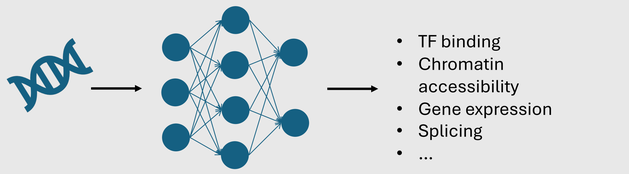
Predicting model uncertainty in genomic deep learning with knowledge distillation

Jessica Zhou[1], Peter Koo[1]
[1]Cold Spring Harbor Laboratory

Introduction

Deep neural networks (**DNNs**) are a powerful tool for predicting the regulatory activity of genomic sequences.



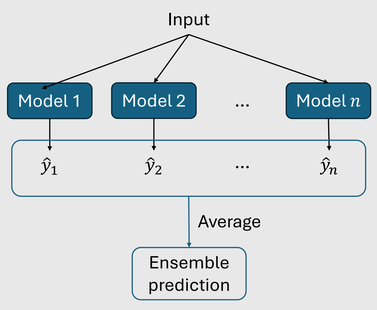
- TF binding
- Chromatin accessibility
- Gene expression
- Splicing
- ...

DNNs may not generalize well to genomic sequences that differ significantly from the training data (**out-of-distribution data**).

We used **ensemble distribution distillation** to improve genomic sequence DNNs and estimate predictive uncertainty.

Ensemble distribution distillation

Ensembles of DNNs have been shown to yield improved performance over individual models.

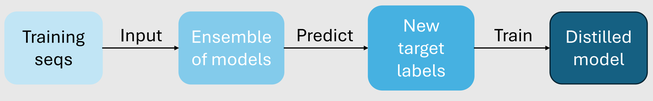


1. Multiple replicates of a model are trained from different random initializations.
2. Each model is used to make a prediction.
3. The predictions are **averaged** to yield an ensemble prediction.

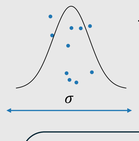
Benefits
- Improved predictive performance
- Better interpretability
- Can quantify epistemic uncertainty

Limitations
- Greater computational overhead
- Slow for large-scale inference tasks

Ensemble distribution distillation involves training a single, new model (the **distilled model**) on the distribution of the ensemble's predictions.



Uncertainty estimation

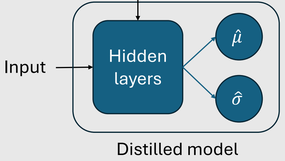


The disagreement (e.g. **standard deviation**) across the ensemble of predictions reflects **epistemic uncertainty**, or model uncertainty due to lack of knowledge (i.e. training data).

$$\mu_{ensemble} = \frac{\sum_{i=1}^{n} \hat{y}_i}{n}$$

$$\sigma_{ensemble} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - \mu_{ensemble})^2}{n}}$$

The mean ($\mu_{ensemble}$) and standard deviation ($\sigma_{ensemble}$) of the ensemble's predictions are used to train the distilled model.

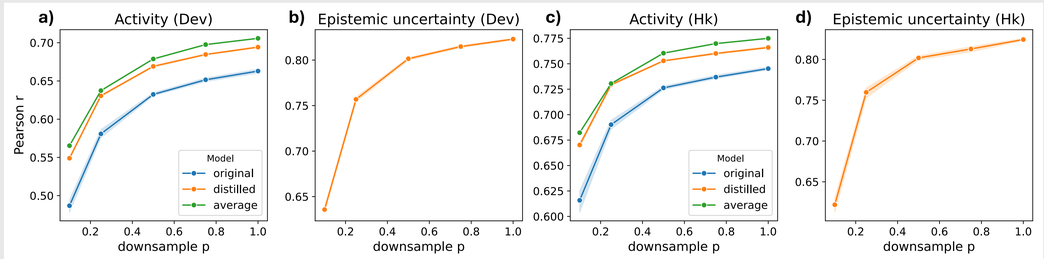


The distilled model predicts both the original task ($\hat{\mu}$) and corresponding epistemic uncertainty ($\hat{\sigma}$).

This does not reflect **aleatoric uncertainty**, or uncertainty arising from inherent noise in the data (e.g. sequencing errors, biological variability).

Ensemble distribution distillation: DeepSTARR

We first demonstrated that ensemble distribution distillation 1) preserves the advantages of ensembles, and 2) captures epistemic uncertainty by applying the approach to DeepSTARR[1], a CNN that predicts regulatory activity from DNA sequences.



a,c) Predictive performance of **a)** Dev and **b)** Hk enhancer activity output heads for 10 DeepSTARR models in ensemble (blue); the ensemble (green); and 10 distilled DeepSTARR models. ***Distilled models match performance of ensemble and perform well in low-data regimes.***
b,d) Predictive performance of **b)** Dev and **d)** Hk epistemic uncertainty output heads for 10 distilled DeepSTARR models. ***Distilled models can predict epistemic uncertainty with good accuracy.***

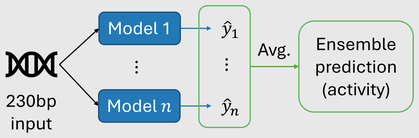
[1]De Almeida et al., *Nat Genet.* (2022)

Capturing aleatoric uncertainty with lentiMPRA

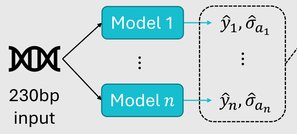
Aleatoric uncertainty arises from inherent variability in the data itself. To capture this, we trained ResidualBind[2] models to predict regulatory activity on data from a lentiMPRA experiment with multiple biological replicates.

Method

1. Train an ensemble of ResidualBind models that predict regulatory activity ($\hat{y}_i$).

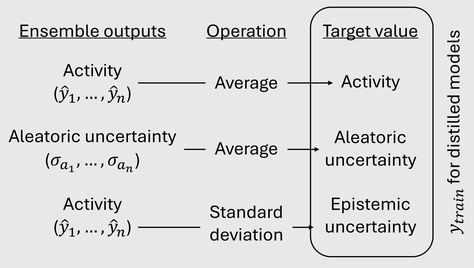


2. Train an ensemble of ResidualBind models that predict regulatory activity ($\hat{y}_i$) and aleatoric uncertainty ($\hat{\sigma}_{a_i}$).

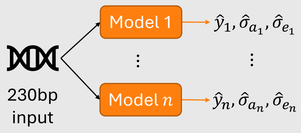


Aleatoric uncertainty is calculated as the standard deviation across biological replicates

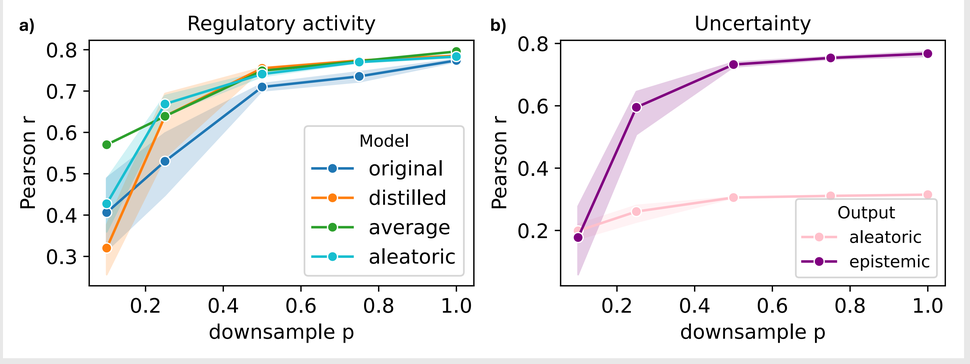
3. Get training data for distilled models using predictions from models with aleatoric uncertainty estimation.



4. Train **distilled** ResidualBind models that predict regulatory activity ($\hat{y}_i$), aleatoric ($\hat{\sigma}_{a_i}$) and epistemic uncertainty ($\hat{\sigma}_{e_i}$).



Results



a) Predictive performance for regulatory activity for 10 models in ensemble (blue); the ensemble (green); 10 models with aleatoric uncertainty estimation (cyan); and 10 distilled models for K562 (orange). ***Distilled models match performance of ensemble and perform well in low-data regimes.***
b) Predictive performance for aleatoric and epistemic uncertainty for 10 distilled models. ***Distilled models can predict epistemic uncertainty well but struggle with aleatoric uncertainty.***

[2]Koo et al., *Methods Mol Bio* (2023)

Takeaways

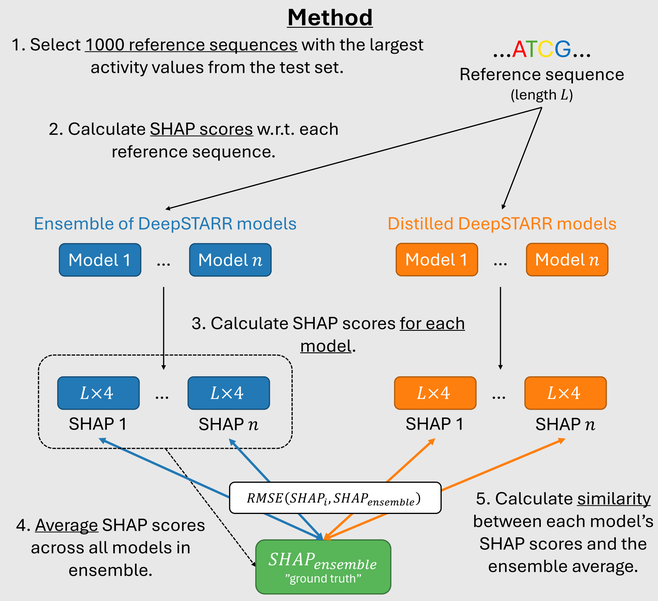
- Ensemble distillation offers the predictive performance improvements of ensembles withy only a single model
- Ensemble distribution distillation can capture epistemic uncertainty
- Ensemble distillation yields more robust attribution maps
- Aleatoric uncertainty is difficult to predict but is valuable to consider alongside epistemic uncertainty
- Ensemble distillation is a promising method for transparency and portability of large genomic DNNs

Attribution analysis: DeepSTARR

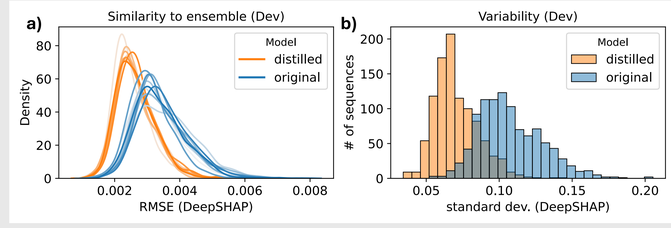
Ensembles produce more robust and consistent maps of feature importance. We evaluated this quality in our distilled models by analyzing attribution scores.

Method

1. Select 1000 reference sequences with the largest activity values from the test set.
2. Calculate SHAP scores w.r.t. each reference sequence.
3. Calculate SHAP scores for each model.
4. Average SHAP scores across all models in ensemble.
5. Calculate similarity between each model's SHAP scores and the ensemble average.



Results

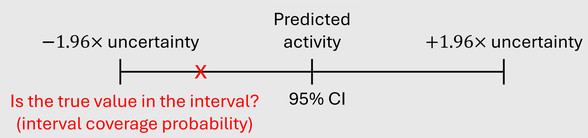


a) Similarity to ensemble SHAP scores for 10 DeepSTARR models in ensemble and 10 distilled DeepSTARR models for the Dev enhancer activity output head. ***Attribution maps for distilled models resemble the ensemble more closely.***
b) Variability across attribution maps from 10 DeepSTARR models in ensemble vs. 10 distilled DeepSTARR models for the Dev enhancer activity output head. ***Attribution maps from distilled models are more consistent.***

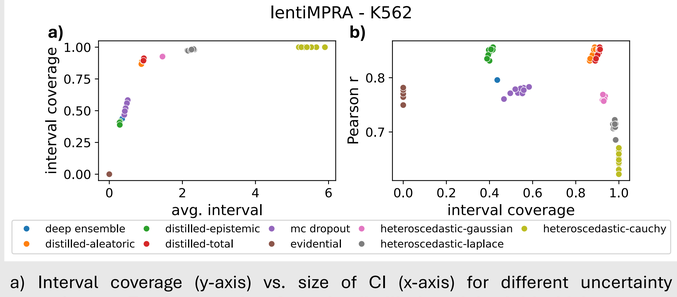
Calibrating uncertainty estimates

To examine the practical use of our uncertainty estimates, we evaluated how often confidence intervals based on the distilled lentiMPRA model predictions contained the experimental activity values.

Method



Results



a) Interval coverage (y-axis) vs. size of CI (x-axis) for different uncertainty estimates. ***Total uncertainty from the distilled models yields high coverage probability and outperforms aleatoric or epistemic uncertainty alone.***
b) Predictive accuracy (y-axis) vs. interval coverage (x-axis) for different uncertainty estimates. ***Distilled models balance predictive accuracy and interval coverage.***

Cold Spring Harbor Laboratory