263	32751	830115	1.80%	E-MTAB-1509
	Human	34253	26749	615953	1.16%	E-MTAB-1509
	Rhesus Macaque	11600	9985	214440	0.40%	E-MTAB-1509
	Dog	44749	32816	780102	1.77%	E-MTAB-1509
	Rat	50851	37010	900363	1.84%	E-MTAB-1509
FOXA1	Mouse	66728	38683	1071971	2.32%	E-MTAB-1509
	Human	36454	27406	651070	1.22%	E-MTAB-1509
	Rhesus Macaque	30546	22421	532725	1.00%	E-MTAB-1509
	Dog	24316	18151	436461	0.99%	E-MTAB-1509
	Rat	59983	37940	993292	2.02%	E-MTAB-1509
HNF4A	Mouse	135057	54343	1762041	3.82%	E-MTAB-1509
	Human	50611	34022	856878	1.61%	E-MTAB-1509
	Rhesus Macaque	32331	21628	535077	1.01%	E-MTAB-1509
	Dog	69264	37839	1049132	2.38%	E-MTAB-1509
	Rat	52694	33640	891098	1.82%	E-MTAB-1509
HNF6	Mouse	57255	38899	966248	2.09%	E-MTAB-1509
	Human	17021	14378	311320	0.59%	E-MTAB-1509
	Rhesus Macaque	9425	8238	174525	0.33%	E-MTAB-1509
	Dog	9283	7687	168142	0.38%	E-MTAB-1509
	Rat	22686	18058	407416	0.83%	E-MTAB-1509

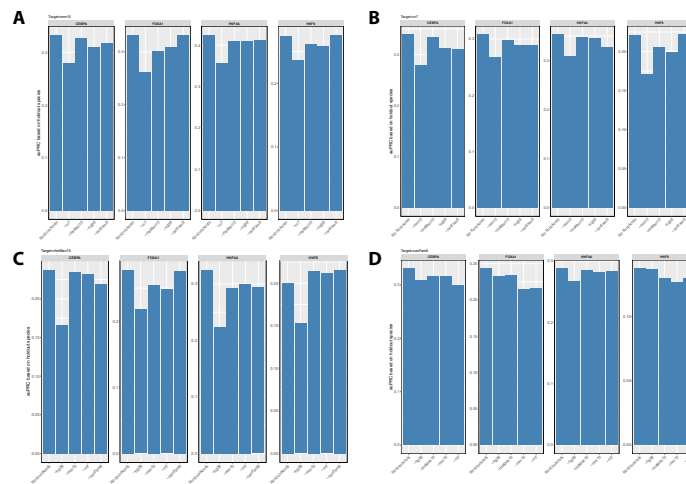


Fig. 11: We include per species holdout under each species (other than human) as the target species.