Look At The Loss: Detecting False-Positive Feature Interactions Learned by NNs on Genomic Data

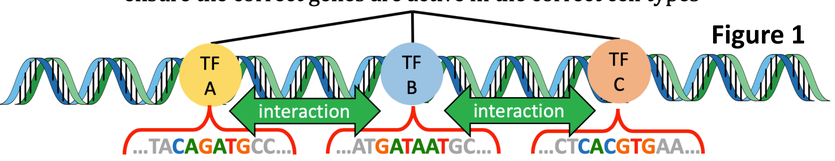
Mara Finkelstein*, Avanti Shrikumar*†, Anshul Kundaje†
(*co-first authors, †co-corresponding authors)

Abstract

Co-operative binding of proteins called Transcription Factors (TFs) to DNA modulates gene expression. Neural networks have been used to find candidate pairs of TFs with non-additive interaction effects. We design a simulated dataset to study the tendency of such networks to learn false positive interactions. We find that popular network architectures are highly susceptible to learning false-positive interactions that have comparatively large magnitudes - however, the learned interactions may not improve loss. We introduce a statistical test for whether a learned interaction significantly improves prediction loss. Combined with checking for consistency across different architectures, this test reliably distinguishes between true & false interactions in our simulated data.

Proteins called "Transcription Factors" (TFs) bind to DNA to ensure the correct genes are active in the correct cell types

Figure 1



TFs recognize short patterns in DNA called "motifs" that are 6-20 positions wide

Simulation

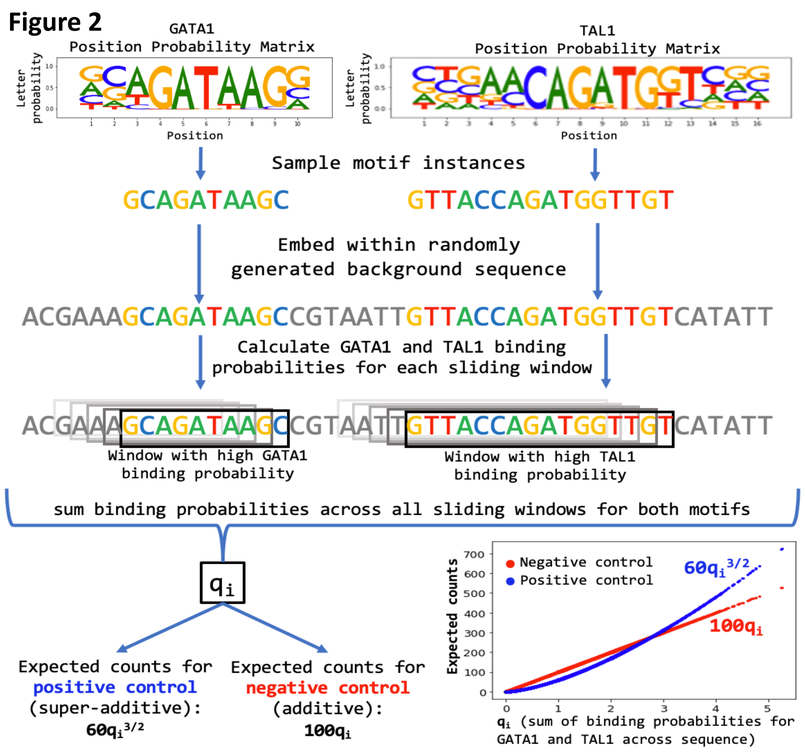
- We designed a regression task where strength of TF binding is a function of 2 motifs called GATA1 and TAL1 (see **Fig. 2**)
- Strength of binding (the output) is measured in integer-valued "counts" (the units of typical experiment), sampled from Poisson
- Two sets of labels for the simulated sequences:
  - Neg. control: binding output is additive func. of the motifs
  - Pos. control: binding output is super-additive in the motifs
- Ideal model trained on negative control dataset would predict no interaction effect between the motifs

How Interactions Are Computed

- Setting: model that accepts one-hot encoded DNA sequence and predicts binding strength as output. "Knocking out" a motif means replacing the motif sequence with random sequence that is a poor motif match
- s_{GT} := sequence containing both GATA1 and TAL1, s_G := seq. with TAL1 "knocked out", s_T := seq with GATA1 "knocked out", $s_\emptyset$:= seq with both TAL1 & GATA1 "knocked out", $f(s)$:= model prediction on sequence s.

Main effect M_G of GATA1 := $f(s_G) - f(s_\emptyset)$
Main effect M_T of TAL1 := $f(s_T) - f(s_\emptyset)$
Joint contribution $J_{G,T}$ of both := $f(s_{GT}) - f(s_\emptyset)$
Interaction effect $I_{G,T}$:= $J_{G,T}$ - (M_G + M_T)

Figure 2

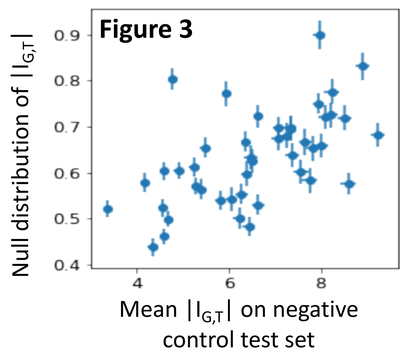


Model Architectures And Training

- Applied variance-stabilizing Anscombe transform $g(x) = 2\ sqrt(x + 3/8)$ to counts, followed by MSE loss (greatly improved model fitting; transforming counts e.g. with log transform is common in genomics). Note: interactions were computed in original counts space
- Trained 3 different types of CNN architectures with different #layers, hidden units and filter widths
- Each architecture was trained with 3 different L1 regularization weights
- Each of the 9 combos of arch & regularization was trained with 5 seeds.
- Result: 45 models **each** for pos. & neg. control data (90 total)
- Trained an additional 90 models to explore effect of sequence padding

Results

Figure 3



(1) Strength of interaction effect is not a reliable indicator of ground-truth interaction

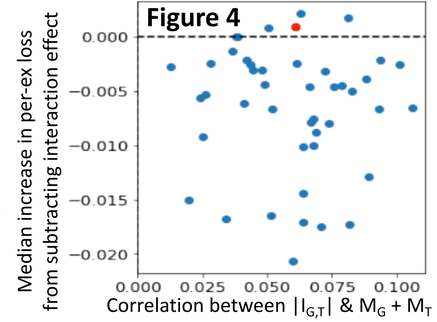
We can compare the mean magnitude of $I_{G,T}$ on the negative control to interaction effects computed between random positions in shuffled sequences (a popular choice of null distribution). Mean interactions magnitude in negative control greatly exceeds null distribution

This occurs because the magnitude of the interaction effect is positively correlated with magnitude of main effects (see Fig. 4)

(2) Model loss on negative control frequently improves when interaction effect is subtracted

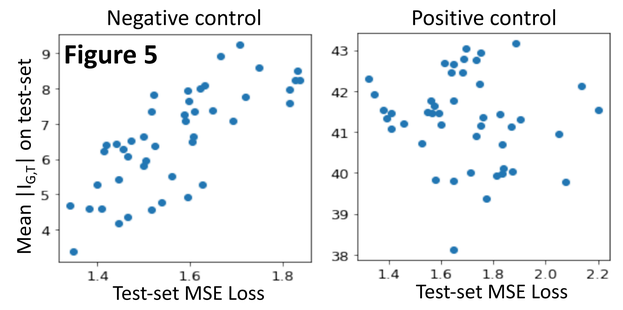
We can compute predictions with interaction subtracted by replacing $f(s_{GT})$ with $f(s_{GT}) - I_{G,T}$. Doing this frequently improved the loss on the negative control (esp. when "valid" padding was used). **Fig 4** shows results with "valid" padding.

Figure 4



Note that in cases where interactions helps the negative control loss, model uses $I_{G,T}$ to compensate for mis-predicted $M_G + M_T$

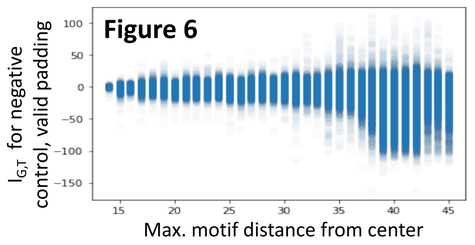
Figure 5



Consistent with this, model loss was positively correlated with mean $|I_{G,T}|$ for the negative control dataset, but not the positive control dataset (**Fig 5**).

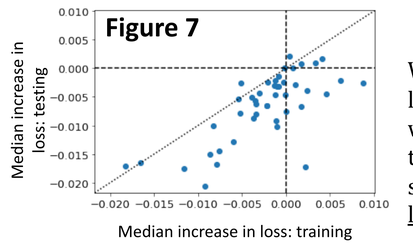
We found that inductive bias led models trained with 'valid' padding to learn strong false-positive interactions at the flanking sequences (**Fig 6**). This trend was not observed for 'same' padding.

Figure 6



We also observed that the increase in MSE loss from subtracting the interaction effect was consistently higher for training than testing data on the negative control (**Fig 7**), suggesting overfitting may play a role in learning fake interactions

Figure 7



(3) On held-out data, testing for consistent, significant improvement in loss due to including interaction effect can separate true from false interactions

We used a one-sided, paired Wilcoxon test to check if MSE($f(s_{GT})$) > MSE($f(s_{GT}) - I_{G,T}$). All 45/45 of models trained on pos. control data had sig. beneficial interactions (p-vaue threshold=0.05, both 'valid' & 'same' padding), while for neg. control, only 1/45 models with 'valid' padding (red dot in **Fig 4**) & 19/45 models with 'same' padding had sig. beneficial interactions. Even when the paired test was significant, unpaired test was always non-significant on negative control because the overall difference was weak (e.g. **Fig. 8**, corresponding to the model in red in Fig 4).

Figure 8