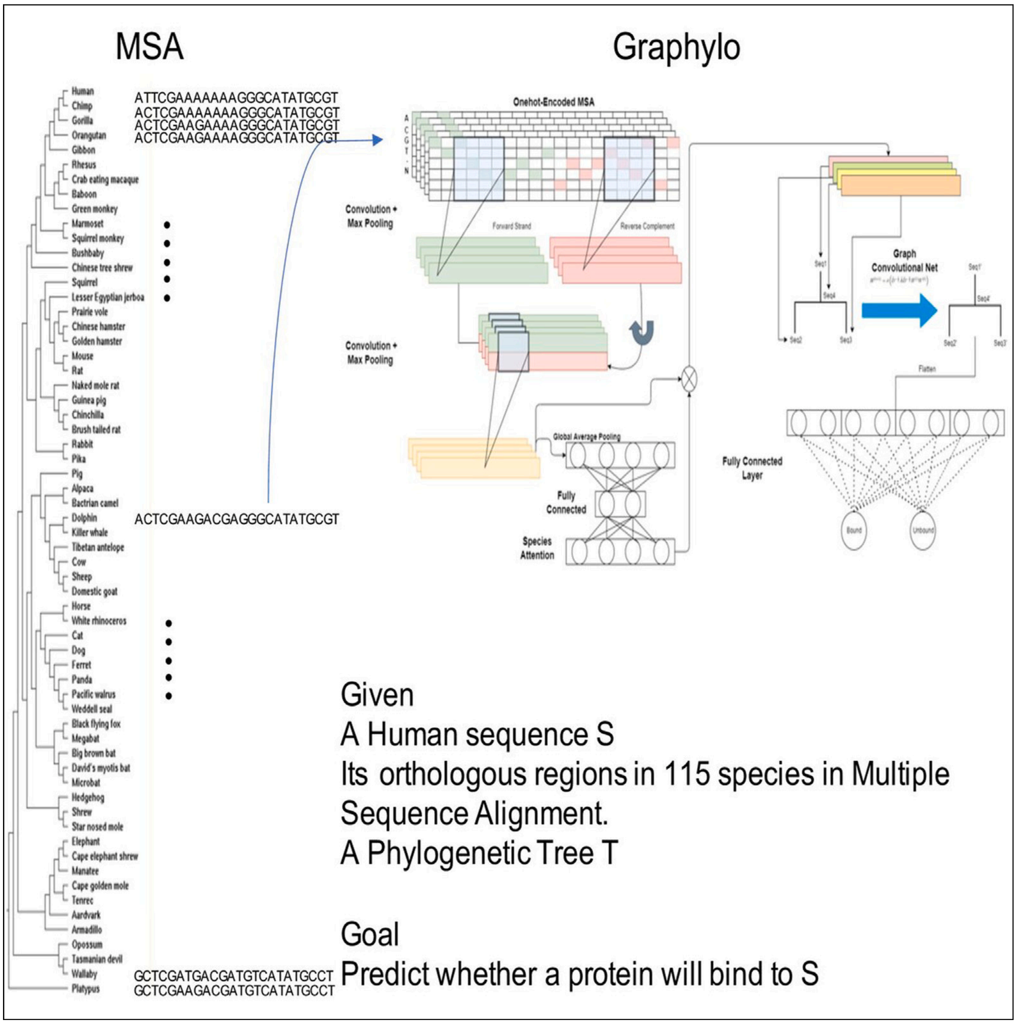
iScience

CellPress
OPEN ACCESS

Article

Graphylo: A deep learning approach for predicting regulatory DNA and RNA sites from whole-genome multiple alignments



Given
A Human sequence S
Its orthologous regions in 115 species in Multiple Sequence Alignment.
A Phylogenetic Tree T

Goal
Predict whether a protein will bind to S

Dongjoon Lim,
Changhyun Baek,
Mathieu
Blanchette

dong.lim@mail.mcgill.ca (D.L.)
blanchem@cs.mcgill.ca (M.B.)

Highlights

Graphylo merges convolutional, graph convolutional networks in genome alignments

It shares info across species, leveraging deep multi-genome alignments effectively

It extracts orthologous/ ancestral sequences without assuming fixed positions

Its ability to learn evolutionary constraints offers nuanced insights into mutations

Lim et al., iScience *27*, 109002
February 16, 2024 © 2024 The Authors.
https://doi.org/10.1016/j.isci.2024.109002

iScience

CellPress
OPEN ACCESS

Article

Graphylo: A deep learning approach for predicting regulatory DNA and RNA sites from whole-genome multiple alignments

Dongjoon Lim,[1,*] Changhyun Baek,[1] and Mathieu Blanchette[1,2,*]

SUMMARY

This study focuses on enhancing the prediction of regulatory functional sites in DNA and RNA sequences, a crucial aspect of gene regulation. Current methods, such as motif overrepresentation and machine learning, often lack specificity. To address this issue, the study leverages evolutionary information and introduces Graphylo, a deep-learning approach for predicting transcription factor binding sites in the human genome.

Graphylo combines Convolutional Neural Networks for DNA sequences with Graph Convolutional Networks on phylogenetic trees, using information from placental mammals' genomes and evolutionary history. The research demonstrates that Graphylo consistently outperforms both single-species deep learning techniques and methods that incorporate inter-species conservation scores on a wide range of datasets. It achieves this by utilizing a species-based attention model for evolutionary insights and an integrated gradient approach for nucleotide-level model interpretability. This innovative approach offers a promising avenue for improving the accuracy of regulatory site prediction in genomics.

INTRODUCTION

Transcription factors (TFs) regulate gene expression by binding to specific genomic locations and interacting directly or indirectly with the transcriptional machinery to modulate gene expression levels. TF binding sites (TFBS) are critical components of gene regulatory networks. TFBS are defined by short sequence motifs (6–15 bp), often with relatively low information content and significant degeneracy at specific positions. Hence TF binding is determined not only by the relatively simple rules of affinity for a given sequence pattern but also by a number of other factors including (1) the epigenetic state of the DNA region (e.g., chromatin accessibility, DNA methylation, 3D genome organization) and (2) the presence of other TFs bound to neighboring sites and resulting in cooperative or competitive interactions.

Similarly, interactions between RNA binding proteins (RBPs) and their RNA targets play a vital role in post-transcriptional regulation, including pre-RNA splicing, mRNA sub-cellular localization, stability, and translation.[1–3]

Chromatin immunoprecipitation followed by sequencing (ChIP-Seq)[4] identifies genomic locations bound by a given TF in a given cell type. It involves isolating cross-linked chromatin complexes with an antibody, followed by high-throughput sequencing of the DNA fragments. CLIP-Seq (cross-linking immunoprecipitation and high-throughput RNA sequencing)[5] experiments and their variants, including PARCLIP,[6] HITS_CLIP,[7] and ICLIP,[8] enable the identification and characterization of RNA binding sites, shedding light on the regulatory roles of RBPs in post-transcriptional processes. These experiments have a limited resolution (typically 100–200 bp), although more recent variants (e.g., ChIP-exo[9]) yield a finer resolution. Obtaining these experimental data can be challenging, but it serves as valuable training data for computational models aimed at predicting transcription factor binding sites (TFBS) or RBP-RNA interaction *in silico*. These sequence-analysis models are important because they can be used to gain nucleotide-level resolution when it comes to identifying regulatory functional sites. These methods can also yield insights into the sequence specificity of TFs and the architecture of multi-TFBS regions such as promoters and enhancers. Computational predictors are also key to interpreting the potential consequences of sequence variants.[10]

TFBS prediction was one of the first problems considered in bioinformatics, with early work on position-weight matrices (PWMs) dating back to.[11] Traditional approaches for TFBS have mostly relied on motif overrepresentation approaches.[12,13] However, in recent years, machine learning approaches based on neural networks trained on ChIP-Seq peak datasets have successfully predicted TFBS. These approaches leverage a variety of neural network architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs effectively extract useful data from noisy sources, whereas RNNs can detect patterns in time-series data that depend on nearby context. One of the earliest deep learning models for TFBS prediction was DeepBind,[14] which uses CNNs to identify the sequence specificity of DNA and

[1]McGill University, Montreal, QC, Canada
[2]Lead contact
*Correspondence: dong.lim@mail.mcgill.ca (D.L.), blanchem@cs.mcgill.ca (M.B.)

https://doi.org/10.1016/j.isci.2024.109002

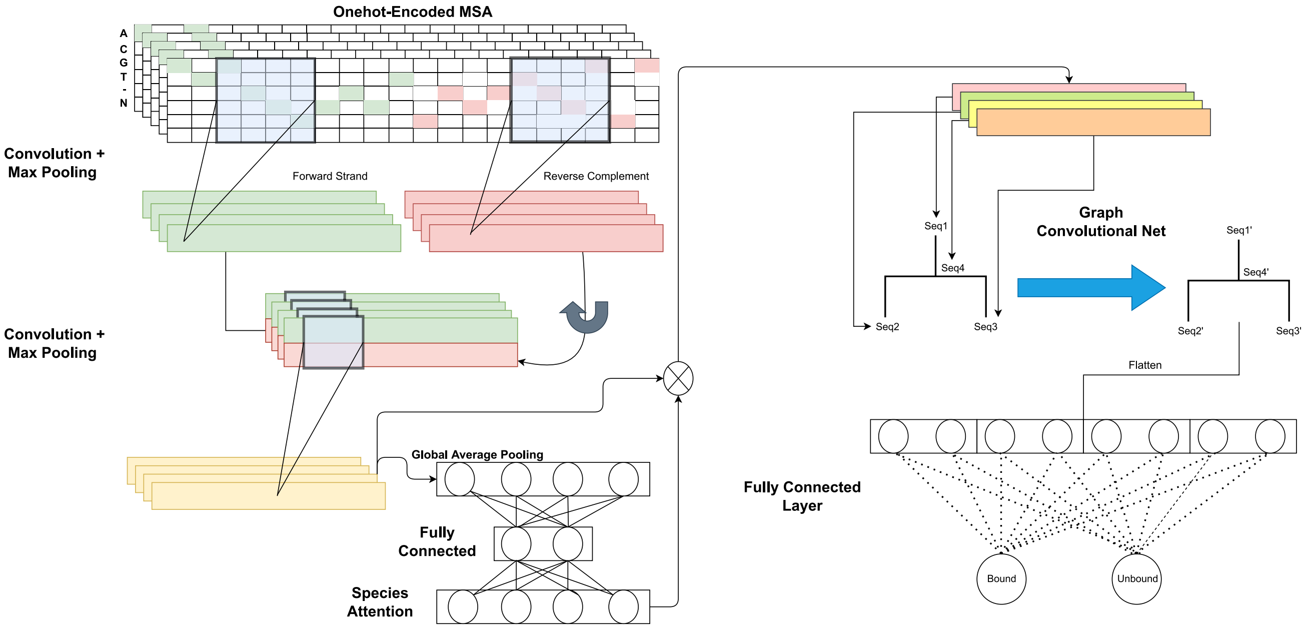


Figure 1. Overview of Graphylo's architecture

For a given 201 bp human sequence and its orthologous and inferred ancestral sequences in other placental mammals, the forward strand and the reverse complement are one-hot encoded and fed into the shared weight CNN layer with 2D filters separately. The output of the reverse complement CNN is flipped horizontally and is stacked with the output from the forward strand CNN. The stacked output is fed into the second CNN layer where the filter is expected to have the potential to capture binding sites on both strands. The output of the second CNN layer is used to generate a species attention vector that serves as a scaling factor for the CNN output. Scaled representation is subsequently fed to the Graph Convolutional layer where the weighted sum is taken from the output that corresponds to species that are connected by edges. The output of the GCN is then flattened out and fed to the fully connected layer to make the final prediction of whether the TF binds to the input sequence or not.

RNA binding proteins.[15] More recent models take advantage of combining CNNs with RNNs or attention mechanisms to account for long-range dependencies and are trained on longer input sequences. FactorNet[16] is a model that has performed well in the ENCODE-DREAM *in vivo* TFBS Prediction Challenge.[17] This model uses a siamese CNN stacked with a long short-term memory (LSTM) network.[18,19] Similar models exist for predicting RNA-RBP interactions.[20–23]

Purely sequence-based approaches have been shown to lack specificity.[16] Evolutionary information can yield important clues about sequence function[24] and has long been combined with other types of sequence-based analyses to improve the detection of functional sites. Phylogenetic footprinting[25,26] is an approach that seeks to detect functional TFBS by combining motif discovery with an analysis of inter-species sequence conservation. Multiple whole-genome alignments (MWGAs) are commonly used to compare the genomes of different species in order to identify similarities and differences between them. Multiz[27] and Cactus[28] have proved effective at aligning large numbers of mammalian genomes, including both coding and non-coding regions. These alignments can be used to accurately infer ancestral genomic sequences.[29,30] Methods like PhastCons[31] and PhyloP[32] use these alignments to measure the level of inter-species sequence conservation of a given region or site. High levels of sequence conservation are evidence of negative selection and hence of function. Machine learning and deep learning models have been shown to be able to take sequence conservation information into account to successfully identify regulatory regions in the genome.[33–37]

With the potential of linking evolutionary trace with functionality, an increasing number of deep learning models in bioinformatics are starting to utilize multiple sequence alignments as an input to take advantage of deep learning models' ability to represent data naturally. For example, Alphafold2[38] uses multiple alignments of amino acid sequences to predict protein structure. For DNA sequences, there are few methods that used multiple sequence alignments including[33,39] but extracting features from the multiple sequence alignment within the context of evolution remains challenging.

Here, we introduce Graphylo, a deep learning model that operates on a set of orthologous placental mammal sequences and their computationally inferred ancestral sequences. Graphylo combines a CNN for individual sequence representation with Graph Convolutional Networks (GCNs)[40] to integrate information across species along the branches of a given phylogenetic tree (see Figure 1). This allows incorporating functional information without sacrificing the flexibility and scalability of sequence-based models. As far as we know, Graphylo is the first deep learning predictor to utilize multiple DNA sequence alignments for TFBS prediction. Here, we show that Graphylo outperforms state-of-the-art models for TFBS prediction. We also analyze how information from different species contributes to improving prediction accuracy in human.

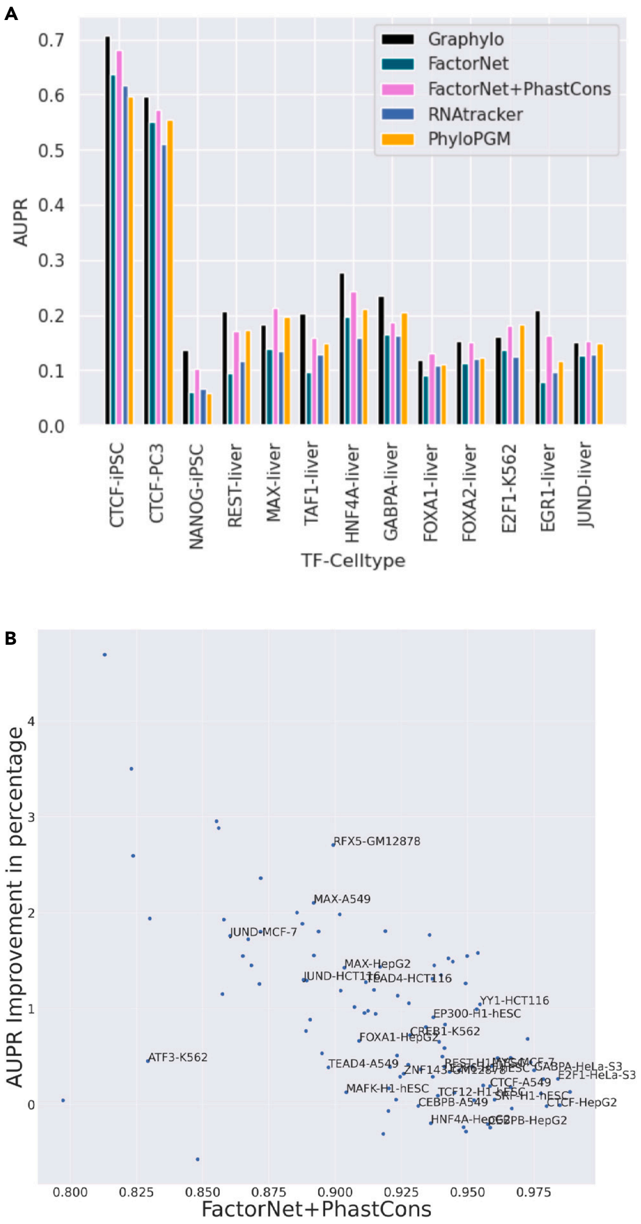


Figure 2. Performance comparison with different models

(A) Performance (area under the precision-recall curve) comparison of Graphylo against other predictors (RNAtracker and FactorNet) on 13 within-cell type ENCODE DREAM benchmarking datasets.

(B) Performance (area under the precision-recall curve) gain of Graphylo compared with the FactorNet+PhastCons model on 124 dataset.

RESULTS

Predictive performance

We evaluated Graphylo against several state-of-the-art machine learning models that operate on single-species data: (1) FactorNet,[16] a top-performing single-species TFBS predictor in a recent DREAM competition[17] but limited to using only sequence information; (2) FactorNet+phastCons, a modified version of FactorNet that uses as input both the human sequence of interest and its interspecies conservation level (average PhastCons score); (3) RNATracker,[41] a tool initially designed for RNA subcellular localization analysis but also effective at TFBS prediction; and (4) PhyloPGM applied to FactorNet prediction values,[39] a probabilistic graphical model that combines predictions made on orthologous and ancestral sequences to improve a base FactorNet's accuracy. Performance was measured using area under the receiver operating characteristic curve (AUROC) for the RBP binding task and the area under the precision-recall curve (AUPR) for the TF binding task on a held-out test set that were not used for training. The choice of AUPR as the primary evaluation metric

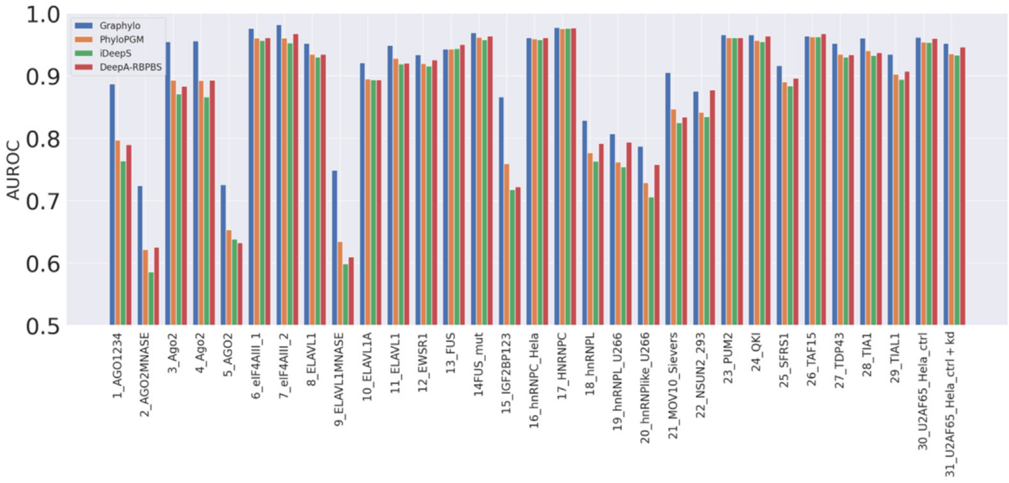


Figure 3. RBP binding site prediction performance comparison

Model performance comparison for the RNA binding protein binding site prediction problem.

for the TF binding task was driven by the significant class imbalance observed in this dataset, with less than 1% positive examples. Considering this class imbalance, we determined AUPR to be an appropriate metric as it focuses on precision and recall, which are crucial measures when dealing with rare positive samples. AUPR provides a more accurate assessment of our model's performance in this specific TF binding task.

Graphylo outperforms all baseline predictors on 9 of the 13 different TF/cell-type pairs introduced in the DREAM challenge and 29 of the 31 different RBP binding datasets (see Figures 2 and 3, as well as Figure 4). This suggests an overall superior performance for Graphylo with Wilcoxon signed-rank test p value of 1.81×10^{-2} and 3.54×10^{-6}, against the second-best predictor for TF prediction (FactorNet+PhastCons) and RBP binding prediction (DeepA-RBPBS), respectively. Note that the results of FactorNet presented here differ from those in their original paper[16] as we only use sequence information, whereas Quang and Xie (2019) used both sequence and chromatin accessibility data (DNase-Seq).[16]

We also evaluated the different predictors on a larger set of 124 balanced datasets derived from ChIP-seq experiments on various TF/cell-type pairs. Graphylo outperforms the two single-sequence predictors FactorNet and RNATracker on 117 and 122 of the 124 datasets (see Figure 5). Graphylo also outperforms FactorNet+PhastCons on 112 datasets. This again supports an overall superior

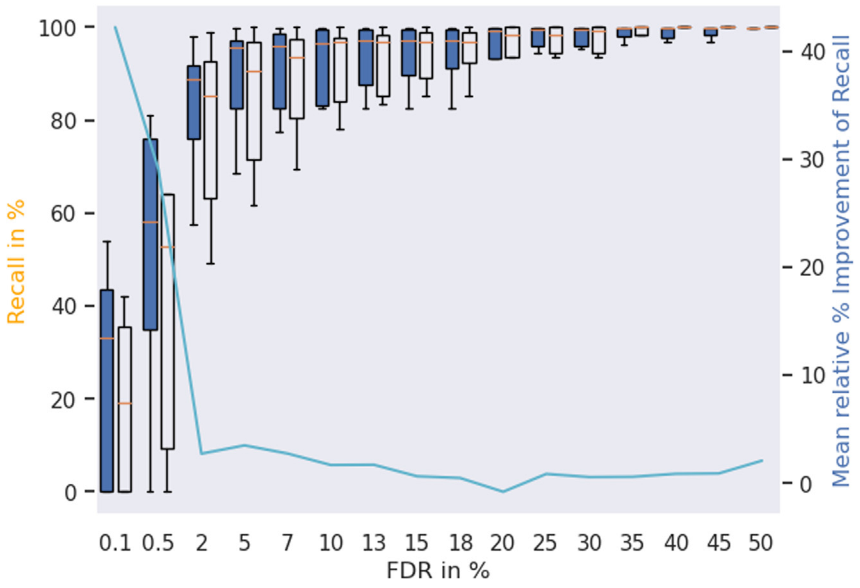


Figure 4. Recall improvements over different FDR thresholds

This figure displays the results of 124 reference training datasets, respectively. This figure shows the mean relative percentage improvement of Graphylo test recall score over FactorNet+PhastCons for 17 different false discovery rate thresholds (blue line). The boxplot shows the summary of recall for Graphylo (blue) and FactorNet+PhastCons (white).

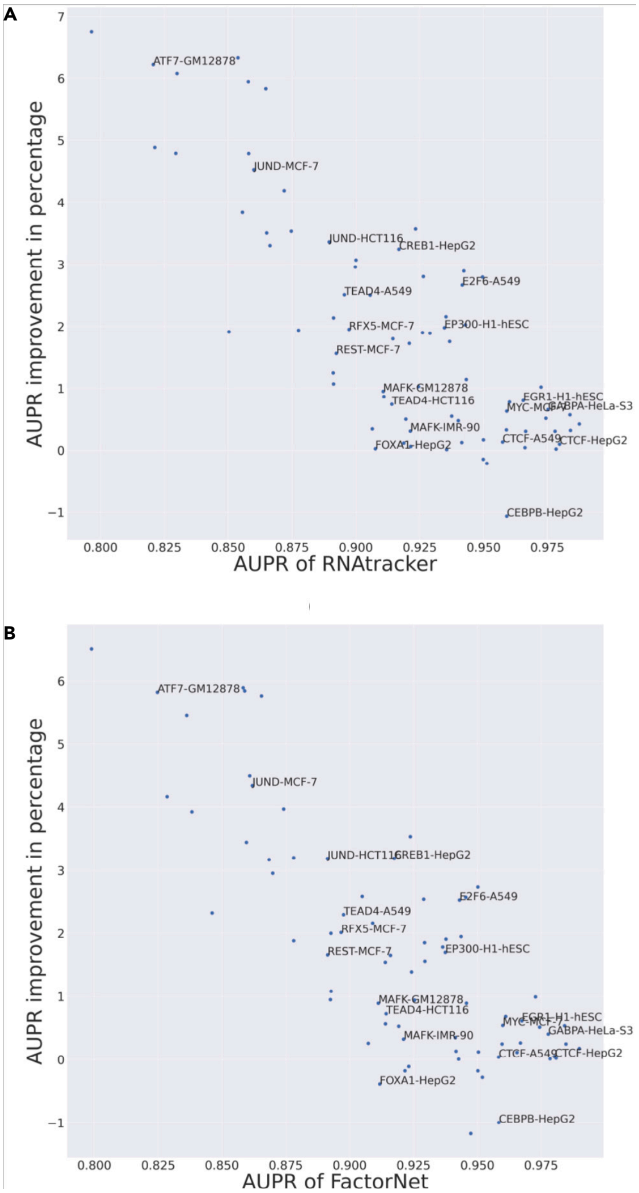


Figure 5. Performance comparison in scattered plot

Graphylo's AUPR improvement over baseline models [RNAtracker (A) and FactorNet (B)] in 124 balanced test sets. Each dot represents a TF/cell-type pair.

prediction performance for Graphylo (Wilcoxon signed-rank test p value of 2.73 × 10^{-6}). Observed from Figures 2A and 2B, the gains provided by Graphylo are particularly striking (up to 6.5%) for TF-cell-type pairs where single-species models perform poorly, while comparatively more modest (though still significant in most cases) for datasets where single-species models are more robust (e.g., for CTCF or GABPA).

We then asked whether the performance gains obtained with Graphylo depend on the level of sequence conservation (PhastCons score) of the human sequence used as input. We observe that Graphylo's benefits compared with FactorNet are highest in regions that exhibit relatively high levels of conservation (Figure 6A, whereas the other two models utilizing evolutionary information (PhyloPGM, FactorNet+PhastCons) do not improve as significantly as a function on sequence conservation levels. Graphylo also exhibits a consistent trend in RBP-RNA binding prediction (Figure 6B, wherein the performance gain is more pronounced in highly conserved regions compared with its baseline counterpart, which is solely trained on human sequence data).

As the highly imbalanced nature of the TFBS prediction task typically leads to a high false discovery rate, it is important to evaluate predictive models based on their recall at low false discovery rates values. Figure 4 shows that Graphylo has a significantly higher recall value at low FDR thresholds compared with its competitor.

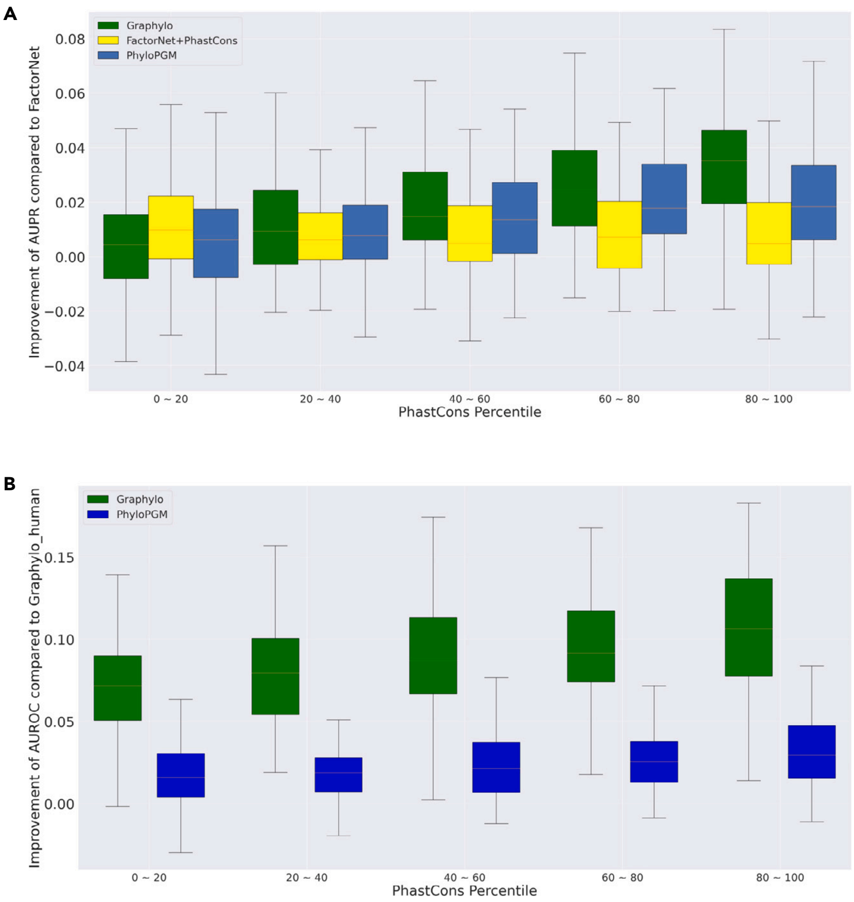


Figure 6. Performance improvement with different conservation level

Performance improvement of different models with different levels of sequence conservation.

(A) Graphylo, Factornet+PhastCons, and PhyloPGM compared with the baseline FactorNet model for TF binding prediction.

(B) Graphylo and PhyloPGM compared with the baseline Graphylo model trained only on the human sequence.

Species attention weights

We used the attention weights learned by Graphylo to gain insights into the role played by different species in the model's predictions. The attention weights allow for the relative importance of different species to be determined, and the values learned during training can be utilized to identify which species have the greatest influence on the final prediction. Figure 7 illustrates the average per-species attention values learned by Graphylo across different TF-cell-type combinations. Although Graphylo is not "told" that its goal is to make a prediction that pertains to the human sequence, it learns to assign the highest attention to that sequence (and to the human-chimp ancestral sequence, which is nearly identical).

The figure illustrates three distinct patterns of attention allocation. The first pattern, located on the left side of the heatmap, showcases datasets where the majority of modern species and ancestors receive notable attention values, indicating that the predictor is utilizing information from the entire tree. The second cluster, situated in the middle of the figure, displays TF-cell-type combinations where attention is primarily focused on the human and human-chimp ancestral sequences, with little to no attention allocated to other species. The final cluster (right portion of the heatmap) reveals datasets where the majority of modern sequences are given attention, whereas most ancestral sequences are disregarded.

Comparison of prediction with ChIP-Seq signals

Figure 8A illustrates the Graphylo predictions for CTCF (one of the TFs for which Graphylo's performance is the best) on various cell lines, compared with ChIP-Seq signal for a small portion of human chromosome 1 that was not used for training. The results confirm that Graphylo's CTCF binding predictions align well with experimental data and that these predictions display a good degree of cell-type specificity. Figures 8B and 8C show similar results for ATF3 and JunD.

Moreover, there are regions with high Graphylo predictions but low ChIP-Seq signal, indicating potential false-positive predictions of binding sites. Nevertheless, these regions could be of interest for further exploration, as the ChIP-Seq experiment may not be able to identify all binding sites perfectly.

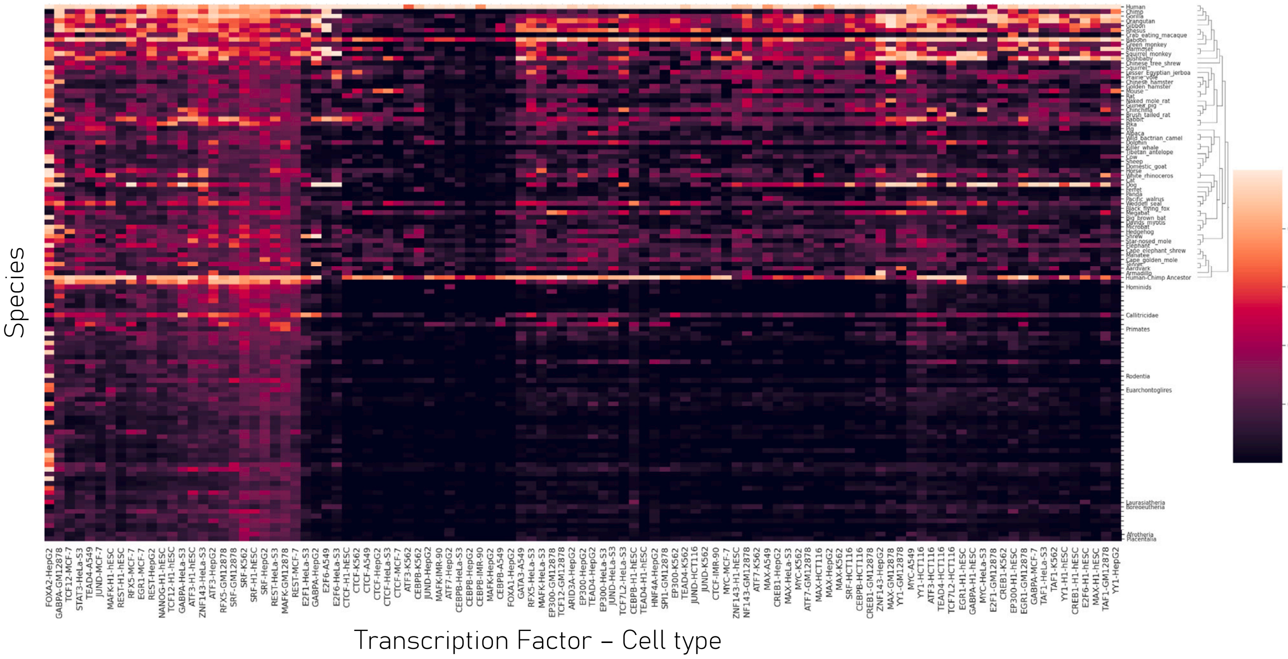


Figure 7. Attention weight visualization

This figure shows the species attention weights learned by Graphylo. The x axis represents TF-cell-type pairs, and the species in the alignment. Each vertical line in the figure represents the averaged attention vectors assigned to different species, over the test set, for a specific TF-cell-type combination. The darker colors indicate lower attention values and the lighter violet colors indicate higher attention values. The attention vectors are clustered with single-linkage hierarchical clustering with Euclidean distance.

Integrated gradients

To understand which portions of the input sequences are important for predicting transcription factor binding sites, we used integrated gradient (IG)[42] values to analyze specific examples (Figure 9). Both panels show the same genome region, which is known to be bound by TF CEBPB in both HCT116 (top) and HepG2 (bottom) and on which the corresponding Graphylo models makes positive predictions. The figure illustrates which nucleotides from the human and orthologous/ancestral sequences had the most impact on the binding prediction. This enables several observations. First, the portions of the input that get assigned the largest IG values are typically in human. This is expected because human sequences are the most relevant to binding predictions for that species. However, the fact that Graphylo, which is agnostic to the fact that the training data were generated for human, successfully assigns importance to human sequences indicates that the model architecture is effective at learning species relevance.

Second, in the CEBPB-HepG2 results, we can see that the consensus motif for the CEBPB binding site (ATTGCGCAAT) is assigned large IG values in humans and its recent ancestors. On the contrary, the largest IG values in HCT116 (a colorectal cancer cell line) are assigned to another portion of the input sequence, which happens to match the consensus binding site of the heterodimer C/EBP:AP-1 (TGACGTCA or TGACTCA). This suggests that the Graphylo model trained on that cell line's data has learned to recognize that heterodimer's signature, rather than the individual CEBPB binding site itself. This is consistent with the fact that in colorectal cancer, AP-1 activity is stimulated through the activation of upper-stream signaling pathways such as mitogen-activated protein kinases or the Wnt/Wingless signaling pathway. Studies have shown that AP-1 can promote the growth and proliferation of colorectal cancer cells as well as increase resistance to chemotherapy.[43,44]

Running time

The amount of time required to train Graphylo depends on the size of the training data and the hyperparameters of the model. The study's results were obtained by training models with a batch size of 256. The training process generally requires 20–50 epochs to reach the minimum validation loss, and the largest dataset (CEBPB-HepG2, with 714,543 training examples) required 28 min per epoch using an Nvidia Quadro RTX 6000 GPU. In terms of inference, predicting transcription factor binding on a single human region of 201 nucleotides takes approximately 20 ms. To make genome-wide predictions with a 50 bp stride, it would take approximately 334 h or 14 days using the same GPU.

DISCUSSION

Graphylo is among the first deep learning approach to fully take advantage of deep multiple genome alignments to enhance the prediction of regulatory functional regions. By combining convolutional neural networks with graph convolutional networks, Graphylo exploits evidence of

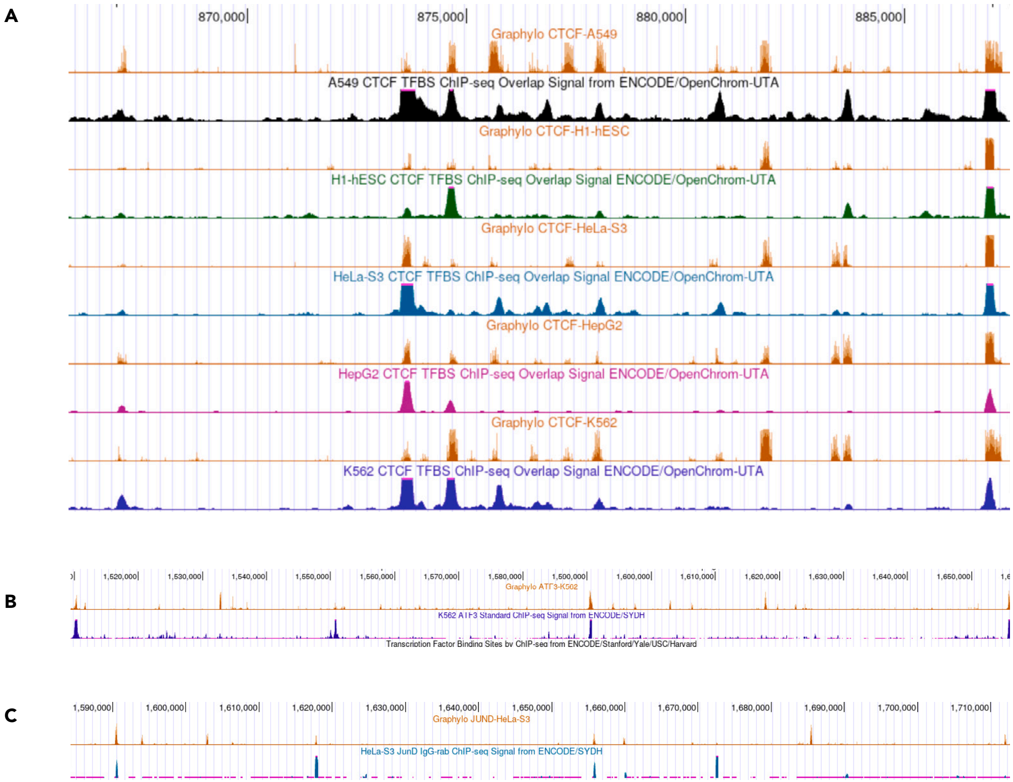


Figure 8. Genome Browser comparison of ChIP-Seq signal and Graphylo predictions

A genome browser-style display of Graphylo's predictions (orange) and ChIP-seq signal data (other colors) for CTCF on various cell lines (A), ATF3 on K562 cells (B), and JunD on HeLa cells (C).

negative selection on TFBS binding to enhance the sequence signal observed in humans. The GCN architecture provides a natural mechanism to represent and share information across related species, in a phylogenetically informed manner. This proves an effective approach to take advantage of deep multi-genome alignments. Our results demonstrate that Graphylo outperforms its competitors by a significant margin, with median AUPR scores improving by over 5% on 13 ChIP-Seq datasets and 1.5% on a balanced set of 124 ChIP-Seq datasets.

Unlike approaches that measure site-specific levels of interspecies conservation (e.g., PhastCons and PhyloP), Graphylo only uses the multiple genome alignments to extract orthologous/ancestral sequences, without assuming that functional sites necessarily remain at the same alignment positions. This makes it more robust to events such as binding site turnover, where one TFBS is replaced by another one nearby through a series of local mutations. Indeed, Graphylo outperforms state-of-the-art single-sequence models such as FactorNet across the entire spectrum of TFBS sequence conservation levels. This suggests that although Graphylo can take advantage of evolutionary sequence conservation when it exists, its performance is not adversely impacted in low-conservation regions, e.g., those corresponding to human- or primate-specific regions.

Although Graphylo has demonstrated promising results, there are several areas for future research. An interesting prospect is combining the Graphylo architecture with approaches that also take input cell-type-specific experimental data such as chromatin accessibility as was done in FactorNet[16] or RNA structural data as was done in DeepA-RBPBS.[22] This would likely improve prediction accuracy.

The integrated gradients and attention values findings indicate that Graphylo tends to place greater emphasis on the leaves of the phylogenetic tree rather than its internal nodes. This may be due to the fact that ancestral genome sequences are inferred from extant species and hence contain little to no new information. Versions of Graphylo where sequence information is only provided at the leaves, but ancestral nodes still exist to enable message passing across the tree, will be worth investigating.

Graphylo has the potential to be applied to other types of sequence function predictions, including other interaction prediction tasks (RNA-protein interactions, micro-RNA targeting, etc.) or higher-level functions (prediction of complex transcriptional or post-transcriptional regulatory regions).

We also anticipate that it may prove powerful at predicting the impact of mutations in these types of regions.[45] Graphylo, by using multiple sequence alignments as input and as a machine learning approach, provides a natural framework to learn evolutionary constraints and patterns of evolution naturally. This can provide a more nuanced understanding of the functional importance of different mutations.

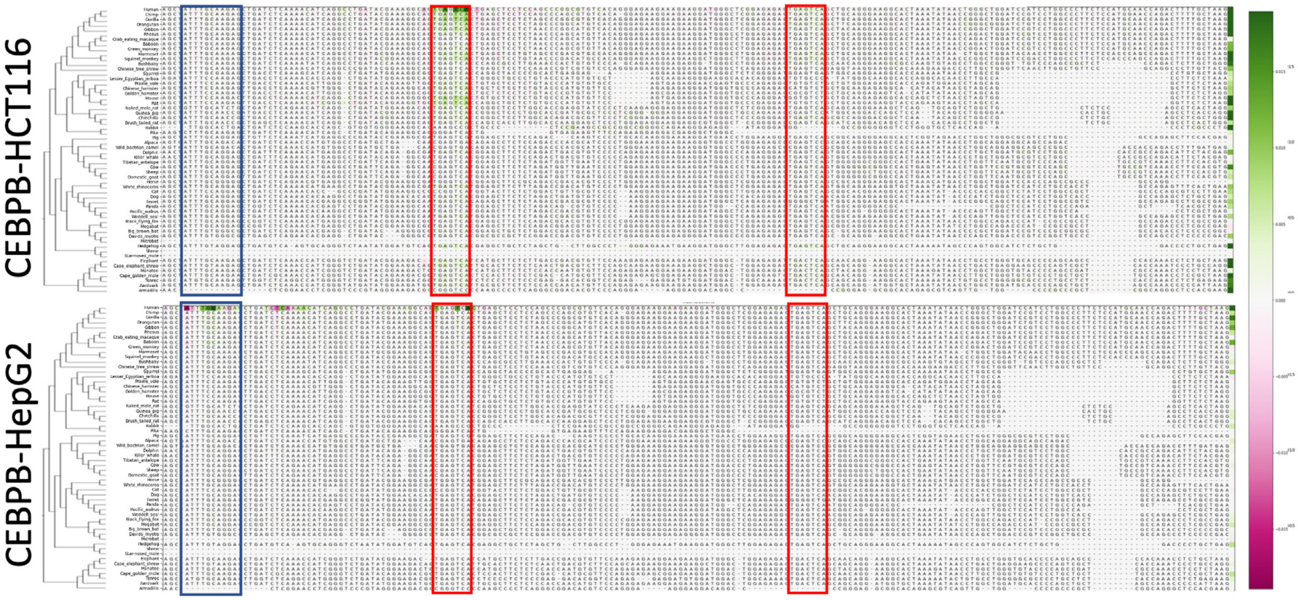


Figure 9. Examples of integrated gradients

The figure illustrates the results of integrated gradient attribution analysis applied to a sequence alignment, for models trained on CEBPB binding ChIP-Seq in two cell types (HCT116 and HepG2). The 58 rows in the alignment depict modern species sequences of length 201. Green nucleotides signify positive integrated gradient attribution scores, indicating their potential contribution to the prediction of TF binding. Conversely, red nucleotides represent negative scores. The blue box in the figure represents the motif match for CEBPB, and the red box represents motif matches for the heterodimer C/EBP:AP-1. The common consensus sequence for CEBPB is “ATGGCGGCAT” and for C/EBP:AP-1 heterodimer is “TGACTCA or TCGCTTCA.”

In conclusion, we believe that models such as Graphylo, which are capable of fully taking advantage of very rich evolutionary data to enhance predictive power, can help address some of the shortcomings of existing single-sequence approaches while yielding valuable biological and evolutionary insights.

Limitations of the study

One disadvantage of Graphylo is that it is quite slow, because it operates on a large set of orthologous sequences. Our work on species attention mechanisms suggests that attention values assigned to certain species or ancestors tend to be very low for some datasets. A sped-up predictor could be obtained by simply excluding those species from the analysis.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Architecture of Graphylo
  - Shared 1D-CNN
  - Species attention
  - Graph convolutional networks
  - Data preprocessing
  - Training Graphylo
  - Probability calibration
  - Baseline approaches
  - Interpretation of Graphylo predictions

ACKNOWLEDGMENTS

We thank Esaie Kuitche Kamala for the insight into evolutionary data, Yanlin Zhang for suggestions on different convolution methods, Elliot Layne for the insight on integrated gradients, Zichao Yan for suggestions on graph convolution network, Faizy Ahsan for the insight into preprocessing ChIP-Seq data and initial discussions about graph neural networks, and all the members of Blanchette Lab for providing comments for the manuscript.

This work was funded in part by a grant from the Natural Sciences and Engineering Research Council of Canada (Discovery program) to M.B. We thank Calcul Quebec and Compute Canada for CPU and GPU resources.

AUTHOR CONTRIBUTIONS

D.L. and M.B. conceived the Graphylo Architecture, conceived the experiments, interpreted the results, and wrote and revised the manuscript. D.L and C.B. carried out the experiments.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 4, 2023
Revised: December 17, 2023
Accepted: January 19, 2024
Published: January 26, 2024

REFERENCES

1. Bandziulis, R.J., Swanson, M.S., and Dreyfuss, G. (1989). RNA-binding proteins as developmental regulators. Genes Dev. *3*, 431–437.
2. Stefl, R., Skrisovska, L., and Allain, F.H.-T. (2005). RNA sequence-and shapedependent recognition by proteins in the ribonucleoprotein particle. EMBO Rep. *6*, 33–38.
3. Corley, M., Burns, M.C., and Yeo, G.W. (2020). How RNA-binding proteins interact with RNA: molecules and mechanisms. Mol. Cell *78*, 9–29.
4. Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of *in vivo* protein-DNA interactions. Science *316*, 1497–1502.
5. Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215.
6. Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129–141.
7. Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genomewide insights into brain alternative RNA processing. Nature *456*, 464–469.
8. König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. *17*, 909–915.
9. Rhee, H.S., and Pugh, B.F. (2012). ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. Curr. Protoc. Mol. Biol. *Chapter 21*. Unit 21.24.
10. Bakhtiari, S., Sulaimany, S., Talebi, M., and Kalhor, K. (2020). Computational Prediction of Probable Single Nucleotide Polymorphism-Cancer Relationships. Cancer Inf. *19*, 1176935120942216.
11. Stormo, G.D., Schneider, T.D., Gold, L., and Ehrenfeucht, A. (1982). Use of the 'Perceptron'algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res. *10*, 2997–3011.
12. McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinf. *11*, 165–211.
13. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineagedetermining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.
14. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.
15. Albawi, S., Mohammed, T.A., and Al-Zawi, S. (2017). Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET) (IEEE), pp. 1–6.
16. Quang, D., and Xie, X. (2019). FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods *166*, 40–47.
17. Kundaje, A., Boley, N., and Kuffner, R.K. (2016). ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge. https://doi.org/10.7303/syn6131484.
18. Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for oneshot image recognition. In ICML deep learning workshop, *2* (Lille).
19. Gers, F.A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. Neural Comput. *12*, 2451–2471.
20. Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., and Zeng, J. (2016). A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Res. *44*, e32.
21. Pan, X., Rijnbeek, P., Yan, J., and Shen, H.-B. (2018). Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genom. *19*, 511.
22. Du, Z., Xiao, X., and Uversky, V.N. (2022). DeepA-RBPBS: A hybrid convolution and recurrent neural network combined with attention mechanism for predicting RBP binding site. J. Biomol. Struct. Dyn. *40*, 4250–4258.
23. Yan, Z., Hamilton, W.L., and Blanchette, M. (2020). Graph neural representational learning of RNA secondary structures for predicting RNA-protein interactions. Bioinformatics *36*, i276–i284.
24. Burd, C.G., and Dreyfuss, G. (1994). Conserved structures and diversity of functions of RNA-binding proteins. Science *265*, 615–621.
25. Blanchette, M., and Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res. *12*, 739–748.
26. Blanchette, M., and Tompa, M. (2003). FootPrinter: a program designed for phylogenetic footprinting. Nucleic Acids Res. *31*, 3840–3842.
27. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. *14*, 708–715.
28. Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., and Haussler, D. (2011). Cactus: Algorithms for genome multiple

sequence alignment. Genome Res. *21*, 1512–1528.

29. Blanchette, M., Green, E.D., Miller, W., and Haussler, D. (2004). Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res. *14*, 2412–2423.

30. Diallo, A.B., Makarenkov, V., and Blanchette, M. (2010). Ancestors 1.0: a web server for ancestral sequence reconstruction. Bioinformatics *26*, 130–131. https://doi.org/10.1093/bioinformatics/btp600.

31. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

32. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

33. Sadri, J., Diallo, A.B., and Blanchette, M. (2011). Predicting site-specific human selective pressure using evolutionary signatures. Bioinformatics *27*, i266–i274. https://doi.org/10.1093/bioinformatics/btr241.

34. Ritchie, G.R.S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nat. Methods *11*, 294–296.

35. Gulko, B., Hubisz, M.J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat. Genet. *47*, 276–283.

36. Huang, Y.-F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat. Genet. *49*, 618–624.

37. Huang, Y.-F., and Siepel, A. (2019). Estimation of allele-specific fitness effects across human protein-coding sequences and implications for disease. Genome Res. *29*, 1310–1321.

38. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

39. Ahsan, F., Yan, Z., Precup, D., and Blanchette, M. (2022). PhyloPGM: boosting regulatory function prediction accuracy using evolutionary information. Bioinformatics *38*, i299–i306.

40. Kipf, T.N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1609.02907.

41. Yan, Z., Lécuyer, E., and Blanchette, M. (2019). Prediction of mRNA subcellular localization using deep recurrent neural networks. Bioinformatics *35*, i333–i342.

42. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In International conference on machine learning (PMLR.), pp. 3319–3328.

43. Ashida, R., Tominaga, K., Sasaki, E., Watanabe, T., Fujiwara, Y., Oshitani, N., Higuchi, K., Mitsuyama, S., Iwao, H., and Arakawa, T. (2005). AP-1 and colorectal cancer. Inflammopharmacology *13*, 113–125.

44. Debruyne, P.R., Bruyneel, E.A., Li, X., Zimber, A., Gespach, C., and Mareel, M.M. (2001). The role of bile acids in carcinogenesis. Mutat. Res. *480–481*, 359–369.

45. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894.

46. Stražar, M., Žitnik, M., Zupan, B., Ule, J., and Curk, T. (2016). Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. Bioinformatics *32*, 1527–1535.

47. 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., et al. (2015). Tensorflow large-scale machine learning on heterogeneous systems. tensorflow.org.

48. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Courna-peau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

49. Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Fried-berg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freelyavailable Python tools for computational molecular biology and bioinformatics. Bioinformatics. *25*, 1422–1423.

50. Grattarola, D., and Alippi, C. (2021). Graph neural networks in tensorflow andkeras with spektral [application notes]. IEEE Computational Intelligence Maga-zine *16*, 99–106.

51. McKinney, W. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference, Vol. 445 (TX: Austin), pp. 51–56.

52. Nair, V., and Hinton, G.E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), J. Fürnkranz and T.H. Joachims, eds. (Israel: Omnipress), pp. 807–814.

53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, pp. 5998–6008.

54. Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141.

55. Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D., et al. (2007). 28- way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res. *17*, 1797–1808.

56. Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. *5*, 1752–1779.

57. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC genome browser database: update 2006. Nucleic Acids Res. *34*, D590–D598.

58. Kingma, D.P., and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, eds. http://arxiv.org/abs/1412.6980.

59. Dal Pozzolo, A., Caelen, O., Johnson, R.A., and Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In 2015 IEEE Symposium Series on Computational Intelligence (IEEE), pp. 159–166.

60. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. (2010). How to explain individual classification decisions. J. Mach. Learn. Res. *11*, 1803–1831.

61. Velleman, D.J. (2005). The generalized Simpson's rule. Am. Math. Mon. *112*, 342–350.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Example data for running Graphylo Code	This paper	https://github.com/DongjoonLim/graphylo
Transcription Factor Binding Chip-seq publicly available dataset	Kundaje et al.[17]	https://doi.org/10.7303/syn6131484
RBN binding Clip-seq publicly available dataset	Strazar et al.[46]	https://github.com/mstrazar/iONMF
Software and algorithms		
Tensorflow	Abadi et al.[47]	www.tensorflow.org
Numpy	Harris et al.[48]	www.numpy.org
Biopython	Cock et al.[49]	www.biopython.org
Spektral	Grattarola et al.[50]	www.graphneural.network
Pandas	McKinney et al.[51]	www.pandas.pydata.org
Graphylo Code	This paper	https://github.com/DongjoonLim/graphylo

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Mathieu Blanchette (blanchem@cs.mcgill.ca).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data availability: This paper uses publicly available Transcription Factor binding data from[17] and RNA binding protein data from[46] listed in the key resources table.
- Code availability: Graphylo is publicly available online at https://github.com/DongjoonLim/Graphylo.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Graphylo is a deep learning model that combines Convolutional Neural Networks and Graph Convolutional Networks to predict transcription factor binding sites. It takes as input a set of orthologous and computationally reconstructed ancestral DNA sequences from various species, including a reference species of interest (e.g. human), as well as a phylogenetic tree representing their evolutionary history. In this section, we describe the architecture of Graphylo, and how it is trained and evaluated.

Architecture of Graphylo

Graphylo is divided into two main parts (see Figure 1): (1) a set of CNN layers that capture local dependencies within each DNA sequence, and (2) a GCN that captures dependencies between the sequences in the phylogenetic tree. The CNN layers analyze the sequence data to extract features, while the GCN processes the phylogenetic tree to incorporate evolutionary information into the model. Together, these two components allow Graphylo to effectively combine sequence data with evolutionary information for improved prediction performance.

Shared 1D-CNN

TFs can bind either on the forward or the reverse strand, and the orientation of binding sites matters in some situations. Existing CNN architectures for TFBS prediction do not adequately capture TFBS orientation while enabling convolution layers to capture in the same field of view motif matches on both strands. Hence we introduce a new type of sequence encoding, where each aligned sequence is one-hot encoded (A, C, G, T, -, N) on both the forward and reverse strands. Those are fed into the shared CNN with 32 2D-filters of size (6 × 11) and ReLU[52] activation function. The output representation is downsampled 2-fold using max-pooling. This results in a downsampled version of the input that is smaller in size and has reduced spatial resolution, enlarging the model's receptive field. The output from the sequence's reverse complement is flipped left to right and stacked with the forward strand output to yield a $64 \times L/2$ representation, where L denotes the length of a sequence.

The stacked outputs are fed into a second 1D-CNN with 32 filters of size (64 × 11). The convolution operation on the stacked outputs from both the forward strand and the reverse complement helps the filter capture meaningful patterns from both strands simultaneously. Learning simultaneously from both strands of DNA is more biologically relevant than the models that separate the learning for each strand and aggregate the result at the end. The output of this CNN layer is again downsampled by max-pooling.

Species attention

To help the GCN learn dependencies between species, we scaled the output of previous CNN layers with separately learned attention weights. Attentions are scalar values designed to represent the relative importance of the input chunk (tokens in natural language processing, channels of CNN in computer vision, etc.) to help the model embed the input data more easily.[53,54] With alignment data, it is logical to give attention to different species in the phylogenetic tree since not every species is equally important in predicting TFBSs in the human genome. We could expect the highest attention weight given to the human sequence and proper attention values given to other species according to what the model naturally learns to focus on.

The species' attention values are generated using simple multilayer perceptrons (MLP). After representations are generated from the CNN layers, global averages (over $\frac{L}{4} \times 32$ values for each species embedding) are taken species-wise 1 to yield a vector of length 115 (number of input species). This vector is then fed into a single fully connected layer with 8 neurons with a sigmoid activation function to generate an attention vector of length 115, where each value represents the attention value of the corresponding species. The attention values are applied to the original CNN representation as a scaling factor.

Graph convolutional networks

To capture the relationships among species, we apply a 4-layered GCN[40] that operates on the phylogenetic tree, with each node's data being the attention-scaled species-wise representation obtained by the CNN. The convolution for the GCN is as follows.

$$S_i^{(l+1)} = \sigma\left(\sum_{j \in N_i} \frac{1}{c_{i,j}} W^{(l)} S_j^{(l)} + b^{(l)}\right)$$

Where $S_i^{(l)}$ is the representation of the species in node i in the l-th layer of the GCN, N_i is the set of neighbours of node i, $c_{i,j}$ is a normalization constant, $W^{(l)}$ is the weight matrix(Trainable) for the l-th layer, and $b^{(l)}$ is the bias term for the l-th layer. The function sigma is an activation function, such as the ReLU function.

This formula defines the convolution operation as a weighted sum of the representations of the neighbours of each node, with the weights determined by the weight matrix $W^{(l)}$. The normalization constant $c_{i,j}$ is used to ensure that the influence of each neighbour is weighted equally. The bias term $b^{(l)}$ and the activation function sigma are used to introduce non-linearity into the model.

This way, the GCN can learn to extract features from the tree-structured data by applying convolutional filters over the phylogenetic tree.

GCNs are known to encounter the issue of over-smoothing[40] when used with many layers. Therefore, we have limited the application of graph convolution to only four layers. This results in message passing done locally rather than over the entire phylogenetic tree, which has 115 nodes and a diameter of 20. Thus, after the GCN layers have processed the graph data, the outputs of all nodes are aggregated using a fully-connected (FC) layer.

The output of the GCN model is a set of embeddings for each node in the graph. These embeddings are then flattened into a single vector of length 1840 (115 nodes * 16 features per node) and fed into a fully-connected (FC) layer with 64 neurons. The FC layer produces two logits, representing the predicted probabilities of the two classes (TF binding or not). The logits are normalized using the softmax function to produce class probabilities, and the cross entropy loss is calculated between the predicted probabilities and the true labels. This loss value is used to perform backpropagation and update the model weights.

Data preprocessing

Sequence alignment and ancestral genome reconstruction

Graphylo is trained from a set of multiple sequence alignment blocks (201 bp) with the phylogenetic tree represented as an adjacency matrix. We isolated DNA sequences of 58 placental mammals from a 100-way alignment of whole vertebrate genomes[27,55] available in the UCSC genome browser. We applied the Ancestors1.0 program[29,30] to infer the maximum likelihood ancestral sequences genome-wide based on a simple context-independent substitution model. This process leaves us with 115 aligned DNA sequences, 58 sequences corresponding to the leaves (modern species) in the phylogenetic tree, and the remaining 57 to the internal ancestral nodes. The alignment and inferred ancestral sequences are available at https://repo.cs.mcgill.ca/PUB/blanchem/Boreoeutherian/.

Alignment refinement

Multiz whole-genome alignments are imperfect. In particular, MultiZ sometimes over-zealously but incorrectly aligns sequences from distant species. We attempted to remove badly aligned species in the alignment with a simple refinement approach. A dubious alignment for a given species' sequence is often characterized by the presence of an unusually large number of gaps in the corresponding alignment row. We tested different thresholds of gaps allowed per 201 bp sequence and removed every sequence containing more than 2 gaps.

Transcription factor binding data

The Graphylo model was trained and evaluated using transcription factor binding data from the ENCODE DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge.[17] This data includes ChIP-Seq experiments on the human genome (hg19/GRCh37), which have been divided into 200 bp bins using a sliding window with a 50 bp stride. Each bin is labelled as ambiguous (A), bound (B), or unbound (U) based on the presence of ChIP-Seq peaks and an estimated Irreproducible Discovery Rate (IDR).[56] Bins without peaks are labelled as unbound, bins with IDRs higher than 0.05 are labelled as ambiguous, and bins with IDRs lower than 0.05 are labelled as bound. The primary dataset we used was 13 "Within-Cell Type" benchmarking data where each cell type had held-out testing chromosomes (1,8,21). In addition to the "Within-Cell Type" benchmarking data, we also analyzed the "Training Cell Types" datasets, which were provided for training models for the across-cell type prediction challenge but not used for leaderboard benchmarks. This includes a total of 124 datasets for different transcription factor (TF)-cell type combinations. Since not all of these datasets included held-out chromosomes for testing, we created a custom dataset by balancing the number of positive and negative samples. Further information about these datasets can be found on the ENCODE DREAM website.

LiftOver[57] was used to map the labelled ChIP-Seq dataset coordinates from hg19 to hg38 assembly positions. For each TF-Cell type combination, we took all the bins labelled as bound to be the positive (+) examples and sampled an equal number of bins labelled as unbound that are located at least 1000 bp away from any positive examples as the negative examples (-). Regions labelled as ambiguous were ignored. The model was two-fold cross-validated, where the dataset is shuffled and divided into two equal numbers of sets, where one set is used for training and validation while the other set is used for benchmarking.

RNA-RBP binding dataset

We used PARCLIP,[6] HITS_CLIP,[7] and ICLIP[8] datasets initially curated by Strazar et al.[46] It comprises the results of 31 RBP binding experiments conducted using different CLIP-Seq protocols. Each experiment contains 8,000 positive examples, representing RNA sequences containing binding sites for a specific RBP, and 32,000 negative examples, which correspond to unbound regions. Each example consists of an RNA sequence that spans 101 nucleotides. The dataset is split into a fixed 3-fold cross-validation scheme.

Training Graphylo

We chose the cross-entropy (sum of negative log-likelihood) as the loss function to minimize. In short, we aim to minimize

$$-\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

where N is the number of samples, yi is the true label of the i^{th} sample, and pi is the predicted probability of the i^{th} sample. Trainable weights include the weights in the CNN filters that will determine how much attribution is assigned to each input pattern, parameters of the fully connected layer in the species attention module that produces a weight for each species that determines how much importance is given to each species in the phylogenetic tree, weights in the GCN that are used for node embedding, and weights in the fully connected layer that helps to make the final prediction. The ReLU activation[52] was applied after every layer. The weights were updated iteratively using the ADAM optimizer[58] with default parameters in TensorFlow.[47] We fed in inputs in batches of size 256. To reduce overfitting, we used early stopping, which ends the training when the validation loss did not decrease for 15 consecutive epochs. There are several hyperparameters of the model that can be optimized. These include the number and size of CNN filters in each CNN layer, the size of node representation in the GCN, and the number of hidden units in the fully connected layer. We performed a grid search on realistic values for each hyperparameter and settled down with 32 CNN filters of the size of (6 × 11), 32 for the hidden weight size of GCN, and 64 perceptrons for the fully connected layer.

Probability calibration

A trained Graphylo will output a probability-like value for each class (Bind, No bind). Although these raw values represent the model's relative confidence in class membership, they should not be interpreted as the actual class assignment as these values are obtained from a balanced training set.

Corrected binding probability estimates are obtained by using the method proposed by Pozzolo and colleagues[59] to address the sampling bias of the training set:

We want to get the probability of $Pr(Bind = True|Data)$ whereas what the model predicts is $Pr(Bind = True|Data, Sampled = True)$. For simplicity, Bind=True is denoted as B, Bind=False is denoted as N, Sampled = True is denoted as S, and original data is denoted as D. From Bayes' theorem,

$$Pr(B|D) = \frac{Pr(S|N)Pr(B|D,S)}{Pr(B|D,S)(Pr(S|N) - 1)+1}$$

where $Pr(S|N)$ is the undersampling ratio which in our case is the ratio of positive examples and negative examples.

Baseline approaches

Most pre-existing computational models built to predict TFs are based on CNN coupled with machine learning techniques that capture context dependencies. One of the top-performing models in the ENCODE DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge data[17] is FactorNet.[16] FactorNet addresses shortcomings of prior simpler models such as DeepBind[14] by coming up with a sophisticated Siamese structure of CNN stacked with LSTM. This unique structure of FactorNet allows the prediction to be invariant of the strand of the input fed in, and the LSTM adds the more extended context-capturing capability to the output of CNN, which is known to be good at capturing local sequence patterns. In this paper, we compare Graphylo with two different versions. The original FactorNet model takes as input both the human sequence and DNase-Seq data; since our setting is purely sequence-based, we set all DNase-Seq values to zero. In the second version, we are replacing the DNase-Seq values with PhastCons conservation scores. By retraining FactorNet on this data, we add evolutionary information to FactorNet and can validate Graphylo's ability to capture different mutations throughout long evolution than a simpler sequence + conservation score-based model.

RNAtracker[41] is a model initially built to predict RNA subcellular localization, mainly driven by trans-regulatory factors called RNA-binding proteins. RNA-binding proteins also depend heavily on the sequence pattern, making RNAtracker a reasonable baseline model for predicting TFs. RNAtracker's architecture also utilizes CNN for capturing local patterns in the sequence, followed by bidirectional LSTM that captures more extended context patterns and the attention mechanism. These baseline models are powerful predictors that take a single sequence as an input. Comparing the performance of these models to Graphylo is a good measurement of the advantages Graphylo is taking from evolutionary information. We also compared Graphylo against PhyloPGM[39] which is a probabilistic graphical model that successfully incorporates predictions from different nodes in the phylogenetic tree.

For RBP-RNA binding prediction, we conducted a comparative analysis of Graphylo with two other baseline models: iDeepS,[21] which employs CNN and LSTM, and DeepA-RBPBS,[22] which shares a similar architecture with other baseline models but incorporates RNA secondary structure information as an additional input to the sequence. Furthermore, we developed a variant of Graphylo that exclusively utilizes human sequence as input, serving as a baseline model to evaluate the impact of integrating evolutionary data.

Interpretation of Graphylo predictions

In this section, we try to interpret what Graphylo learned during the training. In the Integrated Gradients analysis, we examine what feature inputs the model considers important and how the Integrated Gradients scores are distributed along with different species in the alignment. In the attention weights analysis, we look at the relative importance of the input species the model perceives after training.

Integrated gradients

The non-linear transformation of input data of deep learning models allows deep learning to be very flexible and powerful in solving many complicated tasks. Conversely, it makes it hard to understand what exactly and why precisely such learning happens. However, we can observe a few valuable things from trained deep-learning models, one sort being methods using the gradient of the output with respect to the input.

Earlier approaches to calculating attribution of input were multiplying the gradient with the input feature value, assuming that an input value important for the prediction will likely affect the output more than those less important.[60] However, this approach often fails when the input feature is important for the prediction but the gradient is zero resulting in the attribution score of such case as zero. To overcome this limitation, Sundararajan and colleagues developed the Integrated Gradients score.[42] Instead of using simple gradients for attributing the importance of input, this method integrates gradients from the baseline to the input feature, which prevents the sensitivity problem of previous attribution methods.

Integrated Gradients attribution scores are calculated below

$$IG = (x_i - base_i) \times \int_{\alpha=0}^{1} \frac{\partial F(base + \alpha \times (x - base))}{\partial x_i} d\alpha$$

where x is the one-hot encoded sequence alignment, *base* is the baseline input, F is a trained Graphylo model, and α is the interpolation constant.

We have set the baseline feature values to be zero, and the integration was approximated by Riemann Sum with $N = 50$.[61]

Observation of attention weights

Observing the species' attention weights that the model learns is a good proxy for evaluating the importance given to each species by the model. This research aims to predict the candidate TF binding sites for the human genome, so it is natural to assume that the orthologous genome of different species holds different importance in the prediction.