Published in final edited form as:
Nat Methods. 2019 September ; 16(9): 858–861. doi:10.1038/s41592-019-0511-y.

BindSpace decodes transcription factor binding signals by large-scale sequence embedding

Han Yuan[1,2], **Meghana Kshirsagar**[1], **Lee Zamparo**[1], **Yuheng Lu**[1], **Christina S. Leslie**[1,*]

[1]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065

[2]Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY 10065

Abstract

Decoding transcription factor (TF) binding signals in genomic DNA is a fundamental problem. Here we present a prediction model called BindSpace that learns to embed DNA sequences and TF class/family labels into the same space. By training on binding data for hundreds of TFs and embedding over 1M DNA sequences, BindSpace achieves state-of-the-art multiclass binding prediction performance, *in vitro* and *in vivo*, and can distinguish signals of closely related TFs.

Direct measurement of genome-wide transcription factor (TF) occupancy for all expressed factors in a cell type of interest is generally infeasible. Therefore, computational prediction of TF binding to cognate sites at relevant loci – e.g. regions of accessible chromatin or active histone marks – is critically important. Large-scale *in vitro* TF binding experiments[1,2,3,4,5] provide copious data for training binding models. However, each experiment is typically summarized as a position-specific weight matrix (PWM), yielding near-identical PWMs for closely related TFs. Previous supervised methods can accurately discriminate between bound and unbound sequences of individual TFs[6–9] but do not address the multiclass problem of distinguishing between TFs with similar binding signals.

Here we present BindSpace, a novel multiclass method to jointly learn binding models for hundreds of TFs assayed by HT-SELEX by embedding their bound/unbound DNA sequences and TF labels into a common high-dimensional space. We adapted an embedding framework called StarSpace[10] from natural language processing. StarSpace learns to embed words into a semantic space, inducing an embedding of documents, where words with similar meanings embed close to each other (Methods**)**. For multiclass problems, class labels

[*] Correspondence: cleslie@cbio.mskcc.org.

Author Contributions

H.Y. developed and implemented the model, performed all computational experiments, and co-wrote the manuscript. M.K. helped to perform quality control, initial benchmarking, visualization, and formulation of evaluations. L.Z. advised on quality control and experimented with alternative embedding ideas. Y.L. helped with visualization and advised on analyses. C.S.L. supervised the research, contributed to the model design, and co-wrote the manuscript. All authors discussed the results and contributed to the final manuscript.

Competing Financial Interests

We declare no competing financial interest.

(e.g. ‘sports’, ‘politics’) are embedded in the same space, and StarSpace optimizes both word and label embeddings by training on labeled documents, pushing documents towards their true labels and away from incorrect labels. In BindSpace, *k*-mers are analogous to words and DNA sequences to documents, while TFs and TF families serve as class labels (Fig. 1a). BindSpace learns *k*-mer and label embeddings so that probes embed close to the labels of TFs that bind them and away from other labels (>Fig. 1b, Methods). For *in vitro* or *in vivo* TF binding prediction, a test DNA sequence is embedded in BindSpace and assigned the closest TF label. In computational biology, semantic embedding of sequences was previously used for remote protein homology detection[11] and recently proposed for TF binding[12].

To train BindSpace, we combined HT-SELEX binding experiments for 461 mouse and human TFs from previous large-scale studies[5,8]. We applied rigorous quality control (Methods) to obtain 270 experiments for 243 transcription factors for our training set, enabling improved performance (Supplementary Note). The top 2000 enriched probes from each experiment were used as positive examples, yielding over 500K positive training sequences. We randomly sampled universal negatives from HT-SELEX probe libraries and non-accessible genomic regions to obtain ~500K negative training sequences (Methods). Each sequence is represented as a bag of 8-mers containing up to two consecutive wild cards, and each bag is associated with both a TF label (e.g. HOXA2) and TF family label (e.g. Homeodomain) or with a universal negative label. We used two thirds of the HT-SELEX data for training and one third for testing and performed 5-fold cross validation on the training data for hyperparameter tuning (Methods).

BindSpace learns a model where TFs with similar binding specificity are close to each other and their training examples in the embedding space, while more distinct TFs are farther away. Visualizing the embedding space in 2D using t-SNE (Methods) shows the clustering of probes with their TF labels and family labels (Fig. 2a, Supplementary Fig. 1a). The model also learns family substructure such as in the bZIP family (Supplementary Fig. 1b), where human and mouse orthologs embed close to each other (Jdp2 and JDP2, Hlf and HLF, Dbp and DBP), as do C/EBP TFs. To visualize the binding information encoded by TF labels, we embedded 10-mers in BindSpace and aligned the 10-mers nearest to each label (Methods), often yielding informative sequence logos (Supplementary Fig. 2). We also identified the 200 most closely embedded 8-mers for all TF labels (Supplementary Table 1).

We first evaluated our model on held-out HT-SELEX test data and compared against FIMO PWM scoring using published HT-SELEX PWMs curated using a semi-automated algorithm[5,13]; no probe data was held out for estimating these PWMs, presumably giving the PWM-based method an advantage. Nonetheless, BindSpace significantly outperforms PWM-based prediction as evaluated by F1 ($P < 2.94\text{e-}37$, one-sided Wilcoxon signed rank test, Fig. 2b). BindSpace also outperforms one-versus-all LASSO classifiers, trained using the same 8-mer features, and DeepBind, a deep learning method trained on HT-SELEX data[14] ($P < 4.80\text{e-}27$, $P < 1.26\text{e-}34$, respectively, Fig. 2b). To optimize performance for binary one-vs-all evaluation metrics like auPR and auROC that rely on accurate rankings, we developed BindSpace+, where we train regularized logistic classifiers on top of the BindSpace representation of probe data (Methods). BindSpace+ outperforms all competing methods on held-out HT-SELEX data for one-vs-all prediction as evaluated by auPR (Fig. 2b).

Additional comparisons provide insight into BindSpace's performance advantage (Supplementary Fig. 3): binary trained LASSO models outperform PWMs, suggesting an advantage from the *k*-mer representation, consistent with previous studies[7]; one-versus-all LASSO classifiers outperform binary LASSO classifiers, showing an advantage from multiclass training; and BindSpace outperforms one-versus-all LASSO, implying that multiclass training by embedding is superior to standard one-versus-all. Comparing confusion matrices within each TF family across the prediction methods highlights the multiclass performance advantage of BindSpace, although some paralogs, such as the C/EBP subfamily, remain difficult to distinguish (Fig. 2c, Supplementary Fig. 4, Supplementary Note**)**.

We next evaluated BindSpace on independent PBM data sets to confirm that we could distinguish TFs within the same family across *in vitro* platforms. We first downloaded raw PBM intensity data for the BAR15A data set[15], identified 14 homeodomains that overlap with BindSpace, and predicted the binding affinity of these TFs to the top intensity probes from these PBM experiments using BindSpace (Methods). For 8 of the 14 TFs, the PBM intensity correlated best with the BindSpace model for the same TF (compared to 7/14 TFs for LASSO and PWMs, 6/14 TFs for DeepBind), even for paralogs such as POU3F4, POU4F3 and POU6F2 where DeepBind has difficulty (Supplementary Fig. 5). Additional analyses on homeodomains in BAR15A, ETS factors from EMBO[16], POU and PRRX paralogs from Cell08[17], and recent gcPBM data from Shen *et al.*[9] for Ets1 and Elk1 confirm BindSpace's advantage for distinguishing between PBM probe data for TFs in the same family compared to other methods (Supplementary Note).

BindSpace also accurately identified TF binding sites and distinguished between sites of related TFs *in vivo*. For the standard task of distinguishing TF binding sites from flanking sequences or dinucleotide shuffled positive sequences, we used ENCODE ChIP-seq for 39 TFs represented in our model. BindSpace significantly outperformed PWM scoring, LASSO classifiers, and DeepBind for distinguishing the top 5,000 peaks from flanking regions or dinucleotide shuffled sequences by F1 score (Methods, Supplementary Fig. 6). BindSpace+ had good performance but came only in second place in ChIP-seq versus background sequence tasks evaluated with auROC: PWM was the winner for peaks-vs-flanks, DeepBind for peaks-vs-shuffled-sequences (Supplementary Fig. 6). For the more practical task of predicting TF binding versus non-binding at chromatin accessible regions in a given cell type, we processed publicly available ATAC-seq data and used ENCODE ChIP-seq data for 17 TFs in K562 and 11 TFs in GM12878 that had sufficient overlap with ATAC-seq peaks (Methods). BindSpace significantly outperformed all competing methods in discriminating bound vs unbound ATAC-seq peaks on K562 by F1 score, and significantly outperformed LASSO on GM12878, while its performance advantage over PWM and DeepBind was not significant on this smaller number of TFs (Supplementary Fig. 6). There was no significant difference between methods in ranking TF-occupied ATAC-seq peaks from unoccupied peaks as evaluated by auPR (Supplementary Fig. 7).

BindSpace is most noteworthy for its ability to distinguish between the binding of TFs in the same family *in vivo*. We found 115 ENCODE ChIP-seq experiments for TFs represented in BindSpace, including 20 pairs of high quality experiments for paralogous TFs assayed in the same cell type (Methods). For each pair of paralogs TF1 and TF2, we took only peaks unique to TF1 or TF2 and required each method to predict whether the peak was bound by TF1, TF2, or neither. BindSpace outperformed all other methods by F1 score, and BindSpace+ outperformed PWM and DeepBind while tying LASSO when evaluated by auROC (Fig. 3a). To assess multilabel prediction, we next combined all peaks for CEBPB and CEBPG HepG2 ChIP-seq data and annotated each peak as bound only by CEBPB, bound only by CEBPG, or bound by both. For each method, we compared CEBPB and CEBPG prediction scores across groups and indicated predicted labels by Venn diagram (Fig. 3b, Supplementary Fig. 8). Both BindSpace and PWMs could distinguish between the binding of CEBPB and CEBPG on peaks specific to one TF, but BindSpace has much less bias on commonly bound peaks. LASSO and DeepBind had weaker performance at this task, with DeepBind systematically overpredicting CEBPG binding sites. For the paralogs NR2F6 and NR2F1 in HepG2, BindSpace was the only method that could distinguish between NR2F6 and NR2F1 specific peaks (Supplementary Fig. 9).

We also performed multiclass classification on the top 10,000 ChIP-seq peaks across ENCODE experiments to test if we could distinguish binding of the ChIP-ed TF from TFs in the same family. Examining BindSpace predictions for bZIP ChIP-seq in HepG2, BindSpace successfully ranked the ChIP-ed TF above other family members in 5 out of 7 cases (Fig. 3c, Supplementary Fig. 10), compared to 3 out of 7 cases for PWM scoring and LASSO and 2 out of 7 cases for DeepBind. We also restricted to peaks uniquely bound by a single ChIP-ed bZIP TF to compare BindSpace (by F1) and BindSpace+ (by one-vs-all auPR) against other methods for a straightforward multiclass prediction assessment. These results were inconclusive, giving a statistical tie with competing methods due to the small number of TFs, but analysis of confusion matrices again confirmed that BindSpace and BindSpace+ had better calibrated scores (Supplementary Fig. 11). Therefore, while other methods can perform well in binary prediction tasks, their predictions are not well calibrated across TFs, leading to confusion between TFs in the same family. While we focus here on multiclass evaluation, note that BindSpace can detect multiple TF binding sites within an ATAC-seq peak for multi-label prediction (Supplementary Fig. 12, Supplementary Note).

BindSpace is a powerful and scalable machine learning approach to leverage massive *in vitro* binding data sets to decode TF binding signals in DNA sequences. By solving the underlying “many-class” classification problem through a joint embedding space for sequences and TF class labels, BindSpace is often able to distinguish between the binding preferences of closely related TFs, both *in vitro* and *in vivo*. A systematic evaluation across diverse binding prediction tasks (Supplementary Table 2) supports BindSpace’s state-of-the-art performance. Extensions to improve accuracy and completeness include incorporating longer *k*-mers into the vocabulary and other *in vitro* binding platforms in training; preliminary experiments point to the utility of these ideas (Supplementary Note). Finally, we expect that the BindSpace embedding will provide a useful feature representation on which to build more sophisticated models, such as deep learning models, for regulatory genomics.

Online Methods

HT-SELEX quality control

We used public HT-SELEX data sets to train BindSpace and one-vs-all LASSO TF binding specificity models. We combined the HT-SELEX data sequenced in 2013 (ENA accession: ERP001824) and the 2017 re-sequenced libraries (ENA accession: ERP016411), which together include 547 experiments for 461 human or mouse TFs[5,8].

To perform quality control, we computed an enrichment score for every 8mer in probes selected at each cycle of an experiment relative to the initial library (cycle 0), as in a previous study[8]. We first estimated the frequency of every 8-mer in cycle 0 using a fifth-order Markov model[18]. Then, for cycle i, we computed the enrichment score for each 8-mer as the i-th root of (frequency in cycle i)/(estimated frequency in cycle 0).

For each experiment, we then performed quality control filtering based on the following procedure:

1. For experiments with 14bp or 20bp probes, cycle 4 was used to derive enriched probes.

2. For experiments with 30bp and 40bp probes, one of cycle 4, 5 or 6 was used to derive enriched probes, because more cycles of selection are required to enrich for longer probes.

3. Experiments were excluded if the Spearman correlation of 8-mer enrichment scores between cycle 3 and cycle 4 was below 0.8. (The correlation distribution between cycle 3 and cycle 4 for all experiments and examples of poorer correlation are shown in Supplementary Note).

4. Experiments were excluded if the selected cycle did not have good probe enrichment or diversity (fewer than 500 probes with frequency > 10).

5. For the remaining experiments, we removed low complexity probes with a DUST score < 2, where DUST is the BLAST low-complexity masking algorithm[19]. This is based on our observation that some HT-SELEX experiments have non-specific enrichment for low complexity sequences.

6. For remaining experiments with 14bp or 20bp probes, we selected the top 2000 enriched probes (frequency > 10) based on enrichment score. We chose 2000 enriched probes per TF for training based on the observation that further increasing the size of the training data does not improve overall multiclass classification performance while reducing performance for TFs with fewer enriched probes.

7. For remaining experiments with 30bp or 40bp probes, we counted the frequency of unique 20bp sequences, and selected the top 2000 enriched 20bp probes based on enrichment score.

8. Finally, experiments where the most frequent 8-mer in the top 2000 probes occurred less than 100 times were removed. This final filter is applied to remove

any experiments where the top probes do not enrich for any consensus binding sites (Supplementary Note).

After quality control, we ended up with 270 high quality experiments covering 243 TFs. 143 of these experiments overlapped with those selected by quality control in a previous study[8]. To show that our filtering steps improved overall data quality, we measured consistency between replicate experiments or experiments of orthologous TFs. Consistency between experiment A and B was defined as the percentage overlap between the top 100 enriched 8-mers in the respective experiments. We found that after removing low quality experiments and probes, we significantly improved overall data quality ($P < 0.009$, rank sum test, Supplementary Note).

Data preprocessing

The top 2000 enriched probes from each of the 270 experiments that passed quality control were used as examples to train our model. All non-unique probes were removed, and each of the remaining (unique) probes was associated with a TF label and a TF family label. Because 20bp probes have very high diversity, we only had a very small number of non-unique probes (Supplementary Table 3). TF family information was obtained from http://cisbp.ccbr.utoronto.ca. Overall we had a total of 505,194 TF-associated probes.

We then randomly selected 252,597 sequences from probes that only present in initial cycles (random negative sequences), and 252,597 20bp sequences from inaccessible region of the genome in K562 cell lines (genomic negative sequences). We included genomic negatives during training in order to learn to distinguish between TF binding sites and genomic background. After filtering out duplicated sequences, we obtained a total of 505,086 unique negative probes, each of which was associated with a universal negative label.

Training BindSpace

Each HT-SELEX probe input sequence s_i is represented by a bag of 8-mers with up to 2 consecutive wildcards (where the wildcard symbol 'N' matches any nucleotide). A particular 8-mer is considered a token of s_i if it occurs in either s_i or reverse complement of s_i. The total vocabulary of 8-mers with up to 2 consecutive wildcards has a size of 112,800. For a pair of reverse complement 8-mers, only one of the two is included in the vocabulary, similar to the representation used in the SeqGL algorithm[7].

BindSpace is an instantiation of StarSpace, a machine learning algorithm that can be used for a wide range of supervised learning problems. Training examples to StarSpace are structured as left hand side (LHS)-right hand side (RHS) pairs. In BindSpace, the LHS of the i^{th} input is a DNA probe represented by its constituent k-mers $\left(w_{i,1}, \ldots, w_{i,m_i}\right)$, and the RHS consists of the labels associated with this probe $\left(l_{i,1}, \ldots, l_{i,n_i}\right)$ Most probes have two labels, a TF label and family label. For training examples that come from universal negatives or from TFs with unknown family, no family label is assigned.

Overview of StarSpace algorithm

We used the StarSpace default training mode (mode 0) to learn an embedding for a total of 113,074 entities – namely, 112,800 *k*-mers, 243 TF labels, 1 universal negative label, and 30 TF family labels – into a vector space $\mathbb{R}^d$, where d is the dimension of the embedding space. In BindSpace, we use $d = 300$

To explain the embedding of the i^{th} training example and the associated loss used in training, let us denote the embedded vectors of *k*-mers $\left(W_{i,1}, \ldots, W_{i,m_i}\right)$ as $\left(w_{i,1}, \ldots, w_{i,m_i}\right)$ and the embedded vectors of labels $\left(L_{i,1}, \ldots, L_{i,n_i}\right)$ as $\left(l_{i,1}, \ldots, l_{i,n_i}\right)$, where all the vectors $w_{i,j}, j = 1, \ldots, m_i$ and $l_{i,k}, k = 1, \ldots, n_i$, are in $\mathbb{R}^d$.

The embedding of the LHS of the i^{th} example is induced by the embedding of all constituent *k*-mers as follows (this can also be considered to be the embedding for the corresponding training probe):

$$lhs_i = \frac{1}{m_i^p} \sum_{j=1}^{m_i} w_{i,j}$$

Here $p = 0.5$ by default.

Similarly, the embedding of the RHS of the example is induced by the embedding of all its associated labels:

$$rhsP_i = \frac{1}{n_i^p} \sum_{j=1}^{n_i} l_{i,j}$$

To compute the loss associated with this example, we randomly sample K examples with labels different from example i and compute the RHS associated with each:

$$rhsN_{i,k} = \frac{1}{n_k^p} \sum_{j=1}^{n_k} l_{k,j}$$

Here l_{kj} is the embedding of the j^{th} label of negative sample k.

To compute the loss, StarSpace allows for a choice of hinge or softmax and a choice of dot and cosine similarity. We found that hinge loss and dot similarity work best for our data. Thus, the loss function for a given positive example with one random negative would be:

$$Err_{ik} = \max\left(0,\ \text{margin} - lhs_i \cdot rhsP_i + lhs_i \cdot rhsN_{i,k}\right)$$

We used a default value of 0.05 as the margin. Therefore, a loss is incurred unless the similarity of the LHS embedding for the ith example lhs_i, to the embedding of its RHS $rhsP_i$, exceeds the similarity of lhs_i to the embedding of the ‘wrong’ RHS by a margin. The total loss associated with example i using K negative samples is:

$$Err_i = \frac{1}{K}\sum_{k=1}^{K} \max\left(0,\ \text{margin} - lhs_i \cdot rhsP_i + lhs_i \cdot rhsN_{i,k}\right)$$

Subsequently, k-mer embeddings $\left(w_{i,1}, \ldots, w_{i,m_i}\right)$, positive label embeddings $\left(l_{i,1}, \ldots, l_{i,n_i}\right)$, and negative label embeddings $\left(l_{k,1}, \ldots, l_{k,n_k}\right)$ are all updated based on the gradient of this loss using Adagrad algorithm.

We split the whole data set into 2/3 for training and 1/3 for testing with stratified sampling, i.e., each TF class was split into 2/3 for training and 1/3 for testing. We performed 5-fold cross-validation on training data for hyperparameter tuning, and cross-validation performance was measured by average F1 score for all TFs. We then evaluated the performance on test data and compared against other methods. We found that the cross-validation performance no longer improved after 50 epochs, and therefore we stop training at that point.

Finally, we learned a comprehensive BindSpace model with the same hyperparameter settings using all HT-SELEX data. The visualizations and *in vitro* and *in vivo* validation were performed using this model.

Hyperparameter tuning

The hyperparameters we considered in the StarSpace algorithm included: loss, similarity, learning rate, embedding dimension, maximum number of negatives in a batch update, and dropout rate. Preliminary analysis on a small number of examples (15 classes) showed that hinge loss and dot similarity gave the best performance on DNA sequence data.

We sampled 20 sets of hyperparameters from: learning rate = (0.01, 0.05, 0.1, 0.2), dimension = (50, 100, 150, 200), maxNegSamples = (3, 5, 10, 15), and dropout rate = (0, 0.001, 0.01, 0.1). We compared the cross-validation performance of each of these 20 sets by F1 score for each TF class.

Taking the sampled hyperparameter set with best performance (mean F1 score), we then varied each hyperparameter individually while keeping the rest constant. We measured the performance change with respect to the change in each hyperparameter, again by cross-validation performance on training data measured by F1 score for each TF. Based on this analysis, the final hyperparameters that gave the best cross-validation performance were: learning rate = 0.2, dimension = 300, maxNegSamples = 15, dropout rate = 0. We also found that the performance reaches convergence after 50 epochs.

Making predictions with BindSpace

Binary thresholding.—BindSpace predicts similarity scores of a new test example to all class labels. Multiclass prediction can be made based on which label has the highest similarity to the test example. However, when we need to make multi-label predictions (allow a sequence to be positive for more than one TF), we need to determine a threshold for each label. Therefore, we computed the similarity of each label to all training examples not belonging to this class to generate an empirical null distribution of similarity score. For a new test example, if the similarity score significantly achieves $P < 0.05$ relative to this null distribution, we consider the test example to be positive; otherwise, we consider it to be negative.

In vivo evaluation.—When evaluating on ChIP-seq or ATAC-seq data, instead of having 20bp probes, we have 150bp or 300bp genomic sequences, respectively. In order to predict on these large genomic regions, we bin into 20bp windows, embed each window into BindSpace, and determine similarity to TF class labels. Then the similarity score of a TF to a given peak is the maximum similarity score over all 20bp windows. However, when predictions on 20bp bins are considered, BindSpace and BindSpace+ retain the resolution to identify multiple binding sites within an ATAC-seq peak (Supplementary Fig. 12).

Training BindSpace+

Since BindSpace is trained as a multi-class classifier, with prediction based on proximity of embedded sequences to >200 model vectors in a high dimensional latent space, it is not optimized for binary classification tasks as evaluated by auPR or auROC, which rely on accurate one-vs-all or positive-vs-negative *rankings* rather than class assignments. To adapt BindSpace to this binary prediction setting, we trained the BindSpace+ model as follows. We used the previous BindSpace representation of *k*-mers and induced representation of 20bp DNA sequences, but we did not use the model vectors for the labels. Instead, we embedded all training examples into BindSpace's 300 dimensional latent space using the previously learned representation and then trained a one-versus-all logistic classifier with ridge regularization for each TF using the BindSpace representation. For prediction on held-out HT-SELEX probes, we trained BindSpace+ logistic classifiers in one-vs-the-rest fashion. For *in vivo* binary prediction tasks, we trained BindSpace+ models for each TF on positive probes vs randomly sampled negative probes.

PWM-based predictions

The PWM models we use are from 'Jolma2013.meme', downloaded from http://meme-suite.org/db/motifs. They were generated by Jolma *et al.* using a semi-automatic algorithm from the same HT-SELEX data set[5]. We used FIMO with default settings to predict PWM hits[13]. For binary prediction problems, rather than using the default $P < 1e\text{-}4$ threshold, we allowed the full range of P values in order to rank all examples. For multiclass prediction problems, if none of the PWMs satisfied $P < 1e\text{-}4$, we predicted the sequence to be a universal negative; otherwise we predicted the TF label based on the most significant PWM.

LASSO one-versus-all multiclass predictions

We used the R package *glmnet* to train a LASSO logistic regression classifier for each TF in a one-versus-all setting in order to make multiclass predictions[20]. During training, for each TF, we considered the 2000 enriched probes as positives and sampled 20,000 sequences from the rest of the training data as negatives. We represented training sequences using 8-mers features with up to 2 consecutive wildcards. We performed feature selection by computing a binomial Z-score for every 8-mer feature based on its enrichment in positive versus negative training examples and chose the top scoring 10,000 8-mers. For testing, we predicted the class of a given test example by assigning it to the TF model that gave highest posterior probability, or assigning it to be a universal negative if all of the models predicted a posterior < 0.5. We can also make multi-label prediction by allowing a given sequence to be positive for multiple TFs if the posterior > 0.5 for multiple classifiers.

Binary LASSO predictions

We used the R package *glmnet* to train a LASSO logistic regression classifier for each TF in a binary setting. During training, for each TF, we considered the 2000 enriched probes as positives, and randomly sampled 10,000 sequences from cycle 0 of the same experiment as negatives. Training sequences were represented using 8-mer features with up to 2 consecutive wildcards. Feature selection was performed the same way as previously describe for LASSO one-versus-all.

DeepBind predictions

The DeepBind program and models were downloaded from http://tools.genes.toronto.edu/deepbind/. We only used the DeepBind models trained on HT-SELEX data. For making binary decisions, we set the threshold at 0.9 after logistic transform of raw DeepBind scores.

Visualization

We visualized the top 50 probes of each HT-SELEX experiment, TF labels, and TF family labels in a 2D space using an efficient Barnes-Hut implementation of t-SNE[21,22]. To do so, we first computed a similarity matrix of these points (13,500 probes and 274 labels) by dot product. Then we converted this similarity matrix to a dissimilarity matrix and used the R package *Rtsne* with default settings to project these points to 2D, where the dissimilarity is approximated by Euclidean distance[23].

Sequence logos for BindSpace TFs

We embed all 10-mer sequences into BindSpace. For each TF, we sort all 10-mers by their similarity to the TF label in BindSpace (by dot similarity) and used top 200 10-mers (where for reverse complement sequence pairs, only one of the two sequences is included) to generate a PWM. The directions (orientations) of these 200 10-mers are then corrected by “mafft --adjustdirectionaccurately”, and a sequence logo is generated by the “seqlogo” program.

BindSpace correlation with PBM data

We downloaded the original PBM intensity data of ‘BAR15A’, ‘EMBO10’, and ‘Cell08’ from Uniprobe (http://the_brain.bwh.harvard.edu/uniprobe/downloads.php).

BAR15A contains PBM data for 41 TFs, 15 of which overlap with TFs in BindSpace (1 C2H2 zinc finger, 14 homeodomains). For each TF, we took the top 5000 PBM probes with highest intensity and predicted the binding affinity of 14 homeodomain TFs to these 36bp probes using each of BindSpace, PWM, LASSO and DeepBind. Finally we computed Spearman correlations between true probe intensities with BindSpace predicted intensities for the 14 TFs.

For TFs in BAR15A POU family, TFs in BAR15A PAX family, TFs in EMBO10 ETS family, TFs in Cell08 POU family and Cell08 PRRX paralogous TFs, instead of considering only the top 5000 PBM probes with highest affinity, we plotted the Spearman correlation of predicted affinity versus true PBM intensity as a function of number of top probes and compared between all methods.

BindSpace analysis of gcPBM data

We downloaded gcPBM array for Ets1 and Elk1 from GEO (GSE97793). Negative control probes, ELK1 preferential and ETS1 preferential probes were as annotated by Shen *et al.*[9] We looked at all Elk1/Ets1 preferential probes and compared between BindSpace, PWM, LASSO and DeepBind for their ability to distinguish between Elk1/Ets1 preferential probes (Supplementary Note).

We performed an additional multilabel analysis considering the true label of each probe to be one of: Elk1 preferential, Ets1 preferential or common (non-specific). Using each of BindSpace, PWM, LASSO or DeepBind methods, we predicted each probe to be ELK1 specific (ELK1 score > ETS1 score and ELK1 score > cutoff), ETS1 specific (ETS1 score > ELK1 score and ETS1 score > cutoff) or low affinity (ELK1 score < cutoff and ETS1 score < cutoff). See previous description for empirical thresholds for predicting low affinity probes for each method.

ENCODE ChIP-seq

Conservative and optimal IDR thresholded ‘narrowpeak’ files of ENCODE ChIP-seq data were downloaded from the ENCODE portal (https://www.encodeproject.org/) for five major cell lines: GM12878, K562, A549, H1-hESC and HepG2, covering a total of 329 different TFs.

ATAC-seq data processing

K562 ATAC-seq raw reads were obtained from GSE76224[24]. We combined records SRR3822969 and SRR3822972 to create replicate 1. We combined the other two records to create replicate 2. GM12878 ATAC-seq raw reads were obtained from GSE47753[25]. We used only the samples generated from 50k cells. Because replicate 1 is much more deeply sequenced then rest of the replicates, we combined replicates 2, 3, 4 to obtain replicate 2.

We followed the ENCODE ATAC-seq processing pipeline (https://www.encodeproject.org/atac-seq/) for processing K562 and GM12878 ATAC-seq data. For each of the ATAC-seq samples, we trimmed the raw fastq files of adapters with Trimmomatic with default settings. We then aligned to the hg19 genome using Bowtie2 with default settings. After that we removed duplicate reads and adjusted for Tn5 shifts. Peak calling was performed for each replicate using macs2 with --nomodel --shift −37 --extsize 73. Finally, IDR was performed with the idr package and reproducible peaks were called with an IDR cutoff of 0.05. There were a total of 17,271 reproducible peaks in K562 and a total of 76,218 reproducible peaks in GM12878. For both K562 and GM12878, we only evaluated on the top 10,000 peaks with highest accessibility.

ChIP-seq peaks versus flanks evaluation

For this task, we used all available ENCODE TF ChIP-seq data in cell lines, which included a total of 329 different TFs, including 39 represented in BindSpace. For each TF covered by BindSpace, we selected a single experiment that gave the most reproducible peaks, selected the top 5000 peaks with highest IDR score, and took the 150bp region around each peak summit as positive examples. Negative examples (flanks) were taken 300 bp upstream of positive examples. Overlapping regions and sequences containing Ns are removed. For each TF data set of ChIP-seq peaks and flanks, we made binary predictions based on their similarity to the corresponding TF label in BindSpace and measured performance by F1 score. We then compared with predictions made by FIMO scoring using the ‘Jolma2013’ PWMs, LASSO and DeepBind.

ATAC-seq binding versus non-binding evaluation

All ATAC-seq peaks were trimmed to 300bp centered at the peak summit. All TF ChIP-seq peaks were resized to 150bp centered at the peak summit. If an ATAC-seq peak overlaps with a resized TF ChIP-seq peak, we consider that ATAC-seq peak to be bound by the corresponding TF; otherwise, it is unbound by the TF. For this evaluation, we removed any TFs that were bound to < 10% of the ATAC-seq peaks.

Paralogous TF ChIP-seq evaluation

1. There are 20 pairs of ENCODE ChIP-seq experiments for paralogous TFs in the same cell type with more than 5k peaks in each experiment using a conservative IDR threshold. We evaluated BindSpace, PWM, LASSO, and DeepBind for their ability to distinguish between these paralogous TF ChIP-seq peaks. For each pair of ChIP-seq experiments for paralogous TFs, we took top 100 peaks unique to TF1 or TF2 from each ChIP-seq experiment. Each method was required to predict TF1 (when TF1 score > TF2 score and TF1 score > cutoff), TF2 (when TF2 score > TF1 score and TF2 score > cutoff), or neither (when TF1 score < cutoff and TF2 score < cutoff). With various performance metrics (precision, recall, F1), we reported the average performance of TF1 and TF2. We also reported the average of TF1 auPR on ranking TF1 peaks above TF2 peaks and TF2 auPR for ranking TF2 peaks above TF1 peaks.

2. For two example paralogous pairs (HepG2 CEBPB versus CEBPG, HepG2 NR2F6 versus NR2F1), we evaluated the performance by showing the predicted score distributions for BindSpace, PWM, LASSO and DeepBind in a similar way that we evaluated gcPBM data. We split the union of ChIP-seq peaks of paralogous TFs into three groups: TF1 specific, TF2 specific, or common. Using each of BindSpace, PWM, LASSO or DeepBind methods, we predicted each peak to be TF1 specific (TF1 score > TF2 score and TF1 score > cutoff), TF2 specific (TF2 score > TF1 score and TF2 score > cutoff) or low affinity (TF1 score < cutoff and TF2 score < cutoff) and visualized the distribution of scores and class assignments in each of the three groups.

ChIP-seq multiclass evaluation

We resized all ChIP-seq to be 150bp around summit. For each experiment, we took the top 10,000 peaks with most significant IDR and performed multiclass classification of each peak using BindSpace, FIMO PWM scoring, LASSO or DeepBind using TFs in the same family. We evaluated performance by determining the percentage of the top 10,000 peaks that were correctly classified.

Inclusion of 10-mer features

In order to understand whether the choice of 8-mers as our feature set limited our capability to capture long DNA binding signals like CTCF, we trained a new BindSpace model by adding some 10-mer features. Specifically, in addition to 8-mers with wildcards, we picked the top 200 enriched 10-mers from each HT-SELEX experiment, giving a total of 51,380 unique exact match 10-mers. We added these 10-mer features into our vocabulary, which expanded our vocabulary by about 50%. We trained a BindSpace model with the same hyperparamters as previously determined and compared the performance with regular BindSpace on held-out HT-SELEX probes (Supplementary Note).

Comparison with *k*-mer modules

We downloaded *k*-mer modules from Mariani *et al.* 2017[26]. We assigned each module to a TF family if a representative TF of that module has a (human/mouse) family assignment in CIS-BP. In this way, 92 out of the 108 modules could be identified with a total of 22 families represented in BindSpace. We embedded all 92 modules to BindSpace and measured their similarity to all TF family labels (Supplementary Note).

Statistics

We used Wilcoxon signed rank test for pairwise performance comparison between methods.

Data availability

We used only public datasets in this study. Each dataset can be accessed in the original publication referenced.

Code availability

Open source code for BindSpace as well as the trained BindSpace model are freely available for download (https://bitbucket.org/hy395/selex_embed for source code for training BindSpace; https://bitbucket.org/hy395/bindspace for R package to make predictions with the trained model).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Jason Weston for suggesting the StarSpace algorithm and providing access to his code. We also thank William Stafford Noble, Jeffrey Bilmes, and Jacob Schreiber for helpful comments on the project. This work was supported by NIH/NHGRI U01 award HG009395. Support for H.Y. was also provided by the Tri-Institutional Training Program in Computational Biology and Medicine.

References

1. Berger MF et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol 24, 1429–1435 (2006). [PubMed: 16998473]
2. Warren CL et al. Defining the sequence-recognition profile of DNA-binding molecules. Proc. Natl. Acad. Sci 103, 867–872 (2006). [PubMed: 16418267]
3. Gordân R et al. Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. Cell Rep. 3, (2013).
4. Maerkl SJ & Quake SR A systems approach to measuring the binding energy landscapes of transcription factors. Science (80-.). 315, 233–237 (2007). [PubMed: 17218526]
5. Jolma A et al. DNA-binding specificities of human transcription factors. Cell 152, 327–339 (2013). [PubMed: 23332764]
6. Ghandi M, Lee D, Mohammad-Noori M & Beer MA Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLoS Comput. Biol 10, (2014).
7. Setty M & Leslie CS SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. PLOS Comput. Biol 11, e1004271 (2015). [PubMed: 26016777]
8. Yang L et al. Transcription factor family‐specific DNA shape readout revealed by quantitative specificity models. Mol. Syst. Biol 13, 1–14 (2017).
9. Shen N et al. Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential In Vivo Binding. Cell Syst. 6, 470–483.e8 (2018). [PubMed: 29605182]
10. Wu L et al. StarSpace: Embed All The Things! in *AAAI* (2018).
11. Melvin I, Weston J, Noble WS & Leslie C Detecting remote evolutionary relationships among proteins by large-scale semantic embedding. PLoS Comput. Biol 7, (2011).
12. Dai H et al. Sequence2Vec: A novel embedding approach for modeling transcription factor binding affinity landscape. 1–9 (2017).
13. Grant CE, Bailey TL & Noble WS FIMO: Scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018 (2011). [PubMed: 21330290]
14. Alipanahi B, Delong A, Weirauch MT & Frey BJ Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 33, 831–838 (2015). [PubMed: 26213851]
15. Barrera LA et al. Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science (80-.). 351, 1450–1454 (2016). [PubMed: 27013732]
16. Wei GH et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. EMBO J. 29, 2147–2160 (2010). [PubMed: 20517297]

17. Berger MF et al. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. Cell 133, 1266–1276 (2008). [PubMed: 18585359]

18. Slattery M et al. Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. Cell 147, 1270–1282 (2011). [PubMed: 22153072]

19. Morgulis A, Gertz EM, Schäffer AA & Agarwala R A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. J. Comput. Biol 13, 1028–1040 (2006). [PubMed: 16796549]

20. Simon N, Friedman J, Hastie T & Tibshirani R Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J. Stat. Softw 39, (2011).

21. Van Der Maaten L & Hinton G Visualizing Data using t-SNE. J. Mach. Learn. Res 9, 2579–2605 (2008).

22. van der Maaten L Accelerating t-SNE using Tree-Based Algorithms. J. Mach. Learn. Res 15, 3221–3245 (2014).

23. Krijthe JH {Rtsne}: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation. (2015).

24. Litzenburger UM et al. Single-cell epigenomic variability reveals functional cancer heterogeneity. Genome Biol. 18, (2017).

25. Buenrostro JD, Giresi PG, Zaba LC, Chang HY & Greenleaf WJ Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–8 (2013). [PubMed: 24097267]

26. Mariani L, Weinand K, Vedenko A, Barrera LA & Bulyk ML Identification of Human Lineage-Specific Transcriptional Coregulators Enabled by a Glossary of Binding Modules and Tunable Genomic Backgrounds. Cell Syst. 5, 187–201.e7 (2017). [PubMed: 28957653]

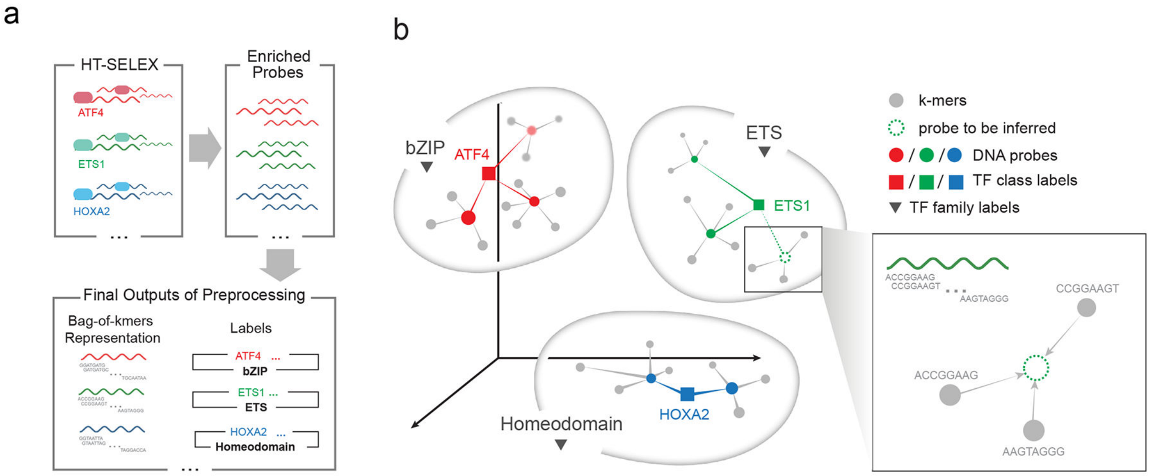

Figure 1. A schematic overview of BindSpace.

BindSpace is an embedding approach that jointly learns binding models for hundreds of TFs.

(a) Enriched DNA probes for TF HT-SELEX experiments are represented as bags of *k*-mers, namely 8-mers with up to two consecutive wildcards (a wildcard represents any of the 4 nucleotides).

(b) BindSpace learns an embedding of *k*-mers (grey dots), DNA probes (colored dots), TF labels (colored squares), and TF family labels (black triangles) in the same high-dimensional space. The embedding of a probe sequence is determined by the embedding of its constituent *k*-mers. Unbound HT-SELEX probes and inaccessible genomic sequences serve as universal negative examples.

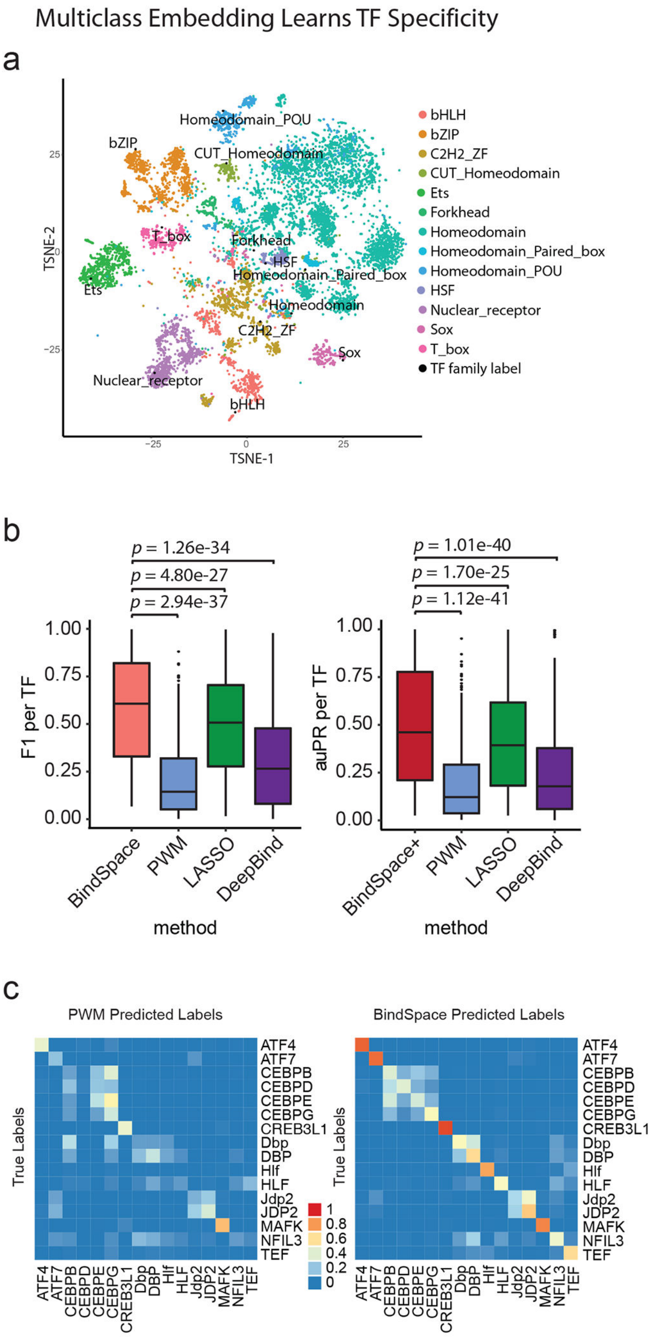

Figure 2. BindSpace accurately predicts TF binding and distinguishes between TF family members *in vitro*.

(a) t-SNE visualization of the embedding space. Colored points represent embedding of enriched probes from HT-SELEX experiments (n=50 probes from each experiment). Black points represent embedding of TF family labels for 13 major families (families with more than three members). A full t-SNE visualization of all TF labels as well as TF family labels is shown in Supplementary Fig. 1.

(b) Left: multiclass prediction performance of BindSpace, PWM, one-versus-all LASSO and DeepBind methods on held-out HT-SELEX probes evaluated by F1 score (n=243 HT-

SELEX TFs, one-sided Wilcoxon signed rank test). Right: prediction performance of BindSpace+, PWM, one-versus-all LASSO and DeepBind on held-out HT-SELEX probes evaluated by one-versus-all auPR per TF (n=243 HT-SELEX TFs, one-sided Wilcoxon signed rank test). Boxplots show median, upper and lower quartiles, and highest and lowest values excluding outliers.

(c) Multiclass prediction performance of PWM scoring (left) and BindSpace (right) as shown by confusion matrices normalized by class support for the bZIP family. Rows are true labels, and columns are predicted labels. For comparison between BindSpace, PWM, LASSO and DeepBind on 13 major TF families, also see Supplementary Fig. 4 and Supplementary Note SN1–2.

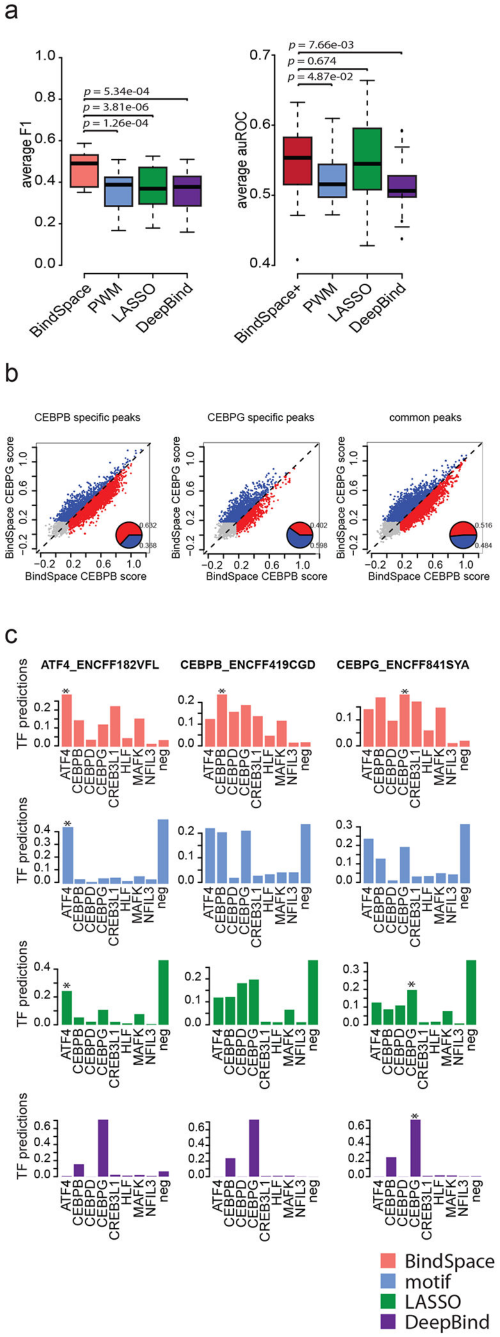

Figure 3. BindSpace predicts binding of TFs and distinguishes between paralogous TF binding sites *in vivo*.

(a) Left: performance of BindSpace, PWM, LASSO and DeepBind on distinguishing ChIP-seq peaks of ENCODE paralogous TF pairs as measured by F1 score averaged over TF1 and TF2 of each pair (n=20, one-sided Wilcoxon signed rank test). Right: performance of BindSpace+, PWM, LASSO, and DeepBind on distinguishing ChIP-seq peaks of ENCODE paralogous TF pairs as evaluated by auROC averaged over TF1 and TF2 of each pair (n=20, one-sided Wilcoxon signed rank test). Boxplots show median, upper and lower quartiles, and highest and lowest values excluding outliers.

(b) BindSpace performance on distinguishing CEBPB specific peaks versus CEBPG specific peaks in HepG2. From left to right, we show the predicted affinities on CEBPB specific peaks (predicted CEBPB score significantly higher than predicted CEBPG score, *P* = 2.1e-56, n = 3045), CEBPG specific peaks (predicted CEBPG score significantly higher than predicted CEBPB score, *P* = 2.14e-40, n = 2583), and common peaks in the scatter plots (no significant different, *P* = 0.317 when alternative is CEBPB>CEBPG, *P* = 0.683 when alternative is CEBPB<CEBPG, n = 5248), one-sided Wilcoxon signed-rank test. Each probe is assigned to be CEBPB specific (red), CEBPG specific (blue) or low affinity (gray). Pie charts show the proportion of peaks predicted to be CEBPB specific versus CEBPG specific in each group. (For performance of PWM, LASSO and DeepBind, see Supplementary Fig. 8)

(c) Multiclass classification performance of BindSpace, PWM, LASSO and DeepBind for three bZIP TFs in the HepG2 cell line. For each plot, we performed multiclass classification using BindSpace on the top 10,000 peaks for bZIP TF ChIP-seq and show the proportion of predicted labels for each model. Experiments where the ChIP-ed TF is predicted more frequent than other members in the same family are indicated by *. See Supplementary Fig. 10 for a comparison of all bZIP family members in HepG2 cell line.