RESEARCH ARTICLE

REVISED Predicting transcription factor binding using ensemble random forest models [version 2; peer review: 2 approved]

Fatemeh Behjati Ardakani [1-3*], Florian Schmidt [1-4*], Marcel H. Schulz[1,2,5]

[1]High throughput Genomics and Systems Biology, Cluster of Excellence on Multimodel Computing and Interaction, Saarland University, Saarbruecken,, Saarland, 66123, Germany
[2]Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarbruecken, Saarland, 66123, Germany
[3]Graduate School of computer science, Saarland University, Saarbruecken, Saarland, 66123, Germany
[4]Computational Systems Biology, Genome Institute of Singapore, Singapore, Singapore
[5]Institute for Cardiovasular Regeneration, Goethe University Frankfurt Am Main, Frankfurt Am Main, Hessen, 60590, Germany

[*] Equal contributors

v2 **First published:** 04 Oct 2018, **7**:1603 (https://doi.org/10.12688/f1000research.16200.1)
Latest published: 02 Sep 2019, **7**:1603 (https://doi.org/10.12688/f1000research.16200.2)

Abstract

Background: Understanding the location and cell-type specific binding of Transcription Factors (TFs) is important in the study of gene regulation. Computational prediction of TF binding sites is challenging, because TFs often bind only to short DNA motifs and cell-type specific co-factors may work together with the same TF to determine binding. Here, we consider the problem of learning a general model for the prediction of TF binding using DNase1-seq data and TF motif description in form of position specific energy matrices (PSEMs).
Methods: We use TF ChIP-seq data as a gold-standard for model training and evaluation. Our contribution is a novel ensemble learning approach using random forest classifiers. In the context of the *ENCODE-DREAM in vivo TF binding site prediction challenge* we consider different learning setups.
Results: Our results indicate that the ensemble learning approach is able to better generalize across tissues and cell-types compared to individual tissue-specific classifiers or a classifier built based upon data aggregated across tissues. Furthermore, we show that incorporating DNase1-seq peaks is essential to reduce the false positive rate of TF binding predictions compared to considering the raw DNase1 signal.
Conclusions: Analysis of important features reveals that the models preferentially select motifs of other TFs that are close interaction partners in existing protein protein-interaction networks. Code generated in the scope of this project is available on GitHub: https://github.com/SchulzLab/TFAnalysis (DOI: 10.5281/zenodo.1409697).

Keywords

ENCODE-DREAM in vivo Transcription Factor binding site prediction challenge, Transcription Factors, Chromatin accessibility, Ensemble learning, Indirect-binding, TF-complexes, DNase1-seq

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
REVISED **version 2** published 02 Sep 2019	✓ report	✓ report
version 1 published 04 Oct 2018	? report	? report

1 **Jan Grau**, Martin Luther University of Halle-Wittenberg (MLU), Halle, Germany

2 **Gary D Stormo**, Washington University School of Medicine, St. Louis, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the RPackage gateway.

This article is included in the Max Planck Society collection.

Corresponding authors: Fatemeh Behjati Ardakani (fbehjati@mmci.uni-saarland.de), Florian Schmidt (fschmidt@mmci.uni-saarland.de), Marcel H. Schulz (mschulz@mmci.uni-saarland.de)

Author roles: Behjati Ardakani F: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Schmidt F**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Schulz MH**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Cluster of Excellence on Multimodal Computing and Interaction (DFG) [EXC248].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Behjati Ardakani F, Schmidt F and Schulz MH. **Predicting transcription factor binding using ensemble random forest models [version 2; peer review: 2 approved]** F1000Research 2019, **7**:1603 (https://doi.org/10.12688/f1000research.16200.2)

First published: 04 Oct 2018, **7**:1603 (https://doi.org/10.12688/f1000research.16200.1)

REVISED **Amendments from Version 1**

In this new version of the manuscript, we assessed and reported the model performance in terms of ROC-AUC and PR-AUC for all analyses.

In addition, we introduced another ensemble approach, which works based on averaging the predictions of the tissue-specific models, as a baseline for comparison between the pooling and RF ensemble classifier.

We also provided a new figure (Figure 7) to explicitly show the top features chosen by the models. Furthermore, we performed an additional experiment on unseen data to show that reducing the feature space to the top 20 features is indeed not affecting model performance negatively (Supplementary Figure 1). In addition to that, we added another experiment on training data illustrating that the ensemble model is able to pick up and to generalize tissue specific TF binding information (Supplementary Figure 2).

Any further responses from the reviewers can be found at the end of the article

Introduction

Transcription Factors (TFs) are key players of transcriptional regulation. They are indispensable to maintain and establish cellular identity and are involved in several diseases[1]. TFs bind to the DNA at distinct positions, mostly in accessible chromatin regions[2], and regulate transcription by recruiting additional proteins. The TFs can alter chromatin organization or, for example, recruit an RNA polymerase to initiate transcription[1]. Hence, to understand the function of TFs it is vital to identify the genomic location of TF binding sites (TFBS). As TFs regulate distinct genes in distinct tissues, these binding sites are tissue-specific[2].

Nowadays, the most prevalent and widely used method to experimentally determine TFBS is through ChIP-seq experiments, which can be used to generate genome-wide, tissue-specific maps of *in-vivo* TF binding. However, ChIP-seq experiments are expensive, experimentally challenging, and require an antibody for the target TF. In this work, target TF refers to the TF of interest, i.e. the TF whose binding sites should be determined. To overcome these limitations, a number of computational methods have been developed to pinpoint TFBS. Most of these methods are based on position weight matrices (PWMs) describing the sequence preference of TFs[3,4]. PWMs indicate, for each position of a TF binding motif independently, how likely the individual nucleotides are to occur at a specified position. Unfortunately, screening the entire genome using a PWM results in too many false positive predictions. Therefore, numerous methods have been proposed to reduce the prediction error by combining PWMs with epigenetics data, such as DNase1-seq, ATAC-seq, or Histone Modifications, reflecting chromatin accessibility. Also, additional features such as nucleotide composition, DNA shape, or sequence conservation can be incorporated into the predictions. Including these additional data sets and information improved the TF binding predictions considerably[5–12]. A non-exhaustive overview is provided in 13. While PWM based models are still the most common means to assess the likelihood of a TF binding to genomic sequences, more elaborate approaches that capture nucleotide dependencies, have been successfully used as well[14,15]. SLIM-models[16] are an example for such approaches. In contrast to other methods, nucleotide dependency profiles inferred by SLIM models can be visually interpreted. Recently, deep learning methods have been used to learn TF binding specificities *de novo* from large scale data sets comprising not only ChIP-seq but also Selex and protein binding microarray (PBM) data[17].

The *ENCODE-DREAM in vivo Transcription Factor binding site prediction challenge*[18] aims to systematically compare various approaches on TFBS prediction in a controlled setup, with the additional complexity of applying the classifiers on the tissues/cell types that were not used for model training. The *challenge* organizers provide TF-ChIP seq data for 31 TFs, accompanied with RNA-seq and DNase1-seq data in 12 different tissues. Using labels deduced from the TF-ChIP-seq data, predictive models for TF binding should be learned and then applied to a set of held-out chromosomes on an unseen tissue. Predictions are computed in bins, covering the entire target chromosomes. The main *challenge* paper will provide a detailed explanation of the *challenge* setup and a comparison across all competing methods. This article is a companion paper to the main *ENCODE-DREAM Challenge* paper, in which we describe our contribution to the *challenge*, delineate the motivation for our work and provide an independent evaluation of our ideas to achieve generalizability across tissues.

We developed an ensemble learning approach using random forest (RF) classifiers, extending the work of Liu *et al.*[12]. Tissue-specific cofactor information was shown to be relevant to accurately model TF binding[12,19]. Thus, we designed our approach to aggregate tissue-specific cofactor data, via an ensemble step, into a generalizable model. Briefly, we compute TF affinities with TRAP[20] for 557 PWMs in DNase-hypersensitive sites (DHSs) identified with JAMM[21]. TF affinities computed by TRAP are inferred from a biophysical model. In contrast to a simple binary classification, e.g. FIMO[22], these scores can capture low affinity binding sites, which were shown to be biologically relevant[23,24]. Here, we show that our ensemble models generalize well between tissues and that they exhibit better classification performance than tissue-specific RF classifiers. Furthermore, we illustrate that only a small subset of TF features is sufficient to predict tissue-specific TFBSs and also show that these TFs are often known co-factors/interaction partners of the target TF.

Methods

Data

Within the scope of the *challenge* participants were provided with ChIP-seq data for 31 TFs, as well as DNase1-seq and gene expression obtained from RNA-seq data for 13 tissues. From the available 31 TFs, 12 were used to assess the model performance in the final round of the *challenge*. As we focus in this article on the generalizability of our models, we use only those TFs that are linked to multiple training tissues. Thus, we consider the TFs listed in Table 1 for model training and general evaluation experiments. Furthermore, we use eight TFs, as provided in Table 2, to evaluate the performance of our models on unseen test data. The *challenge* required that the

Table 1. Number of bins labeled as bound per transcription factor (TF) and tissue, deduced from TF ChIP-seq data.

TF	Number of bins labelled as bound per tissue
ATF7	272,2234 (GM12878), 218,239 (HepG2), 345,775 (K562)
CREB1	164,968 (GM12878), 103,752 (H1-hESC), 178,080 (HepG2), 98,554 (K562)
CTCF	179,672 (A549), 271,097 (H1-hESC), 206,336 (HeLa-S3), 208,868 (HepG2), 215,238 (K562), 305,547 (MCF-7)
E2F1	93,117 (GM12878), 55,391 (HeLa-S3)
EGR1	72,595 (GM12878), 52,733 (H1-hESC), 175,994 (HCT116), 58,793 (MCF-7)
EP300	126,409 (GM12878), 69,247 (H1-hESC), 157,629 (HeLa-S3), 168,173 (HepG2), 137,369 (K562)
GABPA	26,467 (GM12878), 51,666(H1-hESC), 31,202 (HeLa-S3), 60,552 (HepG2), 109,423 (MCF-7), 78,403 (SK-N-SH)
JUND	203,665 (HCT116), 179,999 (HeLa-S3), 183,558 (HepG2), 193,814 (K562), 92,905 (MCF-7), 222,013 (SK-N-SH)
MAFK	34,054 (GM12878), 97,659 (H1-hESC), 62,124 (HeLA-S3), 291,337 (HepG2), 201,157 (IMR90)
MAX	301,615 (A549), 98,327 (GM12878), 224,379 (H1-hESC),
	321,501 (HCT116), 211,590 (HeLa-S3), 317,579 (HepG2), 318,318 (K562), 250,775 (SK-N-SH)
MYC	57,512 (A549), 91,325 (HeLa-S3), 183,627 (K562), 151,748 (MCF-7)
REST	71,251 (H1-hESC), 47,654 (HeLa-S3), 67,453 (HepG2), 59,640 (MCF-7), 48,946 (Panc1), 94,082 (SK-N-SH)
RFX5	161,689 (GM12878), 22,948 (HeLa-S3), 54,961 (MCF-7)
SRF	21,495 (GM12878), 40,201 (H1-hESC), 176,158 (HCT116), 22,593 (HepG2), 18,895 (K562)
TAF1	87,109 (GM12878), 185,027 (H1-hESC), 93,824 (HeLa-S3), 110,385 (K562), 83,276 (SK-N-SH)
TCF12	51,798 (GM12878), 104,834 (H1-hESC), 82,102 (MCF-7)
TCF7L2	100,926 (HCT116), 165,264 (HeLa-S3), 143,025 (Panc1)
TEAD4	66,198 (A549), 103,483 (H1-hESC), 174,716 (HCT116), 125,917 (HepG2), 186,759 (K562)
YY1	136,621(GM12878), 195,489 (H1-hESC), 63,293 (HCT116), 133,943 (HepG2)
ZNF143	197,385 (GM12878), 178,088 (H1-hESC), 48,154 (HeLA-S3), 103,755 (HepG2)

Table 2. Test data used in this article, shown per transcription factor (TF) and tissue.

TF	Tissues
CTCF	PC-3, Induced pluripotent stem cell
E2F1	K562
EGR1	liver
GABPA	liver
JUND	liver
MAX	liver
REST	liver
TAF1	liver

predictions are made in bins of size 200 *bp*, shifted by 50 *bp* each, spanning the whole genome. Except for the held-out chromosomes 1, 8, and 21, all chromosomes are used for model training. We refer to the *challenge* website for a detailed overview on the provided data[18]. Note that we exclude sites labelled as ambiguously bound from this study.

Data preprocessing and feature generation

In order to obtain datasets per tissue and per TF that could be handled in terms of memory consumption and processing time, and also to cope with the large imbalance number of bound and unbound sites, we randomly sampled as many negative sites from the provided ChIP-seq *tsv* files as there were true binding sites per TF. The ChIP-seq labels contained in the balanced and down-sampled *tsv* files are used as the response for training RF models.

Throughout the course of the *challenge*, we have used two distinct ways to generate features for the RF classifiers: (1) with and (2) without considering DHSs. In none of the approaches have we used the provided RNA-seq data nor did we compute DNA shape features. Generally, we computed TF binding affinities with *TRAP*[20] for 557 distinct TFs using the default parameter settings. Within our workflow, we first consider all 557 TFs to determine factors that are predictive for the binding of the target TF. The position specific energy matrices (PSEMs)

used in our computation are converted from position weight matrices (PWMs) obtained from JASPAR[25], UniPROBE[26], and Hocomoco[27]. The code to perform the conversion and to run TRAP is available on GitHub.

We compared two approaches to generate features for the classifier from DNase1-seq data. In the first approach, shown in Figure 1a, we compute tissue-specific DHSs using the peak caller JAMM[21] (version 1.0.7.2). Specifically, we converted the provided DNase1-seq *bam* files to *bed* files using the bedtools[28] *bamtobed* command (bedtools version 2.25.0). For each *bed* file, peaks are computed separately using JAMM's standard parameters and the *–f 1* option. The individual DHS files obtained for one TF are aggregated using the bedtools *merge* command. We decided to take a less conservative approach and merge all peaks identified in individual replicates per TF to ensure that we do not miss any accessible site, all be it this may introduce false positives. Next, TF affinities are calculated in the merged DHS sites using TRAP, and the median DHS signal per peak is computed from the provided *bigwig* files. The computed data are intersected, using a *left outer join* with bedtools, with the binned genome structure required for training (using the bins contained in the *tsv* files mentioned above) and testing (using the provided *bed*-file containing all test regions).

The second approach for computing the features is depicted in Figure 1b. Here, we do not use the information on DHS sites, instead we compute TF binding affinities and the DNase1-seq signal per bin using the bin structure defined by the challenge as explained in the Data section above. We obtain the features genome-wide, without any preselection of the bins. To account for variability between both biological and technical replicates, we calculate the median DNase1-seq coverage across the replicates using the *bedtools coverage* command. Overall, the features for a single bin are composed of the TF affinities in that bin, the DNase1-seq signal in the bin itself together with its left and right neighboring bins.

Ensemble random forest classifier

The Random Forest models, implemented using the randomForest R-package[29] (version 4.6-12), are trained on either of the feature setups explained in the previous section. Training the RF models can be seen as a two-step approach that is independent from the feature setup. Throughout model training, the balance between the *bound* and *unbound* classes is maintained to avoid over-fitting of the RF classifiers and also to ensure an unbiased evaluation of model performance. For fitting the RF classifiers we used 4,500 trees, and at most 30,000 positive and negative, i.e. bound and unbound, samples. This restriction is enforced by the limitations of the *randomForest* R-package. As illustrated in Figure 2a, for a given target TF, we first learn tissue and TF specific RF classifiers using all available features from the input matrix, $T_i \in R^{n \times 557}$; $i \in \{1, \ldots, m\}$, where n is the number of bins forming the training set, and m denotes the number of training tissues for the target TF:

$$RF_i = RandomForest(T_i, Binding(T_i)),$$

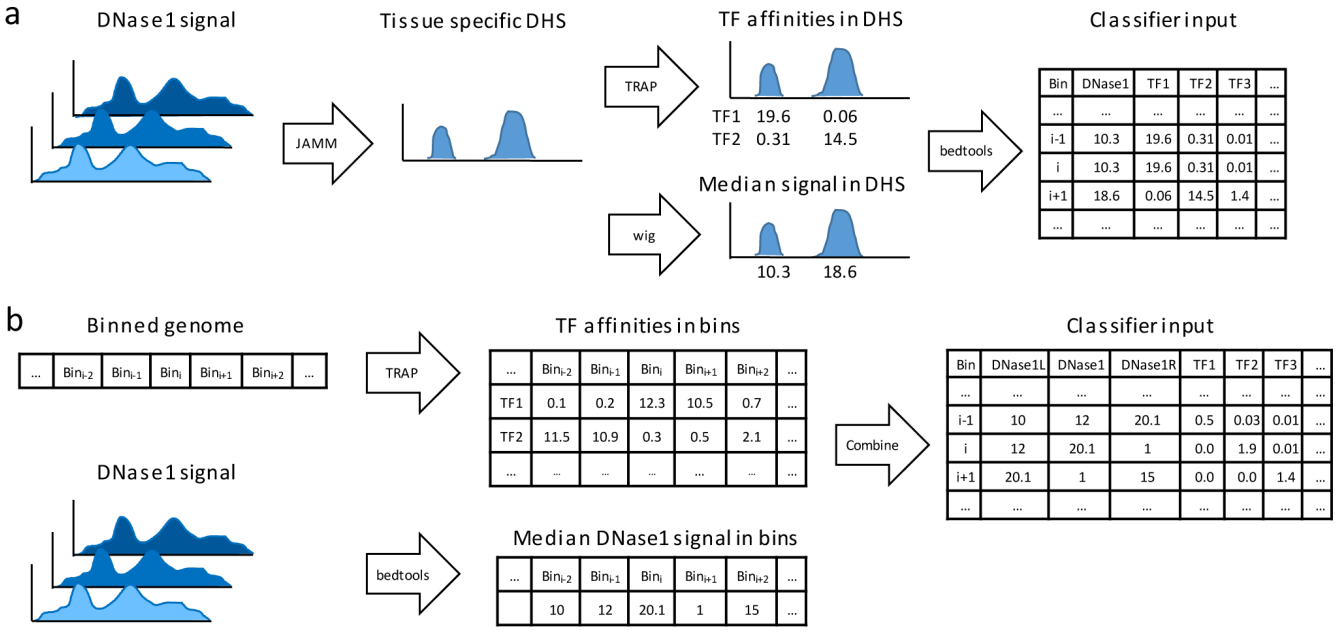

Figure 1. (**a**) Data pre-processing workflow using DNase1-seq Hypersensitive Sites (DHSs). Using JAMM, DHSs are called considering all available replicates for a distinct tissue. Transcription factor (TF) affinities in the identified DHSs are computed using TRAP for 557 TFs, the median signal of DHSs is assessed using bedtools. (**b**) An alternative data pre-processing workflow without DHSs: TF affinities and median DNase1-seq signal are computed per bin.

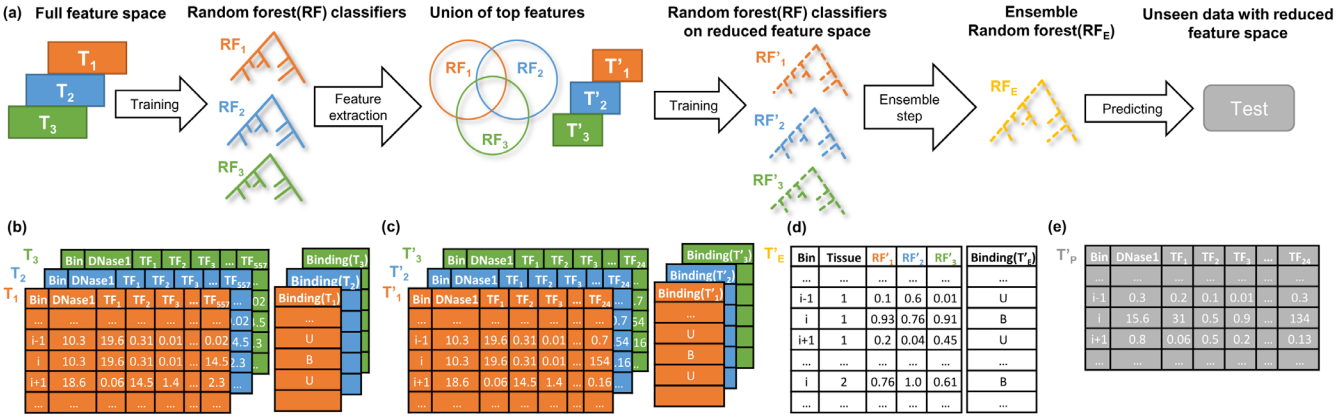

Figure 2. a) An overview of model training for a distinct transcription factor, TF, with multiple training tissues. Using the full feature matrices T_1, T_2, T_3, depicted in (**b**), TF and tissue-specific random forest (RF) classifiers are trained. From those RF classifiers (RF_1, RF_2, RF_3), we determine the union of the top 20 features from each RF. In this example, the union of top TFs is comprised of 24 TFs. Next, we design reduced tissue-specific feature matrices T'_1, T'_2, T'_3, as shown in (**c**) based on the union of the top TF features. Subsequently, tissue-specific RF classifiers (RF'_1, RF'_2, RF'_3) are trained on these reduced feature sets. The tissue-specific RF classifiers are applied to all training tissues and their predictions are aggregated to form the feature matrix T'_E, visualized in (**d**), which is used to train an ensemble model (RF_E). At the testing phase the feature matrix T'_p is fed to the trained ensemble model RF_E to predict the labels for the unseen data (**e**). Note that the column *Tissue* in **d**) is not included in the model but only shown here for illustration purposes. The feature matrices shown represent feature setup (1) using DNase1 Hypersensitive (DHS) sites.

where $Binding(T_i)$ is a vector of length n, holding the binding labels for the target TF in tissue i, and $RandomForest(.,.)$ generates the RF model trained on the features and labels provided by the first and second arguments respectively. An example of the input matrix T_i and the response vector $Binding(T_i)$ is shown in Figure 2b. In the second step, to focus only on essential regulators (c.f. Figure 3a), we shrink the feature space to the union of the top t regulators (t={10,20}) taken over all tissue and TF specific RF classifiers, T'_i, by ranking the predictors according to their *Gini index* (Figure 2c):

$$T'_i = Subset\left(T_i, \bigcup_{j=1}^{m} TopFeatures(RF_j)\right),$$

where $TopFeatures(RF_j)$ denotes the top t features of RF_j and $Subset(.,.)$ generates the reduced feature matrix based on the union of the top TFs. In the following, we refer to a training data set comprised of only one tissue as a *single tissue case* and to a training data set composed of multiple tissues as a *multi tissue case*. Considering the single tissue case, where $i = 1$ we train an RF model, RF'_i, on the reduced feature space and use this as the final model for the respective target TF:

$$RF'_i = RandomForest(T'_i, Binding(T_i)).$$

In the multi-tissue scenario, we retrain tissue-specific RF models on the reduced feature space and apply them across all available training tissues:

$$T'_E = \{prediction(RF'_i, T'_i)\}, i \in \{1,\ldots,m\},$$

where $Prediction\ (RF'_i, T'_i)$ returns the predictions made by RF'_i when applied on T'_i. Thus, T'_E is a $n \times m$ matrix with values between 0 and 1, holding the predictions of the tissue specific RFs trained for the target TF on different tissues. Matrix T'_E is used as input for the ensemble model RF_E. The ensemble model is optimised to predict the binding of the target TF based on the concatenation of predictions obtained from the training tissues. The concatenation of all binding labels is denoted by $Binding(T'_E)$, (Figure 2d):

$$RF_E = RandomForest(T'_E, Binding(T'_E)).$$

By design, the ensemble model incorporates the tissue-specific RF classifiers in a non-linear way to better generalize across all provided training tissues. An example matrix that is used to obtain predictions from an ensemble RF is shown in Figure 2e.

Performance assessment

We assessed model performance in two different scenarios: Firstly, while fitting the RF classifiers, we measure the out-of-bag (OOB) error, which is defined as the mean prediction error for each training sample using trees that were not trained on that sample. The performance on OOB data is computed in terms of the area under the precision recall curve (PR-AUC) and the area under the receiver operator characteristic curve (ROC-AUC) using the PRROC[30] package. The latter contrasts false positive rate against true positive rate, while the former contrasts precision against recall. A ROC-AUC value around 0.5 suggests a random classifier. Note that there exists no random baseline for PR-AUC.

In addition to the curve based measurements, we considered the misclassification rate separately for the *Bound* and *Unbound* classes, denoting the false negative and false positive rate, respectively:

$$Bound(False\ negative\ rate) = \frac{FN}{TP + FN}, Unbound\ (False\ positive\ rate) = \frac{FP}{TN + FP},$$

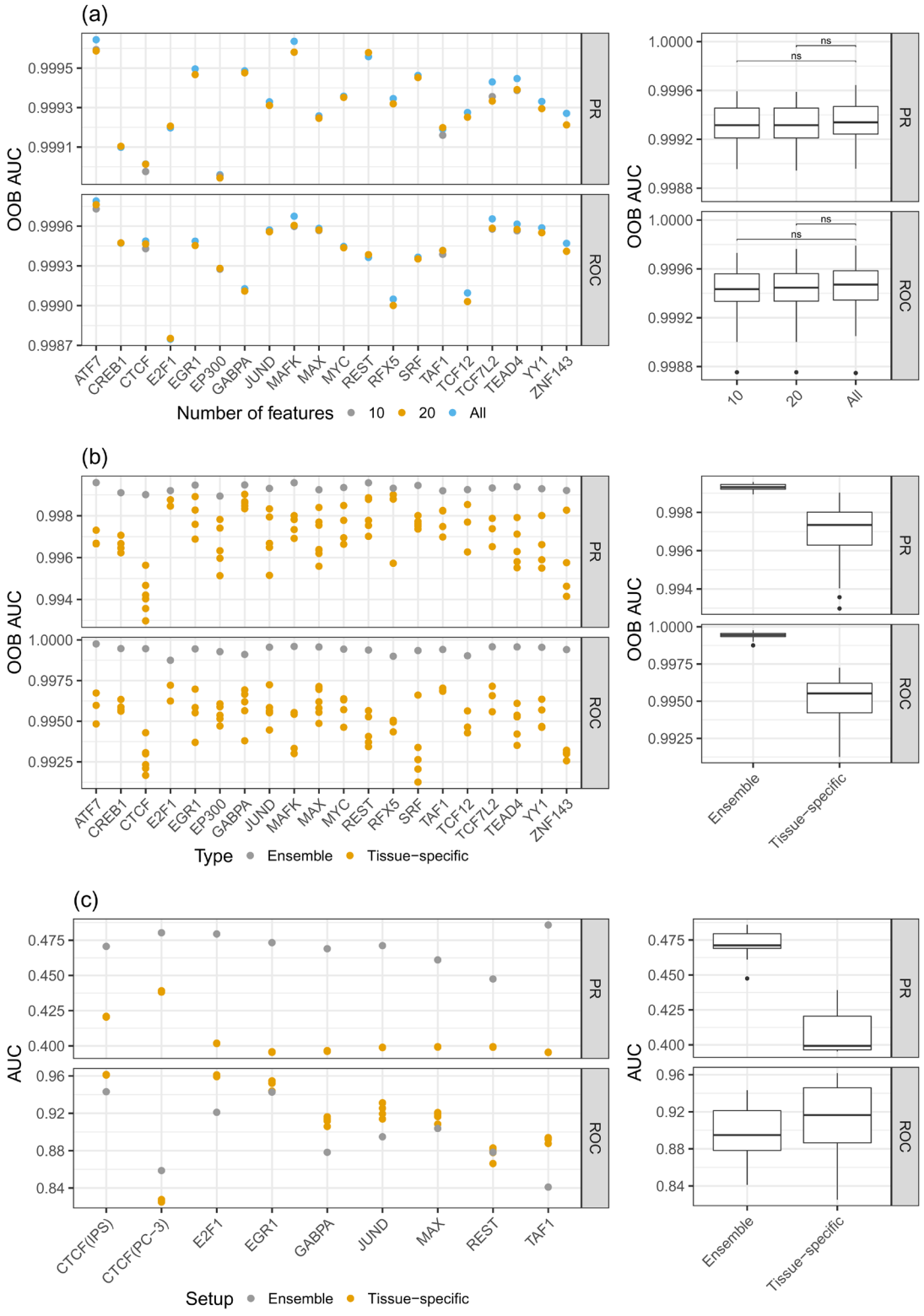

Figure 3. **a**) PR-AUC and ROC-AUC for different sets of features: considering *all* features, the top 10, and the top 20 features. One can see that the difference in model performance between the top 20 and *all* feature cases is only marginal. **b**) Comparison of the out of bag (OOB) error between ensemble models and tissue-specific random forest (RF) classifiers. The ensemble models show superior performance compared to the tissue-specific RF classifiers. **c**) PR-AUC and ROC-AUC computed on unseen test data for ensemble and tissue-specific RF classifiers. Due to the imbalanced nature of the test data, the ROC-AUC values are overly optimistic, as they are biased by the numerous unbound sites. However, the PR-AUC represents a more realistic view on the actual performance of the models. Note that the scale of the y-axes are different for the sub-figures.

where *TP* denotes the bins correctly predicted as bound, *TN* denotes the bins correctly predicted as unbound, *FP* and *FN* represent bins incorrectly predicted as bound and unbound, respectively. Note that, because we use balanced data for training the RF classifiers, the OOB is computed on a balanced data set.

Secondly, we compute the aforementioned performance measurements for a subset of the test data that was used by the *challenge* organizers. As mentioned above, the test data is composed of three held-out chromosomes, which have not been used for training: 1, 8, and 21. Additionally, TF binding is predicted on an unseen tissue, i.e. a tissue that was not used for training. An overview of the test data is provided in Table 2. Note that, in contrast to the training data, the test data is not balanced, i.e. the *Unbound* class is larger than the *Bound* class. Here, we remind the reader that PR-AUC is robust against class imbalance and thus a more appropriate performance metric for the test data than ROC-AUC as well as both false positive and false negative rates. Due to memory limitation of the PRROC package we had to downsample the test data to 100,000 samples, while preserving the original *Bound* to *Unbound* ratio.

Note that, because both suggested feature setups depicted in Figure 1 are evaluated on the same gold standard (the same test data sets), their performance can be contrasted.

Protein-protein-interaction score

By reducing the feature space of the RF models, we assumed to select TFs that are likely to interact with the target TF. To test this hypothesis systematically, we used a protein-protein-interaction score.

We obtained a customized protein-protein-interaction (PPI) probability matrix R as described previously[31], which is derived from a random walk analysis on a protein-protein-association network based on STRING[32] (version 9.05). An entry $R_{i,j}$ represents the probability that protein i interacts with protein j. Note that the probability $R_{i,j}$ is not symmetric by construction, i.e. $R_{i,j} \neq R_{j,i}$. To generate a score describing how likely it is that a subset of proteins P contained in R interact with a distinct TF t, guided by the feature importance the RF models provide, we define the PPI score $S_{t,P}$ as

$$S_{t,P} = -log(\frac{\sum_{p \in P}((R_{p,t} + R_{t,p}) \times GI(p))}{2|P|}), \quad (1)$$

where GI (p) denotes the Gini index values of p obtained from the RF model corresponding to t. Thus, the smaller the value of $S_{t,P}$ the more likely it is that the regulators in P interact with TF t.

Results

In this section, we first show that shrinking the feature space to those TFs essential for training does not affect model accuracy. Next, we demonstrate the benefits of the ensemble learning and how its accuracy is depending on the number of training tissues. We further investigate the top selected TFs by the RF models and find known interaction partners that possess high PPI scores. Finally, we compare the two feature design schemes, described in the *Methods* section, and explore their influences on model performance. If not stated otherwise, all figures presented in the following are based on annotation setup (1), focusing on DHSs.

Reducing the feature space to a small subset does not affect classification performance

Because having a sparse feature space simplifies model interpretation, we reduce the feature space to contain only a few essential features. As explained above, we determined sets of top features using the Gini index, resulting in TF and tissue-specific sets containing either the top 10 or top 20 features. As shown in Figure 3a (Supplementary Figure 3a)[33] the difference in OOB error between the feature set comprised of the top 10 or top 20 features and the full feature space is not significant. Interestingly, on test data we see a slight increase in model performance for the reduced feature space models compared to the full model. This is most likely owing to a better generalizability of the reduced feature space (Supplementary Figure 1)[34]. Due to the performance gain on test data, as well as a substantial improvement in interpretability of the models and in runtime, we decided to use a reduced feature space that consists of the top 20 features per model.

Our results indicate that the most important feature across all TFs is the DNase1-seq signal within the DHSs for feature setup (1). Similarly, in feature setup (2), the DNase1-seq signal within the bins is found to be more important than the TF features (Figure 7).

Ensemble learning improves model accuracy

According to the OOB error shown in Figure 3b (Supplementary Figure 3b)[33], the ensemble RF classifiers outperform the tissue-specific RFs, suggesting the ability of the ensemble model to generalize across tissues. Additionally, we assessed model performance on all test tissues, which are linked to multiple training tissues (Figure 3c). As illustrated in Figure 3c, the PR-AUC is higher for the ensemble models than for tissue specific RFs. Due to the imbalanced nature of the test data, we observe that ROC-AUC is actually in favor of the tissue specific models. However, this is an example for an instance where ROC-AUC is not a suitable performance metric, as it is biased by the high number of negative (i.e. unbound) cases in the test data. The superior performance of the ensemble model is also illustrated by false positive and negative rates, shown in Supplementary Figure 3c[33].

To further demonstrate the applicability of the ensemble approach, we performed a within and across tissue comparison for ensemble and tissue specific RFs. In detail, we learned tissue specific RFs for one TF in all available training tissues as well as one ensemble model. Next, we applied each classifier on each tissue and contrasted their performance (Supplementary Figure 2)[35]. We observe that the ensemble models perform either at least as good, or better, than the tissue specific models applied to the same tissue they were trained on. Further, while we see a decrease in the predictive power of tissue specific models applied across tissues, the performance of the ensemble model remains almost constant.

Overall, we conclude that ensemble learning is a promising approach to deal with the tissue-specificity of TF binding.

Increasing the number of training tissues improves prediction accuracy

Although the results in Figure 3b and 3c suggest that the ensemble methods perform well, it remains unclear what influence the number of training tissues would have on the performance of an RF. To elucidate this, we performed permutation experiments learning multiple RF models using all possible combinations of training tissues that are available for a distinct TF. As this is a computationally demanding task, we performed it for only three, arbitrarily selected, TFs: MAX, TEAD4, and E2F6. Figure 4a (Supplementary Figure 4a)[36] illustrates that the performance on OOB data improves when the number of training tissues increases. Hence, we conclude that the ability of an ensemble RF to generalize across tissues improves with larger number of training tissues.

However, it remains to be shown whether the improved accuracy obtained from the ensemble RF classifiers was in fact because of the ensemble learning. To test this, we designed two additional learning setups. Firstly, we aggregated all tissue-specific data sets into one. In other words, we pooled the training data for one TF across all available tissues into one data set. Then, we used this pooled data set to train a new RF model. Secondly, we examined another ensemble approach, which we consider to be a baseline for our actual ensemble model. In detail, we computed the average of predictions over tissue specific models in order to obtain the final prediction. As depicted in Figure 4b the true ensemble models perform better than both tested alternatives. This shows that the ensemble technique is better suited to capture tissue-specific information than simple data aggregation approaches.

Predictors selected by the RF classifiers are associated to the target TF

As stated before, we hypothesized that the top predictors selected by the RF classifiers represent regulators that either exist in protein complexes with the target TF via direct or indirect binding, or bind directly to DNA in close proximity to the target TF. To investigate this hypothesis, we computed a PPI score $S_{t,P}$ (see *Methods*) for the selected predictors P per TF t and compared it against scores computed for randomly sampled sets of TFs (based on 100 randomly drawn TF subsets). The PPI score $S_{t,P}$ for TF t is small, if t is likely to interact with the factors included in the selected predictor set P. In contrast, the score is high if t is not likely to be interacting with the factors in P. As shown in Figure 5a, except for three TFs (MAX, TAF1, ZNF143), the PPI score of the TFs selected by the RF is better (i.e. smaller) than the scores for the randomly selected set. This indicates that the RF classifiers select features representing regulators that are more likely to be interacting with the target TF, either directly or with indirect contacts.

Figure 5b provides an example of a PPI network focused on the TF *MAFK*. The network was obtained from the STRING database[32], using the settings *highest confidence* and *no more than 10 interactors* to show. The top features selected by the RF classifiers contain all known regulatory proteins in this network, except for *NFE2L2*, shown in red. Among these

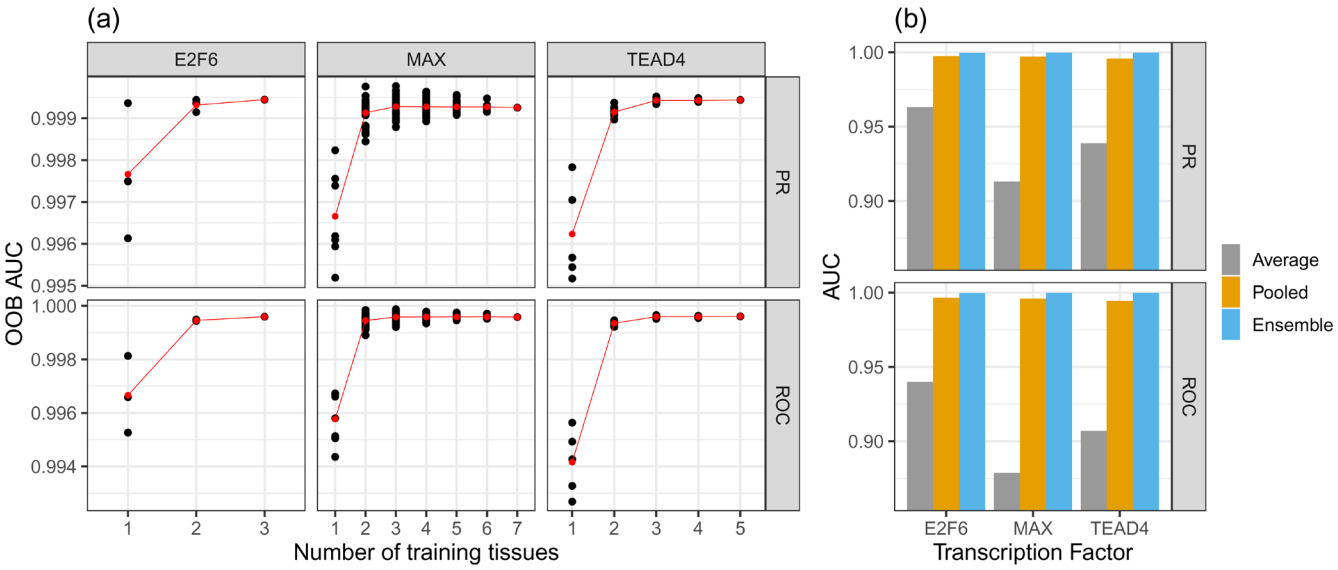

Figure 4. Comparison of tissue number and classifier setups for the three TFs E2F6, MAX, and TEAD4. a) Model performance as a function of number of tissues used for training. The OOB reduces if more tissues are included in the ensemble learning. Red dots represent the mean classification error across all tissue-specific classifiers. The black points represent individual models. **b**) Comparison between two ensemble models: averaging (takes the average of all individual RF predictions) and the RF ensemble model. In addition, one RF classifier was trained on pooled data sets comprised of training data for all available tissues for one target TF. The ensemble models perform better than the models based on aggregated data

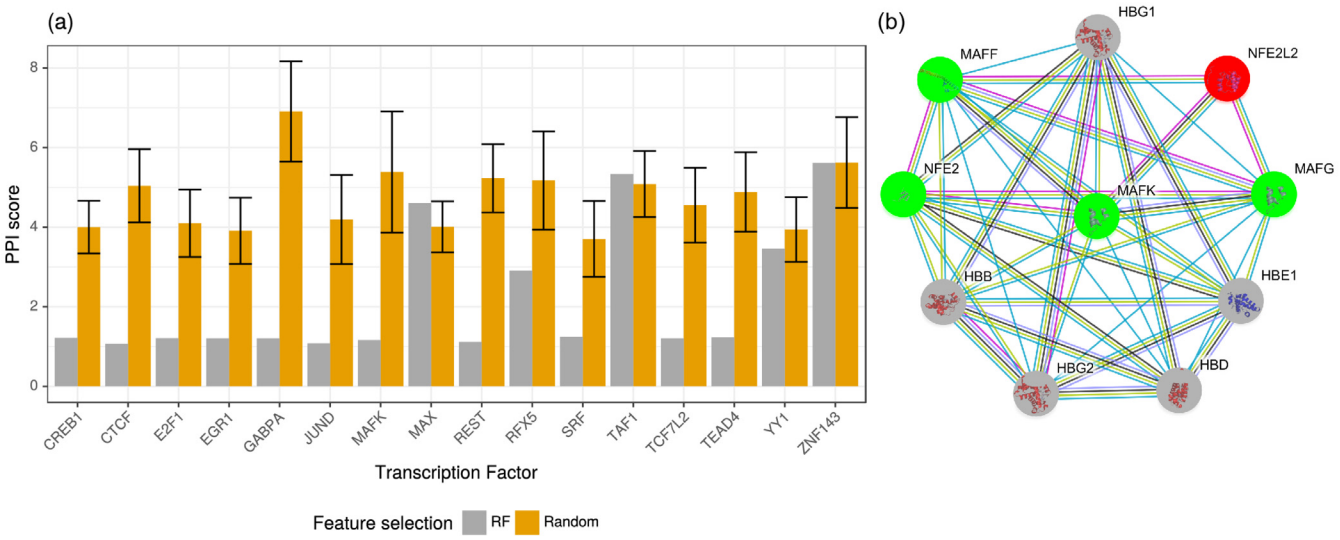

Figure 5. a) Log transformed PPI scores computed for a set of TFs. In the *Random* case, we show the mean PPI score across 100 random draws and its standard deviation. The smaller the PPI score the better. Only for three TFs (*MAX, TAF1, ZNF143*), the randomly sampled PPI score is better than or equal to the score derived for the TFs selected by the RF classifiers. **b**) PPI network obtained from STRING centered around the TF *MAFK*, highlighting proteins that interact with *MAFK* with high confidence. Proteins colored in green were identified as important features in the RF classifiers, proteins shown in grey could not be retrieved by our model, because they are DNA-binding proteins, or we do not have a PWM for them in our set. Regulators shown in red could have been detected by the RF, but were not included in the top set of regulators.

TFs are *MAFK* itself, *MAFF*, *MAFG* and *NFE2* (highlighted in green). The strong interactions among the small *MAF* proteins[37] as well as the dimerization of those with *NFE2*[38] have been reported in the literature before.

Interaction partners shown in grey cannot be identified by our approach as either these are proteins without regulatory functions or we do not have a PWM available for them.

Feature design influences the FP and FN predictions

In the conference round of the *challenge*, we were using feature setup (1), which is based on DNase1 Hypersensitive Sites (DHSs), while in the final round, we switched to design (2), which is purely based on bins. This transition had a strong effect on our performance assessed by the *challenge* organizers. While we improved the recall of our predictions by switching from (1) to (2), the precision decreased. This is reflected by the PR-AUC shown in Figure 6. Due to the unbalanced nature of the test data, which was used for this evaluation, the ROC-AUC values are less conclusive. In Supplementary Figure 5[39], we show the misclassification rates for the *Bound* and *Unbound* classes depending on the two feature designs. As suggested by the PR-AUC, the bin based models (2) outperform the peak based models in the *Bound* case, whereas the peak based models show superior performance in the *Unbound* case. At the same time, bin based models perform poorly in the *Unbound* case, which is probably driven by the strong dependence of the RF classifiers on the DNase1-seq signal. In contrast to that, models based on DHSs perform well in the *Unbound* case, because the search space for TFBSs is limited to only DHSs. This increases the precision of the predictions, but at the same time lowers the recall, which is reflected by the high misclassification rate in the *Bound* case.

In conclusion, according to PR-AUC and the individual error metrics, the peak based approach is the better choice.

Discussion and conclusion

Here, we introduced an RF based ensemble learning approach to predict TFBS *in vivo*. In this article, we did not compare our approach to competitors in the *challenge*, as this is done in the main *challenge* paper. Here, we show the benefits of ensemble learning in a multi-tissue setting and that modelling cofactors is beneficial for the classification.

We show on both test and training data that the ensemble strategy is able to generalize better across tissues, than models trained on only a single tissue (Figure 3). Also the accuracy of the ensemble classifiers increases with an increasing number of available training tissues (Figure 4a). We also illustrate that just using all available training data to learn one RF does not provide as accurate results as an ensemble model (Figure 4b). In this study, we decided to use RF classifiers, because they lead to accurate classification results using non-linear predictions in a reasonable time. Alternative classification approaches, such as logistic regression, or support-vector-machines could have been used too.

RF classifiers have also been proposed recently, independent from the *challenge*[12], as an adequate method to predict TF binding. Although the authors of 12 perform cross cell-type

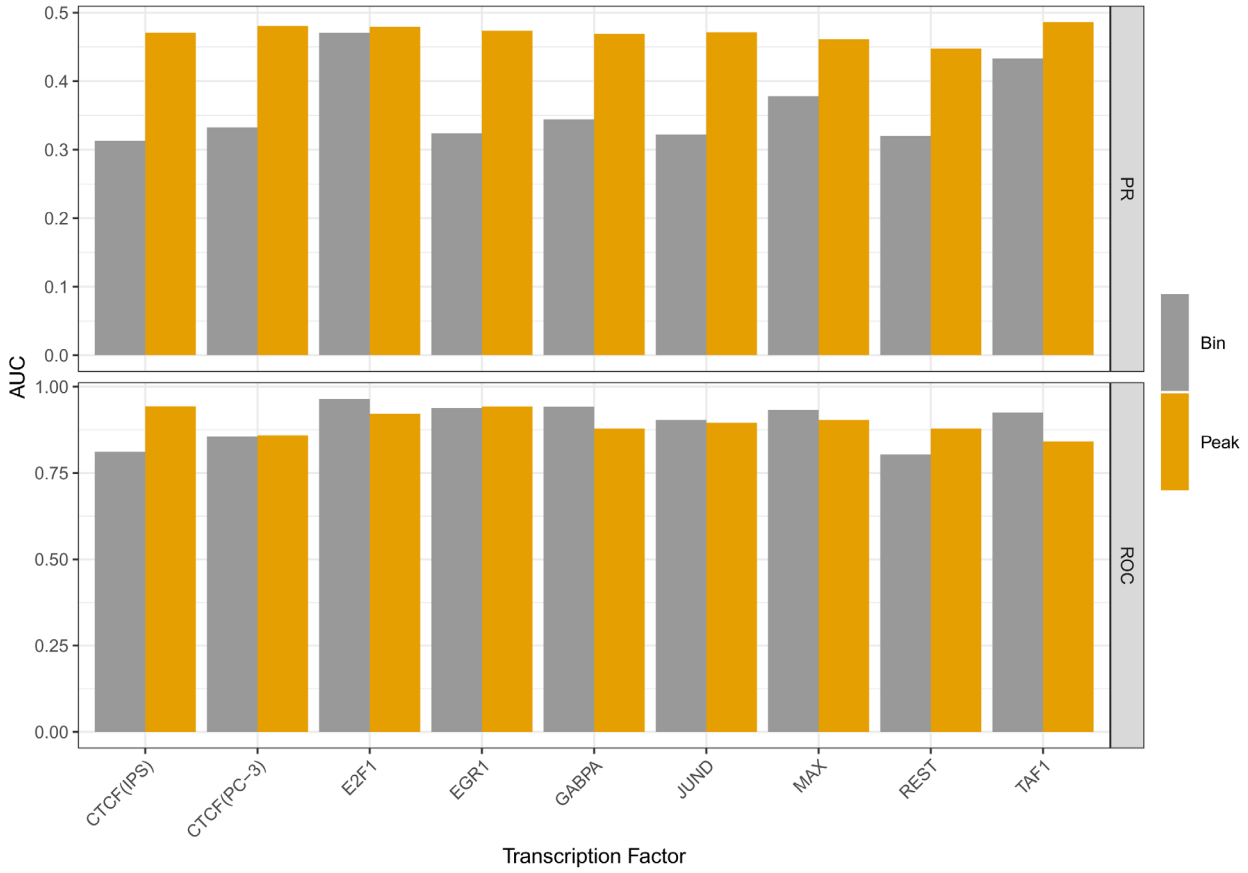

Figure 6. Comparison of PR-AUC and ROC-AUC for both feature setups computed on test data. In terms of PR-AUC, the peak based models clearly perform better than the bin based models. In terms of ROC-AUC it is less clear, however, as the test data is highly unbalanced, ROC-AUC is less reliable than PR-AUC.

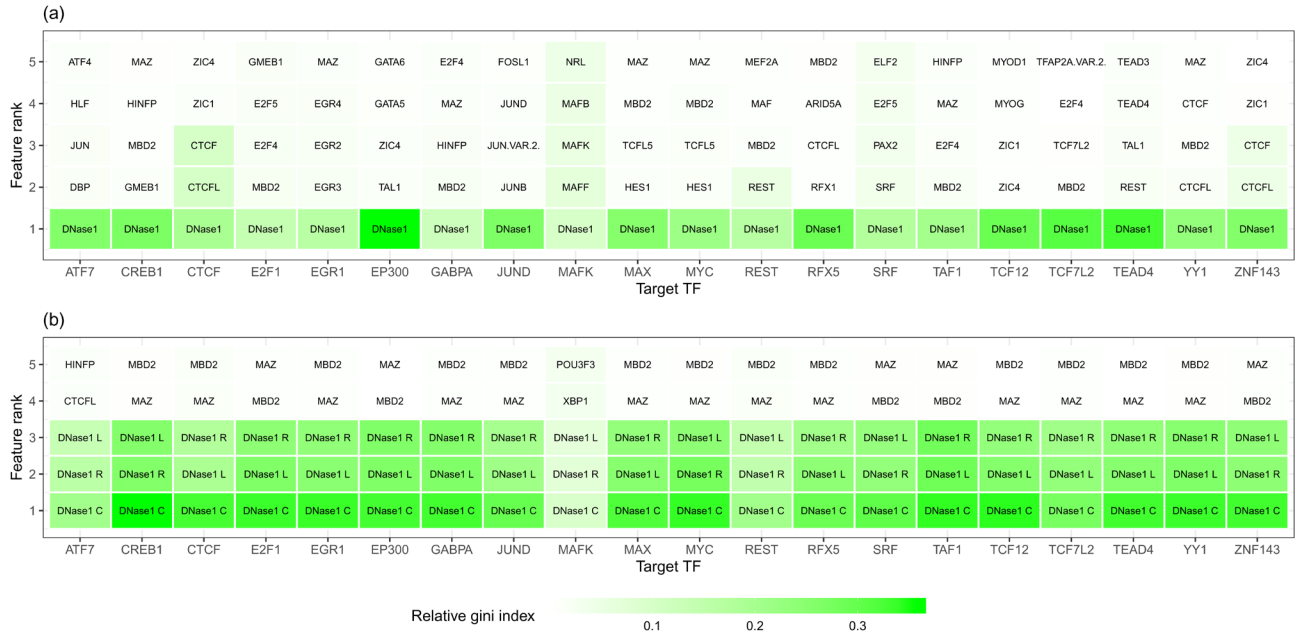

Figure 7. Top 5 features obtained from the average importance ranking of all tissue-specific classifiers for a given target TF shown on the x- axis for the peak setup (**a**) and the bin setup (**b**). Features related to DNase1 are the dominant ones.

predictions, i.e. they predict TF binding in a tissue where the RF was not trained on, they do not use ensemble models as proposed here. However, they did show that it is beneficial for the predictions of a distinct target TF to consider further TFs as predictors, in addition to the target TF itself. This is in agreement with our findings. As shown in Figure 3a, a small subset of features is sufficient to reach similar classification performance as the full feature space. We found that most of these selected TFs are known interaction partners of the target TF, see Figure 5. This is also supported by a recent study illustrating that most TFs bind in dense clusters around genes suggesting a widespread interaction among them[40].

Only for three TFs, we could not find that the predicted TFs lead to a better PPI score than a randomly chosen set. We note that for two of those three, *TAF1* and *MAX*, the performance of the ensemble RF classifiers improved only marginally, or not at all, compared to the tissue-specific classifiers. This suggests that our model does not account for the true interaction partners of those TFs. Indeed, an inspection of the STRING database for *TAF1* revealed that only *TAF1* itself and *TBP* are among the top 20 regulators, which are included in our PWM collection. For the remaining interaction partners, mostly TFs of the *TAF* family, no binding motif is available in the public repositories, thus they are not included in our PWM collection and can therefore not be used by the RF classifiers. Similarly, for *MAX*, only 5 out of 20 high confidence interaction partners are included in our PWM collection. Specifically, no PWM is available for 6 TFs interacting with MAX, while the remaining interacting proteins are not categorized as TFs. Overall, our approach benefits from data availability (Figure 4a). If there are only a few TFs available in our PWM collection, it will be harder to model the co-factor binding behavior of a TF across tissues adequately. Also, the more diverse the co-factor landscape of a TF is between the tissues, the harder it will be to learn a general model. Another crucial aspect with respect to that is the quality of the PWM. During the *challenge*, we realized that the selection of PWMs is crucial for model performance and it is required to compare PWMs obtained from different sources to make sure that one uses the one with highest information content. Nevertheless, instead of using a more recent method to model TF-motifs, we stick to the use of PWMs because they are (1) the most common way to describe the sequence specificity of TFs (2) they are available for a large number of TFs, and (3) they can be interpreted easily.

Switching the feature setup for the RF classifiers from (1) DHS-based to (2) bin-based showed that DHS sites are indispensable to the accurate TFBS predictions (Figure 6). Using only bins, without DHS information, we could improve the recall of TFBS predictions, but only at the cost of poor precision at the same time. The explanation for this behavior is a difference in size of the genomic search space between both feature setups. The bin based models have a low misclassification rate in the *Bound* case, because they do consider the whole genome without neglecting any sites beforehand, thus improving recall. However, our observations suggest that considering only the raw signal does not sufficiently correct for false positive sites, as opposed to use DHSs, which yield an improved misclassification rate in the *Unbound* case compared to the raw signal. It might be possible to overcome the strong biases of the DHS- and the bin-based models, for instance through training yet another ensemble classifier using the predictions of the DHS- and the bin-based models as input. Depending on the application, the model could be optimized for Precision, Recall, or a joint metric like PR-AUC.

In general, both training and evaluating TFBS prediction methods is challenging due to the class imbalance, i.e. there are many more *Unbound* (negative) than *Bound* (positive) binding sites in the genome. This requires both (a) training approaches that avoid over-fitting for one of the two classes and (b) evaluation strategies accounting for this issue. Here, we assess performance in terms of PR-AUC, ROC-AUC as well as misclassification rates separately for both positive and negative classes to deal with potential biases caused by the dominant *Unbound* case.

We note that our current investigation is not meant to construct a genome-wide classifier in which the unbound case is the most abundant. To achieve that, the highly unbalanced training data situation would need to be taken into account, for instance in the loss function of the classifier. Aside from the technical aspects, we show that modelling cofactors is helpful to predict TFBS and that ensemble learning is a promising technique to generalize information across tissues.

Data availability

The raw data used in this study is available online at Synapse after registration and signing of a data usage policy: https://www.synapse.org/#!Synapse:syn6112317.

Extended data

Within the Figshare repository, we provide five additional figures. Links and a brief description of the figures are provided below.

Supplementary Figure 1 (https://doi.org/10.6084/m9.figshare.9361451.v3)

PR-AUC and ROC-AUC for different sets of features: considering *all* features, the top 10, and the top 20 features on several test tissues. One can see that there is a slight advantage for the top20 and top10 model over the full model in these scenarios. The performance is shown for individual tissues in (**a**) and separately for the size of the feature matrices in (**b**).

Supplementary Figure 2 (https://doi.org/10.6084/m9.figshare.9363494.v1)

Within and cross tissue comparisons for ensemble and tissue specific RFs. Model performance is assessed in terms of (**a**) ROC-AUC and (**b**) PR-AUC.

Supplementary Figure 3 (https://doi.org/10.6084/m9.figshare.9364268.v1)

a) Classification error for the *Bound* and *Unbound* classes for different sets of features: considering *all* features, the top 10, and the top 20 features. One can see that the difference in model performance between the top 20 and *all* feature cases is only marginal. **b**) Comparison of the out of bag (OOB) error between ensemble models and tissue-specific random forest (RF) classifiers. Especially in the *Unbound* case, the ensemble models show superior performance compared to the tissue-specific RF classifiers. **c**) Misclassification rate computed on unseen test data for ensemble and tissue-specific RF classifiers. As in **b**) we see that the ensemble models generally outperform the tissue-specific ones. Note that the scale of the y-axis is different for the *Bound* and *Unbound* classes in (**a**) and (**b**).

Supplementary Figure 4 (https://doi.org/10.6084/m9.figshare.9366923.v1)

a) Relation of the OOB error for three TFs (E2F6, MAX, and TEAD4) to the number of tissues used for training. The OOB reduces if more tissues are included in the ensemble learning. Red dots represent the mean classification error across all tissue-specific classifiers. Individual models are represented by the black points. **b**) Comparison between true ensemble models for E2F6, MAX, and TEAD4 and RF classifiers trained on pooled data sets comprised of training data for all available tissues. The ensemble models perform better than the models based on aggregated data.

Supplementary Figure 5 (https://doi.org/10.6084/m9.figshare.9367895.v1)

Comparison of misclassification rate depending on the feature design computed on test data.

Software availability

Code generated as part of this analysis is available on GitHub: https://github.com/SchulzLab/TFAnalysis

Archived code at the time of publication: http://doi.org/10.5281/zenodo.1409697[41]

License: MIT

Acknowledgements

We thank everyone involved in organizing the *ENCODE-DREAM in vivo Transcription Factor binding site prediction challenge* and are grateful for the opportunity to share this article. The PPI scoring matrix used in this study was kindly provided by Sebastian Köhler.

References

1. Vaquerizas JM, Kummerfeld SK, Teichmann SA, *et al.*: **A census of human transcription factors: function, expression and evolution.** *Nat Rev Genet.* 2009; **10**(4): 252–263.
PubMed Abstract | Publisher Full Text

2. Natarajan A, Yardimci GG, Sheffield NC, *et al.*: **Predicting cell-type-specific gene expression from regions of open chromatin.** *Genome Res.* 2012; **22**(9): 1711–1722.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Berg O, von Hippel P: **Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters.** *J Mol Biol.* 1987; **193**(4): 723–750.
PubMed Abstract | Publisher Full Text

4. Stormo GD, Schneider TD, Gold L, *et al.*: **Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. *coli*.** *Nucleic Acids Res.* 1982; **10**(9): 2997–3011.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Pique-Regi R, Degner JF, Pai AA, *et al.*: **Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data.** *Genome Res.* 2011; **21**(3): 447–455.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Luo K, Hartemink AJ: **Using DNase digestion data to accurately identify transcription factor binding sites.** *Pac Symp Biocomput.* 2013; 80–91.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Gusmao EG, Dieterich C, Zenke M, *et al.*: **Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications.** *Bioinformatics.* 2014; **30**(22): 3143–3151.
PubMed Abstract | Publisher Full Text

8. Kähärä J, Lähdesmäki H: **BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data.** *Bioinformatics.* 2015; **31**(17): 2852–2859.
PubMed Abstract | Publisher Full Text

9. Yardımcı GG, Frank CL, Crawford GE, *et al.*: **Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection.** *Nucleic Acids Res.* 2014; **42**(19): 11865–11878.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Cuellar-Partida G, Buske FA, McLeay RC, *et al.*: **Epigenetic priors for identifying active transcription factor binding sites.** *Bioinformatics.* 2012; **28**(1): 56–62.
PubMed Abstract | Publisher Full Text | Free Full Text

11. O'Connor TR, Bailey TL: **Creating and validating cis-regulatory maps of tissue-specific gene expression regulation.** *Nucleic Acids Res.* 2014; **42**(17): 11000–11010.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Liu S, Zibetti C, Wan J, *et al.*: **Assessing the model transferability for prediction of transcription factor binding sites based on chromatin accessibility.** *BMC Bioinformatics.* 2017; **18**(1): 355.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Jayaram N, Usvyat D, R Martin AC: **Evaluating tools for transcription factor binding site prediction.** *BMC Bioinformatics.* 2016.
PubMed Abstract | Publisher Full Text

14. Siebert M, Söding J: **Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences.** *Nucleic Acids Res.* 2016; **44**(13): 6055–6069.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Eggeling R, Gohr A, Keilwagen J, *et al.*: **On the value of intra-motif dependencies of human insulator protein CTCF.** *PLoS One.* 2014; **9**(1): e85629.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Keilwagen J, Grau J: **Varying levels of complexity in transcription factor binding motifs.** *Nucleic Acids Res.* 2015; **43**(18): e119.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Alipanahi B, Delong A, Weirauch MT, *et al.*: **Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.** *Nat Biotechnol.* 2015; **33**(8): 831–838.
PubMed Abstract | Publisher Full Text

18. **ENCODE-DREAM *in vivo* transcritpion factor binding site prediction challenge.** 2017; Accessed: 2018-02-03.
Publisher Full Text

19. Waardenberg AJ, Homan B, Mohamed S, *et al.*: **Prediction and validation of protein-protein interactors from genome-wide DNA-binding data using a knowledge-based machine-learning approach.** *Open Biol.* 2016; **6**(9): pii: 160183.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Roider HG, Kanhere A, Manke T, *et al.*: **Predicting transcription factor affinities to DNA from a biophysical model.** *Bioinformatics.* 2007; **23**(2): 134–141.
PubMed Abstract | Publisher Full Text

21. Ibrahim MM, Lacadie SA, Ohler U: **JAMM: a peak finder for joint analysis of NGS replicates.** *Bioinformatics.* 2015; **31**(1): 48–55.
PubMed Abstract | Publisher Full Text

22. Grant CE, Bailey TL, Noble WS: **Fimo: scanning for occurrences of a given motif.** *Bioinformatics.* 2011; **27**(7): 1017–1018.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Tanay A: **Extensive low-affinity transcriptional interactions in the yeast genome.** *Genome Res.* 2006; **16**(8): 962–972.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Crocker J, Abe N, Rinaldi L, *et al.*: **Low affinity binding site clusters confer hox specificity and regulatory robustness.** *Cell.* 2015; **160**(1–2): 191–203.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Mathelier A, Fornes O, Arenillas DJ, *et al.*: **JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles.** *Nucleic Acids Res.* 2016; **44**(D1): D110–115.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Hume MA, Barrera LA, Gisselbrecht SS, *et al.*: **UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions.** *Nucleic Acids Res.* 2015; **43**(Database issue): D117–122.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Kulakovskiy IV, Vorontsov IE, Yevshin IS, *et al.*: **HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models.** *Nucleic Acids Res.* 2016; **44**(D1): D116–125.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Liaw A, Wiener M: **Classification and regression by randomforest.** *R News.* 2002; **2**(3): 18–22.
Reference Source

30. Grau J, Grosse I, Keilwagen J: **PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R.** *Bioinformatics.* 2015; **31**(15): 2595–2597.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Köhler S, Bauer S, Horn D, *et al.*: **Walking the interactome for prioritization of candidate disease genes.** *Am J Hum Genet.* 2008; **82**(4): 949–958.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Szklarczyk D, Morris JH, Cook H, *et al.*: **The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.** *Nucleic Acids Res.* 2017; **45**(D1): D362–D368.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Behjati F, Schmidt F, Schulz MH: **DREAM Challenge - Predicting TFBS - Supp3.** *figshare.* 2019; [cited 2019Aug11].
Reference Source

34. Behjati F, Schmidt F, Schulz MH: **DREAM Challenge - Predicting TFBS - Supp1.** *figshare.* 2019; [cited 2019Aug11].
Reference Source

35. Behjati F, Schmidt F, Schulz MH: **DREAM Challenge - Predicting TFBS - Supp2.** *figshare.* 2019; [cited 2019Aug11].
Reference Source

36. Behjati F, Schmidt F, Schulz MH: **DREAM Challenge - Predicting TFBS - Supp4.** *figshare.* 2019; [cited 2019Aug11].
Reference Source

37. Kannan MB, Solovieva V, Blank V: **The small MAF transcription factors MAFF, MAFG and MAFK: current knowledge and perspectives.** *Biochim Biophys Acta.* 2012; **1823**(10): 1841–1846.
PubMed Abstract | Publisher Full Text

38. Igarashi K, Kataoka K, Itoh K, *et al.*: **Regulation of transcription by dimerization of erythroid factor NF-E2 p45 with small Maf proteins.** *Nature.* 1994; **367**(6463): 568–572.
PubMed Abstract | Publisher Full Text

39. Behjati F, Schulz MH, Schmidt F: **DREAM Challenge - Predicting TFBS - Supp5.** *figshare.* 2019; [cited 2019Aug11].
Reference Source

40. Yan J, Enge M, Whitington T, *et al.*: **Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites.** *Cell.* 2013; **154**(4): 801–813.
PubMed Abstract | Publisher Full Text

41. SchulzLab, Schmidt F: **Florian411/TFAnalysis: Release for F1000 article (Version 1.0)**. *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1409697

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 30 October 2019

https://doi.org/10.5256/f1000research.22091.r53341

✔ **Gary D Stormo**
Department of Genetics and Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO, USA

The authors have adequately addressed my concerns and questions and I think the revised article is acceptable.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My expertise is in computational and experimental studies of protein-DNA interactions and the regulation of gene expression, which are relevant to this work.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 10 September 2019

https://doi.org/10.5256/f1000research.22091.r53342

✔ **Jan Grau**
Institute of Computer Science, Martin Luther University of Halle-Wittenberg (MLU), Halle, Germany

The authors have carefully addressed all of my previous concerns.

Competing Interests: We have participated in the same challenge (ENCODE-DREAM) as the authors and the data presented here are closely related to that challenge.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 07 January 2019

https://doi.org/10.5256/f1000research.17691.r42084

?

Gary D Stormo
Department of Genetics and Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO, USA

This paper reports the results of this group's entry into the ENCODE-DREAM challenge. The task of the challenge was to learn a model for binding of a target TF based on ChIP-seq data and DHS data from different cell types, and to predict binding on held-out data. They focused on a subset of 12 TFs. There were two types of held-out data, three chromosomes from the same cell types as the training data, and also data from different cell types not used in training. Results are reported as classification errors independently for bound and unbound sites.

This group did not try to learn a model (such as PWM) for the target TF, rather they used existing PWM models, available in databases, for the target TF as well as for 556 other TFs (so 557 in total; when more than one PWM was available for a TF they used the one with the highest information content). They employed a random forest (RF) approach for learning the model, and they compared variations on how the training was performed.

There isn't yet a summary publication of the results of the challenge, so at this time we do not know how this approach compares to others. But there are some results reported that are interesting to know regardless of the ranking of this approach.

One variation they tested was training using all of the features (a DHS score and all of the PWM scores) versus only subsets, and ranking the features to see which are most important. They found that using only the top 20 features was essentially as good as all of them, whereas the top 10 was not. Not surprisingly, the DHS score is the most important feature. They don't state it, but I assume that the PWM for the target TF is the next most important. Is that right? It is also reassuring that the set of other TFs that rank highest in importance are enriched for TFs previously shown to interact with the target TF, indicating that their models are learning something about the coordinated regulation by multiple TFs.

They also compared prediction accuracy on models trained on individual tissue type data, versus a model trained on all of the tissue data merged together, versus an ensemble model obtained from all of the tissue types, with each treated independently. The ensemble models performed significantly better than the others (although I would like to see a separation of results on the different types of test data - see comments below). And the models improve with additional tissue types, although for most TFs the

improvement is marginal beyond three.

Comments and suggestions:

1. Their reporting of results is less informative than it could be. For example, instead of just reporting a classification error for each class (bound and unbound) they could show ROC or PRC curves that indicate those errors for a range of thresholds. Is the reason they don't do that because their program simply returns a binary result, bound/unbound, rather than a probability (or some quantitative score) of being bound? The results as reported highlight the intrinsic tradeoff between false positive and false negative predictions because they vary rather dramatically between different test sets, but don't provide any guidance of how one might balance the two to obtain "optimal" predictions (where optimal may depend on the usage).

2. In Figure 3c they show results on the two types of held-out data, from left out chromosomes from the same tissues as the training data and from data from different tissues. I would like to see those two types of test data reported separately. I can easily imagine that testing on left out chromosomes from the same tissue would provide better predictions, because the same set of additional TFs are utilized within the same tissue, but that on different tissues that might not be the case and that the ensemble approach might be especially useful.

3. I'm a little confused about the differences in the two training methods shown in Figure 1, and I think some clarification is needed. 1a is clear enough, they are just using genomic regions under DHS peaks (in a given tissue), and the training involves those that are bound by the TF (in that same tissue) and those that are not. But in 1b, is the whole genome binned (and what are bin sizes, I didn't see that stated)? And then is the training on that whole genome, so that the unbound training data enormously larger than the bound data (in fact the vast majority of the genome is not under DHS peaks so its relevance isn't clear). And then when testing the models obtained from the binned training, do they make predictions on the whole genome, or only on the DHS regions? They report that training on binned data was better, but it isn't clear to me is the assessments were the same (such as testing on the whole genome versus under the DHS peaks) which may make a difference.

4. The word "inadmissible" occurs twice, once in the Introduction and once in the Discussion. It doesn't seem to be the right word in either case, in fact based on the context I think it is opposite of what they mean. For example, the first occurrence is "(TFs) are inadmissible to maintain and establish cellular identity....". I think "essential" or "required" are more appropriate.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My expertise is in computational and experimental studies of protein-DNA interactions and the regulation of gene expression, which are relevant to this work.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 Jul 2019

Fatemeh Behjati Ardakani, Max Planck Institute for Informatics, Germany

1.Their reporting of results is less informative than it could be. For example, instead of just reporting a classification error for each class (bound and unbound) they could show ROC or PRC curves that indicate those errors for a range of thresholds. Is the reason they don't do that because their program simply returns a binary result, bound/unbound, rather than a probability (or some quantitative score) of being bound?
We agree with the reviewer that ROC and PR curves are meaningful error measures. We did not choose those initially as we believed that the misclassification rate is a more intuitive measure. Our models do allow us to compute ROC and PR curves. In the revised version of the article, we report the area under the precision recall curve (AU-PR) as well as the area under the receiver operator characteristic curve (AUC-ROC). We have moved the misclassification to the Supplement.

The results as reported highlight the intrinsic tradeoff between false positive and false negative predictions because they vary rather dramatically between different test sets, but don't provide any guidance of how one might balance the two to obtain "optimal" predictions (where optimal may depend on the usage).

We thank the reviewer for pointing out to us that the two proposed setups could be combined. It is a thought that did not occur to us. One option would be to combine the predictions obtained using both feature setups in yet another ensemble RF model. The balancing could be controlled by a customized penalization of the model, such that either Precision, Recall, or both are optimized. We addressed this point in the discussion of our article.

==========

2. In Figure 3c they show results on the two types of held-out data, from left out chromosomes from the same tissues as the training data and from data from different tissues. I would like to see those two types of test data reported separately. I can easily imagine that testing on left out chromosomes from the same tissue would provide better predictions, because the same set of additional TFs are utilized within the same tissue, but that on different tissues that might not be the case and that the ensemble approach might be especially useful.

We thank the reviewer for this suggestion. Indeed we see that the ensemble model predicting

tissue X as well as the classifier trained only on chromosomes of tissue X, perform equally well. In contrast when evaluating the classifiers on other cell types, the ensemble method performs better than any other classifier trained on only one tissue. The results are shown in Supplementary Figure 2.

==========

3. I'm a little confused about the differences in the two training methods shown in Figure 1, and I think some clarification is needed. 1a is clear enough, they are just using genomic regions under DHS peaks (in a given tissue), and the training involves those that are bound by the TF (in that same tissue) and those that are not. But in 1b, is the whole genome binned (and what are bin sizes, I didn't see that stated)? And then is the training on that whole genome, so that the unbound training data enormously larger than the bound data (in fact the vast majority of the genome is not under DHS peaks so its relevance isn't clear). And then when testing the models obtained from the binned training, do they make predictions on the whole genome, or only on the DHS regions? They report that training on binned data was better, but it isn't clear to me is the assessments were the same (such as testing on the whole genome versus under the DHS peaks) which may make a difference.

We agree with the reviewer that this is a bit unclear without more detailed information on the challenge setup itself. We have added a description on the training, test, and benchmarking data provided by the challenge organizers to the main text. As stated there,
the challenge's objective was to predict TF binding in bins of size 200bp, shifted by 50bp each. Predictions are computed for all bins in chromosomes 1, 8, and 21, the remaining chromosomes are used for training. To train the models, all bound bins in training chromosomes as well as an equal number of randomly sampled unbound bins have been used. The DNase1-seq signal in these bins is what is used in the setup described in Figure 1b. We believed that using the RFs to learn an association between DNase1-seq signal and TF binding might outperform a peak-calling based method, therefore we have pursued this approach as well.
The models are assessed on the bin level for both setups. In Setup 1 (Fig1a), any bin not overlapping a DHS is predicted as unbound per default, bins overlapping a DHS are subjected to classification. In Setup 2 (Fig1b) each bin is classified. Thus, the setups can be compared.
We have improved the description of Setup 2 (Fig1b) in the main text.

==========

4. The word "inadmissible" occurs twice, once in the Introduction and once in the Discussion. It doesn't seem to be the right word in either case, in fact based on the context I think it is opposite of what they mean. For example, the first occurrence is "(TFs) are inadmissible to maintain and establish cellular identity....". I think "essential" or "required" are more appropriate.

We thank the reviewer for spotting this mistake. We meant to say indispensable.

Competing Interests: No competing interests were disclosed.

Reviewer Report 26 October 2018

https://doi.org/10.5256/f1000research.17691.r39062

?

Jan Grau
Institute of Computer Science, Martin Luther University of Halle-Wittenberg (MLU), Halle, Germany

Transcriptional regulation by transcription factors (TFs) is one of the fundamental steps of gene regulation. Hence, knowing the genome-wide binding regions of a TF is of great interest. Experimentally, those could be determined by ChIP-seq, which, however, is time-consuming and labor-intensive. Hence, computational prediction of cell type-specific, in-vivo transcription factor binding is highly demanded.
In their manuscript "Predicting transcription factor binding using ensemble random forest models", Ardakani, Schmidt and Schulz present a novel method for this purpose, which is based on PWMs describing TF sequence preference, and DNase-seq data capturing chromatin accessibility. This method combines i) learning random forest (RF) classifiers on feature matrices for individual cell types, ii) shrinking feature sets, and iii) learning ensemble classifiers across cell types. The authors illustrate that within their method, peak-based DNase features seem to be favorable compared with bin-based aggregation of DNase-seq coverage. Furthermore, they demonstrate that the ensemble classifier indeed yields an overall improved performance compared with cell type-specific RFs.
As this is a companion paper to the main publication describing the results of the ENCODE-DREAM challenge, I consider a direct comparison to other approaches dispensable in this case.

In general, most of the methods are well described and conclusions are supported by the data.
However, I have a few major and several minor comments regarding choices made by the authors (especially with regard to performance assessment) and the presentation of specific details of methods and results, as detailed in the following.

Major comments:

1. In sub-section "Data" of the Methods section, the authors state that they "focus on these 12 TFs in the scope of this article". However, this is contradicted by the list provided in Table 2 listing only 8 TFs. Results for the same 8 TFs are also shown in Fig. 6, whereas several of the remaining figures (Fig 3a/b, Fig 5) present results for a larger set of TFs, i.e., for TFs not listed in sub-section "Data".

2. The third paragraph of sub-section "Data preprocessing and feature generation" of the Methods section is lacking details. How exactly are "tissue-specific DHSs" called using JAMM? What have been the inputs and input formats? Which peaks are merged and why?

3. Results with regard to feature shrinkage (Fig. 3a) are only shown for OOB Misclassification. As I could imagine over-fitting effects to specifics of the training cell types, I considered an evaluation on the test data highly informative. For instance, I would imagine that we see a decrease in OOB performance when shrinking features to the top 20, whereas on the test data this model achieves a better generalization and, hence, misclassification rate.

4. The authors chose to use misclassification separated by classes, which could also be described as false negative rate and false positive rate, as performance measure for the whole manuscript.

For several reasons, I would consider curve-based measures, especially the (area under the) precision-recall curve the more appropriate measure for this application but also in the context of the ENCODE-DREAM challenge.
First, we face a highly imbalanced classification problem, and the precision-recall curve has been shown to be highly informative in this case[1].
Second, the areas under the ROC curve and precision-recall curve have also been used for performance assessment in the ENCODE-DREAM challenge and choosing the same performance measure in this paper would foster comparison of results to those of the challenge (especially since both use the same test data).
Third, in the discussion of Fig. 6, the authors mention that one choice of DNase data works better for bound regions, while the other works better for unbound regions. Here, we face the typical trade-off between sensitivity and specificity (or false negative rate and false positive rate), where we are unable to decide for one option based on specific, contradictory combinations of the two measures. In the ROC curve, basically (1 - FN/(TP+FN)) would be plotted against FP/(TN+FP), so we would get a broader impression of classifier performance, including the specific points on the curve chosen by the authors.
For these reasons, the area under the ROC curve and the area under the precision-recall curve should be included as performance measures into this study. As the authors illustrate in Fig. 2d, RF classifiers already output continuous scores that could be used for computing these curves. Technically, curves and AUC values could be computed, e.g., using the R packages PRROC or precrec.

5. In sub-section "Ensemble learning improves model accuracy" of the Results section, I agree with the authors that the ensemble classifier performs better than the individual RFs. However, currently it remains unclear if this can really be attributed to "ensemble learning" or just to averaging effects. Hence, I would suggest to include a simple averaging over the predictions of individual RFs (those, for which the predictions are also input of RF_E) as a simple baseline model (in addition to the single RF learned on the pooled data).
In addition, for MAX, the authors might also include results for the test data in addition to what is shown in Figure 4.

Minor comments:

6. In the Introduction, second paragraph, the authors state that "Most of these methods are based on position weight matrices (PWMs) describing the sequence preference of TFs," giving a reference to the publication of the 2016 update of the Jaspar database. While Jaspar indeed provides PWM models, I do not consider this an appropriate reference for the definition of PWMs in general. Specifically, I would suggest to cite the seminal works of Berg & von Hippel [2] and of Stormo[3] instead.

7. In the Introduction, second paragraph, the authors state "PWMs indicate [...] which nucleotide is most likely to occur". From my perspective, this description is not fully accurate. The most likely nucleotide is also represented by consensus sequences. PWMs give a specific weight (or log-probability,...) for each of the nucleotides and not only for the most likely one.

8. I appreciate that the authors reference our work regarding dependency models (Slim models) in the second paragraph of the introduction. However, there are several other approaches for modeling dependencies in TF binding sites. I would encourage the authors to broaden the scope of their references by including, e.g.[4-5].

9. In the third paragraph of the introduction, the authors refer to "the main ENCODE-DREAM Challenge paper". I am aware that this paper has not yet been published, but encourage the authors to update their publication including a reference to that paper when available.

10. In the second paragraph of sub-section "Data preprocessing and feature generation" of the Methods section, it is mentioned that TF binding affinities are computed for 557 distinct TFs. After reading the complete paper, I understood (hopefully correctly) that all 557 TFs are used for all RFs (before shrinking the feature space) regardless of the training TF. If my understanding is correct, the authors might consider to include an explicit statement about this fact already at this stage of the manuscript.

11. In the first paragraph of sub-section "Ensemble random forest classifier" of the Methods section, the authors state that "the balance between the bound and unbound classes is maintained to avoid over-fitting". For me, it remains unclear how exactly this helps to avoid over-fitting. For my understanding, over-fitting typically refers to an over-adaption to specifics of the training data, which do not generalize well to other data sets, leading to a poor performance on unseen (test) data. However, the class imbalance is inherent to the problem and should be (roughly) the same for training and test cell types. Please clarify.

12. In the first paragraph of sub-section "Ensemble random forest classifier" of the Methods section, right before the second formula, the shrunken feature space is described to be the union of top 20 regulators. However, later in the Results section, the authors also consider a case where features per RF are restricted to the top 10 ones (Fig 3a). Hence, I would suggest a generic description, here.

13. The third formula of sub-section "Ensemble random forest classifier" of the Methods section refers to an index i, where (for my understanding), according to the previous definition, i should be in {1}, in this case. If that is indeed the case, I would suggest to replace "i" by "1" in the formula and explicitly state that this is the only index i can be.

14. The fourth formula of sub-section "Ensemble random forest classifier" of the Methods section is partly broken. Specifically, the element sign refers to the set of indexes, which does not seem reasonable to me. I rather think this should refer to the matrix resulting from prediction(RF_i',T_i') Please fix.

15. In Figure 2 (b), (c) and (e), the labels in the table cells are hardly legible in printout. Either increase the thickness of letters or chose a different color.

16. For Figure 2e, it remains unclear from the caption what is shown. It seems to be the input matrix derived from test data, in analogy to the training matrices shown in Figure 2b? Is this the input of each RF? Of RF' (as features might have been shrunken)? Or of RF_E?

17. The fifth formula of sub-section "Ensemble random forest classifier" of the Methods might profit from a bit of additional explanation. Specifically, it took me a while to understand (if I'm right) that for T_E', the outputs of all individual RFs are concatenated row-wise, while "Binding(T_E')" denotes the concatenation of training labels.

18. In the first paragraph of sub-section "Performance assessment" of the Methods section, I wondered what the index "i" refers to. Is this the same index i as before (i.e., an index for the training cell types)? If not, what exactly is "sample i"?

19. In sub-section "Protein-protein-interaction score" of the Methods section, I would have appreciated a

bit more motivation before describing the method itself.

20. In sub-section "Reducing the feature space to a small subset [...]" of the Results section, I would not fully agree with the authors that the difference in error between the full model and the model based on top 20 features is "marginal". I would even assume that a statistical test of the difference between the data behind the two boxplots in Fig. 3a would be significant.

21. In sub-section "Reducing the feature space to a small subset [...]" of the Results section, I did not find the last two sentences (regarding importance of DNase-based features) to be supported by the data shown in the manuscript.

22. In section "Data availability", the authors provide a link to the synapse page of the ENCODE-DREAM challenge. However, the data are accessible only after registration and signing a data usage policy.

23. Typos & Grammar:
- first paragraph of "Data preprocessing and feature generation": "down sampled" should be "down-sampled"
- second paragraph of "Data preprocessing and feature generation": "the course of challenge" should be "the course of the challenge"
- third paragraph of "Data preprocessing and feature generation": "data is intersected" should be "data are intersected"
- 7th paragraph of "Discussion and conclusions": "Bound(positive)" should be "Bound (positive)"
- Reference 15: "transcritpion" should be "transcription"

References

1. Saito T, Rehmsmeier M: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.*PLoS One*. 2015; **10** (3): e0118432 PubMed Abstract | Publisher Full Text
2. Berg O, von Hippel P: Selection of DNA binding sites by regulatory proteins. *Journal of Molecular Biology*. 1987; **193** (4): 723-743 Publisher Full Text
3. Stormo G, Schneider T, Gold L, Ehrenfeucht A: Use of the 'Perceptron' algorithm to distinguish translational initiation sites inE. coli. *Nucleic Acids Research*. 1982; **10** (9): 2997-3011 Publisher Full Text
4. Siebert M, Söding J: Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*. 2016; **44** (13): 6055-6069 Publisher Full Text
5. Eggeling R, Gohr A, Keilwagen J, Mohr M, Posch S, Smith AD, Grosse I: On the value of intra-motif dependencies of human insulator protein CTCF.*PLoS One*. 2014; **9** (1): e85629 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: We have participated in the same challenge (ENCODE-DREAM) as the authors and the data presented here are closely related to that challenge.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 Jul 2019

Fatemeh Behjati Ardakani, Max Planck Institute for Informatics, Germany

1. In sub-section "Data" of the Methods section, the authors state that they "focus on these 12 TFs in the scope of this article". However, this is contradicted by the list provided in Table 2 listing only 8 TFs. Results for the same 8 TFs are also shown in Fig. 6, whereas several of the remaining figures (Fig 3a/b, Fig 5) present results for a larger set of TFs, i.e., for TFs not listed in sub-section "Data".

We thank the reviewer for spotting this. There was indeed an error in Table 1 and some TFs were missing. We have corrected Table 1 to list all TFs considered in Figures 3 and 5. In Fig.6, as well as in Fig.3c, we show results on test data from the challenge, therefore there are fewer TFs than in the remaining Figures. As we only look at the multi-tissue cases, for which there are more than one training tissue per TF available, we use only a subset of the available challenge data.

==========

2. The third paragraph of sub-section "Data preprocessing and feature generation" of the Methods section is lacking details. How exactly are "tissue-specific DHSs" called using JAMM? What have been the inputs and input formats? Which peaks are merged and why?

We have clarified these points in the main text. We converted the provided DNase1-seq bam files to bed files using the bedtools bamtobed command. For each bed file, peaks are computed separately using JAMM's standard parameters and the –f 1 option. The individual DHS files obtained for one TF are aggregated using the bedtools merge command. We decided to take a less conservative approach and merge all peaks identified in individual replicates per TF to ensure that we do not miss an accessible site, all be it this may introduce false positives.

==========

3. Results with regard to feature shrinkage (Fig. 3a) are only shown for OOB Misclassification. As I could imagine over-fitting effects to specifics of the training cell types, I considered an evaluation on the test data highly informative. For instance, I would imagine that we see a decrease in OOB performance when shrinking features to the top 20, whereas on the test data this model achieves a better generalization and, hence, misclassification rate.

We appreciate the suggestions and performed the same experiment as shown in Figure 3a using the challenge's test data (Supplementary Figure 1). As expected, we find a slight decrease in terms of OOB performance for the top10 and top20 models compared to all features, whereas on test data we see that both the top10 and top20 models perform slightly better than models considering all features. However, we note that the differences are not significant.

==========

4. The authors chose to use misclassification separated by classes, which could also be described as false negative rate and false positive rate, as performance measure for the whole manuscript.

We have mentioned these more established names in the main manuscript. However, we decided to stick to the already used nomenclature, as we believe that it is more comprehensible in the context of the TF binding predictions.

For several reasons, I would consider curve-based measures, especially the (area under the) precision-recall curve the more appropriate measure for this application but also in the context of the ENCODE-DREAM challenge.
First, we face a highly imbalanced classification problem, and the precision-recall curve has been shown to be highly informative in this case1.
Second, the areas under the ROC curve and precision-recall curve have also been used for performance assessment in the ENCODE-DREAM challenge and choosing the same performance measure in this paper would foster comparison of results to those of the challenge (especially since both use the same test data).
Third, in the discussion of Fig. 6, the authors mention that one choice of DNase data works better for bound regions, while the other works better for unbound regions. Here, we face the typical trade-off between sensitivity and specificity (or false negative rate and false positive rate), where we are unable to decide for one option based on specific, contradictory combinations of the two measures. In the ROC curve, basically (1 - FN/(TP+FN)) would be plotted against FP/(TN+FP), so we would get a broader impression of classifier performance, including the specific points on the curve chosen by the authors.

For these reasons, the area under the ROC curve and the area under the precision-recall curve should be included as performance measures into this study. As the authors illustrate in Fig. 2d, RF classifiers already output continuous scores that could be used for computing these curves. Technically, curves and AUC values could be computed, e.g., using the R packages PRROC or precrec.

We agree with the reviewer that curve based scores like PR and ROC are better to assess the performance of our models. As suggested, we used the PRROC package to compute AUC values for PR and ROC curves and use these measures throughout the article. It is worth noting that the PR values deliver a more realistic impression on model performance than ROC or the misclassification rate on the highly unbalanced test data sets, which are enriched with negative cases, i.e. unbound sites. We moved the previous figures using the misclassification rate to the Supplement.

==========

5. In sub-section "Ensemble learning improves model accuracy" of the Results section, I agree with the authors that the ensemble classifier performs better than the individual RFs. However, currently it remains unclear if this can really be attributed to "ensemble learning" or just to averaging effects. Hence, I would suggest to include a simple averaging over the predictions of individual RFs (those, for which the predictions are also input of RF_E) as a simple baseline model (in addition to the single RF learned on the pooled data).

We agree with the reviewer's comment, and as suggested, we added another model averaging over the predictions of the tissue specific RFs as a baseline for our ensemble models. As shown in Figure 4b, the averaging leads to a worse performance than simply pooling the information across all samples into one model, indicating that the ensemble step does indeed combine tissue specific information in a more sophisticated way than a simple average.

In addition, for MAX, the authors might also include results for the test data in addition to what is shown in Figure 4.

In the interest of clarity and homogeneity of the analysis, we refrained from performing the analysis for MAX additionally on test data.

Minor comments:

6. In the Introduction, second paragraph, the authors state that "Most of these methods are based on position weight matrices (PWMs) describing the sequence preference of TFs," giving a reference to the publication of the 2016 update of the Jaspar database. While Jaspar indeed provides PWM models, I do not consider this an appropriate reference for the definition of PWMs in general. Specifically, I would suggest to cite the seminal works of Berg & von Hippel 2 and of Stormo 3 instead.

We agree with the reviewer and have changed the citation.

==========

7. In the Introduction, second paragraph, the authors state "PWMs indicate [...] which nucleotide is most likely to occur". From my perspective, this description is not fully accurate. The most likely nucleotide is also represented by consensus sequences. PWMs give a specific weight (or log-probability,...) for each of the nucleotides and not only for the most likely one.

This is true. We adapted the wording in the main text to avoid the confusion.

==========

8. I appreciate that the authors reference our work regarding dependency models (Slim models) in the second paragraph of the introduction. However, there are several other approaches for modeling dependencies in TF binding sites. I would encourage the authors to broaden the scope of their references by including, e.g.4-5.

We appreciate the hint and have included the suggested literature.

==========

9. In the third paragraph of the introduction, the authors refer to "the main ENCODE-DREAM Challenge paper". I am aware that this paper has not yet been published, but encourage the authors to update their publication including a reference to that paper when available.

Yes, we will cite this paper once it is available. Up to the submission of this revised version of our article, the challenge paper has not yet been published.

==========

10. In the second paragraph of sub-section "Data preprocessing and feature generation" of the Methods section, it is mentioned that TF binding affinities are computed for 557 distinct TFs. After reading the complete paper, I understood (hopefully correctly) that all 557 TFs are used for all RFs (before shrinking the feature space) regardless of the training TF. If my understanding is correct, the authors might consider to include an explicit statement about this fact already at this stage of the manuscript.

Yes, the modelling is performed exactly in that way. We have improved the wording to make this more pronounced at the specified position in the main text.

==========

11. In the first paragraph of sub-section "Ensemble random forest classifier" of the Methods section, the authors state that "the balance between the bound and unbound classes is maintained to avoid over-fitting". For me, it remains unclear how exactly this helps to avoid over-fitting. For my understanding, over-fitting typically refers to an over-adaption to specifics of the training data, which do not generalize well to other data sets, leading to a poor performance on unseen (test) data. However, the class imbalance is inherent to the problem and should be (roughly) the same for training and test cell types. Please clarify.

Yes, the term “overfitting” might have been inaccurate when class imbalance was considered. We mean the class imbalance effects on training would've been attenuated by keeping the balance between bound an unbound in our training data. The class distribution of the test data, however, would not be a problem given that the models are fairly and reliably trained.

==========

12. In the first paragraph of sub-section "Ensemble random forest classifier" of the Methods section, right before the second formula, the shrunken feature space is described to be the union of top 20 regulators. However, later in the Results section, the authors also consider a case where features per RF are restricted to the top 10 ones (Fig 3a). Hence, I would suggest a generic description, here.

We thank the reviewer to point out this inconsistency. We have replaced the numeric value by a generic description.

==========

13. The third formula of sub-section "Ensemble random forest classifier" of the Methods section

refers to an index i, where (for my understanding), according to the previous definition, i should be in {1}, in this case. If that is indeed the case, I would suggest to replace "i" by "1" in the formula and explicitly state that this is the only index i can be.

The observation is correct. We adapted the text accordingly.

==========

14. The fourth formula of sub-section "Ensemble random forest classifier" of the Methods section is partly broken. Specifically, the element sign refers to the set of indexes, which does not seem reasonable to me. I rather think this should refer to the matrix resulting from prediction(RF_i',T_i') Please fix.

We thank the reviewer for spotting this mistake. We have corrected it.

==========

15. In Figure 2 (b), (c) and (e), the labels in the table cells are hardly legible in printout. Either increase the thickness of letters or chose a different color.

We have increased the font size.

==========

16. For Figure 2e, it remains unclear from the caption what is shown. It seems to be the input matrix derived from test data, in analogy to the training matrices shown in Figure 2b? Is this the input of each RF? Of RF' (as features might have been shrunken)? Or of RF_E?

Indeed, in Figure 2e the input matrix for the test instances is shown. The matrix is used as input for the individual classifiers T′1, T′2, T′3, which is the classifiers learned on the reduced feature space. We have improved the figure legend to better address this point.

==========

17. The fifth formula of sub-section "Ensemble random forest classifier" of the Methods might profit from a bit of additional explanation. Specifically, it took me a while to understand (if I'm right) that for T_E', the outputs of all individual RFs are concatenated row-wise, while "Binding(T_E')" denotes the concatenation of training labels.

We reformulated the text for better clarity.

==========

18. In the first paragraph of sub-section "Performance assessment" of the Methods section, I wondered what the index "i" refers to. Is this the same index i as before (i.e., an index for the training cell types)? If not, what exactly is "sample i"?

We have removed the index. It was not required at this point.

==========

19. In sub-section "Protein-protein-interaction score" of the Methods section, I would have appreciated a bit more motivation before describing the method itself.

We have added a sentence for motivation.

==========

20. In sub-section "Reducing the feature space to a small subset [...]" of the Results section, I would not fully agree with the authors that the difference in error between the full model and the model based on top 20 features is "marginal". I would even assume that a statistical test of the difference between the data behind the two boxplots in Fig. 3a would be significant.

We performed a statistical test on the difference of PR-AUC and ROC-AUC for both the OOB error as well as the test data (Figure 3a and Sup. Fig1, respectively). The differences are not significant for any of those instances.

==========

21. In sub-section "Reducing the feature space to a small subset [...]" of the Results section, I did not find the last two sentences (regarding importance of DNase-based features) to be supported by the data shown in the manuscript.

We appreciate that the reviewer pointed us to the lack of evidence required for this statement. We have added another Figure (Figure 7) to the main paper illustrating the feature importance of the RFs, which supports the statement made in the section mentioned above.

==========

22. In section "Data availability", the authors provide a link to the synapse page of the ENCODE-DREAM challenge. However, the data are accessible only after registration and signing a data usage policy.

We thank the reviewer for pointing this out to us. We have added it to the main manuscript.

==========

23. Typos & Grammar:
- first paragraph of "Data preprocessing and feature generation": "down sampled" should be "down-sampled"
- second paragraph of "Data preprocessing and feature generation": "the course of challenge" should be "the course of the challenge"
- third paragraph of "Data preprocessing and feature generation": "data is intersected" should be "data are intersected"
- 7th paragraph of "Discussion and conclusions": "Bound(positive)" should be "Bound (positive)"
- Reference 15: "transcritpion" should be "transcription"

We thank the reviewer for spotting the typos, we have corrected them.

Competing Interests: No competing interests were disclosed.