Supplementary Data: Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences

Matthias Siebert[1,2] and Johannes Söding[1,*]

[1] Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

[2] Gene Center, Ludwig-Maximilians-Universität München, Feodor-Lynen-Strasse 25, 81377, Munich, Germany

[*] **For correspondence:** soeding@mpibpc.mpg.de

Supplementary Methods

Gibbs binding energy, log-odds scores, and learning model parameters

We show here the following: The statistical physics approach to learning a model for the sequence-dependent binding energy of a protein to DNA that best explains the observed binding data is equivalent to the purely statistical approach – followed by most work in the field of motif discovery – of maximising the likelihood of the training sequences to have been generated by the motif model and background sequence model.

Suppose we have measured binding sites of a transcription factor, e.g. in vivo by a ChIP-seq experiment or in vitro using HT-SELEX. Our goal is to describe the Gibbs free energy $\Delta G(\mathbf{x})$ for any potential binding site sequence $\mathbf{x} = (x_{1:W}) \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^W$, where W is the number of bases with an influence on the binding specificity. This will allow us to make predictions for arbitrary sequences about where and with what relative strength the factor binds. The following treatment will also generalise to complex, multipartite motifs.

We denote by $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^W$ the N measured binding site sequences and by $p_{\mathrm{bg}}(\mathbf{x})$ the probability distribution of sequences $\mathbf{x} \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^W$ in the background set from which the binding sites were selected. As examples, in ChIP-seq $p_{\mathrm{bg}}(\mathbf{x})$ is learned from a mock immunoprecipitation measurement and in HT-SELEX from the input sequence library prior to the selection step.

According to Boltzmann's law, the probability of a genomic site with sequence $\mathbf{x}$ to be bound by the transcription factor divided by the probability of $\mathbf{x}$ not to be bound is

$$\exp\left(-\frac{\Delta G(\mathbf{x}) - \mu}{k_{\mathrm{B}}T}\right) = \frac{p(\mathrm{bound}|\mathbf{x})}{p(\mathrm{not\ bound}|\mathbf{x})} = \frac{p(\mathrm{bound}|\mathbf{x})}{1 - p(\mathrm{bound}|\mathbf{x})}, \tag{S.1}$$

with the chemical potential μ that depends on the factor concentration but not on $\mathbf{x}$. Solving for $p(\mathrm{bound}|\mathbf{x})$ yields the well-known behaviour for unsaturated binding,

$$p(\mathrm{bound}|\mathbf{x}) = \left(1 + \exp\left(\frac{\Delta G(\mathbf{x}) - \mu}{k_{\mathrm{B}}T}\right)\right)^{-1}. \tag{S.2}$$

We parameterise the dependence of $\Delta G(\mathbf{x})$ on the binding site sequence $\mathbf{x}$ by defining the following normalised probability distribution:

$$p_{\mathrm{motif}}(\mathbf{x}) := \frac{p_{\mathrm{bg}}(\mathbf{x}) \exp(-\Delta G(\mathbf{x})/k_{\mathrm{B}}T)}{\sum_{\mathbf{y}} p_{\mathrm{bg}}(\mathbf{y}) \exp(-\Delta G(\mathbf{y})/k_{\mathrm{B}}T)}, \tag{S.3}$$

where the sum in the normalisation constant runs over all possible binding site sequences $\mathbf{y} \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^W$. Abbreviating the denominator as const. and solving for $\Delta G(\mathbf{x})/k_{\mathrm{B}}T$,

$$-\frac{\Delta G(\mathbf{x})}{k_{\mathrm{B}}T \log 2} + \mathrm{const.} = \log_2 \frac{p_{\mathrm{motif}}(\mathbf{x})}{p_{\mathrm{bg}}(\mathbf{x})} =: S(\mathbf{x}), \tag{S.4}$$

we find that the binding strength of a site $\mathbf{x}$ (as quantified by the negative Gibbs energy of binding in units of $k_{\mathrm{B}}T \log 2$) is, up to a constant, equal to the log-odds score $S(\mathbf{x})$. Once we know $p_{\mathrm{motif}}(\cdot)$ we can compute $S(\mathbf{x})$ and the relative binding strength $\Delta G(\mathbf{x})/k_{\mathrm{B}}T$ for any potential binding site sequence $\mathbf{x} = (x_1 \ldots x_W)$.

To learn the distribution $p_{\mathrm{motif}}(\cdot)$, and therefore also $\Delta G(\mathbf{x})$, from the measured binding sites, we need the likelihood,

$$p(\mathbf{x}_1 \ldots \mathbf{x}_N | \mathbf{p}_{\mathrm{motif}}) = \prod_{n=1}^{N} p(\mathbf{x}_n | \mathrm{bound}, \mathbf{p}_{\mathrm{motif}}) \tag{S.5}$$

of the binding sites given the model parameters $\mathbf{p}_{\mathrm{motif}}$. The probability $p(\mathbf{x}_n|\mathrm{bound}, \mathbf{p}_{\mathrm{motif}})$ for pulling out a sequence $\mathbf{x}_n$ from an underlying distribution of possible sequences $p_{\mathrm{bg}}(\mathbf{x})$ can be

found using Bayes' theorem,

$$p(\mathbf{x}_n|\text{bound}, \mathbf{p}_{\text{motif}}) = \frac{p(\text{bound}|\mathbf{x}_n, \mathbf{p}_{\text{motif}})\, p_{\text{bg}}(\mathbf{x}_n)}{\sum_{\mathbf{y}} p(\text{bound}|\mathbf{y}, \mathbf{p}_{\text{motif}})\, p_{\text{bg}}(\mathbf{y})}. \tag{S.6}$$

Here, $p(\text{bound}|\mathbf{x}_n, \mathbf{p}_{\text{motif}})$ is the probability that $\mathbf{x}_n$ is bound by the factor (eq. S.2).

We now assume, as commonly done, to be in a regime of unsaturated binding, $p(\text{bound}|\mathbf{x}) \lesssim 0.1$ [1]. We can then approximate eq. (S.2) as $p(\text{bound}|\mathbf{x}) \approx \exp\left(-(\Delta G(\mathbf{x}) - \mu)/k_{\text{B}}T\right)$. Inserting this expression into the likelihood and using eq. (S.3) yields

$$p(\mathbf{x}_1 \dots \mathbf{x}_N|\mathbf{p}_{\text{motif}}) = \prod_{n=1}^{N} p_{\text{motif}}(\mathbf{x}_n). \tag{S.7}$$

This equation for the likelihood applies to any choice of models for the binding site and background sequences in the regime of unsaturated binding. It is remarkable because it shows that the statistical physics approach to learning a binding energy model that explains the observed binding data leads to the same likelihood as the purely statistical approach that has been followed in most studies. (See Djordjevic *et al.* [2] for an interesting approach to learn a PWM model that does not assume unsaturated binding but instead makes the simplifying assumption of zero temperature.)

PWM model and maximum a-posteriori approach

Without loss of generality, the probabilities for motif and background sequences can be written

$$\begin{aligned} p_{\text{motif}}(x_{1:W}) &= \prod_{j=1}^{W} p_j(x_j|x_{1:j-1})\,, \\ p_{\text{bg}}(x_{1:W}) &= \prod_{j=1}^{W} p_{\text{bg}}(x_j|x_{1:j-1})\,, \end{aligned} \tag{S.8}$$

with appropriate conditional distributions. The position weight matrix (PWM) model assumes independence between nucleotides at different positions for the motif and background probabilities, $p_j(x_j|x_{1:j-1}) \approx p_j(x_j)$ and $p_{\text{bg}}(x_j|x_{1:j-1}) \approx p_{\text{bg}}(x_j)$, resulting in

$$p_{\text{motif}}(\mathbf{x}) \approx \prod_{j=1}^{W} p_j(x_j)\,, \quad p_{\text{bg}}(\mathbf{x}) \approx \prod_{j=1}^{W} p_{\text{bg}}(x_j)\,. \tag{S.9}$$

The log-odds score in (S.4) can therefore be written

$$S(\mathbf{x}) = \sum_{j=1}^{W} \log \frac{p_j(x_j)}{p_{\text{bg}}(x_j)} = \sum_{j=1}^{W} s_j(x_j), \tag{S.10}$$

showing that this approximation implies independent contributions to the binding energy $\Delta G(x)$ from each nucleotide in the binding site.

To find $p_j(a)$, we insert the left equation in (S.9) into eq. (S.7), take the logarithm, multiply the summed terms by $1 = \sum_{a=\text{A}}^{\text{T}} \mathbb{I}(x_i^n = a)$ (with indicator function $\mathbb{I}(\cdot)$), change the order of the summations, and abbreviate the number of times we observed a base a at position j by $n_j(a) := \sum_{n=1}^{N} \mathbb{I}(x_j^n = a)$, which yields

$$\log p(\mathbf{x}_1 \dots \mathbf{x}_N|\mathbf{p}_{\text{motif}}) = \sum_{j=1}^{W} \sum_{a=\text{A}}^{\text{T}} n_j(a) \log p_j(a). \tag{S.11}$$

We can maximise this log likelihood under the constraints $\sum_{a=\text{A}}^{\text{T}} p_j(a) = 1$ using the method of Lagrange multipliers, which yields the maximum likelihood solution $p_j(a) = n_j(a)/N$.

This maximum-likelihood (ML) estimate has a serious drawback: particularly for small N it tends to overtrain the model. If, for instance, a nucleotide A has not been observed at position j in any of $N = 4$ binding sites, it would be given a zero probability, $p_j(A) = 0$, even though quite obviously from such few sequences we cannot conclude that A will *never* occur in position j of other binding sites.

We can improve our estimation using the *maximum a-posteriori* (MAP) approach, in which one maximises the posterior probability instead of the likelihood. According to Bayes' theorem, the posterior probability is proportional to the likelihood times the prior probability $p(\mathbf{p}_{\mathrm{motif}})$,

$$p(\mathbf{p}_{\mathrm{motif}}|\mathbf{x}_1 \dots \mathbf{x}_N) \propto p(\mathbf{x}_1 \dots \mathbf{x}_N|\mathbf{p}_{\mathrm{motif}})\, p(\mathbf{p}_{\mathrm{motif}}). \tag{S.12}$$

As prior, we can choose a product over Dirichlet distributions, $p(\mathbf{p}_{\mathrm{motif}}) = \prod_j \mathcal{D}ir(p_j(\cdot)|\alpha_0\, p_{\mathrm{bg}}(\cdot)) \propto \prod_j \prod_a p_j(a)^{\alpha_0\, p_{\mathrm{bg}}(a)-1}$, with pseudocount parameters $\alpha_0\, p_{\mathrm{bg}}(a)$. The optimisation of the posterior probability under the constraints $\sum_{a=\mathrm{A}}^{\mathrm{T}} p_j(a) = 1$ then leads to the MAP solution

$$p_j(a) = \frac{n_j(a) + \alpha_0\, p_{\mathrm{bg}}(a)}{N + \alpha_0}\,. \tag{S.13}$$

The Dirichlet prior in effect adds fractional pseudocounts $\alpha_0 p_{\mathrm{bg}}(a)$ to the observed counts $n_j(a)$. The MAP solution interpolates linearly between the maximum likelihood solution $n_j(a)/N$ and the prior distribution $p_{\mathrm{bg}}(a)$.

Bayesian Markov model learning

The inhomogeneous Markov model (iMM) of order k retains from equation (S.8) the dependence on the k previous positions,

$$p^{(k)}_{\mathrm{motif}}(x_{1:W}) = \prod_{j=1}^{W} p^{(k)}_j(x_j|x_{j-k:j-1})\,, \tag{S.14}$$

where for simplicity of notation x_i with indices $i \leq 0$ are ignored. PWMs are therefore iMMs of order 0. The conditional probabilities are commonly estimated by adding pseudocounts proportional to monomer background frequencies, as in PWMs,

$$p^{(k)}_j(x_j|x_{j-k:j-1}) = \frac{n_j(x_{j-k:j}) + \alpha_k\, p_{\mathrm{bg}}(x_j)}{n_{j-1}(x_{j-k:j-1}) + \alpha_k}. \tag{S.15}$$

Here $n_j(x_{j-k:j}) = \sum_n \mathbb{I}(x^n_{j-k:j} = x_{j-k:j})$ denotes the number of times $x_{j-k:j}$ occurs in the bound sequences $\mathbf{x}_1, \dots, \mathbf{x}_N$ ending at position j. The total number of pseudocounts α_k controls the balance between counts and pseudocounts. Analogously to the case of the simple PWM model, the expression in eq. (S.15) is the MAP solution obtained with the prior

$$p(\mathbf{p}^{(k)}_{\mathrm{motif}}|p_{\mathrm{bg}}) = \prod_{j=1}^{W} \prod_{x_{j-k:j-1}} \mathcal{D}ir\left(p^{(k)}_j(\cdot|x_{j-k:j-1})\middle|\, \alpha_k p_{\mathrm{bg}}(\cdot)\right). \tag{S.16}$$

We can dramatically improve this estimate by noting that $p^{(k)}_j(a|x_{j-k:j-1})$ will be much better approximated by $p^{(k-1)}_j(a|x_{j-k+1:j-1})$ than by $p_{\mathrm{bg}}(a)$. Therefore a much better choice for the prior is

$$p(\mathbf{p}^{(k)}_{\mathrm{motif}}|\mathbf{p}^{(k-1)}_{\mathrm{motif}}) = \prod_{j=1}^{W} \prod_{x_{j-k:j-1}} \mathcal{D}ir\left(p^{(k)}_j(\cdot|x_{j-k:j-1})\middle|\, \alpha_k p^{(k-1)}_j(\cdot|x_{j-k+1:j-1})\right). \tag{S.17}$$

As shown below, this prior leads to the following MAP solution for the model parameters,

$$p^{(k)}_j(x_j|x_{j-k:j-1}) = \frac{n_j(x_{j-k:j}) + \alpha_k\, p^{(k-1)}_j(x_j|x_{j-k+1:j-1})}{n_{j-1}(x_{j-k:j-1}) + \alpha_k}\,, \tag{S.18}$$

in which the pseudocounts for order k are derived from the model of order $k-1$. To prove this, we find the maximum of the logarithm of the posterior probability (eq. S.12) plus terms with Lagrange multipliers $\lambda_{\mathbf{x}'}$ for the optimisation constraints, using abbreviations for the context $\mathbf{x}' := x_{j-k:j-1}$ and the shortened context $\mathbf{x}'' := x_{j-k+1:j-1}$:

$$\log p(\mathbf{x}_1 \ldots \mathbf{x}_N|\mathbf{p}_{\rm motif}) + \log p(\mathbf{p}_{\rm motif}) + \sum_{j=1}^{W} \sum_{\mathbf{x}'} \lambda_{\mathbf{x}'} \left(1 - \sum_{a=\rm A}^{\rm T} p_j^{(k)}(a|\mathbf{x}') \right) \underset{\mathbf{p}_{\rm motif}}{\longrightarrow} \max . \qquad \text{(S.19)}$$

The first term can be derived in a way similar to eq. (S.11), yielding $\sum_j \sum_a \sum_{\mathbf{x}'} n_j(\mathbf{x}', a) \log p_j^{(k)}(a|\mathbf{x}')$. The second term follows from eq. (S.17) and $\log \mathcal{D}ir(\mathbf{p}|\boldsymbol{\alpha}) = \sum_a \alpha_a \log p(a) + \text{const}$, which results in $\sum_j \sum_{\mathbf{x}'} \sum_a \alpha_k p_j^{(k-1)}(a|\mathbf{x}'') \log p_j^{(k)}(a|\mathbf{x}') + \text{const}$. Setting the derivative of equation (S.19) with respect to one of the parameters, $p_j^{(k)}(a|\mathbf{x}')$, to zero gives

$$\frac{n_j(\mathbf{x}', a)}{p_j^{(k)}(a|\mathbf{x}')} + \frac{\alpha_k p_j^{(k-1)}(a|\mathbf{x}'')}{p_j^{(k)}(a|\mathbf{x}')} - \lambda_{\mathbf{x}'} = 0. \qquad \text{(S.20)}$$

Solving for $p_j^{(k)}(a|\mathbf{x}')$ and normalising it (which yields the value of $\lambda_{\mathbf{x}'}$) completes the proof.

We can rewrite this result to show that it interpolates between the maximum likelihood solution for order k, $n_j(x_{j-k:j})/n_{j-1}(x_{j-k:j-1})$, and the order-$(k-1)$ probability, $p_j^{(k-1)}(x_j|x_{j-k+1:j-1})$:

$$p_j^{(k)}(x_j|x_{j-k:j-1}) = w \frac{n_j(x_{j-k:j})}{n_{j-1}(x_{j-k:j-1})} + (1-w)\, p_j^{(k-1)}(x_j|x_{j-k+1:j-1}), \qquad \text{(S.21)}$$

with an interpolation weight w that is a saturating function of the frequency of the context, $n_{j-1}(x_{j-k:j-1})$,

$$w = \frac{n_{j-1}(x_{j-k:j-1})}{n_{j-1}(x_{j-k:j-1}) + \alpha_k}. \qquad \text{(S.22)}$$

In our derivation of equations (S.21, S.22), the dependence of w on $n_{j-1}(x_{j-k:j-1})$ was dictated in a natural way by our Bayesian viewpoint, and the only real choice we had was how to set the values of α_k. Our BaMMs are a special case of interpolated Markov models [3, 4], which differ in the way the interpolation weights are chosen. In the past, various ad-hoc heuristics have been proposed. Salzberg *et al.* [5], who introduced interpolated Markov models to computational biology, used an interpolation scheme, in which w is 1 if $n_{j-1}(x_{j-k:j-1}) > 400$ and below that threshold w grows from 0 to 1 with increasing $n_{j-1}(x_{j-k:j-1})$ and increasing significance with which the hypothesis can be rejected that $(x_{j-k:j-1}\, a)$ is distributed according to the order-$(k-1)$ probability $p_j^{(k-1)}(a|x_{j-k+1:j-1})$. In the scheme of Ohler *et al.* [6], w does not only depend on $n_{j-1}(x_{j-k:j-1})$ but on the numbers of occurrence of the k suffixes $x_{j-l:j-1}$ with $l = 1, \ldots, k$. A drawback is that even for very high $n_{j-1}(x_{j-k:j-1})$ the weights for orders lower than k do not approach 0, as they should at the expense of a higher order. Most importantly, all previous schemes require the motif sequences to be aligned, while our BaMMs allow for the de-novo discovery of unaligned motifs enriched in a set of sequences using an EM-type algorithm.

EM algorithm for BaMM-based de-novo motif discovery

We are given sequences $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and we want to discover motifs enriched in them. We do not know the positions of the potential motifs, however. We simplify the situation by using a zero-or-one-occurrence-per-sequence (ZOOPS) model [7], which assumes that each sequence carries one or no motif. When interpreted from the statistical physics viewpoint, the ZOOPS model does *not* actually assume that *at most one* motif is present per sequence. Rather, binding can occur anywhere on the sequences, and several sites per sequence can contribute to the binding, but not more than a single factor can bind at the same time per sequence. If several binding sites are present in one sequence, they can only contribute with a combined weight of up to one. In a

multiple-occurrence-per-sequence (MOPS) model, the combined weight of these sites would be the expected occupancy of the sequence, which can be more than one. In cases where we expect a strongly varying number of binding site occurrences per sequence, a MOPS model might therefore give better results. But this model requires a forward-backward computation at each M-step of the EM algorithm, which would considerably slow down the motif learning.

To learn the parameters of the BaMM, $\mathbf{p}_{\text{motif}}$, from unaligned binding sites, we derive here an Expectation Maximisation (EM) algorithm that maximises the posterior probability (eq. S.12) using the Dirichlet prior in eq. (S.17). The EM algorithm alternates between the E-step, in which it estimates the probabilities r_{ni} of a motif to be present at each position i of each of the training sequences n given the current estimate of model parameters $\mathbf{p}^{(k)}_{\text{motif}}$, and the M-step, in which it updates the model parameters $\mathbf{p}^{(k)}_{\text{motif}}$ given the current estimate of the r_{ni}.

In the ZOOPS model, the probability for a sequence $\mathbf{x}_n = x^n_{1:L}$ of length L with a motif of W nucleotides starting at position $z_n = l$ is

$$p\left(\mathbf{x}_n|z_n=l,\tilde{\mathbf{p}}^{(k)}_{\text{motif}}\right) = \prod_{i=1}^{l-1} p^{(K')}_{\text{bg}}(x^n_i|x^n_{i-K':i-1}) \prod_{i=l}^{l+W-1} p^{(k)}_{i-l+1}(x^n_i|x^n_{i-k:i-1}) \prod_{i=l+W}^{L} p^{(K')}_{\text{bg}}(x^n_i|x^n_{i-K':i-1}). \tag{S.23}$$

We use an inhomogeneous Markov model of order K for the motif and a homogeneous Markov model of order $K'=2$ for the background sequences. The hidden variable $z_n \in \{0,\dots,L-W+1\}$ gives the position of the single motif in sequence n. If no motif is present, which we signify by $z_n = 0$, we have simply $p(\mathbf{x}_n|z_n=l) = \prod_{i=1}^{L} p_{\text{bg}}(x^n_i|x^n_{i-K':i-1})$.

In the E-step, we update the *responsibilities* r_{ni} given the current parameter estimate of $\tilde{\mathbf{p}}_{\text{motif}}$, which according to Bayes' theorem is

$$r_{ni} = p(z_n=i|\mathbf{x}_n,\,\tilde{\mathbf{p}}^{(K)}_{\text{motif}}) = \frac{p\left(\mathbf{x}_n|z_n=i,\tilde{\mathbf{p}}^{(K)}_{\text{motif}}\right)\,p(z_n=i)}{\sum_{i'=0}^{L-W+1} p\left(\mathbf{x}_n|z_n=i',\tilde{\mathbf{p}}^{(K)}_{\text{motif}}\right)\,p(z_n=i')}\,. \tag{S.24}$$

We choose a flat positional prior, $p(z_n=0) = 1-q$ and $p(z_n=i) = q/(L-W+1)$ for $i>0$. The hyperparameter q specifies the prior probability for a sequence to contain a motif. It was found to have little influence on the results and was set to 0.9 throughout this work. This choice prevents false positive training sequences that do not contain any instance of the binding site to negatively affect the quality of the predicted motif model.

In the M-step, we update the parameters $\tilde{\mathbf{p}}^{(k)}_{\text{motif}}$ for $k=0,\dots,K$ given the responsibilities r_{ni}. This is done by maximising the auxiliary function

$$Q(\tilde{\mathbf{p}}_{\text{motif}}|r_{ni}) = \sum_{n=1}^{N}\sum_{i=0}^{L-W+1} r_{ni}\log\left(p(\mathbf{x}_n|z_n{=}i,\tilde{\mathbf{p}}^{(k)}_{\text{motif}})\,p(z_n=i|q)\right) + \log p\left(\mathbf{p}^{(k)}_{\text{motif}}|\mathbf{p}^{(k-1)}_{\text{motif}}\right), \tag{S.25}$$

with the prior given in eq. (S.17). We can maximise this function analytically under the constraints $\sum_a p(a|x_{1:k}) = 1$ for all $x_{1:k} \in \{\text{A},\text{C},\text{G},\text{T}\}^k$ by the method of Lagrange multipliers, which leads after some algebra to equation (S.18), but with a new, probabilistic definition for the counts:

$$n_j(x_{1:k}) := \sum_n r_{ni}\,\mathbb{I}(x^n_{i+j-k:i-j-1} = x_{1:k}). \tag{S.26}$$

We run the M-updates for all model orders from lowest to highest order and update the pseudocounts by the just updated model probabilities from the order below. We iterate the EM-steps until convergence of the model parameters.

Initialization of the EM algorithm We integrated code from our motif discovery tool XXmotif [8] into BaMM!motif and initialize the EM algorithm for all models studied in this work by setting the responsibilities $r_{ni} = 1$ for the motif instances returned by XXmotif, using options

`--reverseComp --XX-localization --XX-localizationRanking --XX-K 2 --mergeMotifsThreshold LOW --maxPValue 0.05 --minOccurrence 0.05`. BaMM!motif also allows the user to directly initialize BaMMs by supplying model parameters or a set of aligned binding sites.

Higher-order sequence logos

We can quantify the information in our learned motif using the relative entropy between the probability distributions for motif sequences and for background sequences and split it up into a sum of terms over model orders:

$$
\begin{aligned}
H\left(\mathbf{p}_{\mathrm{motif}}|\mathbf{p}_{\mathrm{bg}}\right) &= \sum_{\mathbf{x}=x_{1:W}} p_{\mathrm{motif}}(\mathbf{x}) \log_2 \frac{p_{\mathrm{motif}}(\mathbf{x})}{p_{\mathrm{bg}}(\mathbf{x})} \\
H\left(\mathbf{p}_{\mathrm{motif}}|\mathbf{p}_{\mathrm{bg}}\right) &= \sum_{j=1}^{W} \sum_{\mathbf{x}=x_{1:W}} p_{\mathrm{motif}}(\mathbf{x}) \log_2 \frac{p_j^{(k)}(x_j|x_{j-k:j-1})}{p_{\mathrm{bg}}^{(K')}(x_j|x_{j-K':j-1})} \\
H\left(\mathbf{p}_{\mathrm{motif}}|\mathbf{p}_{\mathrm{bg}}\right) &= \sum_{j=1}^{W} \left(\sum_{a=\mathrm{A}}^{\mathrm{T}} p_j(a) \log_2 \frac{p_j(a)}{p_{\mathrm{bg}}(a)} + \sum_{a,b} p_j(b,a) \log_2 \left(\frac{p_j(a|b)}{p_j(a)} \frac{p_{\mathrm{bg}}(a)}{p_{\mathrm{bg}}(a|b)} \right) \right. \\
&\qquad \left. + \sum_{a,b,c} p_j(c,b,a) \log_2 \left(\frac{p_j(a|c,b)}{p_j(a|b)} \frac{p_{\mathrm{bg}}(a|b)}{p_{\mathrm{bg}}(a|c,b)} \right) + \ldots \right). \qquad \text{(S.27)}
\end{aligned}
$$

The right hand-side on the last line splits the relative contribution up into one terms per order k. The sequence logo of order k visualises the contributions of all $(k+1)$-mers $(a, b, \ldots)$ at each position j to these terms. Note that some oligomers contribute negatively; the information contribution of one column is the sum of its positive and negative contributions. For simplicity we use $K' = 0$ for the sequence logos, which eliminates the ratio of background probabilities in 1'st and higher orders.

Supplementary Datasets

ChIP-seq datasets

We evaluated BaMM!motif on human transcription factor ChIP-seq datasets published by The ENCODE Project Consortium [9]. The March 2012 data freeze of the encyclopedia of DNA elements (ENCODE) comprises 708 IDR optimal blacklist-filtered SPP [10] peak sets. The irreproducible discovery rate (IDR) framework verifies the reproducibility of ChIP-seq peaks identified from replicate experiments by computing a quantitative reproducibility score [11, 12]. Peaks that overlap blacklisted regions were removed. These regions were empirically identified by the ENCODE Data Analysis Consortium (DAC) to show anomalous unstructured high signal in next-generation sequencing experiments independent of cell line and experiment type. We restricted our analysis to 87 RNA polymerase (RNAP) II-associated sequence-specific transcription factors characterised by Wang *et al.* [13] (441 datasets) and nine additional sequence-specific transcription factors (ATF1, ATF2, Elk1, FoxM1, IRF4, SREBP2, STAT5A, TCF3, ZnF217) from subsequently conducted experiments (13 datasets).

Positive sequences were compiled from the top 5 000 peak regions (sorted best to worst according to their signal value) or all peak regions if less than 5 000 peaks were available. Sequences were extracted ±100 bp around peak summits using Biopieces (`www.biopieces.org`). Background sequences were sampled from the trimer frequencies observed in positive sequences to ensure similar sequence compositions in both sequence sets. The length and number of background sequences was the same as the length and 100 times the number of positive sequences, respectively.

In order to initialise BaMMs, we ran XXmotif [8] using non-default options `--reverseComp --XX-localization --XX-localizationRanking --XX-K 2 --mergeMotifsThreshold LOW`, and filtered the results by requiring motifs to lie localised to peak summits (`--maxPValue 0.05`) and

to occur in at least 5 % of sequences (`--minOccurrence 0.05`). The motif instances that XXmotif used to calculate its top ranked PWM in each data set were employed to initialise BaMMs. Optionally, we added two or four uniformly initialised positions to both 5′- and 3′-ends of the models. The search space in training and test sequences was guaranteed to be identical and independent of the number of model positions. For instance, in order to compare the performance of models that describe the binding sites of the same transcription factor in the same data set but differ in their number of positions, we adjusted the search space in the benchmark test by extending training and test sequences of the longer model accordingly. In 446 (of all 454) datasets, corresponding to 94 transcription factors, XXmotif found at least one motif in all four cross-validation folds, independent of the length of training sequences. The remaining eight peak sets of the transcription factors c-Myc (2 datasets), E2F1 (1), ELF1 (1), PAX5 (1), PGC1A (1), and STAT1 (2) were excluded from the benchmark test.

To assess the performance of XXmotif, iMMs, and BaMMs in discriminating bound from unbound sequences, we carried out a four-fold cross-validation, that is, we trained on 75 % of data, tested on the 25 % held out data, and pooled results from the four holdout sets. In the process, we calculated the maximum log-odds score over all possible motif positions for each positive and background test sequence and evaluated the partial area under the receiver operating characteristic (ROC) curve (pAUC) up to a false positive rate (FPR) of 5 %. Since the pAUC summarises the part of the ROC curve that is most relevant to practical applications, it is preferable over the area under the entire ROC curve (AUC). In Figure 3 C,D, an improvement was significant at a confidence level of $0.0625 = 1/2^4$ if the BaMM model obtained higher pAUCs on all four test sets.

To evaluate the ability of BaMMs to predict *in vitro* binding affinities measured by competitive electrophoretic mobility shift assay (EMSA) for the mouse embryonic stem cell (mESC) transcription factor Klf4, we learned models from *in vivo* ChIP-seq data [14]. Positive sequences were compiled from the top 5 000 peak regions (sorted best to worst according to their signal value) by extracting ±50 bp around peak region midpoints. To initialise BaMMs, we ran XXmotif with the parameter setting used in the ENCODE ChIP-seq benchmark test.

EMSA datasets

We used the competitive EMSA experiments for the mESC transcription factor Klf4 from Sun *et al.* [15], comprising dissociation constant (K_d) measurements for 33 sequences with single mutations and 25 sequences with multiple mutations to the 10 bp consensus binding site of Klf4. These dissociation constants were divided by the dissociation constant of the sequence with median K_d (single mutant sequences) or with K_d closest to the mean K_d (multiple mutant sequences), and the logarithms of the resulting ratios were computed. Prediction scores are calculated as log ratios of odds scores. Since $K_d(\mathbf{x})$ is proportional to $\exp(\Delta G(\mathbf{x})/k_\mathrm{B}T)$ and according to eq. (S.4) also to $\exp(-S(\mathbf{x}))$, the predicted log ratios are equal to $-S(\mathbf{x}) + \mathrm{const}$. We compared Klf4 BaMMs of increasing order by means of the Pearson correlation between measured and predicted log ratios.

The competitive EMSA scores for 64 double-stranded oligonucleotide probes containing a potential FoxA2 binding site were taken from Levitsky *et al.* [16]. We calculated Spearman correlations between measured EMSA scores and log ratios predicted by FoxA2 BaMMs of increasing order. Spearman correlations to predictions from other methods and models were determined by Alipanahi *et al.* [17].

RNAP I/II core promoter sequences

We analysed sequences around *Drosophila melanogaster* transcription start sites (TSSs) measured by Brown *et al.* [18] using cap analysis of gene expression (CAGE) [19]. Filtered bedGraph-formatted CAGE datasets were pooled. Before clustering TSSs, the genomic distribution of TSSs was smoothed using a 41 bp uniform kernel function. Clusters were defined by genome intervals in which the smoothed distribution of TSS counts was found to lie entirely above the genome-wide average. The mode of the distribution was used as the representative TSS in each cluster. Subsequently, the clusters were filtered by three criteria. First, clusters with less than

five TSS counts were excluded. Second, clusters had to exhibit more TSS counts compared to any other cluster within 150 bp (regarding representative TSSs). Third, clusters had to overlap or lie close to FlyBase-annotated TSSs [20], by requiring a representative TSS to be located within 250 bp upstream of an annotated TSS or within a 5′ untranslated region (UTR). The clustering resulted in 15 971 TSSs assigned to 11 536 unique genes. Genes can thus be regulated by multiple core promoters defined by distinct TSSs.

In order to assign TSS clusters to a broad and narrow transcription-initialising core promoter class, the peakedness of the TSS distributions was quantified with a TSS width score by calculating the mean absolute deviation from the median TSS location as

$$\text{TSS width} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \text{median}(X)| \,, \tag{S.28}$$

where N is the number of TSSs within the cluster, x_i is the position of the i'th TSS, and median(X) is the median TSS position within the cluster. The distribution of TSS widths shows a local minimum at a value of five. Therefore, clusters with a TSS width smaller than five were classified as narrow peak (NP) and the remaining as broad peak (BP) core promoters. This resulted in 7 262 NP and 8 709 BP core promoters assigned to 5 576 and 7 235 genes, respectively. Note that 1 275 genes have core promoters from both classes.

In addition to NP and BP core promoters, we modeled core promoter sequences of ribosomal protein (RP) genes, which are known to differ from NP and BP core promoters in their architecture [21]. The RPG database [22] maintains 87 RP genes from *D. melanogaster*. Except for RpS27A, we could assign at least one core promoter to each RP gene, six of which had two associated core promoters. We thus obtained 92 RP gene core promoters, 60 and 32 belonging to the NP and BP class, respectively. Note that RP gene core promoters were not excluded from NP and BP core promoter sequences.

The core promoter encompasses the region that lies approximately ±50 bp around the TSS [23]. Therefore, we initialised core promoter models of all three classes from 101 bp sequences centered at representative TSSs. We computed the 9'th percentile of TSS width scores for each core promoter class, resulting in 3.64 (NP), 22.71 (BP), and 10.64 (RP), and learned NP, BP, and RP gene core promoter models within 4 bp, 23 bp, and 11 bp using positive sequences of length 109 bp, 147 bp, and 123 bp, respectively, centered at representative TSSs. Background models were learned from the trimer frequencies within 250 bp of representative TSSs.

We assessed the performance of iMMs and BaMMs in predicting TSS locations using a four-fold cross-validation procedure. To calculate the precision (fraction of true in all predictions) of the models in predicting the correct positions of TSSs, we determined the position with highest log-odds score in each test sequence, extracted ±250 bp around representative TSSs. If the position was within 4 (NP), 23 (BP), and 11 (RP) bp of the representative TSS, the prediction was judged as correct, else as false. The window size of each class corresponds to the 9'th percentile of its TSS cluster width scores (see above). Notably, while test sequences provide an identical search space (401) for all core promoter classes, the precision of random predictions is different for NP (0.02), BP (0.12), and RP (0.06) core promoters. We picture the distributions of maximum log-odds score positions, that is, the predictions of signal locations, as enrichments compared to predictions from a random predictor.

RNAP II polyadenylation site sequences

We use the major transcript isoform (mTIF) annotations from Pelechano *et al.* [24], obtained after clustering transcript isoforms (TIFs) from *S. cerevisiae* grown in yeast extract peptone dextrose (YPD). After selecting mTIFs covering one intact open reading frame (ORF) and summing up the sequencing reads of mTIFs with identical polyadenylation (pA) site, we selected the pA site(s) with the maximum number of sequencing reads per gene. We excluded pA sites with less than five sequencing reads. In total, we selected 4 228 pA sites from 4 173 distinct genes. 51 and 2 genes are represented by two and three pA sites, respectively.

The sequence region that surrounds pA sites shows nucleotide preferences within 70 bp upstream to 30 bp downstream of pA sites. Therefore, we modeled pA sites over the length of 101 bp covering this region. To provide a biologically relevant length of test sequences, we determined the length of 3′ UTRs from measured pA sites and *S. cerevisiae* ORF annotations from the Saccharomyces Genome Database [25]. Since more than 90 % of 3′ UTRs are shorter than 300 bp, test sequences were extracted from 220 bp upstream to 180 bp downstream of pA sites. This corresponded to 301 potential pA site positions within 150 bp of measured pA sites, from which the correct pA site is to be predicted in the benchmark tests. We trained background models from the trimer frequencies within 220 bp upstream to 180 bp downstream of pA sites.

Pelechano *et al.* defined mTIFs by clustering the transcripts with each of their 5′- and 3′-end sites co-occurring within 5 bp [24]. On this account, we determined the precision of pA site predictions by considering predictions within 5 bp of measured pA site locations as correct. Hence, the precision of random pA site predictions would be 0.04. In other respects, the benchmark test is identical to the evaluation procedure performed for RNAP I/II core promoter sequences.

RNAP pause site sequences

Larson *et al.* [26] measured 19 960 and 9 989 RNAP pause sites in *Escherichia coli* and *Bacillus subtilis*, respectively, using nascent elongating transcript sequencing (NET-seq), and found approximately one pause site per 100 bp across well-transcribed genes on average. We extracted test sequences that correspond to a search space of 101, centered at the pause sites. To prevent overtraining caused by overlapping training and test sequences, we excluded pause sites within 54 bp of another pause site with higher relative peak height. This reduced the number of *E. coli* and *B. subtilis* pause sites to 11 648 and 6 809, respectively. Background sequences were randomly sampled from the *E. coli* and *B. subtilis* genomes using the NCBI Reference Sequence (RefSeq) accession numbers NC_000913.3 and NC_000964.3, respectively, totaling to 100 times the number of positive sequences.

In *E. coli*, Larson *et al.* [26] identified a 16 bp consensus pause sequence, 10 bp upstream to 5 bp downstream of the pause index (the 3′-end of the transcript). We additionally incorporated the 2 bp immediately flanking the identified 16 bp consensus and learned the resulting 20 bp models by varying the model order. Likewise, we learned 20 bp BaMMs of *B. subtilis* RNAP pause sites. Longer models did not further improve benchmark test results.

Except for considering pause sites predicted to lie within 0 bp of measured sites to be correct, a random prediction could therefore locate pause sites with a precision of 0.01, we resort to the benchmark test described for RNAP I/II core promoter and RNAP II pA site sequences.

PAR-CLIP datasets

We modeled the binding of 25 messenger ribonucleoprotein (mRNP) biogenesis factors from *S. cerevisiae* to mRNA using published PAR-CLIP datasets [27, 28]. After sorting PAR-CLIP crosslink sites by occupancies (number of uracil to cytosine base transitions over RNA-seq counts) and excluding crosslink sites located in tRNA transcripts, we focused on the top 2 000 protein-RNA crosslink sites, which correspond to uracil nucleosides, and extracted 25 nt positive sequences encompassing the central crosslink site. In order to learn to discriminate factor binding sites in the transcriptome, 20 000 uracil-centered sequences (of the same length) were randomly sampled from the *S. cerevisiae* transcriptome using mRNA annotations from Pelechano *et al.* [24] and employed as background sequences both in learning and testing the models of all RNA-binding proteins. Note that background sequences may also contain true RNA-binding motifs.

We assessed the performance of iMMs and BaMMs in discriminating between uracil nucleosides with and without crosslink analogous to the evaluation procedure conducted in the ENCODE ChIP-seq benchmark test.

References

[1] Riley, T. R., Lazarovici, A., Mann, R. S., and Bussemaker, H. J. (2015) Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. *eLife,* **4**.

[2] Djordjevic, M., Sengupta, A. M., and Shraiman, B. I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res,* **13**, 2381–2390.

[3] Jelinek, F. and Mercer, R. L. (1980) Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice* pp. 381–397.

[4] Ristad, E. and Thomas, R. G. (1997) Nonuniform Markov models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* pp. 791–794.

[5] Salzberg, S. L., Delcher, A. L., Kasif, S., and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res,* **26**, 544–548.

[6] Ohler, U., Harbeck, S., Niemann, H., Nöth, E., and Reese, M. G. (1999) Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics,* **15**, 362–369.

[7] Bailey, T. L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* Menlo Park, CA, USA: AAAI Press pp. 28–36.

[8] Hartmann, H., Guthöhrlein, E. W., Siebert, M., Luehr, S., and Söding, J. (2013) P-value-based regulatory motif discovery using positional weight matrices. *Genome Res,* **23**, 181–194.

[9] The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature,* **489**, 57–74.

[10] Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol,* **26**, 1351–1359.

[11] Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (2011) Measuring reproducibility of high-throughput experiments *The Annals of Applied Statistics,* **5**, 1752–1779.

[12] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res,* **22**, 1813–1831.

[13] Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M., and Weng, Z. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res,* **22**, 1798–1812.

[14] Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L., and Ng, H.-H. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell,* **133**, 1106–1117.

[15] Sun, W., Hu, X., Lim, M. H. K., Ng, C. K. L., Choo, S. H., Castro, D. S., Drechsel, D., Guillemot, F., Kolatkar, P. R., Jauch, R., and Prabhakar, S. (2013) TherMos: estimating protein-DNA binding energies from in vivo binding profiles. *Nucleic Acids Res,* **41**, 5555–5568.

[16] Levitsky, V. G., Kulakovskiy, I. V., Ershov, N. I., Oshchepkov, D. Y., Makeev, V. J., Hodgman, T. C., and Merkulova, T. I. (2014) Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-seq data. *BMC Genomics,* **15**, 80.

[17] Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol,* **33**, 831–838.

[18] Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., Wan, K. H., Yu, C., Zhang, D., Carlson, J. W., Cherbas, L., Eads, B. D., Miller, D., Mockaitis, K., Roberts, J., Davis, C. A., Frise, E., Hammonds, A. S., Olson, S., Shenker, S., Sturgill, D., Samsonova, A. A., Weiszmann, R., Robinson, G., Hernandez, J., Andrews, J., Bickel, P. J., Carninci, P., Cherbas, P., Gingeras, T. R., Hoskins, R. A., Kaufman, T. C., Lai, E. C., Oliver, B., Perrimon, N., Graveley, B. R., and Celniker, S. E. (2014) Diversity and dynamics of the Drosophila transcriptome. *Nature,* **512**, 393–399.

[19] Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., and Hayashizaki, Y. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci USA,* **100**, 15776–15781.

[20] dos Santos, G., Schroeder, A. J., Goodman, J. L., Strelets, V. B., Crosby, M. A., Thurmond, J., Emmert, D. B., Gelbart, W. M., and the FlyBase Consortium (2015) FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res,* **43**, D690–D697.

[21] Lenhard, B., Sandelin, A., and Carninci, P. (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet,* **13**, 233–245.

[22] Nakao, A., Yoshihama, M., and Kenmochi, N. (2004) RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res,* **32**, D168–D170.

[23] Duttke, S. H. C., Lacadie, S. A., Ibrahim, M. M., Glass, C. K., Corcoran, D. L., Benner, C., Heinz, S., Kadonaga, J. T., and Ohler, U. (2015) Human promoters are intrinsically directional. *Mol Cell,* **57**, 674–684.

[24] Pelechano, V., Wei, W., and Steinmetz, L. M. (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature,* **497**, 127–131.

[25] Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S., and Wong, E. D. (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res,* **40**, D700–D705.

[26] Larson, M. H., Mooney, R. A., Peters, J. M., Windgassen, T., Nayak, D., Gross, C. A., Block, S. M., Greenleaf, W. J., Landick, R., and Weissman, J. S. (2014) A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science,* **344**, 1042–1047.

[27] Baejen, C., Torkler, P., Gressel, S., Essig, K., Söding, J., and Cramer, P. (2014) Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition. *Mol Cell,* **55**, 745–757.

[28] Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Söding, J., and Cramer, P. (2013) Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell,* **155**, 1075–1087.

Supplementary Figures

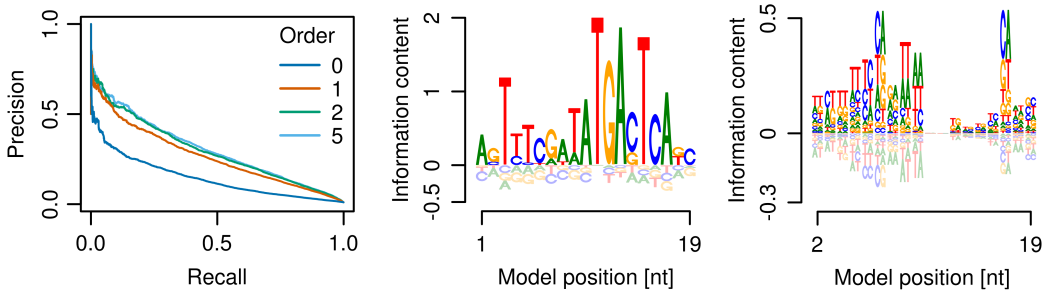

Figure S1. Modelling nucleotide dependencies in BATF binding motifs improves motif discovery and prediction. BATF models learned from ChIP-seq sites in GM12878 cells. Predictive performance (left) for BaMMs of increasing order. 0'th-order (middle) and 1'st-order (right) sequence logos of 2'nd-order BaMM.

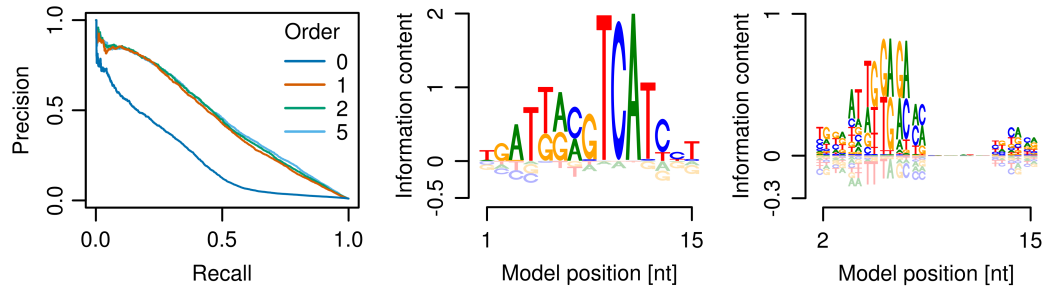

Figure S2. Modelling nucleotide dependencies in c-Jun binding motifs improves motif discovery and prediction. c-Jun models learned from ChIP-seq sites in HepG2 cells. Predictive performance (left) for BaMMs of increasing order. 0'th-order (middle) and 1'st-order (right) sequence logos of 2'nd-order BaMM.

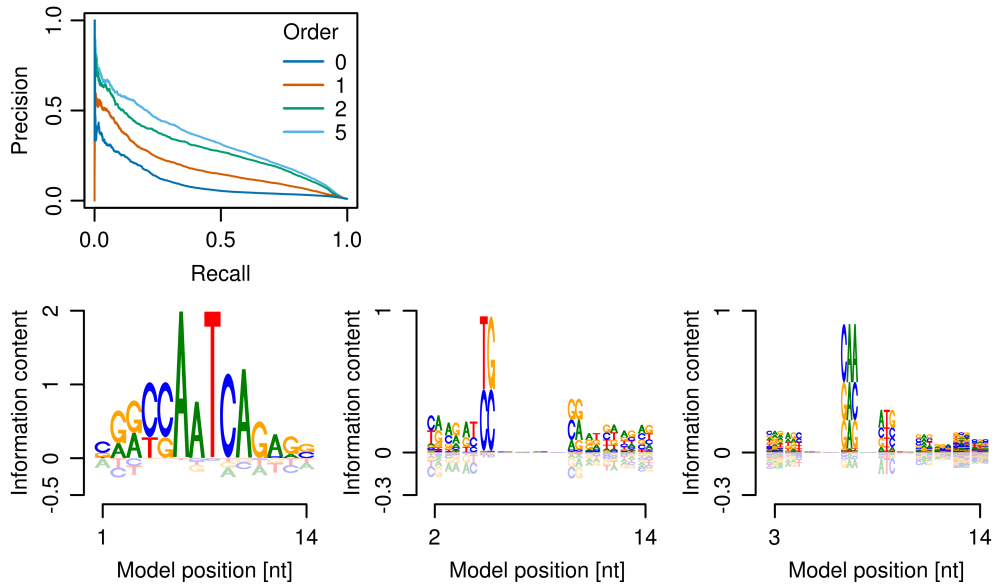

Figure S3. Modelling nucleotide dependencies in c-Fos binding motifs improves motif discovery and prediction. c-Fos models learned from ChIP-seq sites in K562 cells. Predictive performance (top) for BaMMs of increasing order. 0'th-order (left), 1'st-order (middle) and 2'nd-order (right) sequence logos (bottom) of 2'nd-order BaMM.

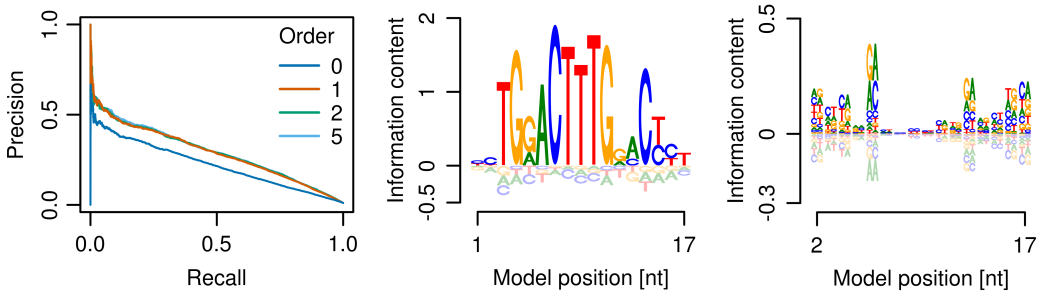

Figure S4. Modelling nucleotide dependencies in Hnf4a binding motifs improves motif discovery and prediction. Hnf4a models learned from ChIP-seq sites in HepG2 cells. Predictive performance (left) for BaMMs of increasing order. 0'th-order (middle) and 1'st-order (right) sequence logos of 2'nd-order BaMM.

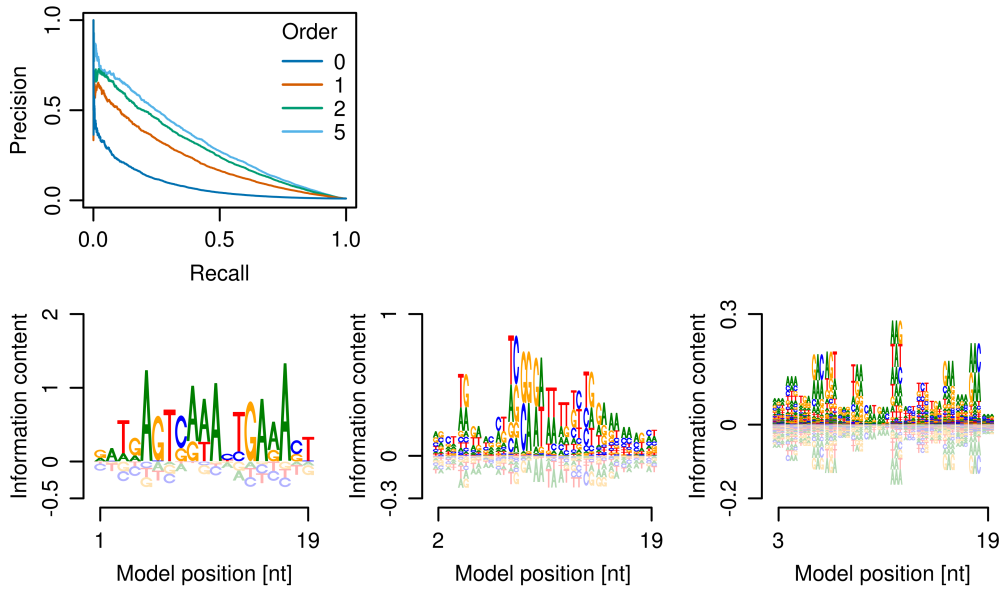

Figure S5. Modelling nucleotide dependencies in IRF4 binding motifs improves motif discovery and prediction. IRF4 models learned from ChIP-seq sites in GM12878 cells. Predictive performance (top) for BaMMs of increasing order. 0'th-order (left), 1'st-order (middle) and 2'nd-order (right) sequence logos (bottom) of 2'nd-order BaMM.

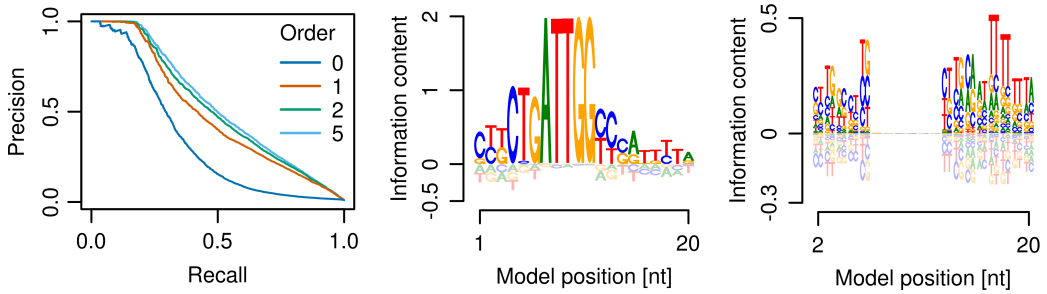

Figure S6. Modelling nucleotide dependencies in NF-YB binding motifs improves motif discovery and prediction. NF-YB models learned from ChIP-seq sites in K562 cells. Predictive performance (left) for BaMMs of increasing order. 0'th-order (middle) and 1'st-order (right) sequence logos of 2'nd-order BaMM.

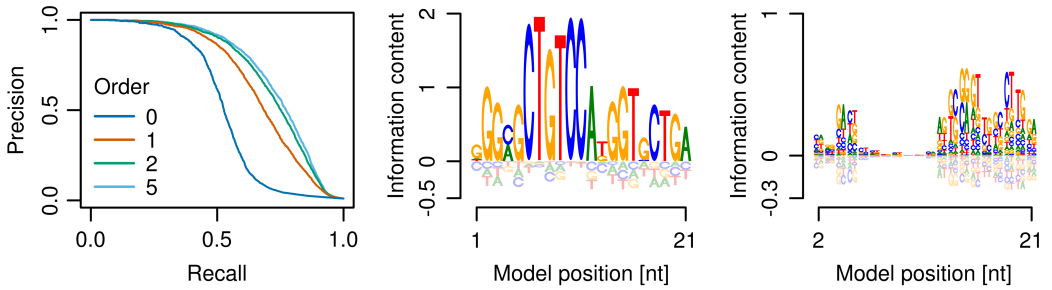

Figure S7. Modelling nucleotide dependencies in NRSF binding motifs improves motif discovery and prediction. NRSF models learned from ChIP-seq sites in PFSK-1 cells. Predictive performance (left) for BaMMs of increasing order. 0'th-order (middle) and 1'st-order (right) sequence logos of 2'nd-order BaMM.

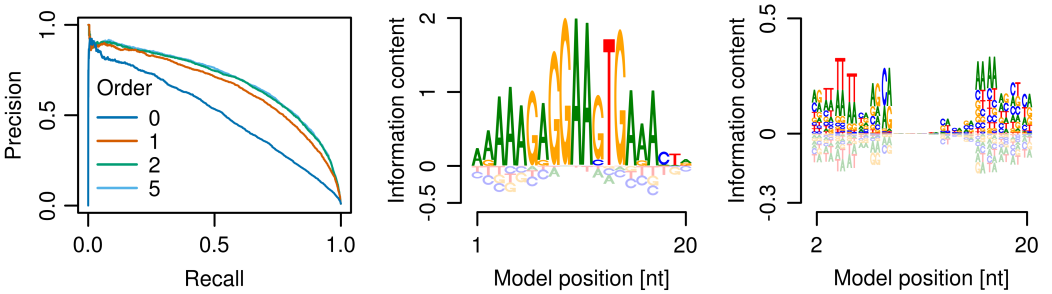

Figure S8. Modelling nucleotide dependencies in PU.1 binding motifs improves motif discovery and prediction. PU.1 models learned from ChIP-seq sites in GM12891 cells. Predictive performance (left) for BaMMs of increasing order. 0'th-order (middle) and 1'st-order (right) sequence logos of 2'nd-order BaMM.

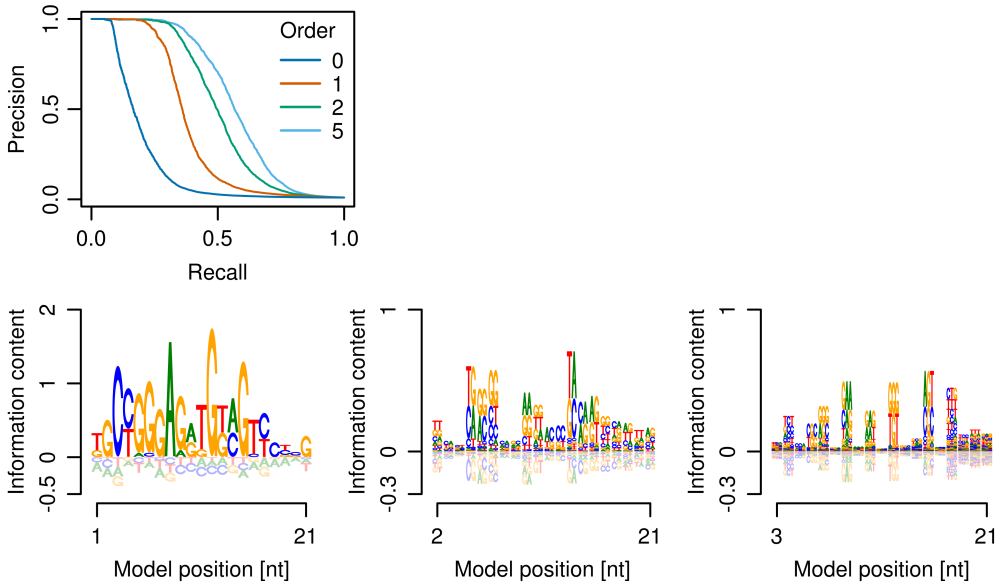

Figure S9. Modelling nucleotide dependencies in ZnF143 binding motifs improves motif discovery and prediction. ZnF143 models learned from ChIP-seq sites in the H1-hESC line. Predictive performance (top) for BaMMs of increasing order. 0'th-order (left), 1'st-order (middle) and 2'nd-order (right) sequence logos (bottom) of 2'nd-order BaMM.

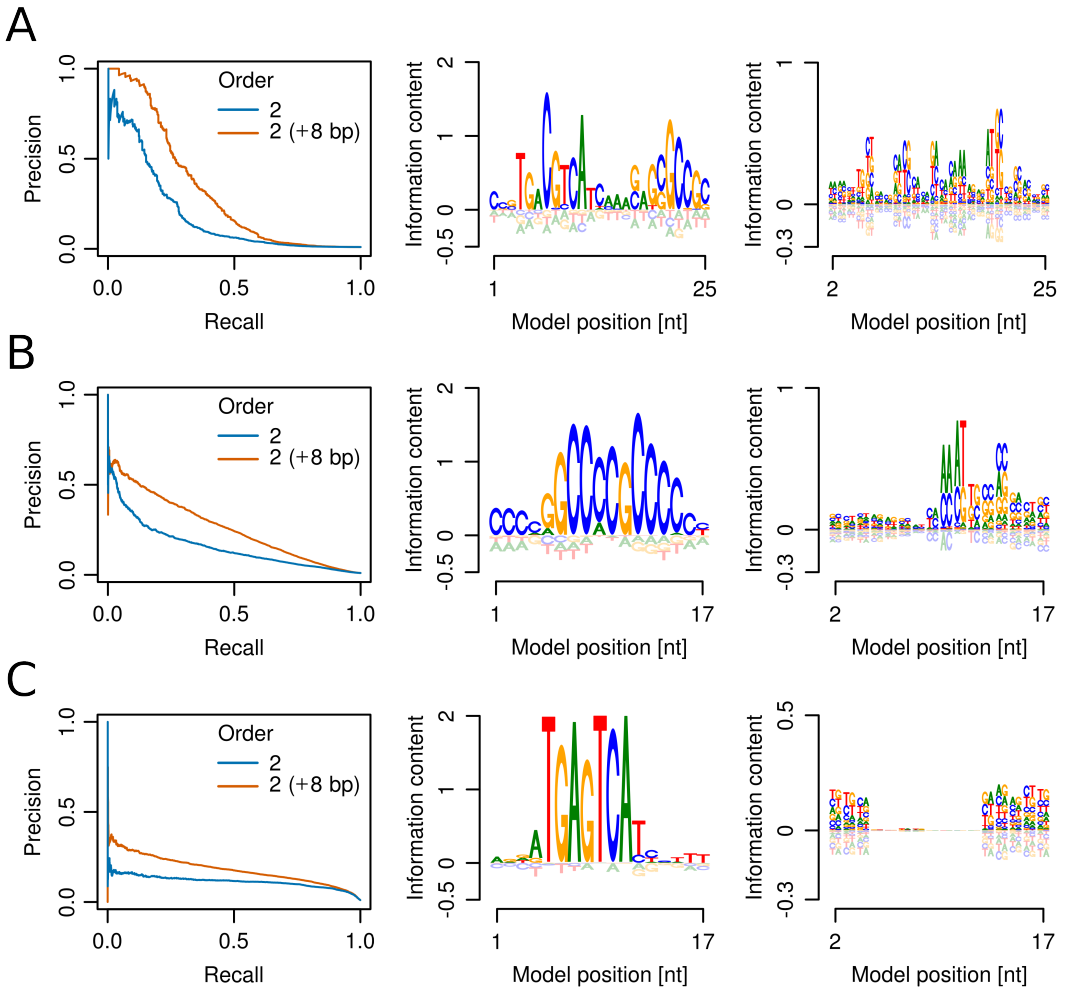

Figure S10. Nucleotides flanking the core binding sites of transcription factors contribute to the specificity of higher-order models. (**A**) GR models learned from ChIP-seq sites in HepG2 cells. Predictive performance (left) for 2'nd-order 8-bp-extended and unextended BaMMs. 0'th-order (middle) and 1'st-order (right) sequence logos of 2'nd-order 8-bp-extended BaMM. (**B,C**) Same as **A** but showing (**B**) IRF1 models learned in K562 cells and (**C**) c-Fos models learned in Mcf-10a cells.

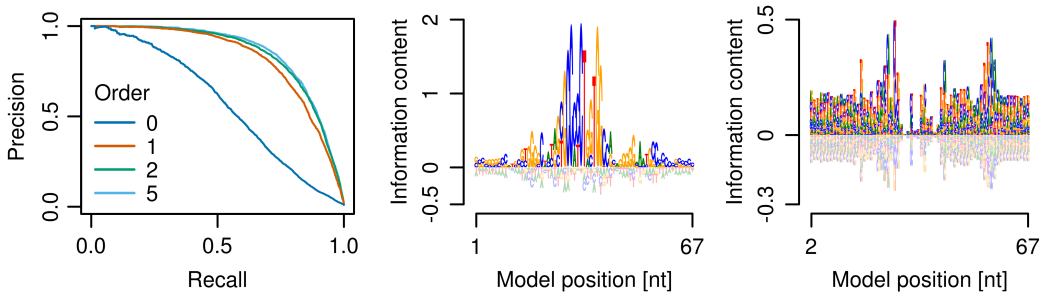

Figure S11. Nucleotides flanking the core CTCF binding site contribute to the specificity of higher-order CTCF models. CTCF models extended by 25 bp on either side, learned from ChIP-seq sites in Mcf7 cells. Predictive performance (left) for BaMMs of increasing order. 0'th-order (middle) and 1'st-order (right) sequence logos of 2'nd-order BaMM.

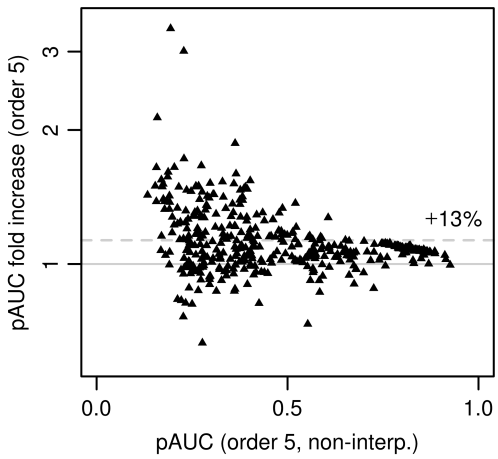

Figure S12. Robustness of BaMM learning at predicting transcription factor-DNA binding motifs. Factor of increase in performance (on log scale) of 8-bp-extended 5'th-order BaMMs versus non-interpolating (non-interp.) iMMs on 446 ChIP-seq datasets for transcription factors from ENCODE. For iMMs we set $\alpha_0 = 1$ and $\alpha_k = 5$ for $k \geq 1$, as this produced the best overall performance of iMMs. Dashed line: mean fold increase.

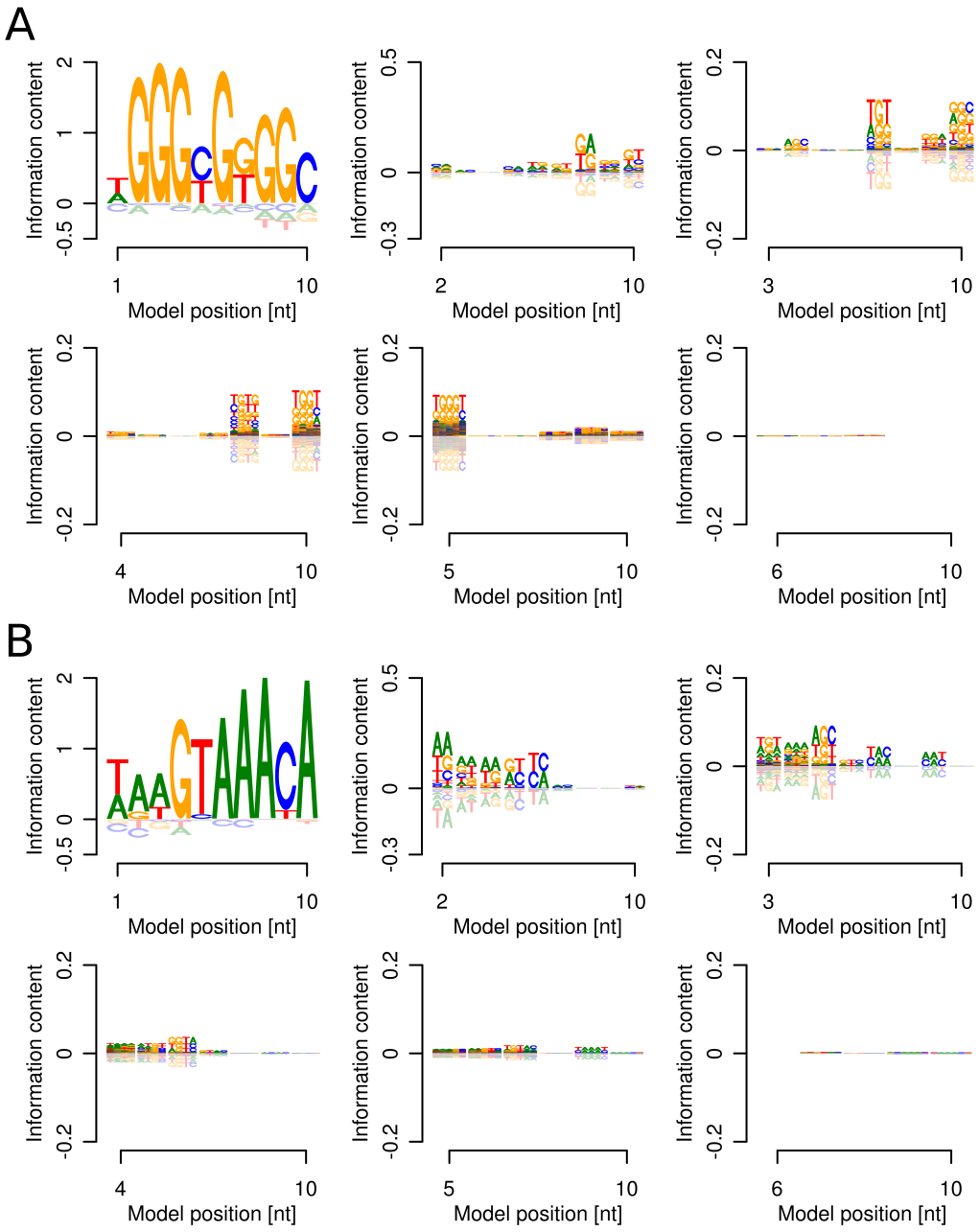

Figure S13. Higher-order sequence logos of pioneer transcription factor BaMMs. Sequence logos of 5'th-order BaMMs for (**A**) Klf4 and (**B**) FoxA2, shown from 0'th up to 5'th order (top left to bottom right).

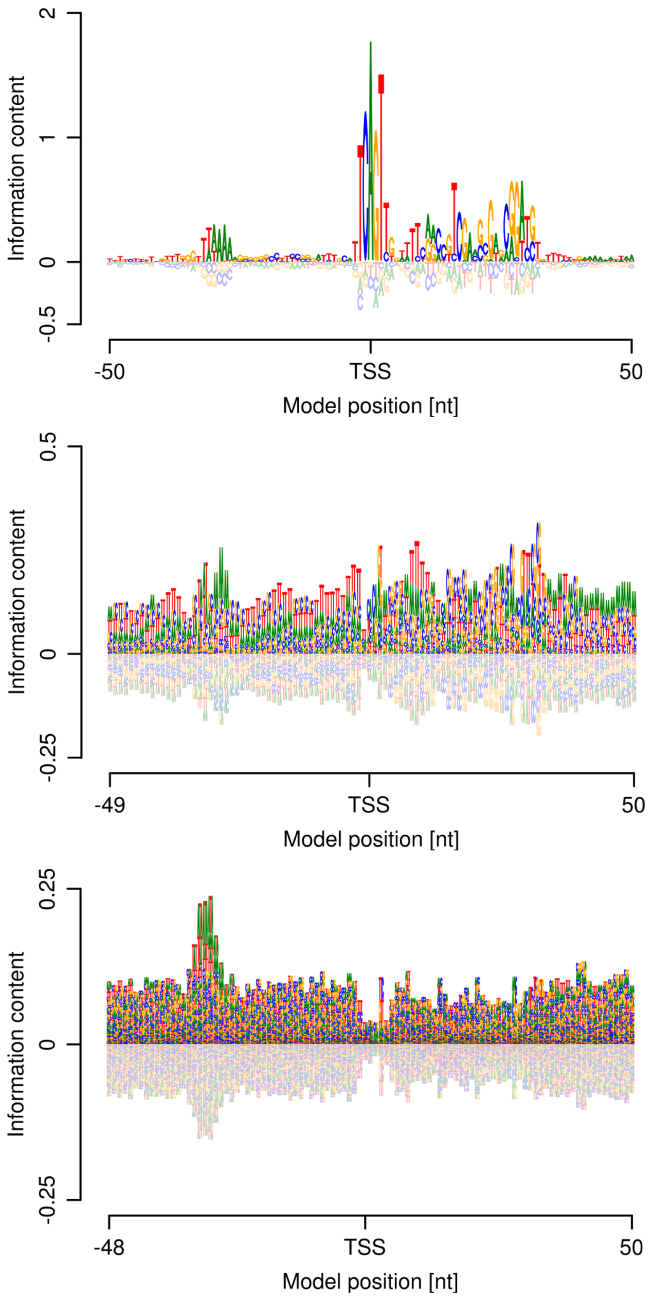

Figure S14. Higher-order sequence logos of NP core promoters from D. melanogaster. 0'th-order (top), 1'st-order (middle) and 2'nd-order (middle) sequence logos of 2'nd-order BaMM.

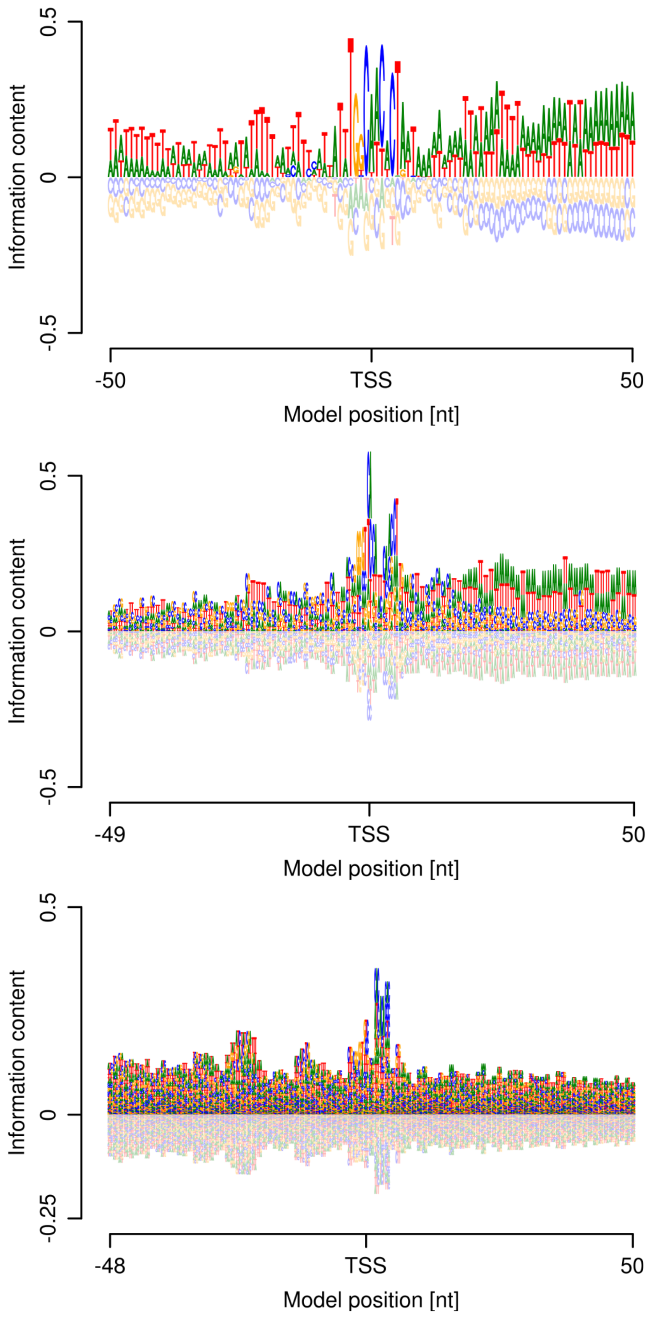

Figure S15. Higher-order sequence logos of BP core promoters from D. melanogaster. 0'th-order (top), 1'st-order (middle) and 2'nd-order (middle) sequence logos of 2'nd-order BaMM.

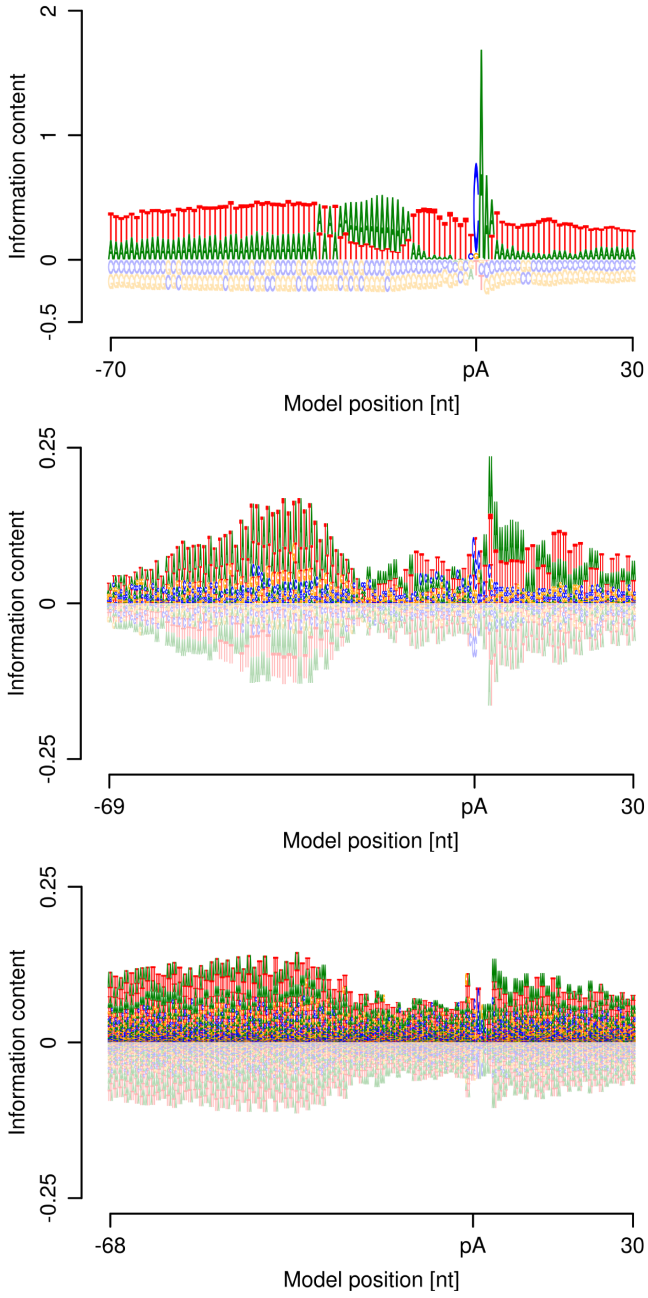

Figure S16. Higher-order sequence logos of pA sites from S. cerevisiae. 0'th-order (top), 1'st-order (middle) and 2'nd-order (middle) sequence logos of 2'nd-order BaMM.

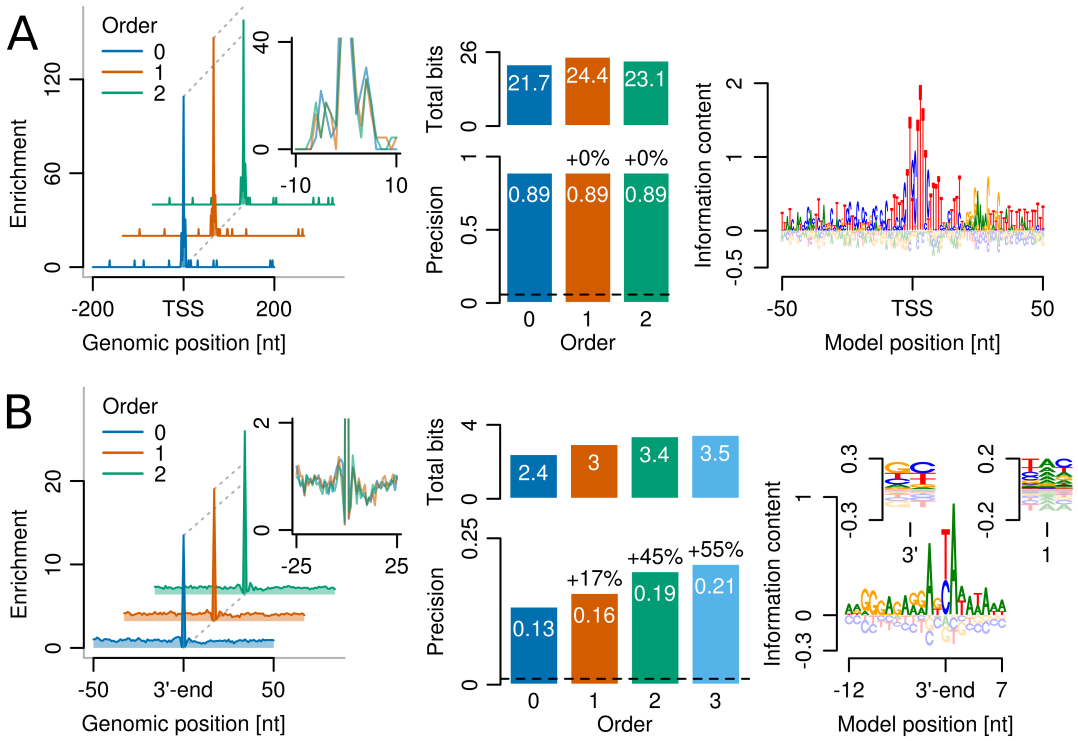

Figure S17. Performance of higher-order BaMMs at predicting RP gene core promoter and RNAP pause site motifs. (**A**) Same as Figure 5A but for TSSs of 92 RP gene core promoters from *D. melanogaster*. Correct predictions are defined to lie within 11 bp of measured TSSs. 0'th-order (right) sequence logo of 0'th-order BaMM. (**B**) Same as Figure 5D but for RNAP pause sites from *B. subtilis*. Correct predictions are within 0 bp of measured pause sites.

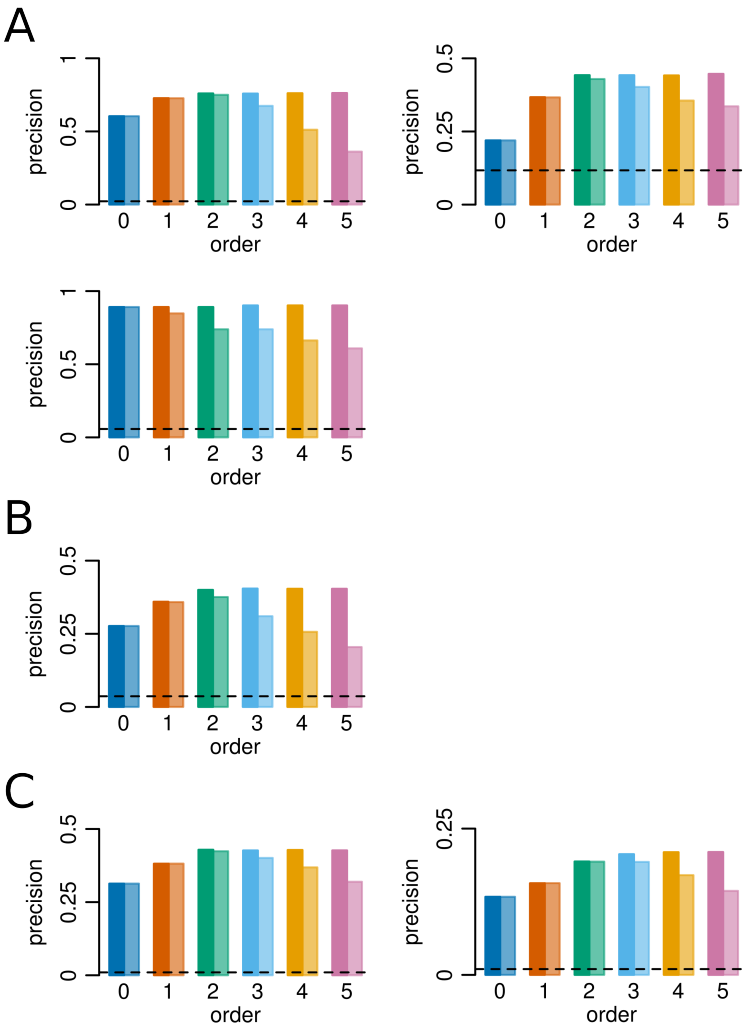

Figure S18. Robustness of BaMM learning at predicting complex, multipartite motifs. Same as precision barplots in Figure 5 and S17 but for BaMMs (dark bars) and iMMs (light bars) of increasing order. Precision of models for (**A**) NP (left), BP (right), and RP gene (bottom left) core promoters from *D. melanogaster*, (**B**) pA sites from *S. cerevisiae*, and (**C**) RNAP pause sites from *E. coli* (left) and *B. subtilis* (right). For iMMs we set $\alpha_0 = 1$ and $\alpha_k = 5$ for $k \geq 1$, as this produced the best overall performance of iMMs.

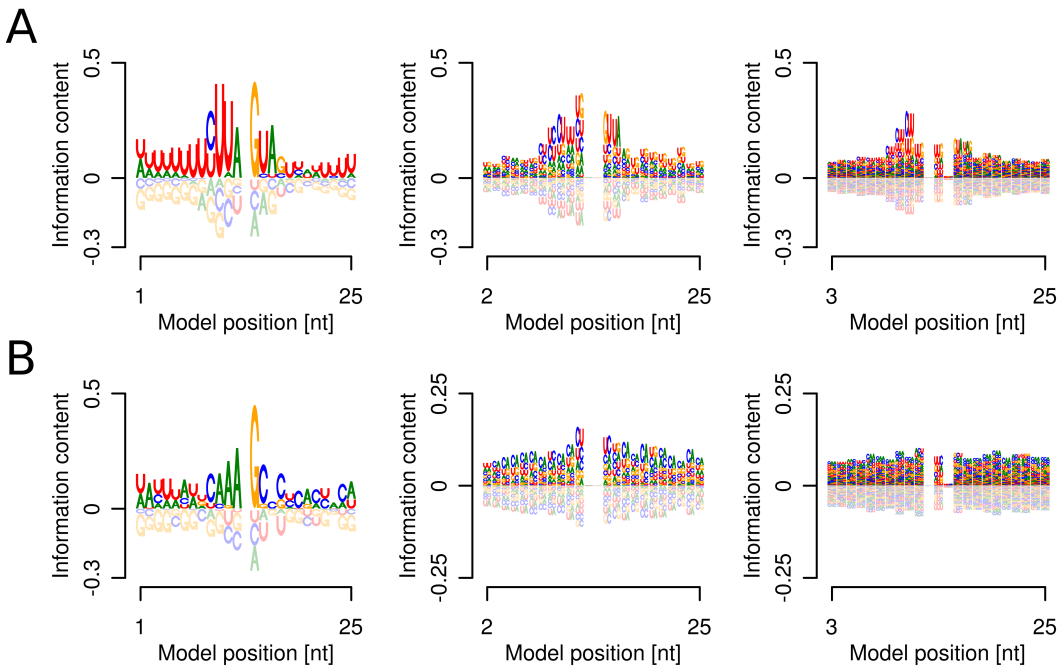

Figure S19. Higher-order protein-RNA binding specificity models. (**A**) 0'th-order (left), 1'st-order (middle), and 2'nd-order (right) sequence logos of 2'nd-order BaMM for Nab3 (central crosslinked U was removed from the 0'th-order logo). (**B**) Same as **A** but for Yra1.

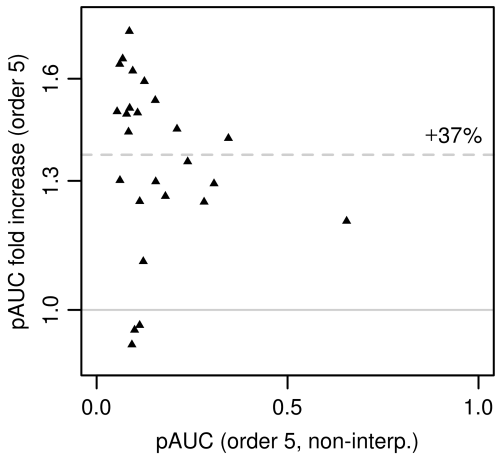

Figure S20. Robustness of BaMM learning at predicting mRNP biogenesis factor-RNA binding motifs. Factor of increase in performance (on log scale) of 5'th-order BaMMs versus non-interpolating (non-interp.) iMMs for 25 mRNP biogenesis factors from *S. cerevisiae* measured by PAR-CLIP. For iMMs we set $\alpha_0 = 1$ and $\alpha_k = 5$ for $k \geq 1$, as this produced the best overall performance of iMMs. Dashed line: mean fold increase.

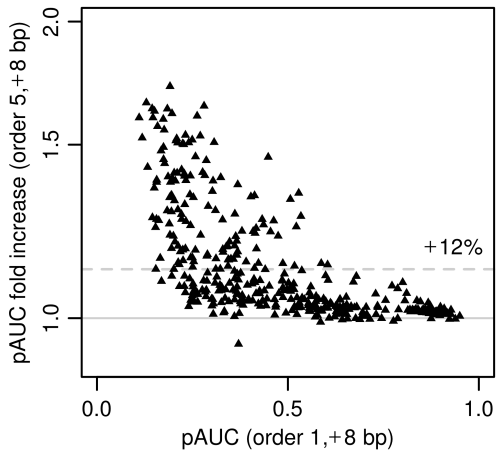

Figure S21. 5'th-order BaMMs improve 97 % of 1'st-order BaMMs trained on 446 ENCODE ChIP-seq datasets. Same as Figure 3C but showing the performance increase of 5'th-order 8-bp-extended BaMMs versus 1'st-order 8-bp-extended BaMMs. The average performance increase is 12 % (dashed line).

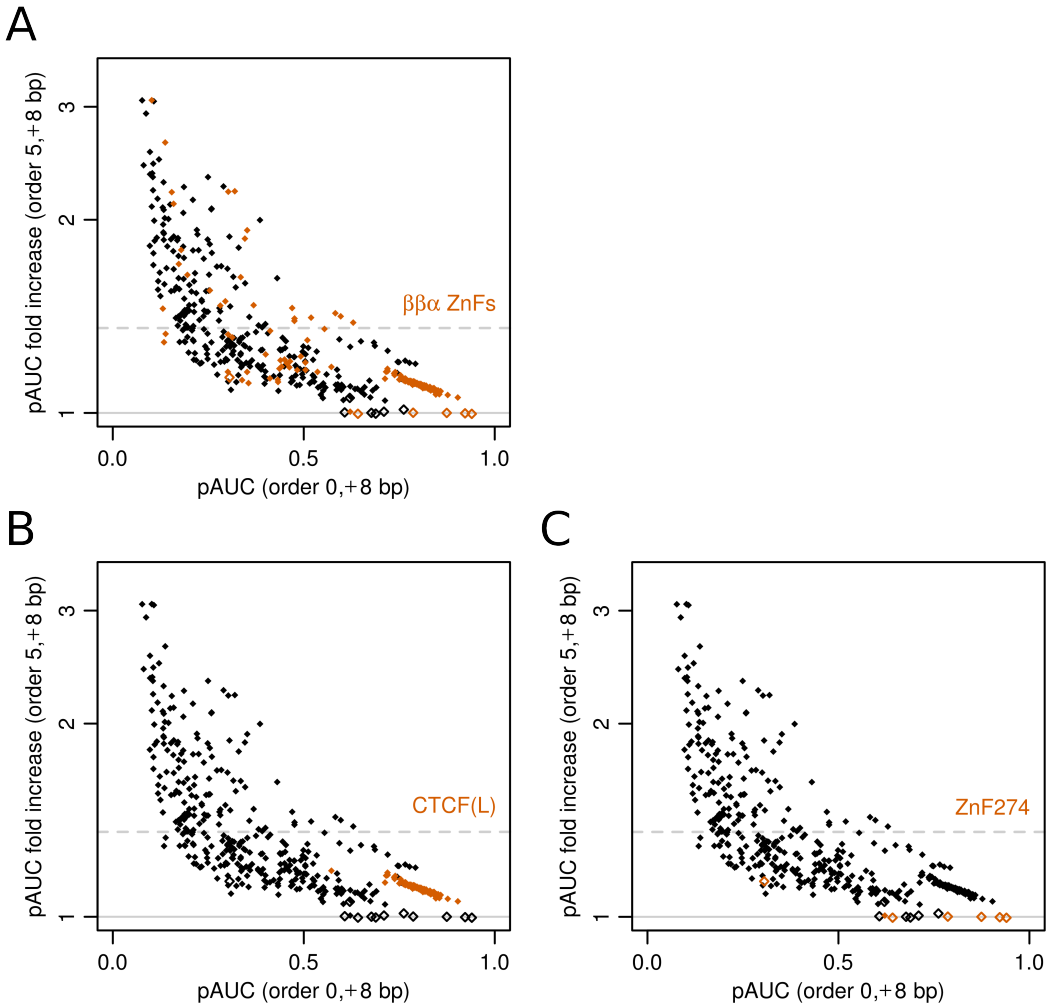

Figure S22. ***ββα*-type ZnF transcription factors profit as much as other factors from higher orders when models are relatively short (core binding sites extended by 2×4 bp).** (**A**) Same as Figure 3C but highlighting all *ββα*-type ZnF transcription factor datasets (`http://v1.factorbook.org`). (**B**) Same as **A** but highlighting only CTCF and CTCFL datasets. (**C**) Same as **A** but highlighting only ZnF274 datasets.