Dirichlet variational autoencoders for *de novo* motif discovery from accessible chromatin

Meghana Kshirsagar*
Microsoft, AI for Good Research Lab, Redmond, WA, USA.

Han Yuan
Calico Life Sciences, South San Francisco, CA, USA

Juan Lavista Ferres
Microsoft, AI for Good Research Lab, Redmond, WA, USA.

Christina Leslie*
Memorial Sloan Kettering Cancer Center, New York, NY, USA

November 29, 2021

Abstract

Determining the cell type-specific and genome-wide binding locations of transcription factors (TFs) is an important step towards decoding gene regulatory programs. Profiling by the assay for transposase-accessible chromatin using sequencing (ATAC-seq) reveals open chromatin sites that are potential binding sites for TFs but does not identify which TFs occupy a given site. We present a novel unsupervised deep learning approach called BindVAE, based on Dirichlet variational autoencoders, for jointly decoding multiple TF binding signals from open chromatin regions. Our approach automatically learns distinct groups of k-mer patterns that correspond to cell type-specific *in vivo* binding signals. Latent factors found by BindVAE generally map to TFs that are expressed in the input cell type. BindVAE finds different TF binding sites in different cell types and can learn composite patterns for TFs involved in co-operative binding. BindVAE therefore provides a novel unsupervised approach to deconvolve the complex TF binding signals in chromatin accessible sites.

*To whom correspondence should be addressed

Introduction

The advent of the assay for transposase-accessible chromatin using sequencing (ATAC-seq)[1] and, more recently, its single-cell counterpart, scATAC-seq[2], have brought about the current ubiquity of chromatin accessibility data across numerous human and mouse cell types, tissue samples, and disease states. Chromatin accessibility maps diverse genomic elements, including regulatory elements such as gene promoters and intronic and intergenic enhancers that are occupied by transcription factors (TFs) as well as structural elements such as CTCF and cohesin binding sites that may anchor 3D chromatin loops. The DNA sequence signals underlying regions of open chromatin are therefore complex: while a single assay allows us to create an atlas of tens of thousands of accessible "peaks" in a given cell type, we expect that dozens of TFs occupy overlapping subsets of these peaks due to the presence of their cognate binding sites or those of co-factors. A key problem in regulatory genomics is interpreting the regulatory information encoded in chromatin accessible peaks, namely TF binding sites and "regulatory grammars" of TFs that bind at different locations within the same peak. A longer-term and challenging goal is to decode how genetic variants, both germline and somatic, can alter this regulatory information by leading to loss or change of accessibility and disruption of gene regulation.

Traditional methods for identifying TF binding sites in chromatin accessible regions involve performing searches and enrichment analyses with a library of known TF motifs, each encoded as a position-specific weight matrix (PWMs). These standard approaches are useful in finding strong signals but are confounded by the problem of redundant or missing motifs, the near-identity of motifs for closely related factors, and the inherent limitation of using weight matrices to define binding sites when more subtle binding sequence signals may be present. De novo motif discovery can be underpowered when the sequence signal is complex; for example, if an important TF binds a small fraction of accessible sites, enrichment-based motif discovery may fail to identify the corresponding binding motif.

To address the limitations of PWMs, a range of supervised machine learning methods using k-mer representation have been used to train sequence models to predict or decipher chromatin accessibility. The first such methods were k-mer based SVMs[3,4], which accurately discriminate between accessible sites and negative (flanking or random genomic) sequences but are more difficult to interpret in terms of constituent TF signals; feature attribution methods have recently been introduced to extract explanatory sequence patterns from gapped k-mer SVM models[5]. Other approaches include SeqGL[6], which trains a group lasso logistic regression model on ATAC-seq data, where k-mer groups correspond to TF binding patterns; BindSpace[7], a latent semantic embedding method for TF SELEX-seq data that enables multi-class identification of the TF signals in genomic sequences; and a topic model based approach on discovering combinatorial binding of TFs from ChIP-seq data[8].

In parallel work, a range of deep learning models have been applied to chromatin accessibility and other epigenomic data sets. Popular methods use a one-hot encod-

ing of DNA sequence and train convolutional neural networks (CNNs)[9–12] to predict epigenomic signals. While these methods have made impressive strides, there is still an interpretability issue, especially for chromatin accessibility data, which contains numerous binding patterns for distinct motifs, as opposed to TF binding data (e.g. ChIP-seq, ChIP-nexus, CUT&RUN), where one might hope to identify a smaller number of binding patterns for the targeted TF as well as its co-factors. Even in this latter setting, a complex process of extracting sequence patterns through feature attribution and aggregating them into motifs may be required for interpretation[13].

In this work, we develop a deep learning approach based on Dirichlet variational autoencoders (VAE) for modeling chromatin accessibility data, using a k-mer representation of genomic sequences as input (**Figure 1a**). VAEs are a family of machine learning models that learn probability distributions with latent variables. Similar to autoencoders, they learn representations of the input data by compressing the input via a 'bottleneck' layer in the neural network. A VAE achieves this compression in a *probabilistic* manner, whereby the encoder transforms the input x into *parameters*, describing a probability distribution, which it then samples from, to get the latent representation z. The decoder then reconstructs the input from the latent representation z, with the goal of making the output x' as close as possible to the input x, by minimizing the reconstruction error. In the Dirichlet VAE or 'topic model' setting, we assume that the input bag of k-mers from a peak are generated by multiple 'topics'[1], which can correspond to binding signatures of TFs as well as other sequence signals in the data. The VAE formulation of latent variable models uses advances in neural network learning and enables efficient training on large data sets using backpropagation of gradients.

In this work, (a) we show that the Dirichlet VAE model captures a useful representation of chromatin accessible elements, where the k-mer distributions encoded in the latent space can often be interpreted as binding patterns for TFs; (b) we present an algorithm to interpret the latent space that uses HT-SELEX probes (c) we show that our model learns co-operative binding signals that involve multiple TFs; and (d) we find that our model learns different TFs for distinct cell types in our experiments with GM12878 and A549. Finally, most of the deep learning approaches for learning from DNA sequences have been supervised prediction based models. Unsupervised deep learning methods have been under-explored in the literature and our work paves the way for further methods development in this direction.

[1] we refer to individual components of the latent space as 'latent dimension', 'topic' or 'latent factor'

Results

BindVAE: a Dirichlet variational autoencoder to deconvolve sequence signals

Each input example to BindVAE is the bag of DNA k-mers in one chromatin accessible region as shown in **Figure 1a**. We describe our k-mer representation in detail in **Methods**. The generative model underlying the VAE is based on the observation that each peak is a combination of DNA sequence patterns from the following categories: (a) binding sites for one or more TFs; (b) low complexity regions; (c) genomic background; and (d) cleavage bias from the enzyme used in DNA fragmentation (or tagmentation). We thus surmise that the representation z, learned for each peak, should have latent dimensions (topics) that correspond to these categories and assume we that the membership of the peak in these categories, follows a Dirichlet distribution. That is, $z_i \sim Dirichlet(\alpha)$. Each category (topic) in turn is represented as a multinomial distribution over k-mers. In **Figure 1a**, the latent dimension topic-1 (red color), parameterized by θ_1, contributes k-mers that capture the binding preferences of TF_1, while the blue-colored topic-2 might contribute k-mers representative of the genomic background. The dimension of our latent space or the width of the bottle-neck layer, which we call 'M', is 100. In our analysis, we mainly consider the 100-dimensional latent representations of inputs and the learned decoder parameters θ, which guide the reconstruction of the input from the latent vector z. Details on VAEs, Dirichlet VAEs, and the training approach we use, hyper-parameter tuning etc. are provided in **Methods**.

In the following sections, we present qualitative and quantitative analysis of the VAE models learned on ATAC-seq peaks from: GM12878, which is a cell line derived from human B cells and A549, a basal epithelial cell line. We show the sequence motifs learned for TFs by summarizing the various dimensions of the latent space; we project DNA sequences from other sources such as *in vitro* assays and ChIP-seq in the latent space to interpret the latent space; we locate co-operative binding signals, and correlate learned TFs to expression data.

BindVAE learns diverse k-mer binding patterns

To summarize the DNA sequence patterns captured by BindVAE, we visualize the weights of the top 20 k-mers for each latent dimension. These weights are obtained from the decoder parameters of BindVAE: $\theta \in \mathbb{R}^{M \times D}$, where M is the size of the latent representation and D is the number of k-mers in our input. **Figure 1b** shows this visualization where, the x-axis has unique k-mers, a total of 1068 over the 100 dimensions, and the y-axis shows the latent dimensions $i \in [0, 100]$. Each entry in the shown matrix indicates if the k-mer j appears in the top 20 for dimension/topic m, and the value shown is θ_{ij}. Based on the diagonal-heavy structure of the matrix, we can say that the model learns diverse patterns. The off-diagonal blocks show that similar

patterns are learned for several dimensions and correspond to TFs from the same family. Overall, we see two types of latent factors: ones that capture unique patterns such as dimension #0 and #72 and ones that capture highly redundant patterns. The latent factors of the former type can be thought of as forming a basis, with each axis roughly corresponding to a different binding pattern[2].

Motifs discovered *de novo* by our model

While Figure 1b shows that diverse patterns are learned, we asked if these patterns represent coherent TF binding patterns. To answer this, we summarized the patterns learned by the model by doing a motif analysis for each latent dimension. Ideally, we would like to obtain motifs directly from the parameters of our model; i.e. using the distribution $\vec{\theta_m}$ for a latent factor m. However, the use of relatively short k-mers (8-mers) to represent the input, and the use of wildcards, limit the length and the accuracy of PWMs directly obtained from $\vec{\theta_m}$.

Given the caveats above, we adopted a post-processing procedure that can use higher-order k-mers to get more accurate binding patterns. First, we generate all possible 10-mer DNA sequences (total of 524,800 10-mers) and score them by running inference using the model trained on the ATAC-seq peaks. For a given latent dimension, we obtain the top-200 10-mer sequences that scored the highest. We then construct a PWM from these 200 sequences using MEME[14] and render it using SeqLogo. These results are tabulated in **Figure 1c** for GM12878, with the third column showing the learned motifs. The second column shows the name of the TF assigned to the latent dimension by our procedure described in Algorithm 1 (discussed in the next section) and the last column shows the corresponding CIS-BP[15] motif. For brevity, we show the top few motifs in **Figure 1c**, that were selected based on the p-values assigned by **Algorithm 1**.

We observe that the motifs learned by BindVAE are similar to the motifs from CIS-BP, which are based on *in vitro* and *in vivo* studies of individual TF binding. Since our input representation relies on 8-mers, the model is biased towards learning shorter motifs more accurately. We also find that TFs with long motifs are split across multiple dimensions, for instance `RUNX3` is split across dimensions #60 and #39, and `CTCF` is split across dimensions #11 and #4. Overall, our results illustrate that the VAE-based model learns binding preferences of representative TFs, *de novo*.

Mapping the TFs to dimensions

Based on our observations above and from **Figure 1c**, we map the individual latent dimensions $i \in [1 \dots M]$ from the bottleneck layer of BindVAE to TFs, to facilitate further analysis. We use the procedure outlined in **Algorithm 1** (please refer to

[2]For an illustration of the specific top k-mers, Figure 2c,d show the 8-mers from θ_{72} and θ_{79}

Methods) to do this, where we compare the latent representations of one TF's HT-SELEX probes to that of probes from all other TFs. A TF t, is mapped to a dimension m, generating the mapping $m \rightarrow t$, if m ranks t's probes higher than probes from other TFs.

We find that our algorithm produces many-to-many mappings between TFs and dimensions. Homologous proteins or multiple members of a subfamily are mapped to the same dimension due to the similarity of their binding preferences. For instance, in the GM12878 model, the T-box family of TFs are assigned to the same latent dimension. Some TFs might be assigned to several dimensions for two reasons: (1) they have a long motif or (2) some homolog of that TF appears in the data but does not exist in our set of 296 TFs, which is limited by the HT-SELEX experiments. Another consequence of this limitation is that some latent dimensions are not assigned to any TF even if there is an enrichment in the DNA sequence pattern that they capture. We analyze some of these dimensions and find that they represent genomic background.

Mapping peaks to TFs

We next 'assign' each ATAC-seq peak in the input data to a single TF for downstream analysis and visualization. Given the M-dimensional latent representation z_i of the ith peak, and the mapping $\mathbf{F}$ from dimensions to TFs, the assigned TF is given by: $\mathbf{F}(\arg\max_{d \in [1 \cdots M]} z_{id})$, i.e., the TF corresponding to the latent dimension that has the highest value. In general, we find that each peak's representation is spread across multiple latent dimensions. For example, a 30bp region of a peak from our GM12878 dataset in **Figure 2a** contains TFs from two different families.

Projecting HT-SELEX probes into the latent space

From the previous sections we can conclude that BindVAE learns diverse and coherent patterns. In the following sections we explore what this entails for input DNA sequences of various lengths and types.

We project HT-SELEX probes into the latent space by doing inference on the probes using our model learned on GM12878 ATAC-seq peaks and visualize the results. Since HT-SELEX experiments use short probes that are 20bp long, and capture *in vitro* binding affinities, they have very distinct patterns across different TFs. The heatmap in **Figure 2b** shows the latent space for 10,000 probes, one probe per row, with the rows ordered by the HT-SELEX experiment, i.e. the TF that each probe was enriched for. There are 200 enriched probes selected per TF experiment. We show 42 dimensions, out of the 100-dimensional latent space ($M = 100$), chosen based on whether a dimension contains any signal over the 10,000 probes. We see that a block diagonal structure emerges in this matrix, which indicates that several latent dimensions are orthogonal to each other and show distinct binding patterns. Note that each red square block corresponds to 200 probes. Further, TFs from a family share the same latent space, due to the similarity in their DNA binding sites. For instance

dimension #67 shows signal for several forkhead box (FOX) proteins: FOXJ2, FOXL1, FOXO4. Similarly, dimension #30 contains T-box proteins' DNA binding preferences for the following TFs: TBR1, TBX19, TBX20, TBX21 and EOMES. In addition to T-box proteins, we see that dimension #30 also captures binding preferences for ZSCAN4, which is a C2H2 zinc finger. Dimension #21 encodes bHLH (basic helix-loop-helix) binding preferences for TCF3 and TCF4.

Top 8-mers learned for NFKB1 and HNF1B

We illustrate the quality of the learned posterior distribution p_θ, which is defined by the encoder parameters θ, by showing the top 8-mers for two TFs in **Figure 2c,d**. We sort the decoder parameters θ_m for each latent dimension m, and show the 8-mers corresponding to the top 50 components. The box on the top shows the 8-mers from dimension $m = 72$ which is assigned to NFKB1 by our mapping procedure in **Algorithm 1**, and the bottom plot shows $m = 79$ and the mapped TF HNF1B. For ease of interpretation, we align the fifty 8-mers using Clustal Omega[16] and also show the motif corresponding to the assigned TF. Both NFKB1 and HNF1B have longer motifs of size ∼13 and HNF1B has a dimer motif. We see that the top 8-mers learned by our model can be partitioned into two groups based on their distinct patterns, with one group matching the beginning of the motif (pattern GGGGA for NFKB1) and the other matching the end (TTCCCC for NFKB1).

Cooperative binding signals in GM12878

We find that multiple latent dimensions show k-mer patterns from diverse TF families, possibly indicating the presence of co-operative binding sites in the ATAC-seq peaks. Since the co-occurrence of multiple binding patterns might merely suggest that binding sites of two different TFs are present in a peak and not necessarily cooperative binding, we use CAP-SELEX data[17] to validate this hypothesis. Jolma et al.[17] developed CAP-SELEX, an *in vitro* assay for studying interactions between pairs of DNA-bound TFs, using DNA sequences of length 40bp.

We analyze two latent dimensions: #60 in **Figure 3a** and #67 in **Figure 3b**, which show co-operative binding signals for MYBL1-MAX and FOXJ3-TBX21 respectively. In order to show the presence of co-operative binding patterns, we compare the scores attained upon inference on enriched CAP-SELEX probes from *pairwise* TF experiments to those obtained by enriched HT-SELEX probes from *individual* TF experiments. We show the distribution of these scores (along the y-axis), and the x-axis shows the source TF (or TF pair) experiment for each distribution. We also show the distribution of scores for probes from all other TFs, which shows that the pairwise signals are significantly higher than average. This is also indicated by the p-value (shown in brackets below each TF label) assigned by our procedure in **Algorithm 1**. For example, the p-value of mapping dimension #67 to the TF pair FOXJ3-TBX21 is $2e^{-62}$, while that for 'other TFs' is 1.0.

We further analyze the motif captured by dimension #67 for the TF pair: `FOXJ3-TBX21` in **Figure 3c**, by constructing a PWM from 15-mers ranked high by this dimension. Since the number of all possible 15-mers is prohibitively large, we sample 10% of them and run inference on these. We show the motif obtained by this process on the right side of **Figure 3c**. On the left, we show the motif constructed from the PWM published by Jolma et al.[17] for `FOXJ3-TBX21` co-operative binding.

Accessibility patterns predicted for GM12878 and A549

In **Figure 3d** we show the proportion of binding sites that we predict across various TFs in the two cell types in our study. Each bar represents one TF (or latent dimension or topic m) and the height of the bar, which we call the '*accessibility score*' is obtained by summing the latent representations (or topic proportions) z_{im} for all peaks i: $\sum_{i \in \text{all peaks}} z_{im}$ for each topic m. We only plot the latent dimensions m that were successfully mapped to TFs using the procedure described in Algorithm 1. The bars are colored by cell type, blue for GM12878 and orange for A549, overlapping areas appearing in grey. The bars are sorted in increasing order with respect to accessibility scores of TFs from GM12878. We see that for different cell types, the model learns different accessibility patterns.

Comparison with HOMER

HOMER[18] is an unsupervised motif discovery algorithm that uses differential enrichment to find motifs. It compares peaks to background DNA sequences, and tries to identify patterns that are specifically enriched in the peaks relative to the background. We compare our unsupervised VAE-based method to HOMER on *de novo* motif detection in GM12878. **Figure 3e** shows the overlap in the TFs found by both approaches and how many of these are expressed in GM12878. Since the TFs detected by BindVAE are restricted by our post-processing algorithm that relies on HT-SELEX data, we present the HOMER results in a similar configuration for a fair comparison: 'selex restricted', where we use the HT-SELEX PWMs published by Jolma et al.[19] for the 296 TFs that we consider.

We find that HOMER finds a total of 99 motifs that are mapped to known TFs within our HT-SELEX set of 296 TFs. BindVAE finds 122 motifs from latent dimensions mapped by **Algorithm 1**. Looking at expressed TFs, 60/99 ($\approx 60\%$) found by HOMER and 73/122 ($\approx 60\%$) found by BindVAE are expressed. HOMER finds fewer TFs overall, than those found by BindVAE many of which are expressed in GM12878. Analogously, 74 out of the 104 learned TFs ($\approx 71.1\%$) are expressed in A549. We provide a table of the learned and expressed TFs in the **Supplementary Table 2**.

Projecting ChIP-seq peaks into the latent space

In **Figure 4a,b**, we show a two-dimensional projection using Uniform Manifold Approximation and Projection (UMAP)[20] of the learned representations for ChIP-seq peaks from GM12878, for three transcription factors with distinct binding preferences. There are 1000 peaks from each ChIP-seq experiment, and each peak is colored by the TF. **Figure 4a** shows proteins from different families: `CTCF`, a C2H2 zinc finger, `MAFK` a bZIP TF, and `ELK1` that contains an ETS domain. Here, we see that our model learns distinct embeddings for TFs that have different binding preferences, however embeddings of CTCF peaks are more distributed due to the ability of its zinc finger domains to bind to heterogenous DNA sequences. **Figure 4b** shows `FOS` a bZIP TF, `IRF4` an IRF family transcription factor and `SPI1` that contains an ETS domain. We find that embeddings of peaks from `FOS` are distinctly clustered while those from `IRF4` and `SPI1` largely overlap, possibly because `IRF4` binds DNA weakly but cooperative binding with factors such as SPIB in B cells increases binding affinity[21]. We notice a similar pattern of overlap with `BATF` and `JUND` which are both bZIP TFs that form heterodimers while binding to the DNA[22].

To analyze whether our model learns meaningful representations, we use ChIP-seq as a source of ground truth and verify whether known binding sites for a given TF are transformed to the same latent dimension by our model. We look at the intersection of the TFs in our HT-SELEX set (296 TFs) and those for which we have reliable ChIP-seq data, which gives us 12 TFs. For each of these TFs, we find all ATAC-seq peaks that have an overlap of at least 50bp with any ChIP-seq peak and plot the latent dimension that was assigned to that TF. In **Table 1** we show the number of overlapping peaks between the ChIP-seq experiment and our GM12878 dataset. We show the resultant matrix for these 12 TFs by depicting it as a heatmap in **Figure 4c**. The heatmap shows a centered log ratio (CLR) transform of the latent representations, with rows representing peaks and columns representing TFs. Our approach gives us a total of $\approx$38,496 peaks across the 12 TFs which we sort by their membership, with peaks that belong to `RUNX3` being shown at the top of the heatmap as it has the largest number of mapped peaks.

Discussion

Supervised deep learning methods for the prediction of TF occupancy data and and chromatin accessibility are numerous, ranging from the early deep convolutional neural network based pipelines such as DeepSEA[10] and Basset[9] to more recent approaches usually mirroring advances in deep learning methods for natural language processing, such as the LSTM-based DanQ[23], Basenji using dilated CNNs[11], DeepSite[24], and DNA-BERT[25]. While these models have produced highly accurate predictions of TF occupancy, the interpretation of models requires detailed feature attribution (e.g. Shrikumar et al.[13]) over a large input window (such as 500bp to 1Mbp), and in general

TF-name	Peaks overlapping with ATAC-seq data	Peaks from column-1 that are mapped to TF by Bind-VAE	Precision	Recall
TCF3	9703	5107	0.17	0.53
MAFK	782	436	0.15	0.56
ELK1	4484	3773	0.15	0.84
NRF1	4322	3874	0.21	0.90
RFX5	3226	1506	0.11	0.47
NR2C2	410	158	0.15	0.39
TBX21	14820	5865	0.25	0.40
POU2F2	14972	4593	0.16	0.31
SPI1	16177	10496	0.33	0.65
RUNX3	33873	25403	0.51	0.75
MAX	9110	6793	0.21	0.75
ETS1	7285	5373	0.28	0.74

Table 1: Statistics of chip-seq data used to plot **Figure 4c** showing the number of overlapping peaks.

these methods do not generalize across cell types. Interpretation of supervised deep sequence models has been more successful where models are trained on high-resolution TF occupancy data[12], since the underlying motif grammar is less complex than that for chromatin accessibility data. Motif-matching methods such as FIMO[26] have been popular in biological studies due to the wide range of TFs, ease of application, and inherent unsupervised nature where the available PWM can be applied on any DNA sequence. However, these approaches may return hundreds of motif hits for genomic regions of the size of peaks, i.e. 100bp to 200bp.

Given the limitations of supervised models for interpreting chromatin accessibility, we explored an unsupervised deep learning model that learns binding patterns given open chromatin regions derived from ATAC-seq and is thus complementary to existing work. There are currently ≈1500 DNase-seq and ATAC-seq datasets spanning hundreds of cell and tissue types on the ENCODE portal, whereas datasets on TF binding experiments such as ChIP-seq are restricted to a few TFs per cell type, with the exception of a few highly profiled ENCODE cell lines. Hence, unsupervised or semi-supervised approaches may be desirable for decoding the TF binding landscape on less studied cell types. Further, unlike recent language-based deep learning models, our VAE based model is easily scalable for training on joint data from several cell types and can also be run in a multi-processor computational environment without GPU support, albeit at lower efficiency.

The input to BindVAE uses 8-mers with wildcards, which allows us to interpret the learned latent factors. There is a rich literature on the use of k-mers for representing DNA and protein sequences, from early works using oligonucleotide frequencies[27] and k-mer based string kernels using support vector machines[4,28]. We rely on the results from these prior works that show the robustness of this representation.

We found that our VAE based model can learn distinct binding patterns from

ATAC-seq peaks without any TF labels. Of the 102 distinct patterns learned over the latent dimensions, we found specific patterns for some TFs and were able to map the latent factors to unique TFs. In contrast, for others, we found a coarser pattern that corresponds to one of several TFs from a family, such as `T-box` proteins. Paralogous TFs are difficult to learn as separate factors due to the highly similar patterns in their binding sites, which cannot be captured uniquely by a distribution over 8-mers with wildcards in an unsupervised fashion. Our model also learns combinations of patterns for TFs that co-occur within peaks and that are involved in co-operative binding, and analyzing these patterns produced composite motifs. Using higher-order k-mers in our model can improve the coverage over longer motifs. However, this would increase the input dimension of BindVAE significantly, thereby increase computation overhead substantially.

To conclude, in this work we show that Dirichlet VAE based unsupervised learning models can be used to learn meaningful patterns and representations for DNA sequences. Such models can be used in downstream applications either by fine-tuning by adding a layer of supervision in the VAE or, by simply using the embeddings generated for the input DNA sequence. Given the recent advances in single-cell sequencing and the generation of scATAC-seq datasets, a natural question to study is whether our model can be used for clustering single-cell data. Our results show that the BindVAE learns different patterns for distinct cell lines such as GM12878 and A549. However, when training a single model on data from several similar cell types, the differences in the DNA sequence patterns of TF binding sites might not be sufficient to discriminate between similar cell types. Methods that are based on genomic locations of the accessible sites, such as cisTopic[29] have shown promise, and one extension of our work is combining genomic location with k-mer composition as inputs to an unsupervised model.

Methods

Variational autoencoders

Variational autoencoders (VAEs)[30,31] are latent variable models that combine ideas from approximate Bayesian inference (variational inference) and deep neural networks, resulting in a framework that can use backpropagation-based training.

Let $\mathbf{x}$ represent the data and $\mathbf{z}$ be the latent variable. VAEs express the joint distribution $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ where $p(\mathbf{z})$ is a prior distribution over $\mathbf{z}$, i.e $\mathbf{z} \sim p(\mathbf{z})$, and $p_\theta(\mathbf{x}|\mathbf{z})$ is the likelihood function. In the context of neural networks, $p_\theta(\mathbf{x}|\mathbf{z})$ is the probabilistic decoder that generates data $\mathbf{x}$ given latent variables $\mathbf{z}$, with the goal of reproducing $\hat{\mathbf{x}}$ that is close to $\mathbf{x}$. Since estimating the true posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ is often intractable, an approximate posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ is used, which is formulated by the probabilistic encoder in the neural network model. The encoder outputs $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}|\eta)$ where $\eta = MLP(\mathbf{x})$ is computed from the observation

$\mathbf{x}$ by a multi-layer perceptron (MLP). **Figure 1a** shows a neural network depiction of this model.

VAEs optimize the parameters ϕ and θ of the encoder and decoder jointly by maximizing the evidence lower bound (ELBO) using stochastic gradient descent. The ELBO is the variational lower bound on the marginal log-likelihood of the model $\log p_\theta(\mathbf{x})$ and is given by:

$$\begin{aligned}\log p_\theta(\mathbf{x}) &= \sum_{i=1}^{N} \log p_\theta(x_i) \\ &\geq \sum_{i=1}^{N} \mathbb{E}_{q_\phi(z_i|x_i)}[\log p_\theta(x_i|z_i)] - KL(q_\phi(z_i|x_i)\|p(z_i)) \end{aligned} \tag{1}$$

Since our goal is to incorporate k-mer distributions in the model, we use a Dirichlet distribution as a prior on the latent variables instead of the more prevalent Normal distribution used in Gaussian VAEs. Further, Gaussian VAEs also suffer from the posterior collapse issue and are difficult to interpret, as the bottleneck layer z, can take arbitrary values. On the other hand, a Dirichlet distribution will only allow non-negative latent variables; thereby the value taken by each latent dimension m given a particular $\mathbf{x}$ can be considered a 'membership', with larger values indicating stronger membership.

However, VAEs with a Dirichlet prior cannot be trained using the explicit reparameterization trick[30], which allows a Gaussian variable $\mathbf{z} \sim N(\mu, \sigma^2)$ to be reparameterized as $\mathbf{z} = \mu + \epsilon\sigma$ where $\epsilon \sim N(0, 1)$, thus allowing the gradient to be back-propagated through the latent variable z. This is because no such simple variable transformation is possible for the Dirichlet distribution. We thus use the implicit reparameterization gradients-based approach developed by Figurnov et al.[32] which provides unbiased estimators for continuous distributions that have numerically tractable cumulative distribution functions (CDFs). To incorporate a Dirichlet distribution, they use the property that it can be rewritten as a composition of several univariate Gamma variables. We refer interested readers to Table 1 from Figurnov et al.[32] for the equations that show the computation of implicit gradients for backward propagation through a node with a Gamma distribution.

Layers: Our encoder has 3 fully connected layers with 300 hidden units in each layer. The decoder simply maps from the bottleneck layer to the output reconstruction layer via p_θ.

Vocabulary or input space

Unlike several other deep learning models that use a one-hot encoding of the raw DNA sequence, we use k-mer features to capture sequence preferences. We use a window of 200 bp around the peak summit and assume that the TF binding site can be present at any location in this window. Inspired by prior work[6,7] and the wildcard kernel[33],

we use all k-mers of length 8 with up to two consecutive wildcards allowed per k-mer to define the input space. We consider exact-matching k-mers and k-mers with wildcards as distinct features: for example, `TATTACGT`, `TANTACGT`, `TNNTACGT` are all counted separately. Further, an 8-mer and its reverse complement are treated as a single feature that combines the counts of both the 8-mer and its reverse complement. This results in a vocabulary or input space of size 112800.

Parameter tuning and model selection

The hyperparameters of our model are the following: the dimension M, of the latent space/bottleneck layer, number of layers and the width for the MLP of the encoder, Dirichlet prior hyperparameter α, which controls the prior distribution $\vec{\alpha}$ of topics, vocabulary size for the k-mer representation. We tried the following values for M: 10, 50, 100, 200, 500, 1000. For α we tried: $1e^{-3}, 1e^{-2}, 0.1, 1, 10, 20, 30, 50, 100$. Note that $\vec{\alpha} = \alpha \mathbf{1}^{1\times M}$.

We found that increasing α, which controls the prior of the Dirichlet distribution, increases the extent of overlap between the basis vectors defining the latent space (i.e more sharing between the topics). In the extreme, this can lead to the so-called 'averaging affect' that variational autoencoders are known to suffer from, where the model learns an 'average' representation of the data. Further, very large values such as $30, 50, 100$ lead to convergence issues during optimization since the non-negativity constraints on θ are not met. Small values of α such as $1e^{-2}, 1e^{-3}$, due to the nature of the Dirichlet distribution, lead to a peaky prior distribution that tries to enforce each peak to have only one 'active' latent dimension. However, this leads to a lower likelihood as it does not capture the heterogeneous nature of peaks. We find that $\alpha \in [10, 20]$ results in models with a good trade-off between the diversity of the posterior and the likelihood (i.e the loss function). We also find that as α changes, the learned topic distributions vary and result in different TFs being learned based on the prior. We keep α fixed for the initial several epochs (we set a burn-in of 150,000 steps, which is also a tunable parameter) and then optimize over $\vec{\alpha}$ by back-propagating the corresponding gradients.

Increasing the dimension of the latent space from $M = 10$, as expected, leads to the bottleneck layer learning more diverse patterns up to $M = 100$. For higher values such as $M = 200, 500, 1000$ the redundancy across dimensions increases substantially, i.e several θ_is will be similar to each other. We tried various batch sizes and found 128 to be optimal. We used the Adam optimizer with a learning rate of $3e^{-4}$ and terminating optimization upon a maximum of 300,000 steps.

We select the final model based on the number of TFs mapped, i.e. the number of TFs that satisfy the p-value threshold of 0.05, in the procedure outlined in Algorithm 1. Given our observation about complementary sets of TFs being learned as we change the prior through α, we decided to use an ensemble model. We pick the three best models, where we rank the models based on the number of learned TFs, and aggregate the non-redundant TFs from them to get the set of all learned meaningful dimensions.

Model training time: All experiments were run on Microsoft Azure Virtual Machines. With a single gpu, our code takes 6 to 8 hours to train for 300,000 epochs. Inference on test data takes ≈ 1 minute.

Mapping latent dimensions to TFs

We describe the algorithm used for mapping latent dimensions to transcription factors. If the probes of TF t have an enriched presence in latent dimension m as compared to that of all other probes, the label t is assigned to dimension m. The significance of enrichment computed by the Mann-Whitney U test. Note that this procedure can lead to a many-to-many mapping between dimensions $m \in \{1 \dots M\}$ and TFs $t \in T$.

Algorithm 1: Map latent dimensions to TFs

Input: $M \leftarrow$ number of latent dimensions
$\{1 \dots M\} \leftarrow$ set of latent dimensions
$N = 54000$ (number of probes)
$y^s \in \mathbb{R}^N$: TF labels of probes
$z^s \in \mathbb{R}^{N \times M}$: latent representation of probes
T: set of 296 TFs
Output: Mapping $\mathbf{F} : \{1 \dots M\} \rightarrow T$
foreach $t \in T$ **do**
  $z^s_t \leftarrow$ rows of z^s where $y^s = t$
  foreach $m \in \{1 \dots M\}$ **do**
    Let $a_{\cdot m}$ be the mth column vector of z^s_t
    Let $b_{\cdot m}$ be the mth column vector of z^s_{-t}
    $p_{tm} \leftarrow$ p-value of Mann-Whitney U test ($a_{\cdot m} >$ top 5% of values from $b_{\cdot m}$)
    if $p_{tm} < 0.05$ **then**
      $F \cup (m, t)$
    end
  end
end

Datasets

ATAC-seq data from cell lines

We downloaded the publicly available GM12878 ATAC-seq dataset from the GEO database, reference ID: GSE47753[3]. We took only samples generated using 50,000 cells.

[3] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47753

Since replicate 1 has much higher sequencing depth than the other replicates, we combined replicates 2,3 and 4 to obtain a second replicate (renamed replicate 2). We followed ENCODE ATAC-seq processing pipeline (https://www.encodeproject.org/atac-seq/) for ATAC-seq processing. Raw fastq files were adapter-trimmed using Trimmomatic and aligned to hg19 genome using Bowtie2 with default settings. PCR duplicates were then removed using Picard MarkDuplicates and Tn5 shifts are adjusted for. Peak calling was performed for each replicate using macs2 with -nomodel -shiift -37 -extsize 73. Finally, IDR was performed with the idr package and reproducible peaks were called with an IDR cutoff of 0.05 (reference). We identified a total of 76,218 reproducible peaks in GM12878 ATAC-seq dataset. We downloaded the publicly available A549 ATAC-seq dataset from the ENCODE portal[4].

HT-SELEX

We used the filtered HT-SELEX probes from Yuan et al.[7] for training our models. Briefly, this dataset contains HT-SELEX data sequenced in Jolma et al.[19] (ENA accession: ERP001824) and in 2017 (ENA accession: ERP016411), which together constitute 547 experiments for 461 human or mouse TFs. The experiments that were filtered out were those showing: 1) poor consistency of 8-mer enrichment in consecutive HT-SELEX cycles, 2) low number of enriched probes, or 3) low diversity of probe enrichment. For each remaining experiment, the top 2,000 enriched 20-bp probes were selected per experiment. The filtered dataset contains 325 high quality experiments covering 296 TFs.

ChIP-seq data: GM12878

Conservative and optimal IDR (irreproducible discovery rate) thresholded 'narrowpeak' files of ENCODE ChIP–seq data were downloaded from the ENCODE portal[5] for GM12878.

CAP-SELEX

Consecutive affinity-purification systematic evolution of ligands by exponential enrichment (CAP-SELEX)[17] is an approach to identify TF pairs that bind cooperatively to DNA. It is an in vitro assay based on a consecutive affinity-purification protocol coupled with enrichment of bound ligands. For a given pair of TFs, say TF_1 and TF_2, we downloaded the probes from cycle-4 and selected candidate probes for co-operative binding by picking frequent probes where the PWM model for the pair TF_1 and TF_2 (see supplementary material from Jolma et al.[17]) was found. We used the MAST algorithm[34] for motif matching with a high e-value cut-off of 10, due to the relatively short

[4] `https://www.encodeproject.org/files/ENCFF548PSN/`

[5] `https://www.encodeproject.org/`

length of 40bp of the probes. We selected the top 1000 probes (or fewer, as found) and ran inference on them to obtain their latent representations.

RNA-seq expression data

For GM12878, we download gene expression data from the ENCODE portal with reference ids: ENCFF906LSJ and ENCFF630BDD. We consider a gene to be expressed if it has an average expression of 0.05 rpkm or higher over the two datasets. For A549, we download ENCFF203NNS.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data analysed during this study are publicly available and listed within the paper. Our implementation will be made available, along with the processed datasets, upon publication.

Competing interests

MK and JLF are employees at Microsoft. HY is an employee at Calico Life Sciences.

Funding

This research was supported by Microsoft, Calico, and NIH/NHGRI award U01 HG009395 to CL.

Authors' contributions

MK worked on problem formulation, method development and implementation, data processing, generating plots, and writing. HY contributed to problem and analysis discussion, data preprocessing, generating plots, and writing. JLF contributed to problem formulation, method discussion and editing. CL contributed to problem formulation, analysis discussion, designing experiments, and writing.

Acknowledgements

We would like to thank Michael Figurnov for help with the implementation of the VAE.

References

1 Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods. 2013;10(12):1213.

2 Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523(7561):486-90.

3 Arvey A, Agius P, Noble WS, Leslie C. Sequence and chromatin determinants of cell-type–specific transcription factor binding. Genome Research. 2012;22(9):1723-34.

4 Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Computational Biology. 2014;10(7):e1003711.

5 Shrikumar A, Prakash E, Kundaje A. GkmExplain: fast and accurate interpretation of nonlinear gapped k-mer SVMs. Bioinformatics. 2019;35(14):i173-82.

6 Setty M, Leslie CS. SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps. PLoS Computational Biology. 2015;11(5):e1004271.

7 Yuan H, Kshirsagar M, Zamparo L, Lu Y, Leslie CS. BindSpace decodes transcription factor binding signals by large-scale sequence embedding. Nature Methods. 2019;16(9):858-61.

8 Guo Y, Gifford DK. Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. BMC Genomics. 2017;18(1):1-16.

9 Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research. 2016;26(7):990-9.

10 Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods. 2015;12(10):931-4.

11 Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Research. 2018;28(5):739-50.

12 Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nature Genetics. 2021:1-13.

13 Shrikumar A, Tian K, Shcherbina A, Avsec Ž, Banerjee A, Sharmin M, et al. Tf-Modisco v0. 4.4. 2-Alpha. arXiv preprint arXiv:181100416. 2018.

14 Bailey TL, Elkan C, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. Proceedings of the International Conference on Intelligent Systems in Molecular Biology. 1994.

15 Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158(6):1431-43.

16 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011;7(1):539.

17 Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature. 2015;527(7578):384-8.

18 Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell. 2010;38(4):576-89.

19 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152(1-2):327-39.

20 Becht E, McInnes L, Healy J, Dutertre CA, Kwok IW, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology. 2019;37(1):38-44.

21 Li P, Spolski R, Liao W, Wang L, Murphy TL, Murphy KM, et al. BATF–JUN is critical for IRF4-mediated transcription in T cells. Nature. 2012;490(7421):543-6.

22 Rodriguez-Martinez JA, Reinke AW, Bhimsaria D, Keating AE, Ansari AZ. Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. Elife. 2017;6:e19272.

23 Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Research. 2016;44(11):e107-7.

24 Zhang Y, Qiao S, Ji S, Li Y. DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding. International Journal of Machine Learning and Cybernetics. 2020;11(4):841-51.

25 Ji Y, Zhou Z, Liu H, Davuluri R. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. arXiv. 2021.

26 Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017-8.

27 van Helden J, André B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. Journal of Molecular Biology. 1998;281(5):827-42.

28 Leslie C, Eskin E, Noble WS. The spectrum kernel: A string kernel for SVM protein classification. In: Biocomputing 2002. World Scientific; 2001. p. 564-75.

29 González-Blas CB, Minnoye L, Papasokrati D, Aibar S, Hulselmans G, Christiaens V, et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nature Methods. 2019;16(5):397-400.

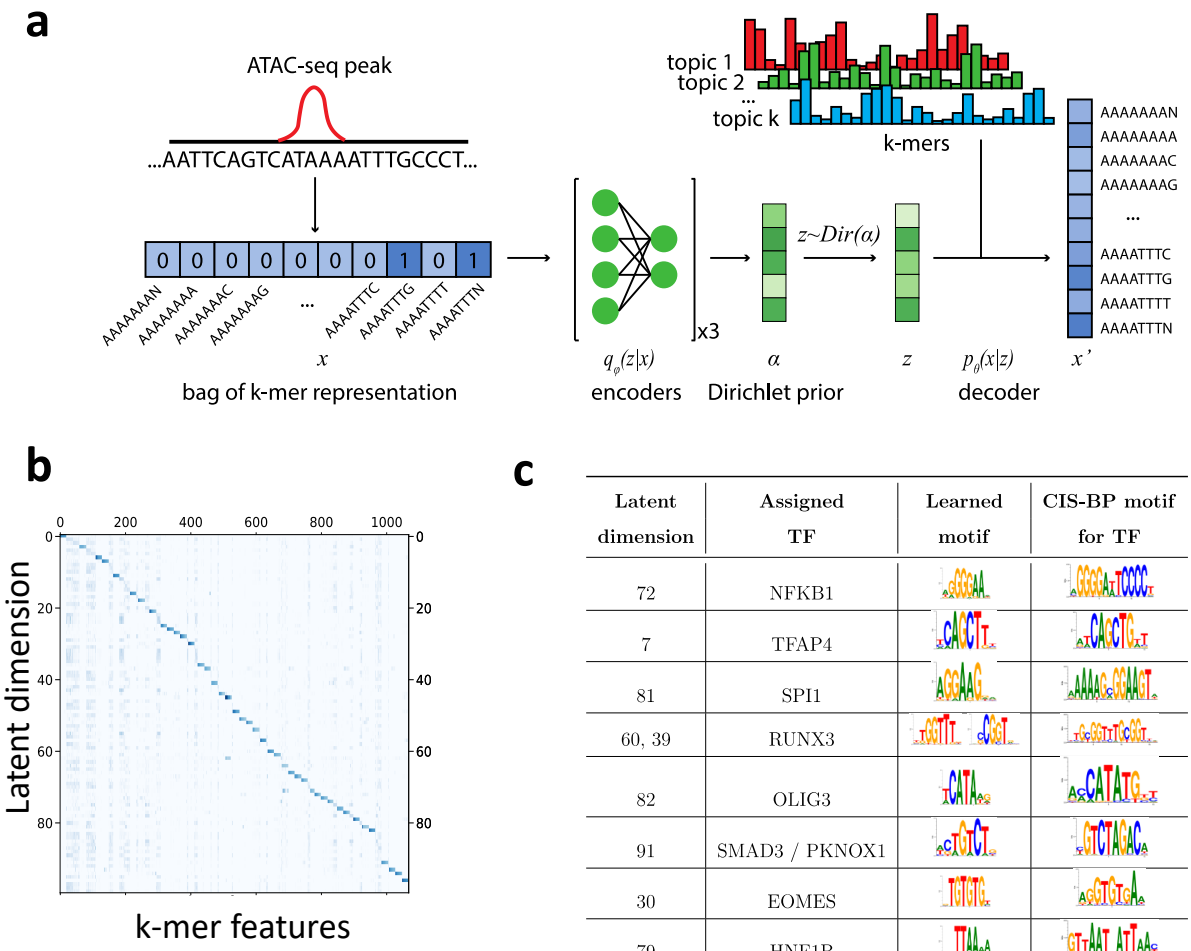
30 Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv preprint arXiv:13126114. 2013.

31 Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. In: International Conference on Machine Learning. PMLR; 2014. p. 1278-86.

32 Figurnov M, Mohamed S, Mnih A. Implicit Reparameterization Gradients. arXiv. 2019.

33 Leslie C, Kuang R, Bennett K. Fast string kernels using inexact matching for protein sequences. Journal of Machine Learning Research. 2004;5(9).

34 Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics (Oxford, England). 1998;14(1):48-54.



Latent dimension	Assigned TF	Learned motif	CIS-BP motif for TF
72	NFKB1		
7	TFAP4		
81	SPI1		
60, 39	RUNX3		
82	OLIG3		
91	SMAD3 / PKNOX1		
30	EOMES		
79	HNF1B		
11, 4	CTCF		
25	PRDM1		
51	RXRB		
54	VDR		
6	GATA4		
52	ETS1		

Figure 1: **(a)** Generative model captured by our VAE **(b)** Matrix showing the top 20 k-mer feature weights learned for each latent dimension (i.e entires of θ sorted by their value). Most dimensions have unique top k-mers **(c)** Motif analysis of each latent dimension shows that different binding specificities are captured by each latent factor in GM12878. Column 2 shows the name of the TF assigned by Algorithm 1 and the last column shows the CIS-BP motif corresponding to this TF.

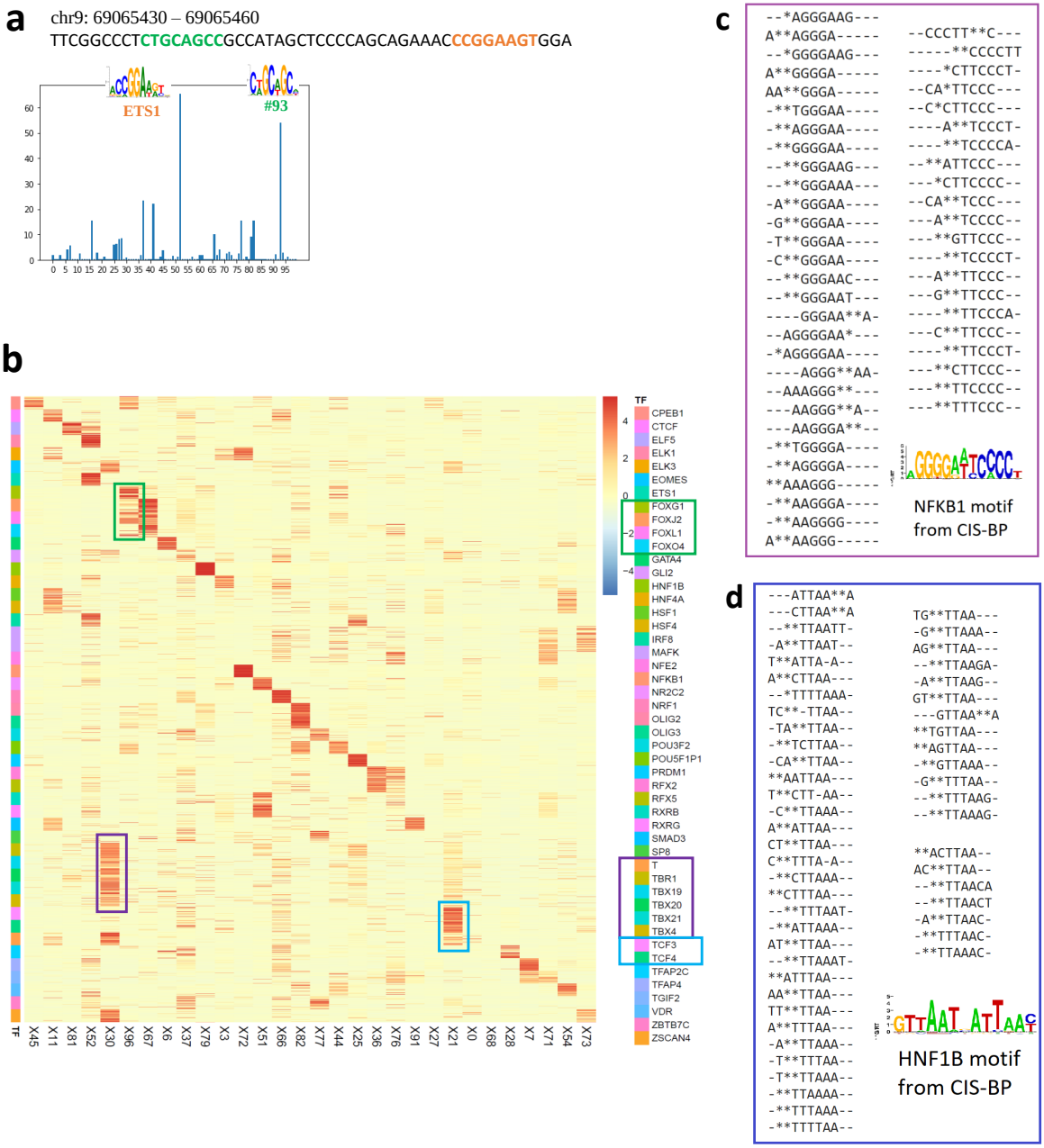


Figure 2: **(a)** TFs and the corresponding 'scores', assigned to a 30bp region of a random peak from our GM12878 dataset. **(b)** Heatmap of the latent space obtained by our model upon doing inference on ≈10,000 SELEX probes from 48 TF experiments. Each row is a SELEX probe, with the probes being colored by their TF. There are 200 probes per SELEX experiment. The green boxes highlight the FOX family of TFs and the corresponding probes scores in dimension 96 (X96). **(c), (d)** Top 50 8-mers from two latent dimensions obtained by sorting the decoder parameters θ_m, for m =72 and 79. The 8-mers have been aligned using multiple sequence alignment and the $*$ symbols show the wildcards.

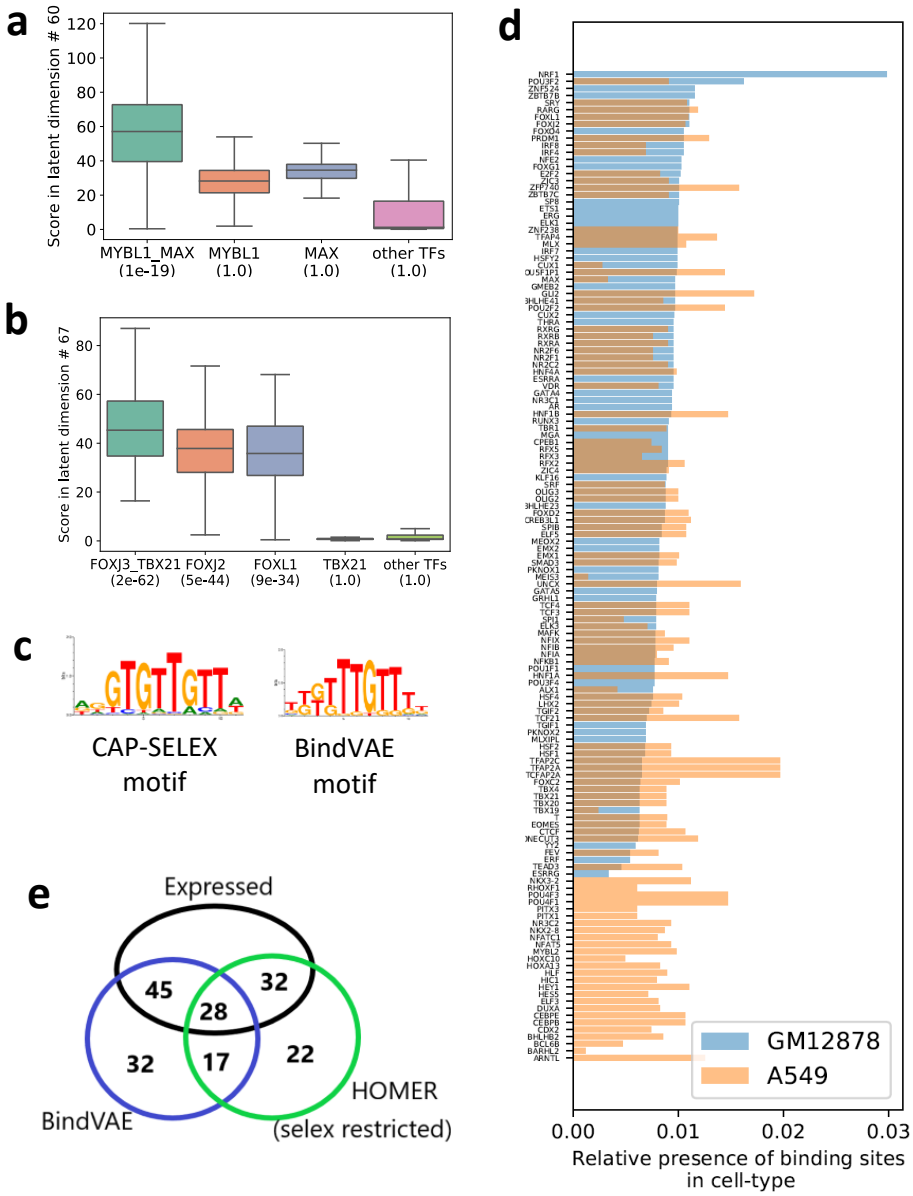


Figure 3: **(a), (b)** Analysis of cooperative binding signals learned by the model trained on GM12878 peaks by scoring HT-SELEX probes from individual TFs and CAP-SELEX probes from pairs of TFs. The y-axis shows the distribution of scores of individual probes for the TF experiments shown on the x-axis. In the brackets below each TF label, we show the p-value of assigning to the latent dimension, that particular TF, using Algorithm 1. **(c)** *left:* CAP-SELEX based motif for the TF pair binding from Jolma et al.[17] and *right:* Motif learned by BindVAE for FOXJ3:TBX21 cooperative binding **(d)** TFs found in each cell type by our model and their 'accessibility scores' obtained by summing over all ATAC-seq peaks showing the possible extent of accessibility in the two cell lines: GM12878 (light blue) and A549 (orange). See text for details **(e)** Venn diagram showing the overlap in the TFs found *de novo* by the VAE and HOMER in GM12878 and the extent of their overlap with all expressed genes. We show HOMER results run with the set of 296 motifs from Jolma et al.[19] (selex restricted) where HOMER only uses the TFs from our HT-SELEX set of 296 TFs.

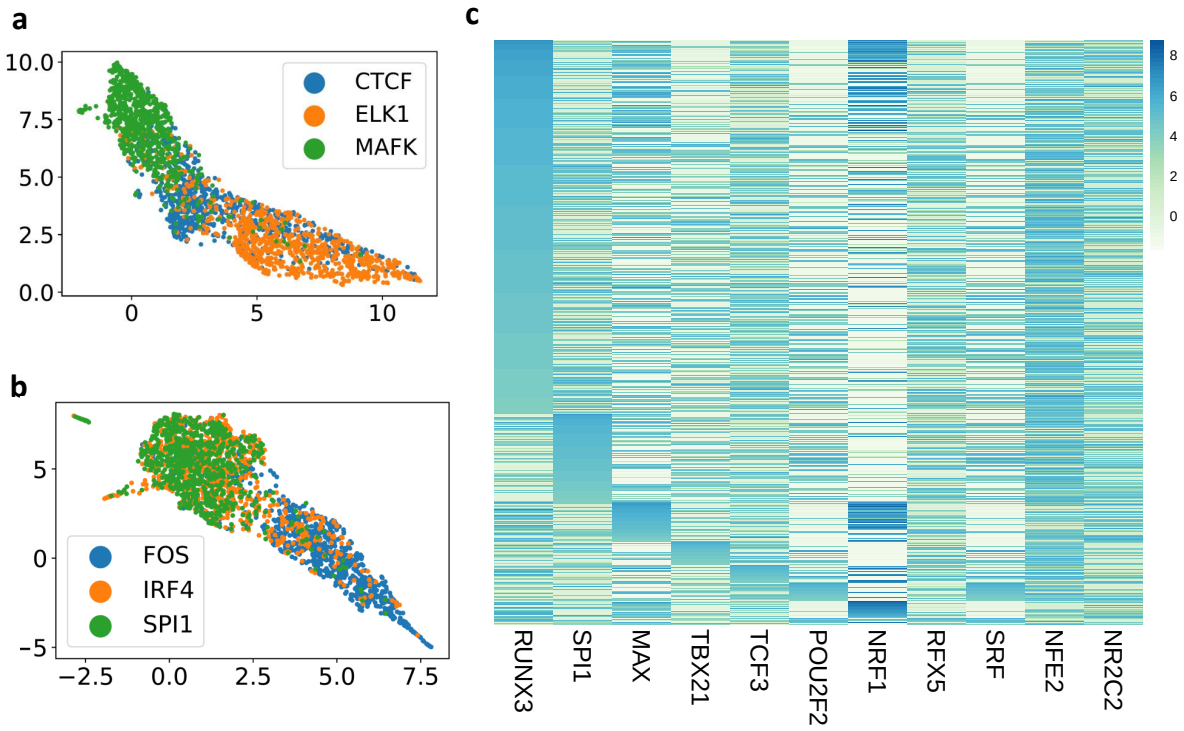


Figure 4: **(a)** UMAP projection of the latent representation for ChIP-seq peaks from `CTCF`, `ELK1` and `MAFK` **(b)** UMAP projection of the latent representation for ChIP-seq peaks from `FOS`, `IRF4` and `SPI1`. For each TF, 1000 peaks are shown. **(c)** Heatmap showing 38,496 ATAC-seq peaks from GM12878 cell line, that overlap with ChIP-seq peaks from 11 different TFs. For a TF t, the peaks have been sorted by the score of the latent dimension that was mapped to t by Algorithm 1.