Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy

W. Xi[1] and M.A. Beer[1,2]

[1]Department of Biomedical Engineering and [2]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD 21205

Abstract

We report an overfitting issue in recent machine learning formulations of the enhancer-promoter interaction problem arising from the fact that many enhancer-promoter pairs share features. Cross-fold validation schemes which eliminate the possibility of overfitting through shared features show markedly reduced ability to predict enhancer-promoter interactions from epigenomic state.

Introduction

While quantitative modeling of cell-specific enhancer activity and variant impact is progressing rapidly, predicting the promoter and gene targets of enhancers remains challenging. We therefore read with great interest the paper by Whalen, et al,[1] which reported a computational model (TargetFinder) using epigenomic features that could predict enhancer-promoter (EP) interactions with high accuracy. In this brief note we report that because many EP pairs share features, the random cross-fold validation scheme used in Ref. 1 is subject to overfitting, and that cross-fold validation schemes which do not allow overfitting predict EP interactions with much lower accuracy.

Results

Whalen, et al,[1] used an F1-score performance measure, $F1=2P\cdot R/(P+R)$, where P=precision and R=recall, and reported that their average F1 across 6 cell lines was 0.83. This typically implies P,R>0.8 and AUPRC,AUROC~0.9, and that most EP pairs are being correctly classified as interacting or not according the epigenomic state of the enhancer (E), promoter (P), and the intervening genomic interval, or window (W), between the enhancer and the promoter. Interestingly, they found that window features (W) were often selected as being most predictive in the model, and that lack of some epigenomic marks in the genomic interval were correlated with positive EP interactions, a compelling hypothesis. Curiously, this high performance was only achieved with gradient boosting using a very large number of trees, and not with a linear SVM, which achieved F1-score≈0.2. In the course of examining their training data to find direct evidence in support of this hypothesis, we noticed that partly due to the courser resolution of the Hi-C interaction data, multiple contiguous enhancers often positively interact with the same promoter, but are all labelled independent positive EP interactions for training, even though they share P and W features. If samples with identical features are not in the same cross-fold validation test set, the shared features could lead to overfitting. In the K562 and GM12878 training data from Whalen, et al,[1] over 92% of EP pairs share a promoter with another EP pair and have a 2-fold or greater positive/negative class imbalance, and are thus subject to significant overfitting through the shared P and W features. In Fig. 1 we show how many promoters share a given number of positive and negative enhancers (positive and negative EP pairs connected to a single promoter). The greater the class imbalance, the more accurately this set of EP pairs can be predicted from overfitting on promoter features alone. This is a serious problem for all pairs labelled red or blue in Fig 1. For example, at position (7,2) in Fig. 1a, there are 27/(7+2)=3 distinct promoters in the training data which each interact positively with 7 enhancers and negatively with two enhancers (two negative EP pairs) in the training data. When one of these positive pairs is in the training set, a specific but non-generalizable rule on that promoter's features would predict the other pairs with accuracy of 78% when they are in the test-set. This overfitting problem becomes especially problematic for all EP pairs near the x

and y axes in Fig 1, and most of the negative samples are particularly susceptible. Enhancers on the same side of the promoter can also be overfit by shared window features.

We reproduced the published results from TargetFinder using the training data for the K562 cell line as shown in Fig 2a), and when we further observed that a nonlinear SVM (rbf) could not achieve the high performance of gradient boosting, and that high F1 is only achieved with 4000 trees, we suspected that this potential overfitting was important. Then we instead used cross-fold validation (CV) test sets sorted by chromosomal position so that no EP pairs with shared features could be in the same test set, and found that F1 dropped dramatically, as shown in Fig 2b. In fact, with sorted chromosomal CV test tests the predictive performance is only slightly better than random (F1=1/11 for this 1:20 ratio of positive to negative interactions). We also tested a cross-fold validation scheme were all EP pairs sharing a promoter were constrained to be in the same test set. In this case, promoter features cannot lead to overfitting, but window features can and do (EPW and W) when a gradient boosting classifier is used. A similar but slightly less dramatic reduction in predictive power is seen when removing or reducing the potential for overfitting in the GM12878 cell line (Fig 2d,e,f). That SVMs are less susceptible to overfitting explains how only trees were able to achieve high F1, and with 4000 trees it is likely that many of the EP pairs are being fit by very specific rules which do not generalize unless features are identically shared by another pair (classic overfitting).

Discussion

We have shown that the high accuracy of TargetFinder[1] in predicting EP interactions is mostly due to overfitting. We consider it extremely likely that subsequent publications using similar or identical training data are also subject to this overfitting problem, and that reports of accurate predictive modelling of EP interactions from local epigenomic state should be re-evaluated.[2–4] On the bright side, there is now considerable room for improvement in models of EP interactions. We suspect that EP interactions indeed may be specified by epigenomic state, but that nonlocal epigenomic state differences between enhancers competing for the same promoter must be directly taken into account. Finally, we note that this overfitting problem is unlikely to be a problem in approaches which learn interactions between non-overlapping chromosomal bins.[5]

References

1. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics* **48,** 488–496 (2016).
2. Cao, Q. *et al.* Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. *Nature Genetics* **49,** 1428–1436 (2017).
3. Yang, Y., Zhang, R., Singh, S. & Ma, J. Exploiting sequence-based features for predicting enhancer–promoter interactions. *Bioinformatics* **33,** i252–i260 (2017).
4. Singh, S., Yang, Y., Poczos, B. & Ma, J. Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks. *bioRxiv* 085241 (2016). doi:10.1101/085241
5. Schreiber, J., Libbrecht, M., Bilmes, J. & Noble, W. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. *bioRxiv* 103614 (2018). doi:10.1101/103614

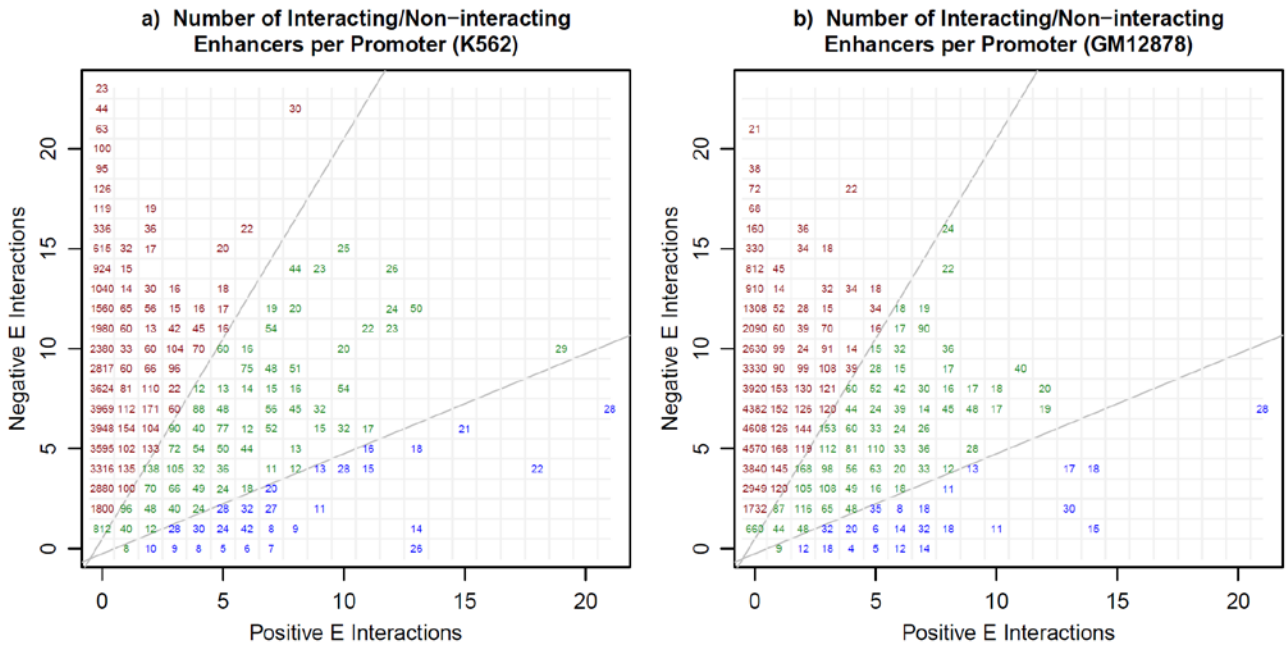


Fig 1. Most of the EP pairs used for training in Ref. 1 share promoters and promoter features with multiple EP pairs. a) For cell line K562, and **b)** for cell line GM12878, for each promoter we show the number of positive and negative EP pairs sharing a promoter. In **a)** for example, there are 27 interactions in position (7,2), which means three distinct promoters each interact positively with 7 enhancers and negatively with two enhancers in the training data. When the number of negative and positive EP pairs interacting with a promoter are imbalanced, they can be predicted correctly by overfitting on promoter features. The blue interactions have a 2-fold or greater imbalance (pos>neg) and the red interactions have a 2-fold or greater imbalance (neg>pos), and both red and blue interactions (92% of the data) can be overfit by shared promoter features.

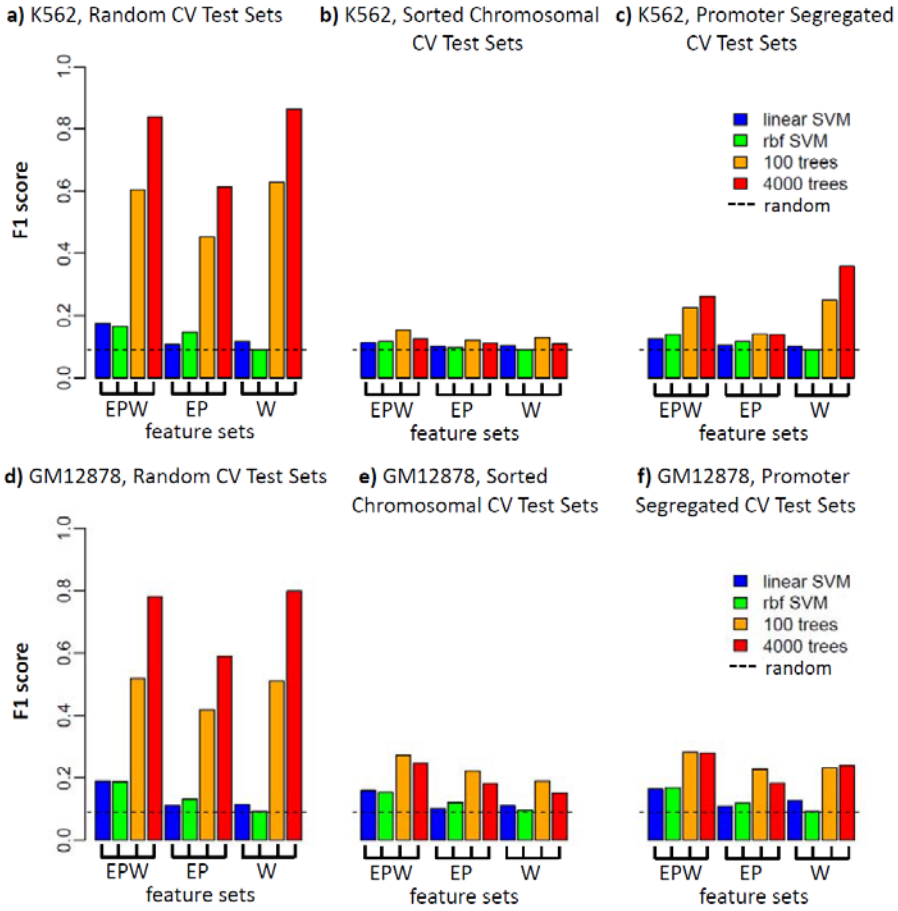
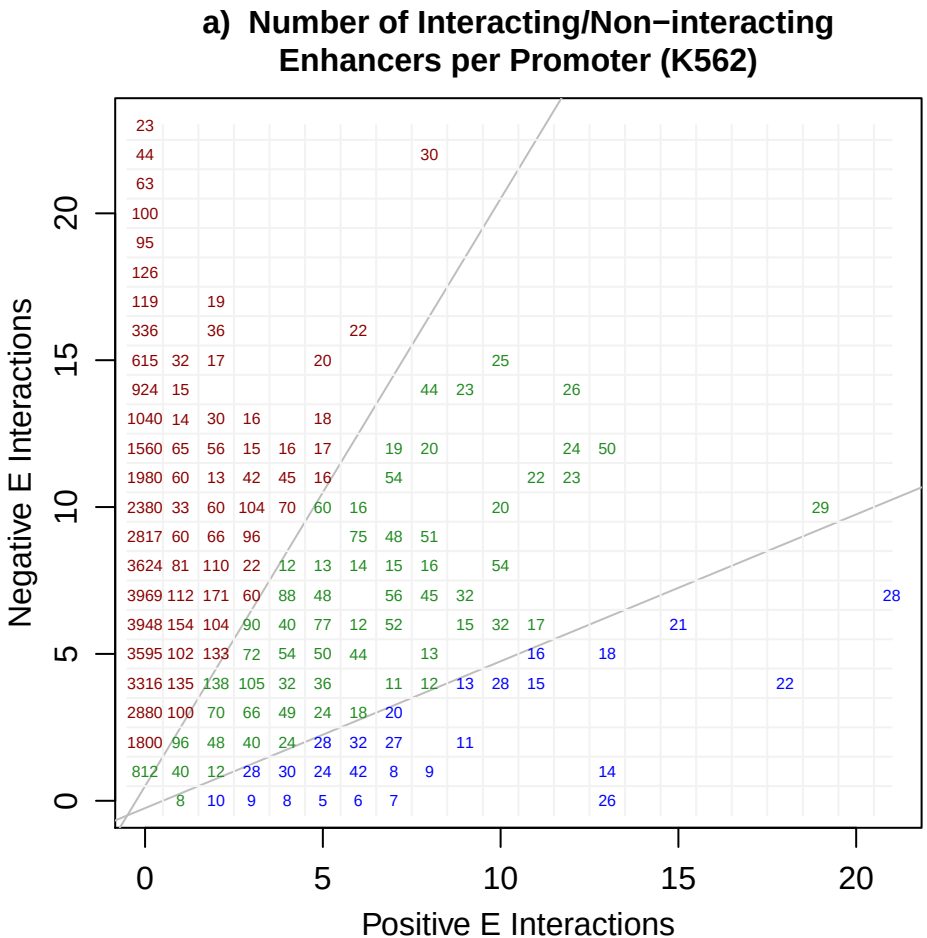
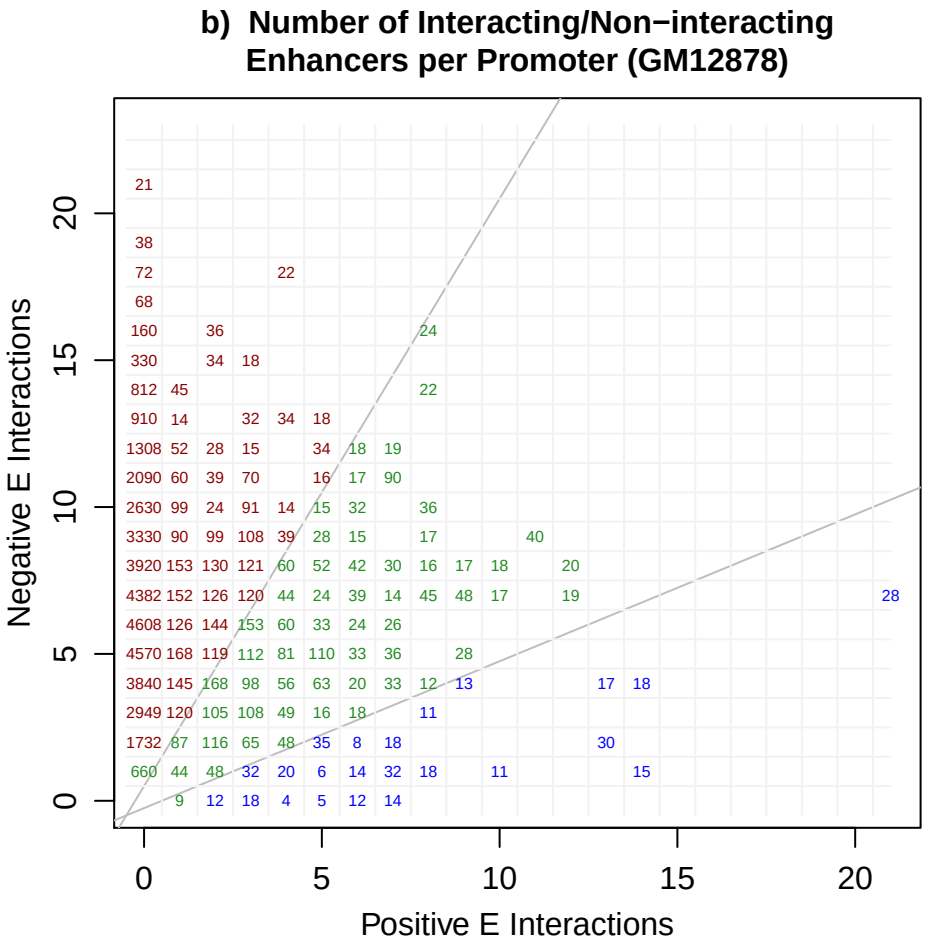
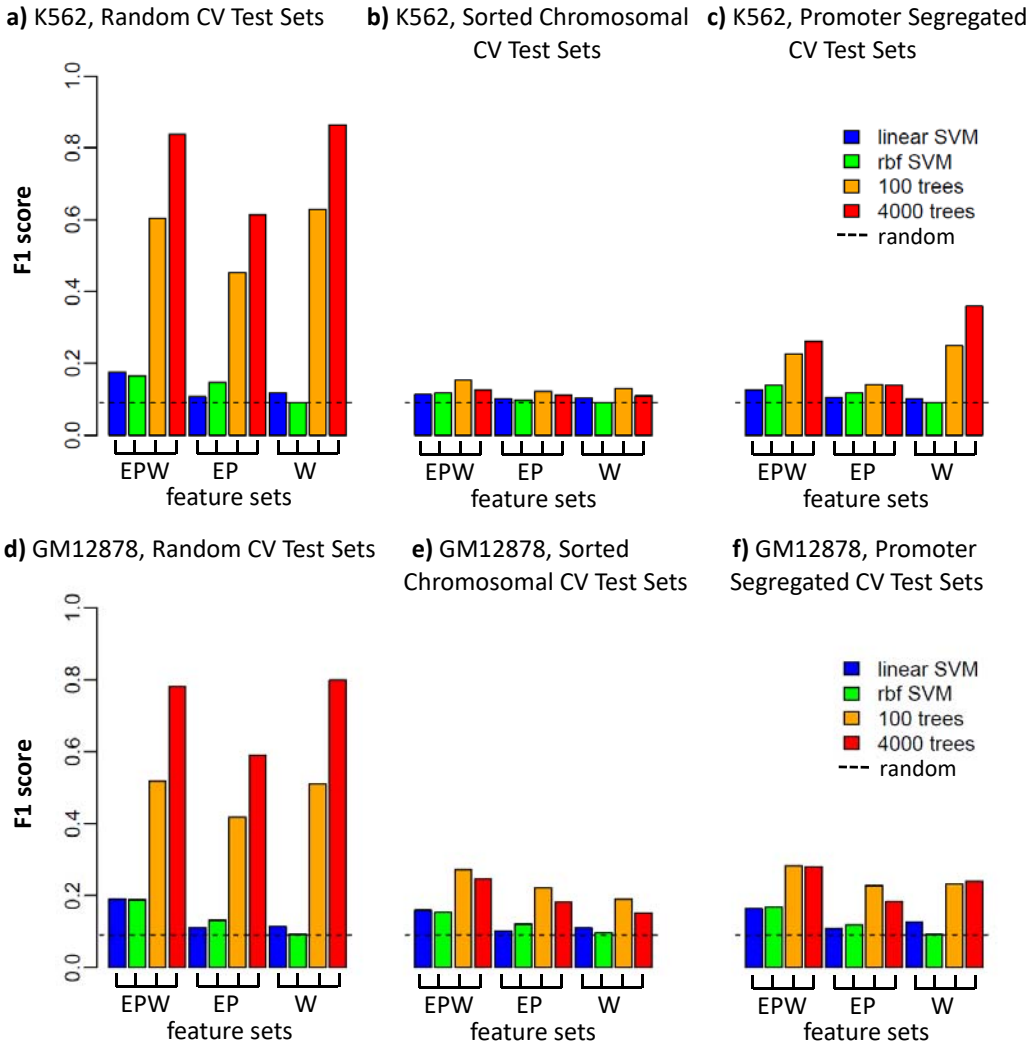


Fig 2. The apparently high predictive power of epigenetic state to identify enhancer-promoter (EP) interactions is largely due to overfitting. a) For cell line K562, using training data from Ref. 1, with random assignment of positive and negative EP pairs to cross-fold validation (CV) test sets, gradient boosting classifiers with a very large number of trees (4000) is able to achieve F1>0.8 with EPW or W features. **b)** If EP pairs with shared epigenomic features are properly forced to be in the same CV test set by sorting by chromosomal position, the predictive power is only slightly better than random, F1=1/11. **c)** If EP pairs sharing the same promoter are forced to be the same test sets, EP overfitting is eliminated but some overfitting still occurs through shared window features (W). **d,e,f)** A similar reduction in performance is observed when removing the potential overfitting in cell line GM12878 with sorted chromosomal CV test sets.


a) Number of Interacting/Non−interacting Enhancers per Promoter (K562)
Negative E Interactions
Positive E Interactions
0
5
10
15
20


b) Number of Interacting/Non−interacting Enhancers per Promoter (GM12878)
Negative E Interactions
Positive E Interactions
0
5
10
15
20


a) K562, Random CV Test Sets
b) K562, Sorted Chromosomal CV Test Sets
c) K562, Promoter Segregated CV Test Sets
d) GM12878, Random CV Test Sets
e) GM12878, Sorted Chromosomal CV Test Sets
f) GM12878, Promoter Segregated CV Test Sets
F1 score
0.0
0.2
0.4
0.6
0.8
1.0
EPW
EP
W
feature sets
linear SVM
rbf SVM
100 trees
4000 trees
random