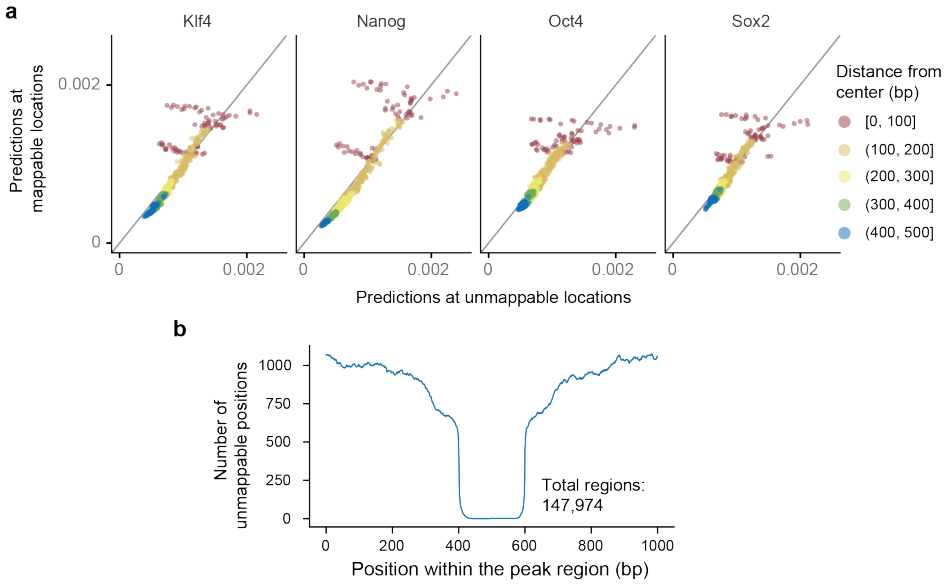
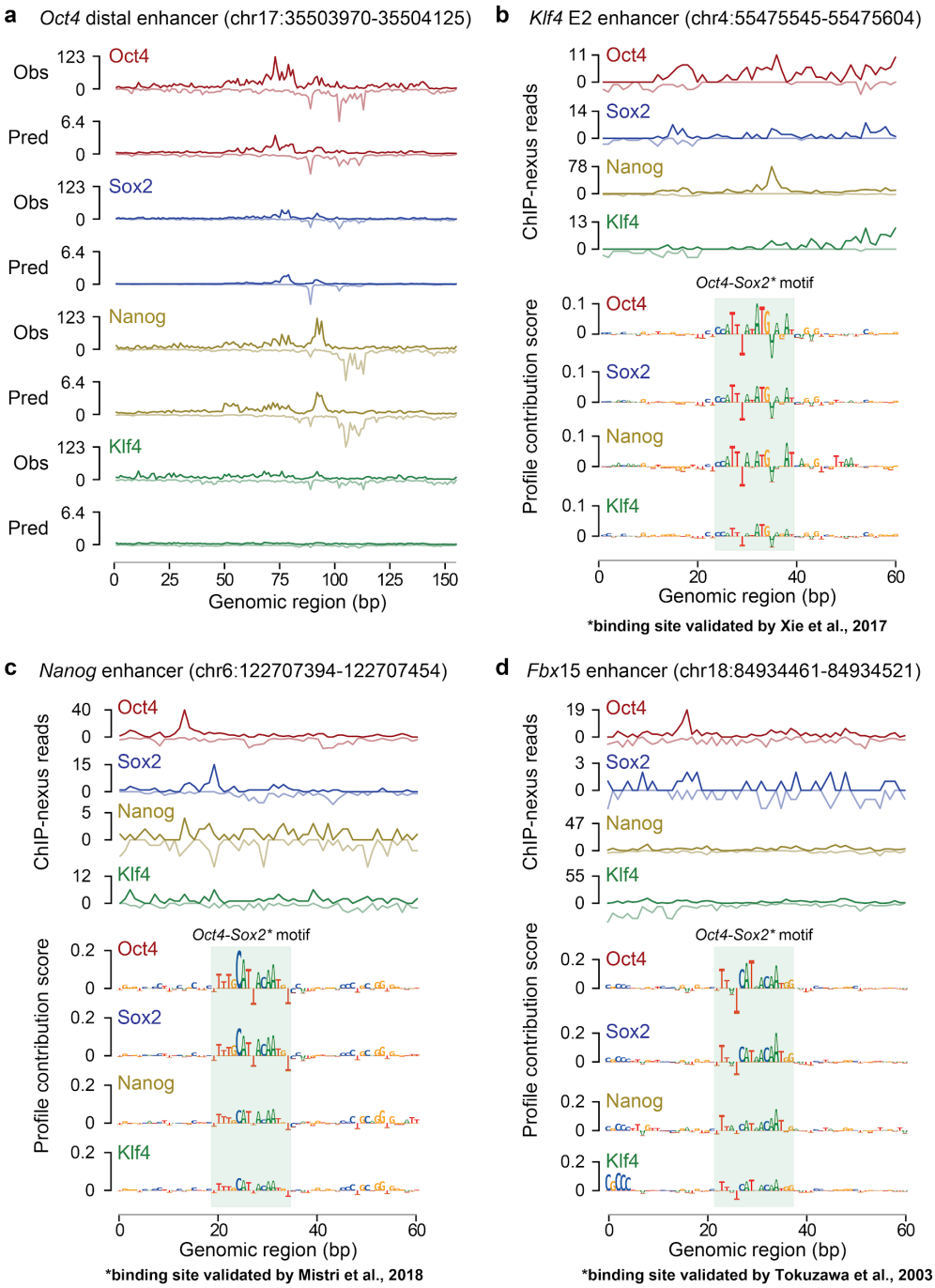
Supplementary Information

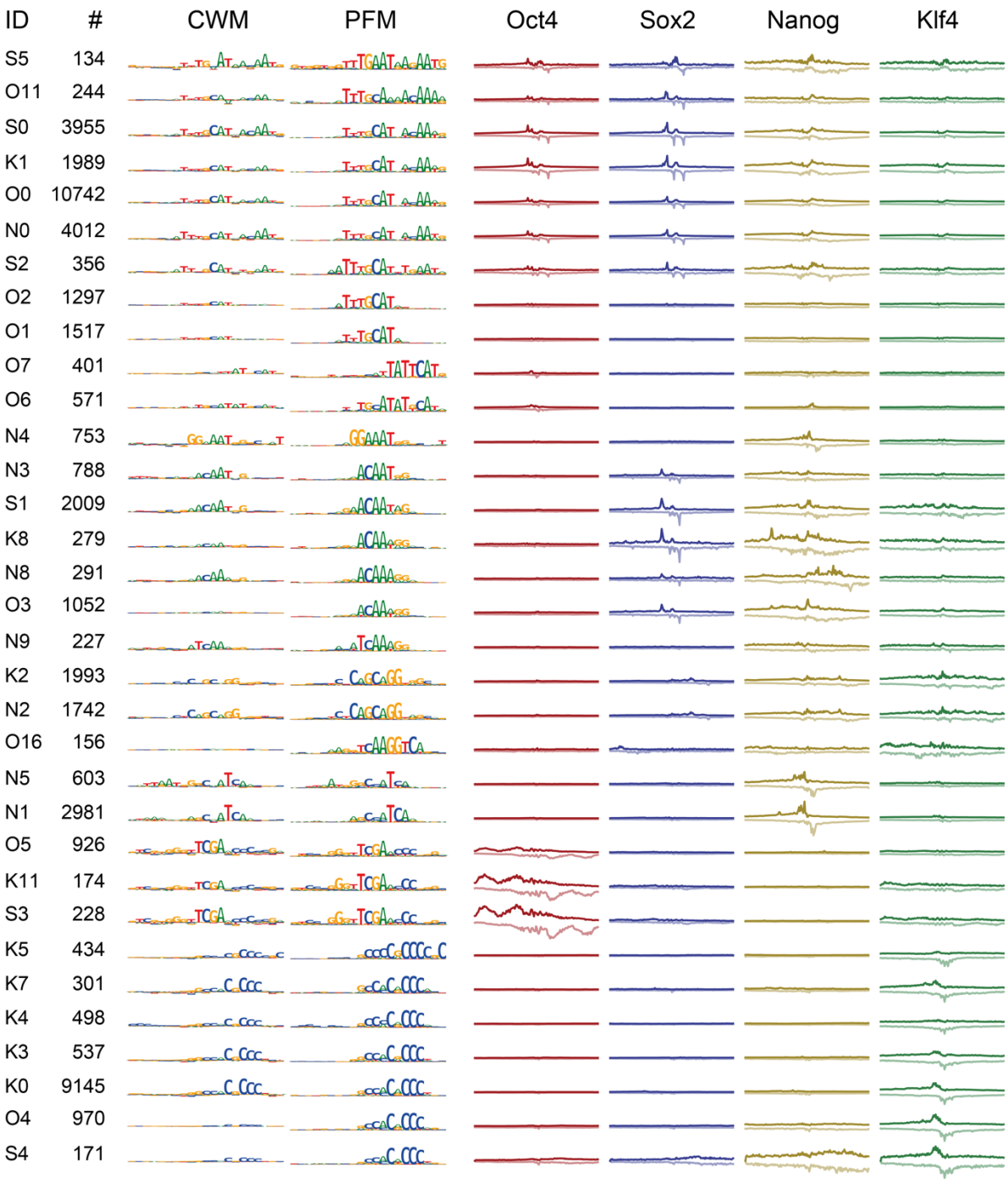
Supplementary Figures



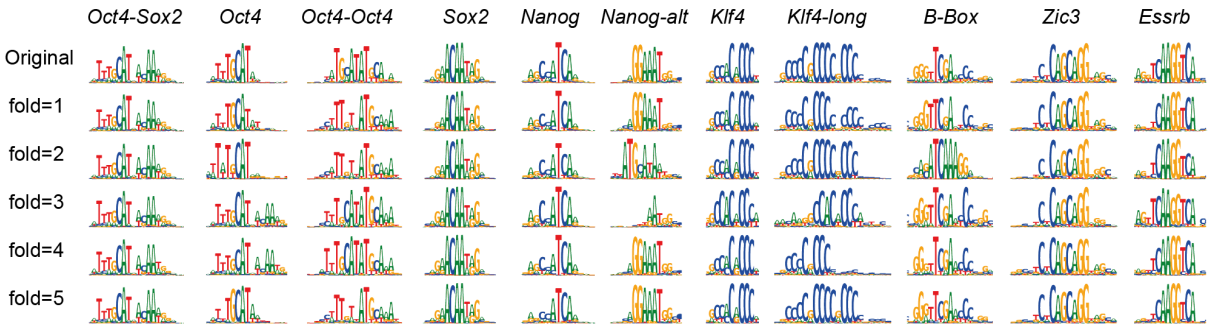
Supplementary Fig. 1: Model predictions exhibit no systematic overfitting to unmappable positions. a) Median model predictions for the positive strand at unmappable (x-axis) and randomly chosen mappable positions (y-axis) stratified by distance from the peak summit (denoted by points and color). Each of the thousand points corresponds to a specific relative position within the 1 kb peak and the color highlights different subregions within the peak regions. Points on the diagonal mean that model prediction at unmappable positions (x-axis) is not systematically different from the mappable positions (y-axis). None of the unmappable positions are predicted to have 0 probability of reads indicating that the model is not overfitting to the false 0 counts at these positions. Points with [0, 100] distance from the center are more scattered since the median is computed over very few points for unmappable positions and because the strongest signal associated with footprints (and hence highest variance) is also observed in these regions. **b)** Number of unmappable positions for different positions within the 1000 bp peak region. Overall, 0.5% of the positions were unmappable.



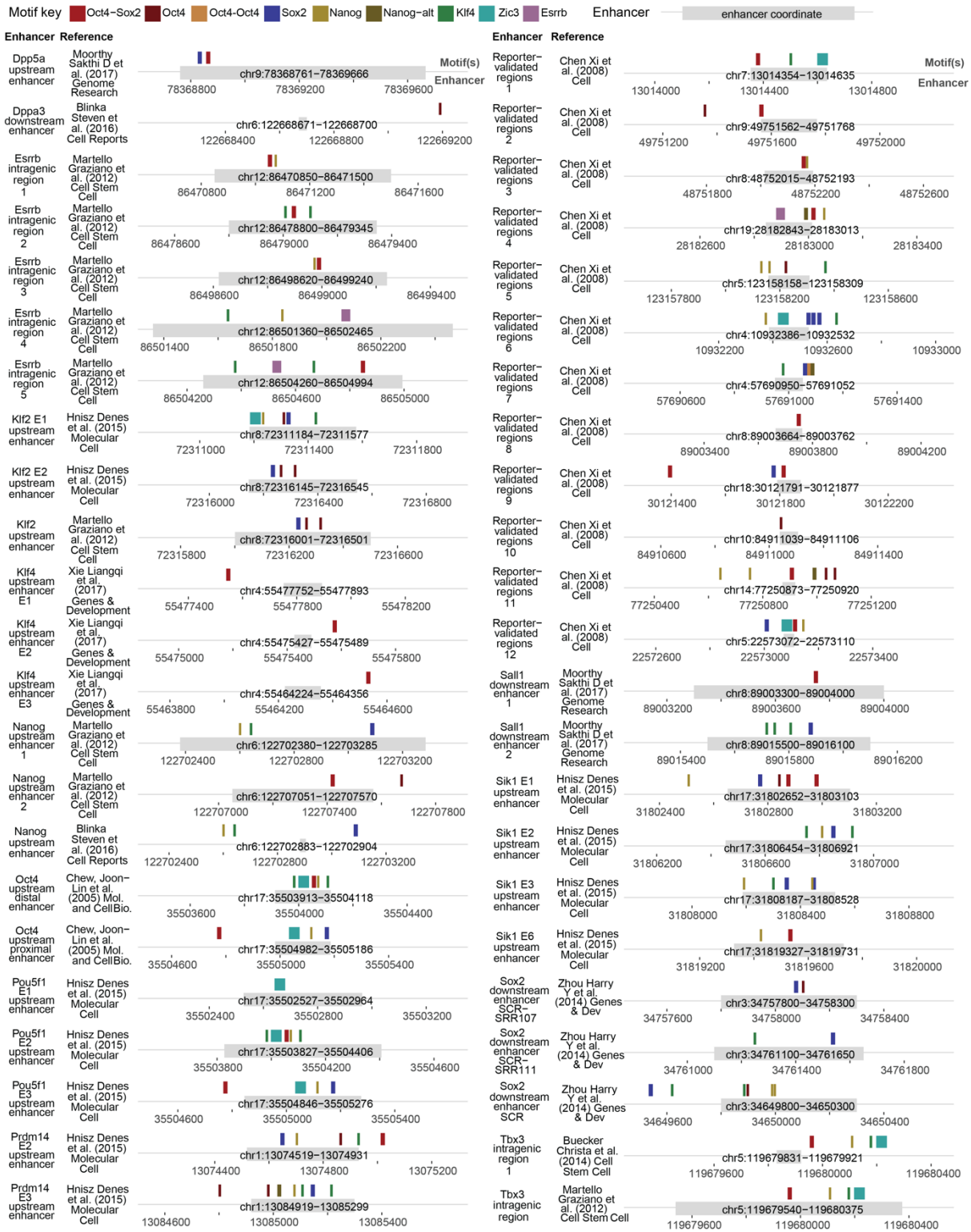
Supplementary Fig. 2: BPNet predictions and sequence contribution scores at enhancers with experimentally validated motifs. a) Observed and predicted ChIP-nexus read counts for the *Oct4* distal enhancer are shown to complement Figure 2b. **b,c,d)** Previously validated binding motifs for *Oct4-Sox2* were re-discovered by BPNet. ChIP-nexus read counts and BPNet contribution scores for three enhancers are shown. **b)** The *Oct4-Sox2* motif site in the *Klf4* E2 enhancer was validated by deleting the site using CRISPR/Cas9[1]. **c,d)** The *Oct4-Sox2* binding motifs in the *Nanog* and *Fbx15* enhancers were confirmed previously using reporter assays of constructs with various motif mutations[2,3].



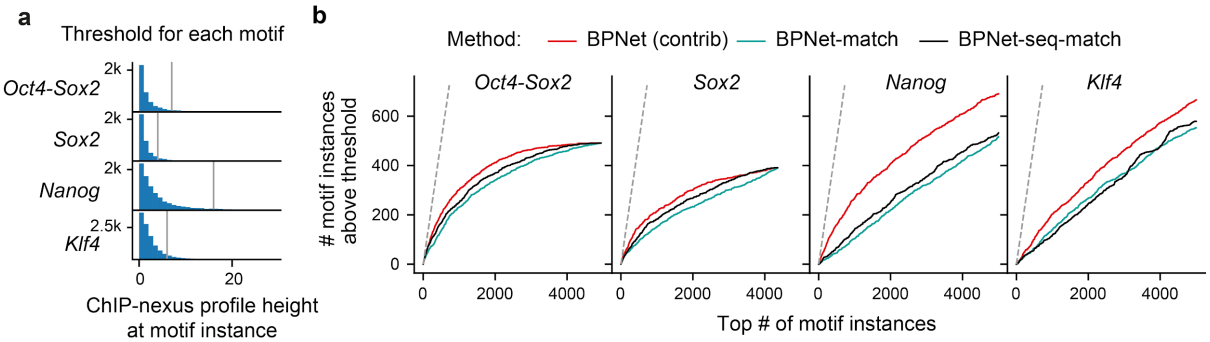
Supplementary Fig. 3: Overview of all short motifs discovered by TF-MoDISco. All 33 discovered short motifs (information content <30 bits) are shown from left to right with: motif ID, number of seqlets supporting the motif, CWM, PFM and average ChIP-nexus read count distribution (180 bp) for each TF. Sequence logos and profile plots share the same y-axis in each column. Motif ID consists of the TF name for which the motif was discovered (O for Oct4, S for Sox2, N for Nanog and K for Klf4) and the order (starting with 0) in which the motif was discovered by the TF-MoDISco run for the TF. The height of the CWM varies depending on how important the motif was for the TF for which it was discovered.



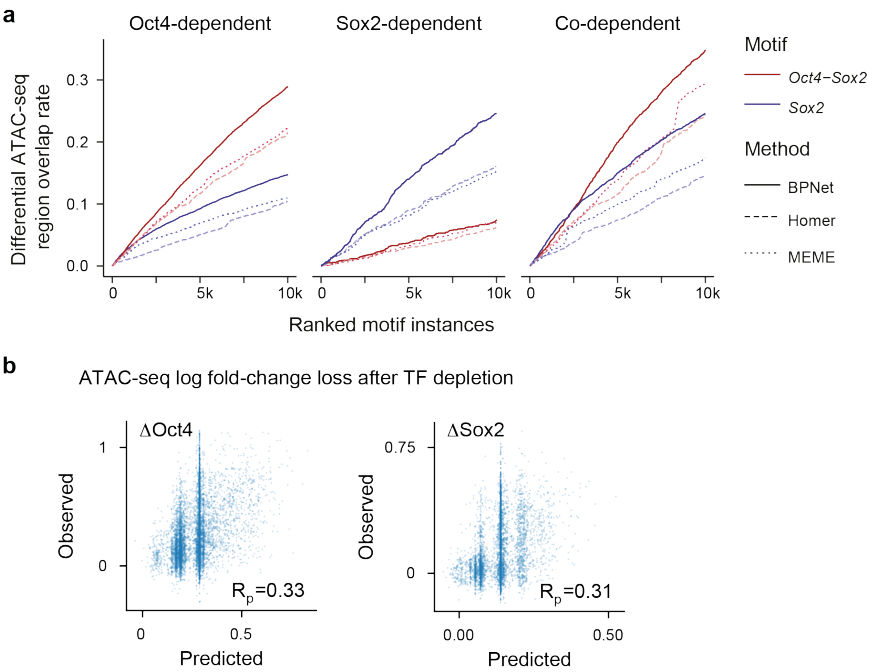
Supplementary Fig. 4: BPNet trained on different chromosome sets (folds) yields similar motifs. The closest motif match from each BPNet/TF-MoDISCo cross-validation run resembles the originally discovered motifs.



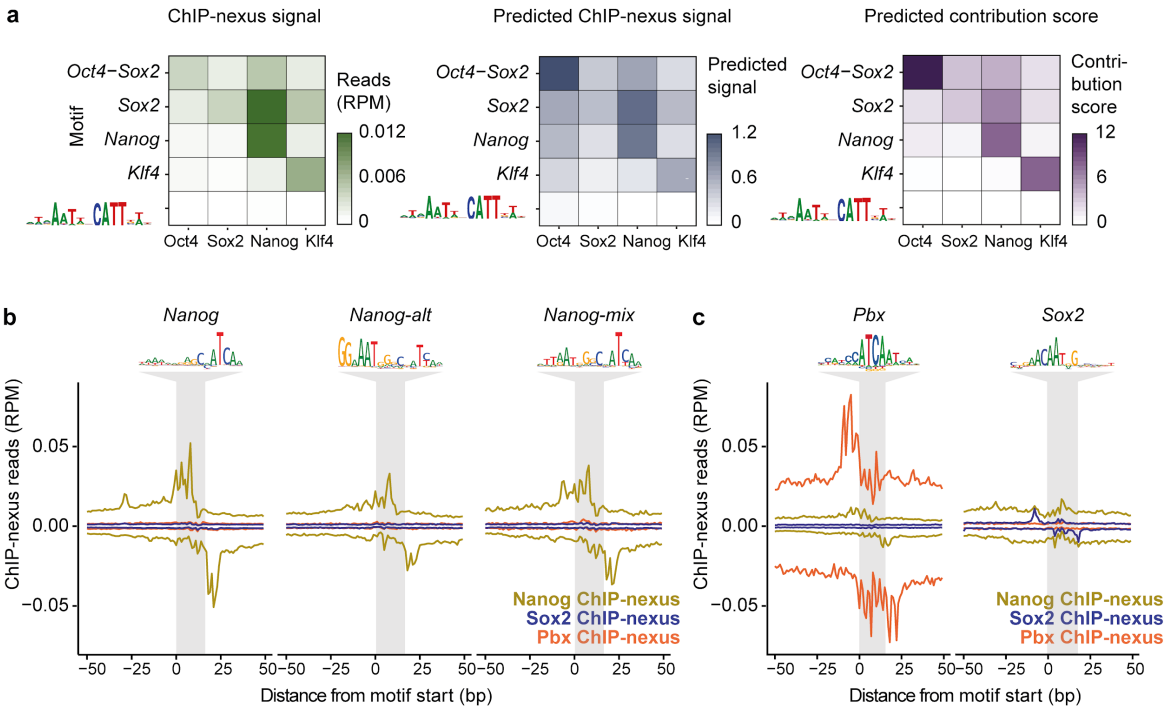
Supplementary Fig. 5: Mapped motif instances across previously identified enhancer regions. The enhancer coordinates (gray boxes) are based on previous publications[1,4–10] and do not necessarily imply that the region is sufficient for enhancer activity. Motif instances mapped by CWM scanning (colored boxes) are shown in the surrounding region.



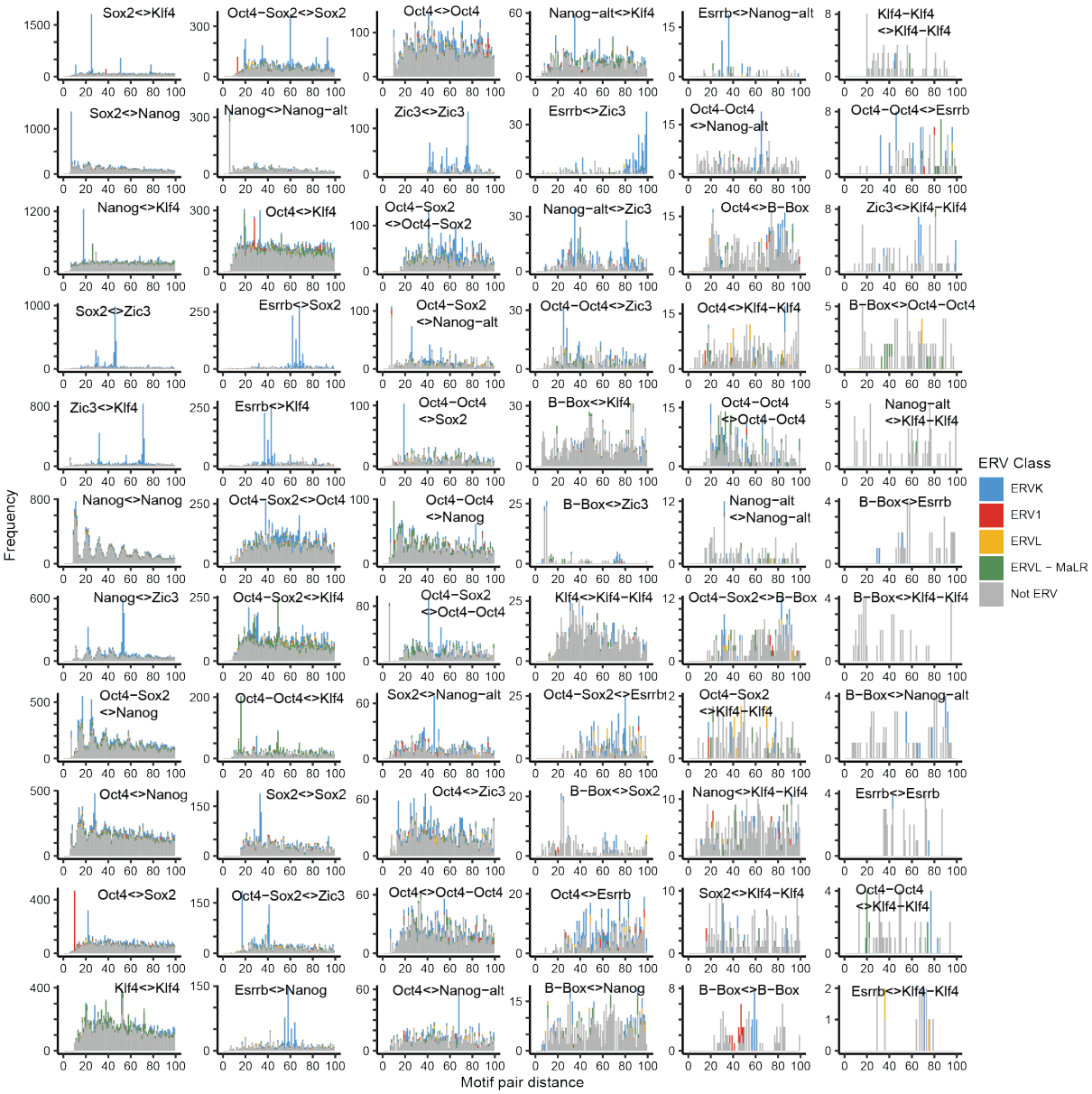
Supplementary Fig. 6: BPNet's contribution scores outperform match-based scores in recovering motif instances supported by ChIP-nexus footprints. a) ChIP-nexus profile height distribution at the reference summit position for BPNet motif instances of different TFs (Supplementary Methods). The vertical gray lines denote the 90th percentile that is used as a very stringent threshold for calling motif instances as having a ChIP-nexus footprint. **b)** Number of top-ranked N motif instances located up to 500 bp away from the ChIP-nexus peak summits showing a ChIP-nexus footprint larger than the threshold defined in (a). The top N motif instances (x-axis) were ranked either by the contribution score magnitude at the motif region (BPNet), match score between the CWM and the contribution scores (BPNet-match), or the match score between the PWM and the sequence (BPNet-seq match). Note that all motifs in this analysis were originally called by BPNet (using both the contribution score magnitude and match), hence motifs with a good sequence match and poor contribution scores (or good contribution score and poor sequence match) were already excluded during the original instance calling.



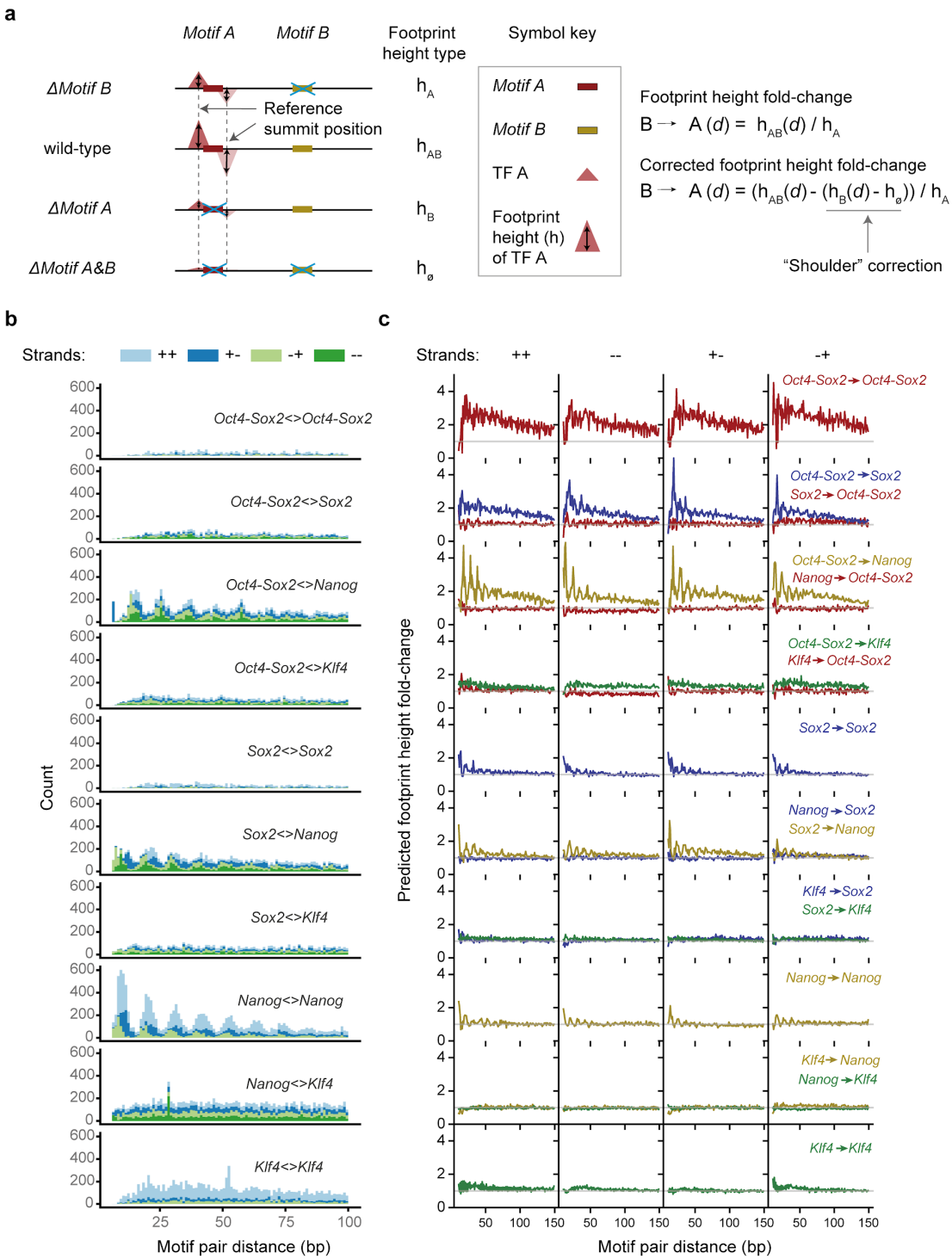
Supplementary Fig. 7: BPNet's contribution scores better predict differential ATAC-seq data than traditional methods. a) Overlap of ranked *Oct4-Sox2* and *Sox2* motif instances (in thousands *k*) with regions that lose ATAC-seq signal in response to either Oct4 or Sox2 depletion as previously defined[11]. In addition to the data shown in Fig. 2g, the results for co-dependent regions are shown, in which ATAC-seq signal is lost both in response to Oct4 or Sox2 depletion. Both *Oct4-Sox2* and *Sox2* motifs are present. Motif instances ranked by BPNet contribution scores also outperformed those obtained by HOMER and MEME (ranked by PWM match scores). **b)** Linear regression model based on motif instance features, rather than the BPNet bottleneck layer (Fig. 2h), of the ATAC-seq log fold-change between Oct4 ON and OFF (*Δ*Oct4, left) or Sox2 ON and OFF (*Δ*Sox2, right). Note that the Pearson correlation (R_P) coefficient for both motif instance based models is only around half of that of the models based on the bottleneck layer. The motif instance features were the number of BPNet motif instances located in the 1 kb ChIP-nexus peaks and their average sequence match scores (Methods). Similar results were obtained for motif instances derived by MEME/FIMO, ChExMix and HOMER. This result indicates that the mapped motif instances are strongly enriched in differentially accessible sites after TF depletion and contribute to the prediction of differential ATAC-seq signal. However, the sequence representation (bottleneck layer) learned by the BPNet model encodes additional information (such as motif syntax) beyond linear, additive effects of motif instances, thereby significantly improving prediction of differential accessibility.



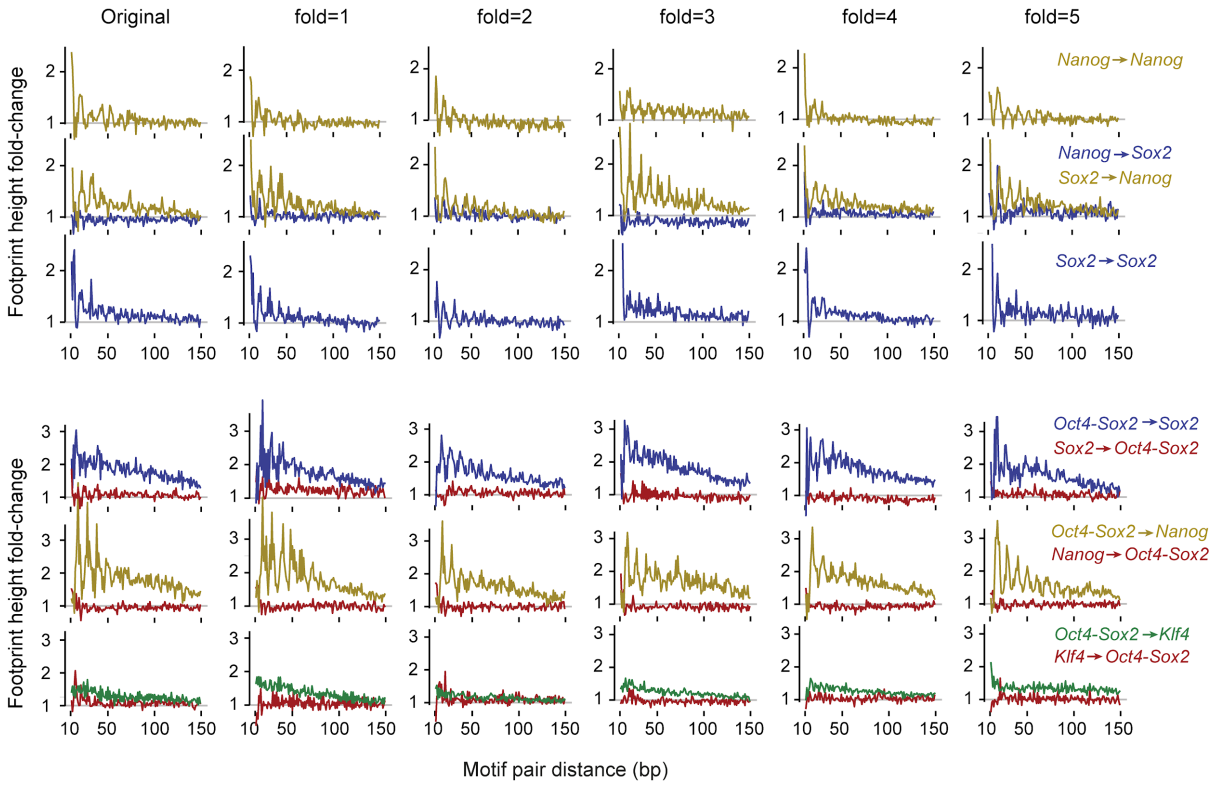
Supplementary Fig. 8: No evidence that Nanog binds with a partner. a) Median Oct4, Sox2, Nanog and Klf4 ChIP-nexus signal, predicted BPNet signal and DeepLIFT contribution scores show no signal across the genomic instances matching the putative *Nanog-Sox* heterodimer motif (RMWMAATWNCATTSW)[12]. The signal for *Oct4-Sox2*, *Sox2*, *Nanog*, and *Klf4* motif instances are shown as control. **b)** Since the *Nanog* motif resembles the known *Pbx* binding motif, we performed Pbx ChIP-nexus experiments to test whether Pbx might be a binding partner for Nanog. However, the average Nanog, Pbx and Sox2 ChIP-nexus binding profiles (positive strand on top, negative strand at bottom) show no detectable footprints for Pbx or Sox2 on the three *Nanog* motifs, arguing against Pbx or Sox2 being stable interaction partners. However, an unknown interaction partner cannot be ruled out. **c)** The average Pbx ChIP-nexus footprint on the known *Pbx* motif from JASPAR[13] (top 1000 based on PWM-match score) confirms that the Pbx ChIP-nexus experiment worked (left). Likewise, Sox2 shows specific binding to its identified *Sox2* motif (right). Note that the y-axis in Read per million (RPM) is the same for all graphs to allow comparisons of signal strength.



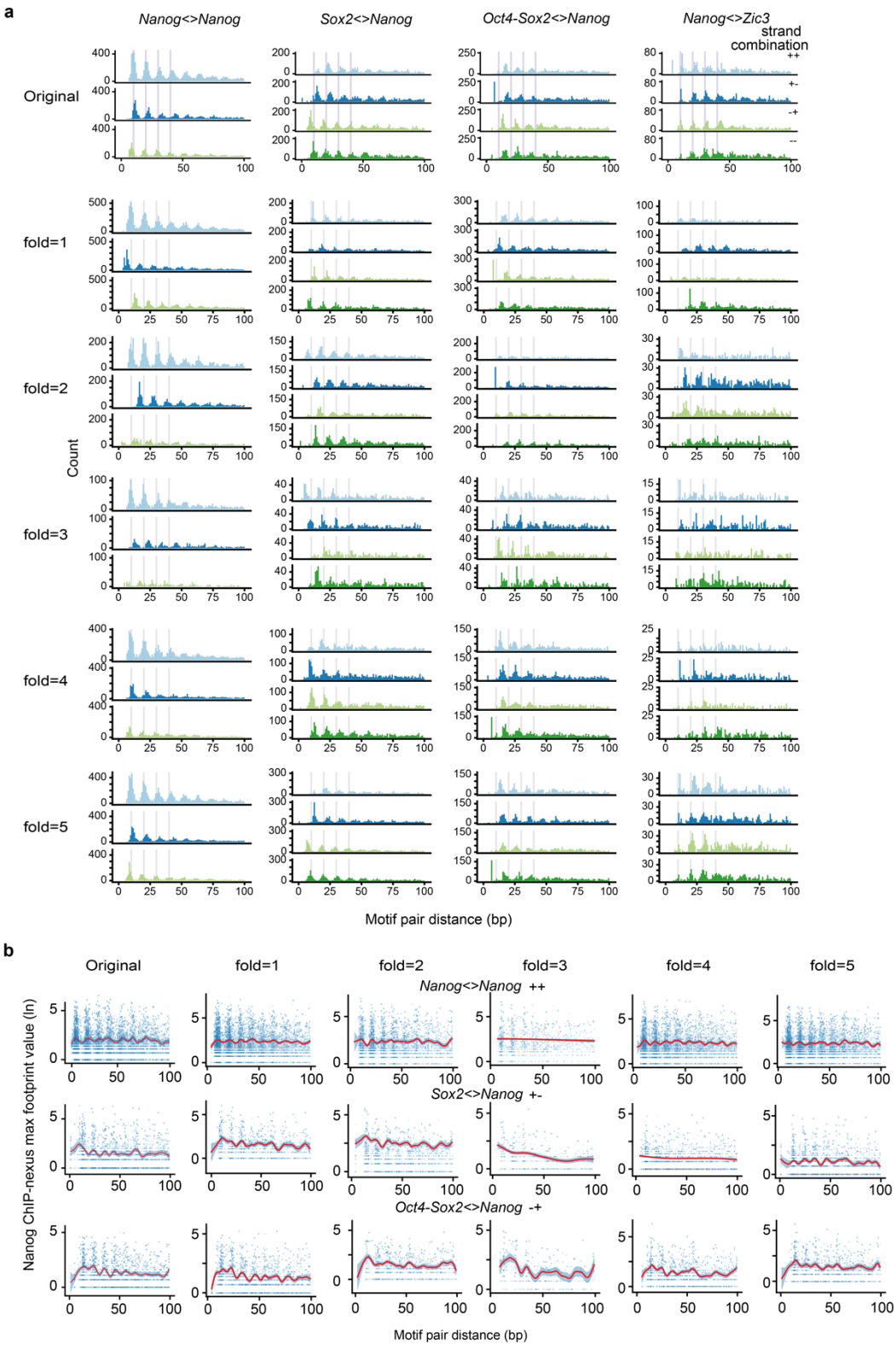
Supplementary Fig. 9: Most over-represented instances of strict spacings between motifs are due to ERVs. Histograms depicting the frequency of center-to-center inter-motif distances across the 11 representative motifs. Colors represent ERV classes which overlap with the corresponding motif pairs.



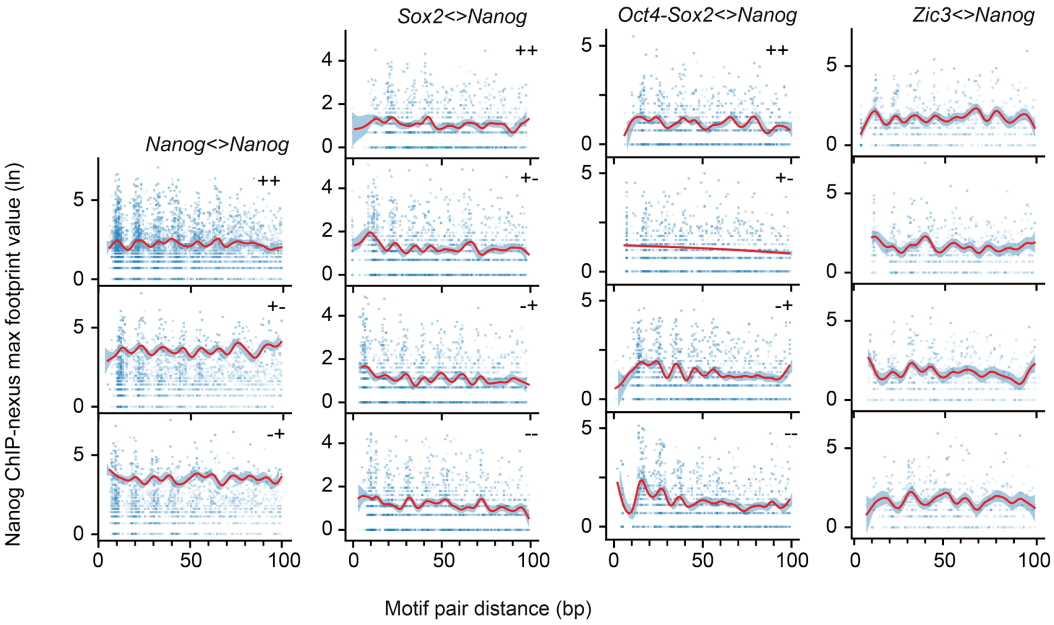
Supplementary Fig. 10: *In-silico* motif interactions for all strand orientations. a) The influence of *Motif B* on the binding of TF A at *Motif A* is quantified by the fold-change of predicted footprint height at the reference summit position when *Motif B* is present or absent nearby (h_{AB} vs h_A). The footprint height fold-change is corrected for the "shoulder" effect of *Motif B* by subtracting the predicted footprint height when only *Motif B* is present in the sequence. **b)** Distance distribution of all CWM-derived motif instance pairs in the genome stratified by motif identity and strand orientation. Note that for homotypic interactions, ++ and -- are the same and are shown as ++. Motif pairs overlapping transposable were filtered out. **c)** *In-silico* analysis of motif interactions on synthetic sequences measuring the predicted footprint height fold-change for all motif pairs across all strand orientations. Note that no clear differences between the possible strand orientations were detected.



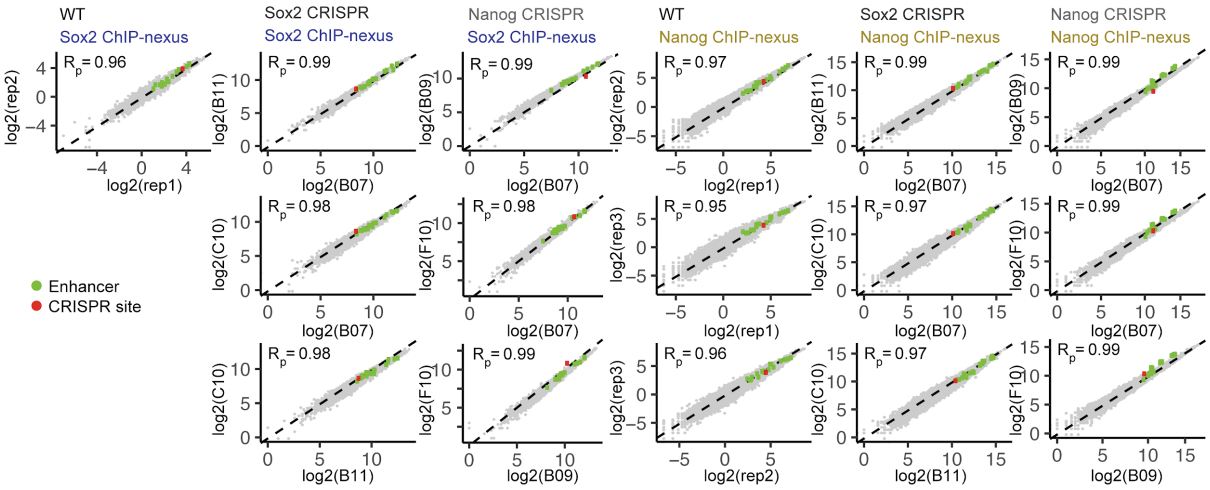
Supplementary Fig. 11: BPNet trained on different chromosome sets (folds) yields similar *in-silico* interactions results. *In-silico* interaction analysis as shown in Fig. 4a using BPNet trained on five different chromosome folds and using the motifs discovered by TF-MoDISCo for each fold. The observed interactions are highly reproducible across all folds.



Supplementary Fig. 12: Nanog motif periodicity analysis for BPNet trained on different chromosome sets (folds). a) Distance distribution of motif pairs (as shown in Figs. 5e-h) for motif instances obtained from BPNet models trained on each of the five different chromosome folds. The motif instances were obtained by scanning the contribution scores with the CWMs obtained from the different chromosome folds as shown in Supplementary Fig. 4. **b)** Nanog ChIP-nexus signal at the reference summit position for motif pairs as shown in Fig. 5i-k for different BPNet models trained on different chromosome folds and the corresponding motifs. Smooth curve fit (B-splines) is depicted as a red line. The fit error bars denoting the 95% confidence intervals are shown as blue ribbon.



Supplementary Fig. 13: Nanog exhibits helical periodicity independent of motif orientation. Nanog ChIP-nexus binding maximum signal across *Nanog* motif pairs (blue dots), where the median of each motif pair distance is depicted as a red line. Nanog on average binds higher when the partner motifs (*Oct4-Sox2*, *Sox2*, *Zic3*) are within the preferred spacing to *Nanog.* This trend is observed for all motif orientations (unless there are not enough data points as observed for one of the *Oct4-Sox2<>Nanog* strand orientations). Smooth curve fit (B-splines) is depicted as a red line. The fit error bars denoting the 95% confidence intervals are shown as blue ribbon. Number of data-points used to estimate 50 smoothing parameters for each plot: Nanog<>Nanog ++ 8930 +- 3922 -+ 3382, Sox2<>Nanog ++ 3299 +- 4011 -+ 3920 -- 3736, Oct4-Sox2<>Nanog ++ 4481 +- 4849 -+ 4947 -- 4945, Zic3<>Nanog ++ 1456 +- 1300 -+ 1359 -- 1214.



Supplementary Fig. 14: Replicates for the CRISPR experiments were highly correlated. ChIP-nexus experiments were performed on wild-type (WT) R1 ESCs in biological replicates (on different days with cells grown at a different passage number): two for Sox2 and three for Nanog. For each CRISPR motif mutation, ChIP-nexus experiments of three monoclonal cell lines were performed and used as replicates: clones B07, B11 and C10 for the mutant *Sox2* motif, and clones B07, B09 and F10 for the mutant *Nanog* motif. Pairwise comparisons between biological replicates across WT and clonal replicates of cell lines clones of CRISPR motif mutations show high Pearson correlations (R_p), as shown in the top left of each scatter plot. Each point in the scatter plots represents ChIP-nexus RPM counts of either Sox2 or Nanog across 151 bp genomic windows, centered on the respective motif. Known enhancer regions from Supplementary Fig. 5 (Enhancers) and the selected mutated genomic region (CRISPR site) remain consistent between replicates.

Supplementary Tables

Supplementary Table 1. List of all ChIP-nexus and ChIP-seq replicate experiments and the associated quality-control metrics including the number of unique de-duplicated reads, highest number of IDR peaks between replicate pairs, number of "optimal IDR peaks", the IDR rescue ratio, and fraction of reads in IDR optimal peaks (FRiP). All samples had uniformly FRiP scores, an estimate for enrichment. For all four TFs, ChIP-nexus samples had a <2 IDR rescue ratio (ratio of the number of IDR optimal peaks from pseudo-replicates to the number from true replicates), an estimate for reproducibility. The IDR optimal peaks from ChIP-nexus data also showed strong overlap with those from ChIP-seq data targeting the same TF.

Supplementary Table 2: Further binary classification metrics corresponding to Extended Data Fig. 4a.

Supplementary Table 3: Sequences of guide RNA and single-stranded oligo donors for CRISPR mutations.

Supplementary Data

Supplementary Data 1. Clustered motifs and their labels. Motifs were obtained by TF-MoDISco ran on BPNet models trained on 6 different datasets: i) seq/profile.peaks-union (ChIP-seq profile model trained on a union of ChIP-nexus/ChIP-seq peaks), ii) seq/binary (binary classification model trained on genome-wide ChIP-seq peaks), iii) seq/profile (ChIP-seq profile model trained in ChIP-nexus peaks), iv) nexus/profile.peaks-union (ChIP-nexus profile model trained on a union of ChIP-nexus/ChIP-seq peaks), v) nexus/binary (binary classification model trained on genome-wide ChIP-nexus peaks), vi) nexus/profile (ChIP-nexus profile model trained in ChIP-nexus peaks). Each motif logo shows the sequence information content of a PFM. The logo title consists of the manually assigned motif label (e.g. TE1, Oct4-Sox2) and the motif ID composed from the model name, the task name and TF-MoDISco motif ID (e.g. seq/profile/Nanog/m0_p13).

Supplementary Videos

Supplementary Videos 1-6. BPNet profile predictions averaged across 128 random sequences with two motifs inserted at different positions. Centers of the motifs are marked by the vertical gray line. Motif distance is shown on the right. For each motif, the predicted profile of the corresponding TF is shown on the y-axis.

Supplementary Note

Supplementary Q&A

Q1. What are the key innovations in the design of the BPNet model compared to previous deep learning models of TF binding?

1) The predominant prediction task for TF binding models in the current literature is a binary classification task where the goal is to discriminate sequences putatively bound by a TF (based on relatively low-resolution ~200 bp peak calls) from background unbound sequences. However, the primary objective of our study is the discovery of cis-regulatory motif syntax of TF binding. Binary peak classification is not the optimal prediction task for this objective. For binary peak classification, the output labels for all bound regions (peaks) are the same irrespective of the subtle changes in motif syntax across these bound sequences. Hence, the classification models tend to focus more on features that discriminate bound from unbound sequences such as motifs as compared to subtle variation in motif syntax across the bound peak sequences that give rise to distinct base-resolution binding profiles. The reason for our model's success in deciphering motif syntax is that we modeled the experimental binding profiles at the highest possible resolution across the peak regions, which allowed the models to learn how motif syntax within bound sequences affects subtle variations in base-resolution binding profiles.
2) BPNet is the first neural network architecture designed to model continuous base-resolution binding profiles from ChIP-nexus/exo and ChIP-seq experiments as a function of DNA sequence. This in contrast to all previous deep learning approaches that model binding data as binary binding events or continuous binding signal summarized at low-resolution (100-200 bp). The BPNet architecture is inspired by the original WaveNet architecture with dilated convolutions for modeling speech signals. The dilated convolutional architectures with residual connections have been previously used for modeling chromatin profiles (e.g. Basenji) and splicing (e.g. SpliceAI). The dilated convolutions allow the model to use a wide receptive field without exploding the number of parameters. We use a fully convolutional model without any pooling operations for the profile prediction because we want to maintain base resolution throughout the model and primarily want to model translational equivariance (captured by convolutional layers) instead of translational invariance (captured by pooling layers). We also avoided pooling layers (particularly max pooling layers) because they result in models that are less robust to specific choices of hyperparameters (e.g. size and stride of pooling) and produce less stable interpretation (e.g. DeepLIFT profiles). The residual connections allow the models to learn more efficiently using features across multiple scales. This architecture class that we chose is simple to implement, memory efficient and trains efficiently. But as we have shown, our models show high prediction performance (on par with replicate concordance for profile prediction), are very robust to different hyper-parameters and allow stable inference of motifs and motif syntax that are validated by experiments.
3) BPNet introduces a new multi-scale loss function to separately optimize the predictions of ChIP-nexus profile shape and total read counts. We use a multinomial negative log-likelihood loss function for profile shape prediction. The multinomial distribution elegantly models the sparse distribution of reads across all the positions in the input sequence accounting for zeros that occur due to low read coverage. Since the model predicts the

expected number of reads at each position based on the estimated multinomial distribution, the model is able to effectively impute and denoise sparse base-resolution profiles.

4) BPNet introduces a new approach to automatically correct binding profiles for assay biases and prevent the model from learning spurious sequence patterns that explain assay biases. Specifically, BPNet can use a control experiment as an auxiliary input. The model is fit to the ChIP-nexus/ChIP-seq binding profiles using both the sequence and the control experiment track. The model automatically learns to regress out the signal that can be explained by the control track. This approach allows us to directly model and predict raw counts and use count-based loss functions such as the multinomial, while implicitly accounting for biases. Previous approaches typically fit their models to binding data corrected via ad-hoc pre-processing methods.

Q2. What are the enhanced model interpretation methods introduced in the BPNet framework to infer motifs and motif syntax?

1) Several neural network feature attribution methods (e.g. *in-silico* mutagenesis, input gated gradients, DeepLIFT, integrated gradients and GRAD-CAM) have been previously applied to deep learning models of TF binding to infer contribution scores for each base (feature) in an input sequence to a scalar binary or continuous binding prediction. Here, we introduce the first approach to infer base-resolution contribution scores for input sequences with respect to profile outputs from sequence-to-profile models. Specifically, we adapted our DeepLIFT method for profile outputs. Previously, these contribution scores have typically only been used to visualize anecdotal examples of sequence motifs in specific case study sequences or to score non-coding genetic variants. Here, we provide the first genome-wide analysis of contribution scores and show convincingly how they can accurately identify predictive motif instances genome-wide, outperforming traditional motif discovery and scanning approaches.
2) The typical approaches used to derive motifs from deep learning models of genomic sequence include visualizing the convolutional filters directly or deriving position weight matrices (PWMs) from subsequences that activate convolutional filters in the first layer. These approaches have several drawbacks since neural networks learn representations in a distributed fashion i.e. no single filter is guaranteed to capture a complete motif and there is significant redundancy in the patterns learned by different filters. Hence, interpreting filters directly often results in deriving incomplete motif patterns and many partially redundant motifs. Another drawback is that deriving PWMs that simply capture base frequencies results in a fundamental loss of information with regard to the predictive base contributions, as well as any interaction effects between bases within and across motifs. The TF-MoDISco algorithm we use here is fundamentally different in that it reconstructs less redundant and complete motif representations from the base-resolution contribution scores instead of individual filters. We show that TF-MoDISco is able to learn several novel motifs missed by other approaches for four highly studied pluripotency transcription factors.
3) We are the first to highlight the fundamental difference between contribution scores and base frequencies to derive motif representations. We introduce the contribution weight matrix (CWM) representation, which is conceptually similar to the PWM but records the average contribution rather than frequency of bases in a motif. Using transposable elements, we show very clearly the advantage of the CWM over the PWM representation.

PWMs highlight the entire transposable element, whereas the CWM representation only highlights the predictive subsequences corresponding to the motifs bound by the TFs. Furthermore, the CWM motif representation for Nanog highlights the helical pattern in the flanks of the motif, whereas the equivalent PWM representation does not.

4) We develop new methods to identify predictive motif instances by scanning contribution score profiles with CWMs (derived using only sequence and the BPNet model without using the measured binding profiles). We show that this new motif scanning approach outperforms PWMs even when restricting the regions to be close to the peak summit or when augmenting the ChIP-nexus summit position for BPNet (Extended Data Fig. 3b). Further, the CWM scanning approach shows the most dramatic improvements for short motifs with complex periodic patterns such as the Nanog motif.
5) Finally, we combine all of these innovations in the *in-silico* oracle approach for discovering motif syntax. We develop two new approaches - one based on using simulated sequences and one based on perturbing real genomic sequences to derive robust, global rules learned by the models of how motif syntax influences transcription factor binding cooperativity. We show that these two approaches complement each other and support each other's findings. We are the first to show that high performance neural networks of regulatory DNA sequence learn *ab-initio* subtle but critically important syntactic patterns, e.g. the pervasive 10.5 bp helical periodicity displayed by Nanog. We are also the first to show that neural networks can learn preferred soft spacing constraints between motifs that are predictive of cooperative binding. The ability of neural networks to learn these higher-order, non-linear patterns has long been known, but to our knowledge, no one has previously shown robustly that deep learning models of genomic sequences can successfully capture these higher-order patterns and that they can be extracted to reveal biologically meaningful information. Not only can we predict these patterns, but we now also validate the extracted syntax experimentally, by performing point mutations using CRISPR/Cas9 and analyzing the change in TF binding with ChIP-nexus experiments. The results clearly confirm that the periodic Nanog binding depends on the *Nanog* and the *Sox2* motifs (but that Sox2 binding does not depend on the *Nanog* motif).

Q3. What is the conceptual difference between classical motif discovery methods like MEME/HOMER and BPNet's motif discovery method?

Traditional motif discovery methods (e.g. MEME, HOMER) are based on identifying statistically overrepresented patterns in bound sequences relative to a background set of sequences. Our method does not rely on the frequency of sequence patterns in a foreground set of sequences relative to a background set. Instead it learns sequence patterns that are predictive of the binding profiles. Frequent patterns that have no predictive value are not learned and rare patterns that have predictive value are learned. Thus, while patterns have to be present across multiple sequences to have predictive value, our ability to discover them is not simply based on over-representation. This is exactly why we can learn several less frequent motifs predictive of ChIP-nexus footprints, while traditional methods that primarily focus on over-represented patterns miss these motifs.

The contrast between frequency of patterns and their predictive contribution to the output is the fundamental difference between the contribution weight matrix (CWM) representation that we introduce, and the classical frequency-based position frequency/weight matrix (PFM/PWM) representation (PWMs are log-odds of PFMs normalized against background frequencies). The case study of transposable elements best highlights this

fundamental difference. The PFMs highlight the entire over-represented retrotransposon sequences. However, the CWMs highlight the specific predictive bases within the retrotransposons, which is a clear advantage over traditional methods.

Q4: How is the BPNet oracle approach for syntax discovery different from classical methods?

Previous methods derive summary statistics of over-represented or evolutionary conserved patterns of motif or TF peak co-occurrence and spacing from bound and unbound genomic sequences[14–16]. While these summaries are useful, they suffer from several issues. First, it is difficult to estimate the marginal effect of each property of cis-regulatory syntax (e.g. spacing between two motifs) due to systematic confounding from other syntactic properties (e.g. presence of other motifs, homotypic and heterotypic motif density) and background sequence composition of genomic sequences. Second, the number of genomic instances that sample each syntactic property may not be sufficient to obtain robust statistics. Finally, it is difficult to make conclusions about the impact of an over-represented syntactic property on cooperative TF binding without systematic perturbation experiments. These issues are typically resolved experimentally by performing *in vitro* binding experiments using libraries of carefully designed synthetic sequences that sample desired properties of interest[17–21].

Our *in-silico* oracle approach that uses designed synthetic sequences mimics this experimental approach since the model is trained to predict experimental *in vivo* binding profiles. However, the model can sample substantially larger numbers of synthetic constructs by smoothly varying syntactic properties while accounting for sequence backgrounds. The *in-silico* mutagenesis experiments on genomic sequences not only provide additional support for conclusions derived from the synthetic sequences, but also reduce the likelihood of making unreliable conclusions about "out-of-distribution" syntax properties that are never found in the genome.

By mimicking the experimental approach *in silico*, the oracle approach allows one to home in on precise hypotheses that can be tested using less scalable *in vivo* experiments such as the CRISPR editing experiments we present. Furthermore, the oracle approach is very general and will be useful to systematically study further syntactic properties and their joint effects in the future (e.g. trade off between affinity, motif density and spacing). These kinds of interactions would be very difficult to infer from explicit parameters in computational models.

Q5. Can we use CWM motifs to identify motif instances in sequences that do not have experimentally measured ChIP-nexus profiles?

Yes. The CWM scanning approach can be used to identify motif instances in any query sequence. The procedure requires two components - the input sequence and the BPNet model. These are the same two components also used for traditional PWM motif scanning (the PWM happens to the "model" in that case). The BPNet model predicts the output binding profile from the raw input sequence using a forward pass. That "predicted output" is then used to infer a contribution score profile across the input sequence using backpropagation (DeepLIFT). The CWM (also derived from the DeepLIFT contribution scores with respect to predicted profiles from all training set sequences) is then used to scan the contribution score profile of the input sequence. Nowhere in this procedure do we use the measured ChIP-nexus

profiles. All components are derived just from the raw sequence using the model and the interpretation methods.

Q6. Can BPNet be used to predict TF binding in new cell types or new species not used in training?

The sequence-only BPNet models cannot be directly used, as is, for cross cell-type prediction. Transcription factors typically have different genomic occupancy profiles across different cellular contexts. The direct binding motifs for most TFs are generally consistent across cell types. However, the higher-order syntactic rules and cooperative interactions with motifs of other TFs vary across cell types. A model that uses only DNA sequence as input and is trained on binding profiles of a TF in a one cell type will learn sequence features that are specific to that cell type. Because the DNA sequence of a genome is the same across different cell types, a sequence-only model of TF binding cannot predict different genome-wide TF binding landscapes in new cell types not used in training. However, the primary use-case for BPNet framework is not prediction of TF binding in new cellular contexts. Rather, we designed BPNet to enable inference of context-specific sequence determinants of TF binding. BPNet models could be extended to take as input sequence *and* cell-type specific information such as chromatin accessibility or histone mark profiles. These multi-modal models trained to account for differences in training and test cell type regulatory syntax are likely to generalize across cell types.

BPNet models for a TF in a specific cell-type in one species could in principle be used to predict binding profiles for the same TF in a well matched cell-type in a closely related species (e.g. human and mouse). However, there can be several issues with this naive cross-species prediction approach. First, the cell types may not be equivalent across species. E.g. mouse ESCs that we use in this work correspond to a “naive” pluripotent state whereas typical human ESCs cultured under the same conditions correspond to a later “primed” states. Further, the genomes of different species often house very different classes of repeat elements. E.g. mouse and human genomes have very different classes of transposable elements to which several TFs bind, which will cause erroneous predictions of a model that is naively applied across species. Finally, a sequence-only model is unlikely to be optimal for genome-wide prediction of TF binding across species. Cross-species prediction would improve significantly by coupling the BPNet sequence model with (i) additional input signals as as chromatin accessibility or histone marks and (ii) the use of domain adaptation methods for training that enable models to generalize to out-of-distribution settings by avoiding domain-specific (i.e. species-specific) features. These are exciting directions for future enhancements.

Q7: What were the rigorous evaluations and independent validations that support the statistical robustness and biological validity of results?

First, we performed extensive quality control analyses on our ChIP-nexus data, ensuring high reproducibility, sensitivity and antibody specificity (Online Methods). Second, we corrected for assay-specific biases by explicitly modeling control datasets (Online Methods) and excluded any discernible influence of mappability artifacts on BPNet’s predictions (Supplementary Fig. 1). Third, we showed that independently trained models using different subsets of the binding data produced highly consistent results, including the motif syntax rules (Supplementary Fig. S4, S11, S12a), thereby minimizing the possibility of artifacts due to memorization or over-

fitting to the training data. Fourth, we found that the derived motif syntax rules were internally consistent. They were inferred from both synthetic and genomic DNA sequences (Fig. 4, Extended Data Fig. 6) and the directionality was consistent with the indirect binding footprints observed in ChIP-nexus data (Fig. 3). Nanog's helical periodicity was also found in the ChIP-nexus data, the raw contribution scores, as well as the spacings between motifs (Fig. 5). Fifth, the sequence representation learned by BPNet from TF ChIP-nexus data transferred seamlessly to accurately predict independent, previously published experiments, i.e. the changes in chromatin accessibility after TF depletion (Fig. 2g-h). Finally, we performed CRISPR-induced point mutations in two binding motifs and showed that the changes in ChIP-nexus profiles are in remarkable agreement with BPNet's predictions and inferred syntax (Fig. 6). These careful controls and evaluations provide confidence in the ability to use BPNet as a generic toolkit for deriving biological insights about syntactic properties of regulatory DNA from ChIP-nexus and ChIP-seq experiments.

Q8: Can motifs, motif instances and motif syntax be learned from BPNet models of TF ChIP-seq data? Can BPNet also be applied to CUT&RUN, DNase-seq or ATAC-seq profiles?

Yes. As we have shown in this paper, the BPNet model, with minor modifications to the architecture, can be successfully trained on base-resolution TF ChIP-seq profiles.

- BPNet shows high predictive performance for ChIP-seq profiles similar to the ChIP-nexus data (on par with replicate concordance for profiles).
- DeepLIFT+TF-MoDISCo applied to the ChIP-seq models is able to recapitulate the majority of the motifs learned by the ChIP-nexus models.
- With respect to accuracy of motif instances, the ChIP-nexus BPNet profile models outperform ChIP-seq BPNet profile models, which outperform ChIP-seq binary models. Hence, modeling profiles directly and improved resolution of the ChIP-nexus profiles collectively contribute to improved motif instance identification.
- The ChIP-seq models are also able to learn subtle syntax features such as helical periodicity and preferred spacing of motifs. E.g. helical periodic Nanog motif spacing is also observed from the DeepLIFT profiles inferred from Nanog ChIP-seq models.

BPNet can be applied to CUT&RUN, DNase-seq and ATAC-seq with minor modifications to the architecture. We plan to release archetype models for these other data modalities in the near future. The fidelity of the BPNet models as well as quality and sensitivity of the interpretation of syntax largely depends on the quality of the data.

Supplementary Text

BPNet and TF-MoDISco outperform PWM scanning methods in motif discovery, mapping of motif instances and syntax discovery

Motif discovery

To evaluate the extent and quality of motifs discovered by the BPNet framework in the light of previous methods, we compared our approach to ChExMix[22], MEME[23–26] and HOMER[27]. We used each of these methods to discover motifs using ChIP-nexus peaks for Oct4, Sox2, Nanog and Klf4 (for specific parameters, see Benchmarking alternative methods in the Supplementary Methods section below). For each of the methods (besides BPNet), we selected the motifs with the closest match to the 11 core TF-MoDISCo motifs discovered using BPNet (Extended Data Fig. 3a). We found that ChExMix, MEME and HOMER individually discovered at most only 6 out of 11 motifs. They collectively found 9 out of 11 motifs. Only *Oct4-Sox2*, *Sox2* and *Klf4* motifs were discovered by all four methods. *Zic3* and *Esrrb* motifs were only discovered by BPNet and HOMER. The *B-box* motif was only discovered by BPNet and MEME. We speculate that these motifs were missed by ChExMix due to the associated footprints being heterogeneous and fuzzy with relatively lower read coverage. MEME/HOMER may have missed some of these since they are not as over-represented as the primary motifs. Moreover, the other methods were limited in their ability to discover long TEs motifs with the default parameters. Although changing the parameters for the other methods may allow the discovery of some TEs, the dependence on these parameters makes it difficult for the methods to discover TEs alongside short motifs in a flexible and robust manner. Moreover, these other methods cannot highlight the short constituent motifs bound by the TFs within the longer TE motifs. These results suggest that the BPNet framework provides substantial improvements in motif discovery compared to ChExMix, MEME and HOMER.

Evaluation of motif instances

To evaluate the quality of the called motif instances in the genome in terms of their false-positive rates, we compared the BPNet approach using CWM scanning to classical position weight matrix (PWM) scanning. To determine the false positive rate of the motif instances in the test chromosomes, we considered motif instances supported by strong ChIP-nexus footprints ('ChIP-nexus profile height' above the 90th percentile) as true binding sites (Extended Data Fig. 3b). Since the number of motif instances depends on the motif scoring threshold, rather than using a fixed threshold, we instead evaluated the true-positive fraction across the ranked list of motif instances. PWM scanning approaches only use the input sequences and the PWMs. CWM scanning uses TF-MoDISCo CWMs, the input sequence and DeepLIFT contribution scores to the predicted (not measured) ChIP-nexus profiles which are derived from the BPNet model. We performed motif instance evaluation only on 1 kb input sequences from the held-out (test) chromosomes, which were not used to train BPNet. Hence, both BPNet CWM scanning and PWM scanning approaches (FIMO and HOMER) do not implicitly or explicitly use the measured ChIP-nexus profiles to call or score instances. ChExMix uses the measured ChIP-nexus profiles to call motif instances. CWM based motif instances were ranked based on their 'motif contribution scores'. PWM based motif instances were ranked based on PWM match log-odds scores. We found that across matched number of ranked instances, CWM motif instances had substantially more overlap with strong ChIP-nexus footprints ('true positives') and thereby exhibited a lower false positive rate compared

to PWM based motif instances from FIMO and HOMER, as well as ChExMix (Extended Data Fig. 3b, top).

Since this evaluation was performed for motif instances in 1 kb sequences centered at ChIP-nexus peak summits, most of the motif instances tend to be located near the center. To show that the improvement of BPNet is not primarily coming from the fact that it could prioritize motifs in the center higher, we computed the contribution scores with BPNet on sequences jittered +/- 200 bp around the peak summit (BPNet-augm). We found that this jittering resulted in negligible drop in performance (Extended Data Fig. 3b), indicating that the superior performance of CWM scanning is not due to an implicit positional bias.

Furthermore, we repeated the evaluation by restricting motif instances to the central 200 bp around the ChIP-nexus peak summits (Extended Data Fig. 3b, bottom). This scenario is commonly used with PWM scanning to reduce the false discovery rate, because true motif instances are more likely to be found closer to the peak summits. Even in this scenario, BPNet's CWM scanning outperformed all other methods for all motifs. The only exception was *Oct4-Sox2*, where several PWM methods performed as well as CWM scanning, likely due to the high information content of this particular motif. The largest improvement of BPNet CWMs over PWM methods was seen for the short *Nanog* motif. Even though the CWM has the same length as the PWM, the base-resolution contribution scores scanned by the CWM are dependent on the entire sequence context of the motif within the 1 kb region and thus can integrate more contextual information relevant for TF binding. E.g. a *Nanog* motif instance can get a higher contribution score if it is present in the vicinity of other ~10.5 bp spaced *Nanog* motif instances. In contrast, the PWM scores sequence matches of each sliding window within the input sequence independently and is unable to account for the influence of surrounding bases and motifs.

Finally, we note that the superior performance of CWM scanning over PWM scanning is highly reproducible when evaluated based on independent ChIP-nexus experiments using a different Nanog antibody (Extended Data Fig. 3b, last column). Hence, our approach of scanning the contribution scores using the CWM (instead of the raw sequence using the PWM) greatly reduces the false positive sites while still following the familiar scanning procedure as with PWMs. These results highlight the advantages of using profile contribution scores and the novel CWM motif representation to identify motif instances associated with ChIP-nexus footprints.

Helical motif syntax for Nanog

Next, we tested whether the ~10.5 bp helical *Nanog-Nanog* motif syntax discovered by BPNet could also be discovered using motif instances from the other methods (HOMER, MEME, ChExMix). We note that ChExMix binding site calls also use the ChIP-nexus footprints in addition to sequence. We found that the methods showed substantially weaker signals of helical periodicity in the *Nanog-Nanog* inter-motif motif distance histograms across a range of distances (Extended Data Fig. 7a). For PWM scanning, the higher false discovery rate of motif instances likely attenuates the detection of Nanog's helical periodicity. For ChExMix, we observed a substantial depletion of helical periodicity for spacing <40 bp compared to BPNet CWM scanning. This depletion at close proximity could be due to two reasons. First, the optimized likelihood of ChExMix is non-convex and hence the global optimum might be difficult to find and may strongly depend on the initial conditions. Second, the key assumption of

ChExMix is that the tag distribution (representing the average profile) associated with a specific motif is constant. However, this assumption is an oversimplification since ChIP-nexus profiles associated with a motif can change their form in the presence of motifs of other cooperatively bound TFs. Altogether, these results demonstrate that compared to previous approaches, the BPNet framework provides substantially improved sensitivity for detecting subtle, closely-spaced, soft motif syntax since CWM scanning yields more accurate motif instances without relying on the measured ChIP-nexus profiles.

BPNet's profile regression yields more motifs and more accurate motif instances than binary peak classification

A frequently used approach for training deep learning models is to treat the TF binding prediction problem as a binary classification task[28,29]. In this approach, the training examples are sequences extracted from contiguous bins in the genome and the sequence label is positive if a TF binding peak overlaps the bin region (and negative otherwise). The purported benefits of such a labeling approach are as follows. First, the assay-specific biases may be already accounted for in the peak-calling process. Second, the resulting machine learning task – binary classification – is well understood. Hence the standard loss function such as binary cross-entropy and the standard evaluation metrics such as the area under precision-recall curve (auPRC) can be used. However, this coarse-grained binary summarization of the binding profiles discards valuable fine-scale information about signal strength and shape of the profiles.

To investigate the benefit of training the BPNet model on the base-resolution ChIP-nexus profiles compared to lower resolution binary labels, we modified the BPNet architecture and replaced the output heads performing profile regression with output heads performing binary classification. The new output heads consisted of weighted global average pooling using spline transformation[30] and a dense layer followed by sigmoid activation. We trained the model on 1 kb input sequences sampled every 50 bp across the training chromosomes. Sequences were labeled positive if the central 200 bp of the sequence overlapped an IDR-optimal peak. The predictive performance on the held-out tuning chromosomes (2, 3 and 4) was 0.25 auPRC on average across the 4 TFs after tuning the optimal learning rate (Extended Data Fig. 4a, Supplementary Table 2). We also observed that chromosome-wide training of the binary classification models took 3 times longer (Extended Data Fig. 4b) than BPNet, which is trained only on ChIP-nexus profiles from 147,974 peak regions. To ensure that the dilated convolutional layers are also appropriate for binary classification, we also trained and evaluated the Basset[31] and factorized Basset[32] architectures. After tuning the dropout rate with random search, we obtained a slightly lower auPRC of 0.24 for both models, suggesting that our original architecture with dilated convolutions was also a good fit for binary classification. Next, we asked whether the predictive performance of the binary classification model could be improved by adding another output head predicting the strand-specific ChIP-nexus profile as originally done by BPNet. Indeed, the classification performance increased for all TFs yielding an average of 0.31 auPRC (Extended Data Fig. 4a, Supplementary Table 2). We conclude that the read profiles indeed provide additional information not captured by the binary labels.

We next asked whether the contribution scores of the profile regression model highlight additional motifs compared to those obtained from the binary classification model. For the

binary models, we computed the DeepLIFT contribution scores for each TF task (pre-sigmoid activation) and ran TF-MoDISco with the same parameters in the same regions as previously done for BPNet. We clustered the discovered motifs based on their PFM similarity and manually assigned motif labels as done in Extended Data Fig. 2d. Using the contribution scores of the binary classification model, TF-MoDISco discovered 9 out of 11 main short motifs found by the profile regression model (Extended Data Fig. 4c, Supplementary Data 1). The 2 missed motifs, *Oct4* monomer and *B-box*, are hence not frequently used by the binary model to predict the presence or absence of the peak as they might co-occur with other more predictive motifs. Interestingly, a higher number of questionable motifs including GC sequence composition bias motifs, ambiguous motifs and degenerate or noisy motifs were discovered from the contribution scores of the binary classification model. This suggests that the contribution scores of the binary classification model are noisier than for the profile regression model. Nevertheless, we note that the high reproducibility of the discovered motifs using two different model architectures trained on different labeling schemes for the same underlying data demonstrates the robustness of TF-MoDISco.

To compare the accuracy of motif instances for the 4 cognate motifs discovered by TF-MoDISco for both models (*Oct4-Sox2, Sox2, Nanog* and *Klf4*), we performed the instance ranking analysis on test set sequences, as for PWM scanning methods in Extended Data Fig. 3b considering sites with strong ChIP-nexus profile heights (footprints) as 'true' binding sites. The contribution scores of both models yielded a similar recall of *Oct4-Sox2* and *Sox2* motifs supported by strong ChIP-nexus footprints (Extended Data Fig. 4d). Strikingly, the contribution scores of motif instances from the BPNet profile model recalled a much higher fraction of *Nanog* motifs with strong footprints as compared to those derived from the binary models (Extended Data Fig. 4d). Since the *Nanog* motif is frequently found in complex homotypic or heterotypic syntactic arrangements with *Sox2*, the ChIP-nexus profile shape contains rich information reflecting these motif arrangements. Since BPNet was trained on ChIP-nexus profiles directly, it is able to learn these subtle patterns and encode them in the contribution scores, thereby resulting in more accurate motif instances even on unseen sequences in the test set. Additionally, CWM scanning of contribution scores from the binary classification model is comparable to PWM scanning (MEME/FIMO), suggesting that BPNet trained on ChIP-nexus profiles is the key for accurate motif maps.

Altogether, we observe that learning to predict the full ChIP-nexus profiles as done by BPNet instead of lower resolution binary classes reduces the training time by three fold, increases the number of discovered motifs with strong seqlet and biological support, reduces the number of questionable motifs and improves the accuracy of the called motif instances. Moreover, the profile predicted by BPNet assesses binding at individual motifs, which offers higher resolution to study the directionality of TF interactions mediated by soft motif syntax as shown in Fig. 4.

BPNet can also be used to model and interpret transcription factor ChIP-seq profiles

The BPNet model together with the interpretation workflow using DeepLIFT and TF-MoDISco can be readily applied to any regulatory profiling experiment such as ChIP-seq, since it does not make any modeling assumptions specific to ChIP-nexus profiles. The major difference between ChIP-seq and ChIP-exo/nexus is the resolution. For ChIP-seq, the 5' ends of the reads mapping to either strand within an enriched peak region are dispersed in a 100-200 bp

window around the primary binding site (peak summit). In contrast, for ChIP-exo/nexus data, the read density is colocalized in the immediate vicinity (+/- 20 bp) of binding events.

To demonstrate that BPNet can also model ChIP-seq profiles, we performed ChIP-seq for 3 out of 4 previously studied TFs (Oct4, Sox2 and Nanog). We processed the data using the ENCODE ChIP-seq pipeline (v 1.3.6) https://github.com/ENCODE-DCC/chip-seq-pipeline2/releases/tag/v1.3.6 and generated the strand-specific 5' read count tracks as for ChIP-nexus. We then optimized multi-task BPNet architectures to predict strand-specific ChIP-seq profiles of the three TFs from their corresponding 1 kb sequences at IDR optimal peak regions across all three factors. We used the same BPNet architecture for ChIP-seq as for ChIP-nexus and determined optimal hyper-parameters by varying each hyper-parameter individually while keeping others the same as for the ChIP-nexus model (learning rate: 0.05, 0.04, 0.02, 0.01, 0.005, 0.004, 0.002, 0.001, 0.0005; deconvolution size: 1, 10, 20, 30, 40, 50, 60, 70, 80, 100; number of layers: 1-12; profile vs total count loss weight λ: 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000). We observed that the BPNet model for ChIP-seq overall required the same hyper-parameters as for ChIP-nexus. The only hyper-parameter that differed was the increased width (50) of the deconvolutional layer (compared to 25 which was optimal for ChIP-nexus). Similar to the ChIP-nexus control experiment PAtCh-Cap, we used the ChIP-seq input control experiment using a non-specific antibody to control for experimental biases (Methods). We also added data augmentation (genomic intervals jittered uniformly by [-200, 200] bp with random reverse complementation). This is more important when ChIP-seq data are trained on peaks only since the shape of the profiles will be fairly constant, hence a constant model can already fit the data well.

To gain intuition about the prediction quality of BPNet compared to replicate experiments, we investigated the known *Zfp281* and *Lefty1* enhancers as done before for ChIP-nexus data. Since the model evaluation was performed in peak regions, we added data augmentation (genomic intervals jittered uniformly by [-400, 400] bp with random reverse complementation) to make sure the model does not simply predict the average ChIP-seq signal centered at the peak. We observed that the predicted profile shapes significantly de-noise the base-resolution, strand-specific 5'-end coverage ChIP-seq profiles. Indeed, the predicted profiles resemble smoothed versions of the ChIP-seq 5'-end coverage profiles (averaging sliding window of 50bp, Extended Data Fig. 9a).

To evaluate the predictive performance of the ChIP-seq BPNet model, we performed a similar analysis as for ChIP-nexus with the difference that we assessed the quality of profile shape prediction by comparing the similarity of BPNet predictions to the ground truths smoothed ChIP-seq profiles on the test set against the similarity of ChIP-seq smoothed profiles between replicate experiments using the multinomial log-likelihood. We found that BPNet outperformed the smoothed replicate experiments in terms of profile shape prediction on almost all TFs except Nanog where both performed similarly (Extended Data Fig. 9b). Consistent with the ChIP-nexus models, the total count predictions from the BPNet ChIP-seq model did not surpass the concordance between replicate experiments (Extended Data Fig. 9c). As previously discussed, this result is expected since the total counts are likely influenced by factors besides local DNA sequence that we do not model, such as chromatin context and distal interactions with other genomic elements. Altogether, we conclude that BPNet generalizes seamlessly to learn accurate models of TF ChIP-seq profiles.

Next, we investigated the DeepLIFT profile contribution scores inferred from the ChIP-seq BPNet model for all three TFs in the well-known *Oct4* enhancer. The contribution scores were computed in the exact same manner as for the ChIP-nexus model. Consistent with the inferences from the ChIP-nexus model, we found that the contribution scores also precisely highlighted the *Oct4-Sox2* motif in the center and the Nanog motif on the immediate flanks (Extended Data Fig. 9d). Hence, the contribution scores derived from a ChIP-seq BPNet model is able to accurately highlight the expected motifs within a well-known enhancer.

We then investigated the globally predictive motifs learned by the ChIP-seq BPNet model. We used TF-MoDISco with the same hyper-parameters as for the ChIP-nexus model. To allow for unbiased comparison of the motifs obtain from ChIP-seq and ChIP-nexus BPNet models, we retrained additional models on ChIP-nexus and ChIP-seq data using a common set of peak regions that were found by either of the two assays i.e. union of ChIP-nexus and ChIP-seq peaks, for the three TFs (Oct4, Sox2, Nanog). We further restricted model interpretation to a common set of peak regions found by both assays i.e. intersection of ChIP-nexus and ChIP-seq peaks for each TF.

Additionally, to evaluate the benefits of a profile regression model for ChIP-seq, we trained a binary classification model on ChIP-seq data in the same manner as done before for ChIP-nexus data.

We observed that TF-MoDISco applied to all the different types of ChIP-seq BPNet models discovered the majority of the expected motifs. However, the ChIP-seq models trained on binary labels found a few additional motifs that appear to be spurious and lacking clear biological significance (Extended Data Fig. 10a, Supplementary Data 1). These results show that ChIP-seq BPNet profile models perform comparably to ChIP-nexus BPNet profile models in terms of motif discovery with fewer spurious discovered motifs compared to models trained on binary labels.

To evaluate the quality of the CWM motif instances obtained by the four models, we first analyzed whether the helical spacings between *Nanog* motifs were discovered with the ChIP-seq profile models and found this indeed to be the case (Extended Data Fig. 10b). We then used the same approach as in Extended Data Fig. 3b and Extended Data Fig. 4d in which we compared the ranked motif instances against co-localized strong ChIP-nexus footprints for each TF. We observed that ChIP-nexus BPNet profile models recalled a higher fraction of motif instances with strong ChIP-nexus footprints for the *Nanog* motif compared to ChIP-seq BPNet models (Extended Data Fig. 10c). Both models performed similarly well for *Oct4-Sox2* and *Sox2* motifs. Additionally, ChIP-seq BPNet profile models yielded more accurate CWM motif instances than ChIP-seq binary classification models trained on the same data as well PWM based instance calling (Extended Data Fig. 10d). In conclusion, with respect to accuracy of motif instances, the ChIP-nexus BPNet profile models outperform ChIP-seq BPNet profile models, which outperform ChIP-seq binary models. Hence, modeling profiles directly and improved resolution of the ChIP-nexus profiles collectively contribute to improved motif instance identification.

Altogether, these results show that the BPNet framework, which includes BPNet training, inference of DeepLIFT contribution scores, CWM motif discovery with TF-MoDISco, and motif

instance identification via CWM scanning, can be readily applied to ChIP-seq data. These results were obtained with very minor hyper-parameter adjustments while explicitly controlling for assay specific biases. It should be possible to easily adapt and apply the BPNet workflow to any other regulatory profiling assays such as CUT&RUN, ATAC-seq and DNase-seq.

Supplementary Methods

BPNet architecture

BPNet architecture (without bias correction) can be implemented in the Keras framework (v2.2.4) as follows:

```
import keras; import keras.layers as kl; from bpnet.losses import multinomial_nll
tasks = ['Oct4', 'Sox2', 'Nanog', 'Klf4']

# body
input = kl.Input(shape=(1000, 4))
x = kl.Conv1D(64, kernel_size=25,
              padding='same', activation='relu')(input)
for i in range(1, 10):
    conv_x = kl.Conv1D(64, kernel_size=3, padding='same',
                       activation='relu', dilation_rate=2**i)(x)
    x = kl.add([conv_x, x])
bottleneck = x

# heads
outputs = []
for task in tasks:
    # profile shape head
    px = kl.Reshape((-1, 1, 64))(bottleneck)
    px = kl.Conv2DTranspose(2, kernel_size=(25, 1), padding='same')(px)
    outputs.append(kl.Reshape((-1, 2))(px))
    # total counts head
    cx = kl.GlobalAvgPool1D()(bottleneck)
    outputs.append(kl.Dense(2)(cx))

model = keras.models.Model([input], outputs)
model.compile(keras.optimizers.Adam(lr=0.004),
              loss=[multinomial_nll, 'mse'] * len(tasks),
              loss_weights=[1, 10] * len(tasks))
```

Relationship between the Poisson log-likelihood, mean-squared error and multinomial log likelihood

We start by writing down the negative log-likelihood for the Multinomial distribution. Let L be the sequence length, N the total number of events (i.e. total number of read counts in the region) and p_i the probability of obtaining the outcome i (e.g. the read gets aligned to position $_i$). Then, the negative log likelihood can be written as

$$NLL_{Mult}(k_1,\ldots,k_L \,|\, N, \mathbf{p}) = -\log \frac{N!}{k_1!\ldots k_L!} \prod_{i=1}^{L} p_i^{k_i}$$

$$= -\sum_{i=1}^{L} k_i \log p_i + M$$

.

Note that we gathered all the terms independent of $p_i \ \forall i$ into the constant *M*. Let's assume the read counts at each genomic location k_i are distributed according to the Poisson distribution. The Poisson log likelihood for the sequence region of length *L* can be written as

$$\begin{aligned}\sum_{i=1}^{L} NLL_{Poiss}(k_i, \boldsymbol{\mu}) &= -\sum_{i=1}^{L} logP_{Poiss}(k_i \mid \mu_i) \\ &= -\sum_{i=1}^{L} \log e^{-\mu_i} \frac{\mu_i^{k_i}}{k_i!} \\ &= \sum_{i=1}^{L} (\mu_i - k_i \log \mu_i) + P \ .\end{aligned}$$

If we replace μ_i with $N_p p_i$, where N_p is the predicted number of total counts and use $\sum_{i=1}^{L} p_i = 1$, $\sum_{i=1}^{L} k_i = N$, we obtain:

$$\begin{aligned}\sum_{i=1}^{L} NLL_{Poiss}(k_i, \boldsymbol{\mu}) &= \sum_{i=1}^{L} (N_p p_i - k_i \log N_p - k_i \log p_i) + P \\ &= N_p \sum_{i=1}^{L} p_i - \log N_p \sum_{i=1}^{L} k_i - \sum_{i=1}^{L} k_i \log p_i) + P_2 \\ &= N_p - N \log N_p - \sum_{i=1}^{L} k_i \log p_i + P_2 \ .\end{aligned}$$

We observe that the second term equals to the multinomial negative log-likelihood. If we set $N_p = e^{\log N_p}$, $N = e^{\log N}$, and perform a Taylor expansion

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2} f''(x)(x-a)^2 + O((x-a)^3)$$

up to the squared term for variable $\log N_p$ around $\log N$, we obtain:

$$\begin{aligned}N_p - N \log N_p &= e^{\log N_p} - e^{\log N} \log N_p \\ &\approx N(1 - \log N) + \frac{N}{2} (\log N_p - \log N)^2 .\end{aligned}$$

This means that we can approximate the Poisson log-likelihood by a sum of mean-squared errors and the multinomial loss function where the predicted log of total counts $\log N_p$ is close to the true total counts $\log N$:

$$NLL_{Poiss}(\mathbf{k} \mid N_p, \mathbf{p}) \approx NLL_{Mult}(\mathbf{k} \mid N, \mathbf{p}) + \frac{N}{2} MSE(\log N, \log N_p) \ .$$

We approximate the expression further by replacing the N in front of MSE with $\alpha \bar{N}$, where $\bar{N}$ is the average (or median) value of N across the dataset and α is the tuning parameter which allows to up or down-weight the importance of total count prediction:

$$NLL_{Poiss}(\mathbf{k} \mid N_p, \mathbf{p}) \approx NLL_{Mult}(\mathbf{k} \mid N, \mathbf{p}) + \alpha \frac{\bar{N}}{2} MSE(\log N, \log N_p) \ .$$

If $\alpha = 1$, the multinomial loss and the mean squared error loss are balanced according to the Poisson log-likelihood.

Performance evaluation of profile predictions

ChIP-nexus profiles contain TF footprints characterized by local spikes with high read counts surrounding a valley (putative TF binding site) with low read counts. Typical measures of similarity such as Pearson or Spearman correlation are not well suited to these types of profiles. To quantify the ability of the model to accurately localize footprint positions, we use a binary classification formulation to evaluate how well the model can distinguish positions with high read counts from lower read counts within each ChIP-nexus profile in the test set regions. Positions with more than 1.5% of the total number of reads in each 1 kb test set region were labeled as belonging to the positive class. Positions with less than 0.5% of total read counts were labeled as belonging to the negative class. These two thresholds were manually determined by visually inspecting the ChIP-nexus profiles in peak regions from the training chromosomes. The number of negative examples far outnumber the number of positive examples. Hence, we used the area under the Precision-Recall curve (auPRC) to evaluate the performance of the predicted read probability profiles relative to these binary labels. To evaluate the predictive performance at lower resolutions, we applied auPRC on binary labels and the predicted profile probabilities summarized in 2-10 bp contiguous bins as follows: a bin was labeled as positive if there was at least one position in the bin with a positive label. If all the labels in the bin were negative, the bin was labeled as negative. Otherwise, the bin was labeled as ambiguous. For the predicted profile probabilities, the maximum value in the bin was used.

We used profiles sampled from replicate experiments to compute a corresponding upper bound for the above-mentioned profile evaluation for each TF. For each TF, replicate experiments were divided into two groups with approximately equal numbers of sequencing reads. Read count profiles from one group were used as ground truth and the read counts profiles from the other group were treated as a predictor similar to BPNet. The roles of the replicate groups were then swapped, and the final predictive performance was averaged across both scenarios. Random baseline was obtained by using shuffled regions for model predictions.

We have used two further measures that evaluate not just spikes in ChIP-nexus data, but the profile similarity across the entire profile (Extended Data Fig. 1b). The first is the Jensen-Shannon (JS) divergence between the observed and predicted profile probability distributions. We note that JS divergence is very sensitive to sparse coverage in observed profiles. Positions with 0 observed reads end up with a 0 probability value, which is simply an artifact of low coverage. The models predict expected profiles (implicit imputation) and the multinomial probabilities are almost never 0. Hence, the observed profiles need to be smoothed (via pseudocounts or via a moving average window) to obtain stable JS divergence estimates for similarity between replicate experiments. We tested both options and found that using a 50 bp moving average window gave the best results for replicate concordance (which we use an upper bound for the model's performance). Note that we do not smooth the ground truth observed profiles or the model predicted profiles before computing JS divergence. We only use smoothing when comparing replicates to each other i.e. we smooth the profiles of one

replicate and use these as “predictors” of the unsmoothed profiles of the other replicate. This is analogous to comparing the expected predicted profiles (implicit smoothing) from the model to the unsmoothed observed signal.

The second measure is the multinomial negative log-likelihood (M-NLL) already used as the profile loss function which evaluates the likelihood that the observed count profile could have been generated by the multinomial probabilities learned by the model. Unlike the JS divergence, the M-NLL explicitly accounts for observed 0s due to low coverage in a principled manner. Hence, no adjustments need to be made to the observed profiles.

5-fold cross validation for analysis of robustness of BPNet models

To evaluate the reproducibility and robustness of our approach, we trained five BPNet models initialized with different random seeds on sequences from five different chromosome sets or folds. For each fold, we held-out the following validation and test set chromosomes:

	Validation chr	**Test chr**
Fold 1	1,8,9	2,3,4
Fold 2	2,3,4	10,11,12,13
Fold 3	10,11,12,13	14,15,16,17,18
Fold 4	14,15,16,17,18	19,5,6,7
Fold 5	19,5,6,7	1,8,9

Fold 1 uses the exact same validation and test chromosomes as the original BPNet. We used the exact same hyper-parameters as for the original BPNet model. TF-MoDISco and motif instance calling based on CWM scanning were run on the contribution scores for each of the five models. For each fold, the motifs most similar to the original 11 core motifs (Fig. 3c) were determined by using the continuous Jaccard similarity metric for the PFM scaled by information content. Figs. 5c, 6e-h and 6i-k were re-generated for all 5 BPNet models trained on different chromosome folds and their corresponding motif instances.

Model prediction analysis in unmappable regions

The mappability of the mm10 reference genome with k-mers of length 24, 36, 50, and 100 bp as previously generated[33] were downloaded from https://drive.google.com/drive/folders/0B1fks4X_Jjn5NDJjaE9TUmxrR28. We then classified positions from the positive strand in the ChIP-nexus peaks into three groups: unmappable, mappable and ambiguous. For unmappable regions, we considered those that were not uniquely mapped by any k-mer of length up to 100 bp (value=0 in the provided uint8). For mappable positions, we considered those that were uniquely mapped with k-mers of lengths up to including 50 bp (value>0 and value<=50 in the provided uint8). The remaining positions (~1%) were considered ambiguous and were excluded from the analysis. These positions were namely uniquely mappable by k-mers of length between 51 and 100 bp, which may or may not be longer than the used ChIP-nexus reads which were 50 bp or 75 bp long.

Motif discovery using TF-MoDISco

The TF-MoDISco algorithm[34] consists of three stages. In the first stage, the total contribution in sliding windows of length 21 (`sliding_window_size`) is computed, both for contribution scores from the real sequences and for contribution scores on the shuffled sequences. The distribution of sliding window scores on the shuffled sequences is used to define a 'null distribution' against which sliding windows from the real sequences that pass a FDR threshold of 0.01 (`target_seqlet_fdr`) are identified. Sliding windows are expanded on either side by 10 bp (`flank_size`) are selected in such a way that no two sliding windows overlap by more than 50%. The segments underlying these expanded sliding windows are termed 'seqlets', and are provided to the next stage for clustering. A total of 145,748 non-overlapping seqlets were identified. We limited the total number of seqlets to 50,000 for each run of TF-MoDISco in order to always satisfy computational memory constraints (250 GB).

In the second stage, seqlets are clustered into motifs. First, a similarity for each pair of seqlets is computed using the seqlet contribution scores. For a given pair of seqlets, different possible alignments of the seqlets are considered, and for every alignment, the similarity of the contribution scores is calculated using a correlation-like metric called continuous Jaccard similarity[34]. The best similarity across all alignments is then taken to be the similarity of the seqlet pair. The similarities of the seqlets are provided to a clustering algorithm, after transforming the similarities in a way that grants robustness to the fact that different clusters can have different densities. The clusters are found using a Louvain community detection algorithm[35] that automatically determines the number of clusters by optimizing graph modularity.

After the clusters have been identified, seqlets within a cluster are aligned to each other, and the coordinates of the seqlets are expanded to fill out any overhangs in the alignment. This kind of seqlet expansion makes it possible to discover motifs that are longer than the sliding window used for seqlet identification in the first stage. A Position Frequency Matrix (PFM) and a Contribution Weight Matrix (CWM) are computed from the aligned seqlets by averaging the base frequencies and the contribution scores respectively. The seqlet coordinates are then re-centered such that the region of highest contribution falls near the middle of the CWM. Because these seqlet coordinates can be slightly different from the original seqlet coordinates, the second stage computation is run for a second time using the seqlets with the new coordinates, for added robustness.

In the third and final stage, heuristics are applied to postprocess the motifs using the default TF-MoDISco settings for version 0.5.1.1. Clusters appearing to consist of two distinct motifs are split apart, following which clusters with highly similar motifs are iteratively merged. After all merging is complete, any clusters with fewer than 60 seqlets are treated as noise and disbanded, with their seqlets reassigned to larger clusters. Finally, motifs are expanded to the length of 70 bp and then trimmed down to their final lengths by removing flanking positions with an information content (IC) of less than 8% of the information of the base with the maximal information content in the motif. Motifs supported by less than 100 seqlets or with an information content smaller than 4 bits were discarded. The PFM information content is defined as:

$$-\sum_{i,j} p_{i,j} \log_2(p_{i,j} / b_j),$$

where $p_{i,j}$ is the PFM value at position i and base j and b_j is the background base probability[36]. We used the following background base probabilities: A=0.27, C=0.23, G=0.23, T=0.27.

Transposable element analysis

RepeatMasker annotations for mm10 were obtained from http://www.repeatmasker.org/genomes/mm10/RepeatMasker-rm405-db20140131/mm10.fa.out.gz and used to compute the overlap of seqlets with transposable elements (TEs). A seqlet was considered to overlap a TE if it was fully contained within at least one element defined in RepeatMasker annotation. Kimura 2-parameter distance[37] between the seqlet sequence and the consensus sequence of the motif was used to sort the seqlets in Extended Data Fig. 5a by mutation counts. This distance metric was re-implemented in Python and is equivalent to `dist.dna` function from R's APE package with the `model='K80'` parameter (https://www.rdocumentation.org/packages/ape/versions/5.2/topics/dist.dna).

Analysis of strict spacing constraints for motif pairs

We obtained and filtered the 11 representative motif instances as described in the previous section using CWM scanning. We discarded *Sox2* sites overlapping the *Oct4-Sox2* motif and removed palindromic motif pair matches. Motif pairs were considered when spaced center-to-center between 6 bp and 100 bp. Each motif pair was checked for overlap with RepeatMasker-annotated ERVK, ERVL, ERVL-MaLR or ERV1 genomic regions. For each motif pair, histograms were generated comparing the inter-motif distance between each motif pair instance and its ERV overlapping class. The frequency of motif inter-motif distance relative to both the motif pair and the ERV overlapping class was computed for motif pairs that occurred more than 500 times across the genome.

Validation of discovered motifs

Protein structure visualizations on composite motifs

The structure of Sox2 and Oct1 bound to DNA in Fig. 3a was rendered in VMD[38] using secondary structure information from STRIDE[39] and surfaces from SURF[40], based on the NMR structure 1O4X[41]. This Sox2-Oct1-DNA model has been used as a homology model to build the *Oct4-Sox2*-DNA complex[41], and is therefore representative of the structure of that complex, though coordinates for that model have not been made available.

Nanog

TF-MoDISco returned three *Nanog* motifs with sharp and specific Nanog binding profiles. While the footprints are similar across these *Nanog* motifs, there were also some differences (Fig. 3b), raising the possibility that Nanog binds with a partner. Since Sox2 had previously been reported to be a Nanog binding partner characterized by a *Nanog-Sox* heterodimer motif (RMWMAATWNCATTSW)[12], the median ChIP-nexus signal, BPNet predicted signal and profile contribution scores were measured for each TF across the CWM-mapped *Oct4-Sox2*, *Sox2*, *Nanog* and *Klf4* instances, as well as PWM-mapped putative *Nanog-Sox* heterodimer motifs (Supplementary Fig. 8a). *Nanog-Sox* heterodimer motifs were found by scanning the *Nanog-Sox* heterodimer PWM[12] across IDR-optimal peak regions used as inputs for BPNet. Additionally, Pbx is a common partner for homeodomain transcription factors and has a reported binding motif with a TCA core that is similar to the identified *Nanog* motifs. In order

to test whether Pbx shows specific footprints on any of the *Nanog* motifs, we generated Pbx ChIP-nexus data in mouse ESCs. However, Pbx did not show binding on the Nanog motifs (Supplementary Fig. 8b). In contrast, Pbx showed strong footprints on its reported motif (1,000 top-scoring genomic PWM-matches to the PH0134.1 *Pbx* motif from JASPAR[13]), showing that the Pbx ChIP-nexus data are of high quality. We also show Sox2 binding as a control, but found no specific footprints of Sox2 on the *Nanog* motifs (Supplementary Fig. 8c).

Zic3 and Essrb

TF-MoDISco returned three short motifs not matching *Oct4, Sox2, Nanog* or *Klf4* binding motifs. Two of these motifs resembled the *Zic3* and *Esrrb* motifs as reported in the literature. To confirm their identity, we performed Zic3 and Esrrb ChIP-nexus experiments and plotted their average binding across the putative *Zic3* and *Esrrb* motif instances discovered by TF-MoDISco and CWM scanning. In both cases, this confirmed their identity. As a control, we also plotted Esrrb binding across the 1,000 top-scoring genomic PWM-matches to the MA0141.1 *Esrrb* motif from JASPAR[13] and observed the same binding footprint.

TFIIIC B-box and tRNAs

One of the three short motifs did not appear to be a known TF motif important in ESCs. We queried the TRANSFAC database[42] using a motif identifier tool called TOMTOM from the MEME Suite[23]. This revealed a match with sequences associated with TFIIIC subunits. Upon further inspection, this motif was revealed to be the *TFIIIC B-box*, a binding site that contributes to the recruitment of TFIIIC binding[43], thus is associated with Pol III transcription. Consistent with this, we found that the *TFIIIC B-box* motif we discovered frequently mapped across tRNA genes in the mouse genome. The tRNA genes were obtained from the tRNAscan-SE predictions stored in GtRNAdb 2.0, release 17.1[44] (http://gtrnadb.ucsc.edu/GtRNAdb_archives/release17/genomes/eukaryota/Mmusc10/mm10-tRNAs.tar.gz). We then classified the *B-box* motifs based on their gene overlap and computed the copy number of the tRNAs overlapping with the mapped *B-box* motif instances based on amino acid anti-codons, separating methionine (Met) and activated methionine (iMet) as two separate amino acid classes.

Smooth curve fitting in Fig. 5i-k

Software package pygam version 0.7.1 (https://github.com/dswah/pyGAM) was used to fit the generalized additive model (GAM) with 50 B-spline basis functions using the pygam.LinearGAM class to the data. Optimal regularization strength `lamb` was determined using Generalized cross validation (GCV) by searching over the following values: [10, 32, 100, 326, 1000, 3162, 10000, 31622, 100000, 316227, 1000000]. Confidence intervals were computed by pygam.LinearGAM.confidence_intervals, which uses the confidence intervals of the linear fit coefficients to estimate the confidence bands of the fitted curve.

Periodicity analysis using Fourier transform

For each TF-MoDISCo motif of a TF, we identified the locations and sequences containing all the seqlets of the motif. We extracted base-resolution DeepLIFT profile contribution scores (w.r.t. to the TF's profile prediction) across 200 bp sequence windows centered at each of the seqlets. We computed the average contribution score at each base across all the extracted contribution score profiles. We subtracted a smoothed version of the average contribution

score (smoothing window of size 10) in order to correct for overall higher contribution scores in the center. We call this the corrected average contribution score profile of a motif.

For each motif, we computed Discrete Fourier transforms of the corrected average contribution score profile using the Numpy (v1.16.1) function `numpy.fft.rfft`. Power spectrum was obtained by taking the squared absolute value of the returned Fourier coefficients. The finite length of the corrected average contribution score profiles, results in discrete frequency values from the Fourier transform. We used half of the difference between adjacent frequency values as an estimate of the error-bars (uncertainty) around the discrete frequency values value (Fig. 5c).

In-silico motif interaction analysis

We studied the interaction between the following motifs discovered by TF-MoDISco:

- *Oct4-Sox2* (pattern 0 from Oct4, consensus=TTTGCATAACAA),
- *Sox2* (pattern 1 from Sox2, consensus=GAACAATGG),
- *Nanog* (pattern 1 from Nanog, consensus=AGCCATCA),
- *Klf4* (pattern 0 from Klf4, consensus=CCACGCCC).

We considered motif instance pairs (A, B) spaced at some distance d <160 bp and compared BPNet ChIP-nexus profile predictions between 4 cases: where either *Motif A* or *Motif B* was replaced by a random sequence, where both were replaced by a random sequence or where both were left intact. Motif instance pairs were either simulated in synthetic sequences or were detected by CWM scanning in sequences underlying ChIP-nexus peaks.

Synthetic sequences

For synthetic sequences, we first created 128 random background sequences of 1 kb in length by sampling the base at each position with equal probability. Next, we replaced the central bases by the consensus sequence of *Motif A* and similarly inserted *Motif B* d bases downstream of *Motif A* (d is the distance between motif centers). We used BPNet to predict the strand-specific ChIP-nexus profile for the primary TF of *Motif A* (e.g. Oct4 for the *Oct4-Sox2* Motif and Nanog for the *Nanog* motif). We averaged the predictions across the 128 random background sequences to obtain the profile P_{AB}. We repeated the same procedure by i) inserting only the *Motif A* in the center (P_A), ii) inserting only the *Motif B* d-bases downstream of the center, and iii) not inserting any Motif and hence only averaging the predictions across random sequences ($P_\emptyset$). We used the predicted profile P_A to determine the predicted summit (maximum) location within 35 bp of the *Motif A* center for each strand. The strand-specific summit location at *Motif A* was then used to determine the profile height in all 4 scenarios averaged across the two strands. We denote the average predicted profile summit height of the 4 different predicted profiles (P_A, P_B, P_{AB} and $P_\emptyset$) by h_A, h_B, h_{AB}, and $h_\emptyset$ correspondingly.

We define the corrected binding fold change by quantifying the influence of *Motif B* on *Motif A* as: $(h_{AB} - (h_B - h_\emptyset)) / h_A$.

A binding fold-change of 1 indicates that profile summit height of TF A is the same whether or not *Motif B* is present in the vicinity of *Motif A*. If the fold-change is higher than one, then the profile summit of TF A is higher compared to the case where *Motif B* is absent. The second term in the numerator $(h_B - h_\emptyset)$ corrects for predicted signal of TF A found near *Motif B* (“shoulder” effects). For homotypic motif interactions, a shoulder is present because ChIP-

nexus motif footprints have a low decaying signal surrounding the summits (Fig. 3e). For heterotypic motif interactions, where the TF bound to *Motif B* is different from TF A, a shoulder may nevertheless be present if TF A is predicted to show an indirect binding footprint at *Motif B* (e.g. Nanog at the *Sox2* motif, Fig. 3e). By correcting for shoulder effects, we make sure that the measured interaction is not due to the indirect binding footprint at the nearby motif (Supplementary Fig. S10a).

We performed the analysis for all motif pairs, strand orientations, and possible inter-motif distances ranging from 11 bp to 160 bp (Supplementary Fig. S10b).

Genomic sequences

To compute the corrected binding fold-change of motif interactions in genomic sequences, we first obtained motifs instance locations in 1 kb ChIP-nexus peak regions using CWM scanning. We discarded motif instances from duplicated peak regions overlapping other peak regions by more than 200 bp as well as motif instances overlapping TEs (discovered by TF-MoDISco and mapped back to the genome using CWM scanning). Also, *Sox2* motif instances overlapping the *Oct4-Sox2* motif were discarded. For each motif pair, 4 model predictions were made:

- P_{AB}: the reference sequence of the whole interval in which the motifs were present
- P_{A}: motif instance B replaced by random sequence
- P_{B}: motif instance A replaced by random sequence
- $P_{\emptyset}$: motif instances A and B replaced with random sequence

We computed the profile heights at *motif A* profile summit locations in the same manner as for the synthetic sequences yielding 4 profile heights: h_A, h_B, h_{AB} and $h_{\emptyset}$. We added "pseudo counts" defined as the 20th percentile of the considered quantity to the shoulder-corrected profile height of the reference sequence: $h_{AB} - (h_B - h_{\emptyset}) + PC_{AB}$ as well as the profile height of the A-only sequence: $h_A + PC_A$. Next, we kept only the motif pairs where the shoulder-corrected profile height of the motif was in the top 20% for both motifs. This ensured that only motif pairs showing a footprint were used. Finally the corrected binding fold-change was computed for each motif instance pair as:

$$(h_{AB} - (h_B - h_{\emptyset}) + PC_{AB}) / (h_A + PC_A) .$$

We note that there are three main differences between the synthetic and genomic sequences. First, in genomic sequences, the background sequences were not random and may contain other motifs. Second, the "perfect" consensus sequence was used for injecting motifs in synthetic sequences, whereas for genomic sequences the motif instance sequences vary and do not necessarily match the consensus. Third, the distribution of inter-motif distances in genomic sequences is not perfectly uniform as for the synthetic case, hence some inter-motif distances might be under-represented.

Co-occurrence likelihood of motif pairs

We obtained and filtered motif instances as described in the previous section using CWM scanning. We discarded *Sox2* sites overlapping the *Oct4-Sox2* motif. To compute whether *Motif A* is located close to *Motif B* more frequently than expected by chance, we counted i) the number of times a motif instance A is close to motif instance B and ii) the number of times motif instance A is close to motif instance B if we shuffle all motif instances between peaks

while maintaining the relative location within the peak. We constructed the following 2-by-2 contingency matrix c_m:

$$c_m = \begin{pmatrix} \#A \text{ not close to } B \text{ (shuffled)}, & \#A \text{ not close to } B \\ \#A \text{ close to } B \text{ (shuffled)}, & \#A \text{ close to } B \end{pmatrix}$$

and applied the Pearson's Chi-square test (chi2_contingency from scipy.stats) to obtain the *p*-value quantifying whether the odds-ratios (A close vs not close to B) between the observed and shuffled motif instances are significantly different. We note that Pearson's Chi-square test is by definition one-sided. Finally, we use the odds-ratio to visualize whether A is closer to B more frequently than expected by chance:

$$\frac{\#A \text{ close to } B}{\#A \text{ not close to } B} \Big/ \frac{\#A \text{ close to } B \text{ (shuffled)}}{\#A \text{ not close to } B \text{ (shuffled)}}.$$

Benchmarking alternative methods

ChExMix

ChExMix[15,22] is a state-of-the-art motif discovery and TF binding event calling method for ChIP-exo and ChIP-nexus data. ChExMix v0.3 with default parameters was run for each TF on the pooled BAM file containing reads of all the replicates for the corresponding TF. The same blacklisted regions (--exclude) as for peak calling in the ChIP-nexus pipeline were used. The following mm10 background file (--back) was used (http://lugh.bmb.psu.edu/software/chexmix/backgrounds/mouse.back).

HOMER

HOMER v2[27] was run on the 1 kb peak regions for each of the 4 TFs profiled by ChIP-nexus with the findMotifsGenome.pl command with the following command line arguments: `-len 12 -size 200`. These specify the motif length (12) and the size of the considered regions around the peak summits (200 bp). Motif instances in the ChIP-nexus peak regions were determined using the findMotifsGenome.pl with the default arguments.

MEME / FIMO

MEME[23–26] version 5.0.2 was run on sequences extracted from the central 50 bp of the peak summits for the top 500 peaks. Command line arguments as specified by the MEMESuite webtool were used: `-dna -revcomp -mod=anr -nmotifs=3 -minw=6 -maxw=50 -objfun=classic -markov_order=0`. FIMO version 5.0.2 was used to determine the motif instances in the mm10 reference genome. We used the defaults for all parameters except for a less stringent *p*-value threshold of 0.001 (the default is 0.0001).

PWM-ChExMix and PWM-BPNet

Motif instances were also determined for PWMs discovered by ChExMix and BPNet. PWM score was computed using the numpy correlate function used to compute the dot-product score between the PWM and the one-hot-encoded DNA sequence. The following background probabilities were used to convert the PFM to PWM: A=0.27, C=0.23, G=0.23, T=0.27.

BPNet-augm

We used a sequence jittering approach to control for any potential implicit biases in the CWM scanning due to the colocalization of the summits of the ChIP-nexus profiles with the center

of the input sequences used to train BPNet. For each of the original 1 kb peak regions in the test set, the predicted ChIP-nexus profiles were used to record the position of the predicted maxima (summit). Note that we never use the experimentally measured ChIP-nexus profiles. For each original 1 kb sequence, we obtained a jittered version by randomly selecting a position within +/- 200 bp of the original predicted summit and extracting the 1 kb sequence centered at this jittered center. We then used the model to infer base-resolution DeepLIFT profile contribution scores for the jittered sequence. We then scanned these contribution scores of jittered sequences with the CWMs.

Evaluation of validity of motif instances using ChIP-nexus profile height

We developed an approach to evaluate the validity of motif instances obtained from different motif discovery methods (TF-MoDISCo, MEME/FIMO, HOMER, ChExMix) and instance calling methods (CWM vs PWM scanning) in the absence of a ground truth set of motif instances. We expect true bound motif instances to exhibit colocalized ChIP-nexus footprints. Hence, we used the strength of the ChIP-nexus signal in the immediate vicinity of motif instances (as described below) as a surrogate measure of validity of motif instances. For all the methods, we removed all *Sox2* instances overlapping the *Oct4-Sox2* motifs. We also performed separate sets of analyses for motif instances located within +/- 100 bp and +/- 500 bp of the ChIP-nexus peak summits.

Given a set of motif instances of a motif (from each of the methods), we extracted the measured ChIP-nexus profiles centered at each motif instance. We computed an aggregate ChIP-nexus footprint of the motif by averaging the ChIP-nexus profile read counts (5' end positions) at each position across all motif instances. We recorded the distance of the maxima of the aggregate footprint on each of the strands from the center (where the motif instances are located). We call these 'reference summit positions' for the motif (Supplementary Fig. 10a). We then compute the 'ChIP-nexus profile height' of each motif instance as the total number of ChIP-nexus read 5' ends aligning to the reference footprint summit offset positions on both strands around the motif instance. It is important to note that we only compare ChIP-nexus profile height scores for motif instances from different methods within test set sequences. Using only the test set sequences ensures that BPNet derived motif instances are not implicitly using information from the measured ChIP-nexus profiles.

For each motif, a motif instance is considered to be supported by a ChIP-nexus footprint if its 'ChIP-nexus profile height' is greater than a predefined threshold. We selected this threshold to be the 90th percentile of the ChIP-nexus footprint height distribution over the motif instances in the test chromosome called by CWM scanning on BPNet contribution scores. This stringent threshold minimizes cross-talk from ChIP-nexus signal originating from nearby motif instances since the ChIP-nexus footprint flanks have typically 10 times lower values than the summit itself.

Analysis of ATAC-seq after induced depletion of Oct4 and Sox2

ATAC-seq data processing

Friman *et al.*[11] profiled chromatin accessibility via ATAC-seq in mouse ESCs before and after induced depletion of Oct4 and Sox2. We downloaded the corresponding paired-end ATAC-seq FASTQ files for both replicates of the Sox2 26h (ON & OFF) and the Oct4 S2iL (ON & OFF) from GSE134680[11]. We processed the ATAC-seq data using the ENCODE ATAC-seq

pipeline (v1.5.3) at https://github.com/ENCODE-DCC/atac-seq-pipeline. We trimmed adapters from the reads in the FASTQ files using cutadapt (v1.9.1) with parameters `-e 0.1 -m 5` [45]. Next, the trimmed reads were aligned to the mm10 reference genome assembly with Bowtie2 (v2.2.6)[46,47] using the parameters `-X2000 --mm -k 5` (report up to 5 distinct, valid alignments). Mapping stats were computed using SAMtools flagstat (v1.2) [48]. Reads were filtered using SAMtools v1.7 view to remove unmapped reads and mates, non-primary alignments, reads failing platform or vendor quality checks, and PCR or optical duplicates (-F 1804) marked using Picard v1.126 MarkDuplicates. Reads mapping to more than 4 locations were discarded. For the remaining reads, the alignment with the best score is retained. The final filtered BAM file was converted to tagAlign format (BED 3+3) using bedtools `bamtobed` (v2.26)[49].

Peaks were called using MACS2 (v2.1.1.20160309) by extending 5'-ends of reads on each strand using a 73 bp window (±36 bp) and then computing coverage of extended reads across both strands (`shift=-36, extsize=73`). Peak calling was performed on filtered, aligned reads from each replicate using a relaxed *p*-value threshold of 0.01 and retaining the top 500,000 peaks as described[50]. Relaxed peak calls were similarly performed on pseudo-replicates, which were obtained by pooling filtered, aligned reads from all replicates for each sample and randomly splitting the pooled reads into two balanced pseudo-replicates. We identified two types of reproducible peaks. First, 'naive overlap peaks' were defined as relaxed peaks obtained from pooled reads that overlapped relaxed peaks from both true replicates or both pooled-pseudoreplicates. Furthermore, we used the Irreproducible Discovery Rate (IDR) framework to obtain more stringent, rank consistent, reproducible peaks across the true-replicates and pseudo-replicates[51]. The larger of these two sets of IDR peaks (in terms of number of peaks) was defined as the "IDR optimal set" of peaks. Peaks overlapping the blacklisted regions from https://www.encodeproject.org/files/ENCFF547MET/ were excluded.

Models for predicting differential ATAC-seq from derived types of sequence features

The Sox2 26h (ON / OFF) ATAC-seq sample and the Oct4 S2iL (ON / OFF) samples were selected from GSE134680[11] to compute the differential ATAC-seq signal (log fold-change) upon the depletion of Oct4 or Sox2 in each of the 1 kb ChIP-nexus peaks. Let $c_{i,OFF}^{Sox2}$ represent the number of ATAC-seq 5' read ends aligned to the *i*-th ChIP-nexus peak 1 kb region when Sox2 (TF, not motif) has been depleted and $c_{i,ON}^{Sox2}$ when Sox2 has not been depleted. Let pc_{OFF}^{Sox2} represent the 10th percentile of the $c_{i,OFF}^{Sox2}$ distribution across all *i*. We define the differential ATAC-seq log-fold change signal as follows:

$$PC_i(Sox2) = \log_{10}\left(\frac{c_{i,ON}^{Sox2} + pc_{ON}^{Sox2}}{c_{i,OFF}^{Sox2} + pc_{OFF}^{Sox2}}\right).$$

For each of the depletion experiments, we trained separate linear models (LinearRegression from scikit-learn version 0.20.1) to predict the differential ATAC-seq log fold change signal at specific regions based on two different sets of sequence features derived from the corresponding 1 kb input DNA sequences. We restricted to regions corresponding to ChIP-nexus peaks of the four TFs that also overlapped IDR optimal ATAC-seq peaks from the wild-type (unperturbed) condition. We used the same train/test split by chromosomes as for the original BPNet training. Regions from test set chromosomes 1, 8 and 9 were used to evaluate

the performance of the model. Regions from validation/tuning set chromosomes 2, 3 and 4 were not used. Regions from the remaining chromosomes were used to train the models. We evaluated the performance of the models using the Pearson and Spearman correlation metrics.

The first set of sequence features represent the complete sequence representation learned by the BPNet model trained on ChIP-nexus profiles of Oct4, Sox2, Nanog and Klf4. For each 1 kb input sequence, we computed the bottleneck activation features from the original BPNet model averaged across the spatial axis followed by a log transformation yielding 64 features.

The second set of features were derived from motif instances mapped to the 1 kb regions. For each 1 kb input sequence, we counted the number of motif instances of each of the 11 main TF-MoDISCo motifs. For each motif, we also computed the mean, maximum and sum of the motif match scores (PWM log odds) across all motif instances in the input sequence. We thus obtained a total of 44 sequence features. If no motif instances were mapped to a region, we set all the feature values for that motif to 0. We computed these features for motif instances mapped by BPNet and all other methods described in the previous section (i.e. ChExMix, MEME/FIMO, and HOMER).

Overlap with motif instances from different methods

Friman *et al.*[11] classified differentially accessible states via peak calling and differential enrichment analysis between the ON and OFF ATAC-seq experiments for Oct4 and Sox2. This yielded regions annotated as Oct4-dependent (OD), Sox2-dependent (SD), or Co-dependent (CD), which we downloaded from GSE134680. We then collected the *Oct4-Sox2* and *Sox2* motif instance sets mapped by BPNet, MEME/FIMO and HOMER as described above. For each motif set, we removed any *Sox2* motif instance that overlapped with an *Oct4-Sox2* motif instance. Next, we ranked each motif instance in decreasing order. BPNet motif instances were ranked by the weighted contribution scores. MEME/FIMO and HOMER motif instances were ranked by their respective motif match scores. We computed the cumulative overlap fraction across each rank step for OD, SD, and CD regions using the following equation:

$$y(i) = \frac{1}{|D|}\sum_{k=1}^{i}\begin{Bmatrix} 1 & \text{if m[k]} \cap \text{D} \neq \varnothing \\ 0 & \text{otherwise} \end{Bmatrix} \quad \text{for i} = 1,2,\ldots, \text{N}\Bigg\}.$$

Given a ranked set of motif instances *m* of length *N* and a set of differentially accessible regions *D* (OD, SD, or CD), we counted the overlap occurrences, up to rank *i*. We then obtained the overlap rate by dividing by the differentially accessible set length |*D*|. When computing the overlap occurrences, if a region in *D* occurred more than once, only the first overlap was counted. The cumulative overlap fractions of ranked motifs were then compared between BPNet, MEME/FIMO and HOMER.

References

1. Xie, L. *et al.* A dynamic interplay of enhancer elements regulates Klf4 expression in naïve pluripotency. *Genes Dev.* **31,** 1795–1808 (2017).
2. Mistri, T. K. *et al.* Dynamic changes in Sox2 spatio-temporal expression promote the second cell fate decision through Fgf4/Fgfr2 signaling in preimplantation mouse embryos. *Biochem. J.* **475,** 1075–1089 (2018).
3. Tokuzawa, Y. *et al.* Fbx15 is a novel target of Oct3/4 but is dispensable for embryonic stem cell self-renewal and mouse development. *Mol. Cell. Biol.* **23,** 2699–2708 (2003).
4. Zhou, H. Y. *et al.* A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev.* **28,** 2699–2711 (2014).
5. Buecker, C. *et al.* Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell* **14,** 838–853 (2014).
6. Martello, G. *et al.* Esrrb is a pivotal target of the Gsk3/Tcf3 axis regulating embryonic stem cell self-renewal. *Cell Stem Cell* **11,** 491–504 (2012).
7. Hnisz, D. *et al.* Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol. Cell* **58,** 362–370 (2015).
8. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133,** 1106–1117 (2008).
9. Moorthy, S. D. *et al.* Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res.* **27,** 246–258 (2017).
10. Blinka, S., Reimer, M. H., Pulakanti, K. & Rao, S. Super-Enhancers at the Nanog Locus Differentially Regulate Neighboring Pluripotency-Associated Genes. *Cell Rep.* **17,** 19–28 (2016).
11. Friman, E. T. *et al.* Dynamic regulation of chromatin accessibility by pluripotency transcription factors across the cell cycle. *elife* **8,** (2019).
12. Gagliardi, A. *et al.* A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J.* **32,** 2231–2247 (2013).
13. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48,** D87–D92 (2020).
14. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489,** 91–100 (2012).
15. Guo, Y., Mahony, S. & Gifford, D. K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.* **8,** e1002638 (2012).
16. Suryamohan, K. & Halfon, M. S. Identifying transcriptional cis-regulatory modules in animal genomes. *Wiley Interdiscip. Rev. Dev. Biol.* **4,** 59–84 (2015).
17. King, D. M. *et al.* Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. *elife* **9,** (2020).
18. Fiore, C. & Cohen, B. A. Interactions between pluripotency factors specify cis-regulation in embryonic stem cells. *Genome Res.* **26,** 778–786 (2016).
19. Weingarten-Gabbay, S. *et al.* Systematic interrogation of human promoters. *Genome Res.* **29,** 171–183 (2019).
20. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30,** 521–530 (2012).
21. de Boer, C. G. *et al.* Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* **38,** 56–65 (2020).
22. Yamada, N., Lai, W. K. M., Farrell, N., Pugh, B. F. & Mahony, S. Characterizing protein-DNA binding event subtypes in ChIP-exo data. *Bioinformatics* **35,** 903–913 (2019).
23. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** W202-8 (2009).
24. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.* **296,** 1205–1214 (2000).
25. Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* **32,** W199-203 (2004).
26. Thijs, G. *et al.* A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17,** 1113–1122 (2001).
27. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).

28. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33,** 831–838 (2015).
29. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12,** 931–934 (2015).
30. Avsec, Ž., Barekatain, M., Cheng, J. & Gagneur, J. Modeling positional effects of regulatory sequences with spline transformations increases prediction accuracy of deep neural networks. *Bioinformatics* **34,** 1261–1269 (2018).
31. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26,** 990–999 (2016).
32. Wnuk, K., Sudol, J., Rabizadeh, S., Szeto, C. & Vaske, C. Predicting DNA accessibility in the pan-cancer tumor genome using RNA-seq, WGS, and deep learning. *BioRxiv* (2017). doi:10.1101/229385
33. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.* **46,** e120 (2018).
34. Shrikumar, A. *et al.* TF-MoDISco v0.4.2.2-alpha: Technical Note. *arXiv* (2018).
35. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. The Louvain method for community detection in large networks. *J of Statistical Mechanics: Theory and Experiment* **10,** P10008 (2011).
36. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188,** 415–431 (1986).
37. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16,** 111–120 (1980).
38. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14,** 33–8, 27 (1996).
39. Heinig, M. & Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **32,** W500-2 (2004).
40. Varshney, A., Brooks, F. P. & Wright, W. V. Computing smooth molecular surfaces. *IEEE Comput. Graph. Appl.* **14,** 19–25 (1994).
41. Williams, D. C., Cai, M. & Clore, G. M. Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1.Sox2.Hoxb1-DNA ternary transcription factor complex. *J. Biol. Chem.* **279,** 1449–1457 (2004).
42. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34,** D108-10 (2006).
43. Moqtaderi, Z. *et al.* Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. *Nat. Struct. Mol. Biol.* **17,** 635–640 (2010).
44. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44,** D184-9 (2016).
45. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17,** 10 (2011).
46. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
47. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
48. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
49. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
50. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22,** 1813–1831 (2012).
51. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5,** 1752–1779 (2011).