Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding

Daniel D. Le[a,1], Tyler C. Shimko[a,1], Arjun K. Aditham[b,c], Allison M. Keys[c,d], Scott A. Longwell[b], Yaron Orenstein[e], and Polly M. Fordyce[a,b,c,f,2]

[a]Department of Genetics, Stanford University, Stanford, CA 94305; [b]Department of Bioengineering, Stanford University, Stanford, CA 94305; [c]Stanford ChEM-H (Chemistry, Engineering, and Medicine for Human Health), Stanford University, Stanford, CA 94305; [d]Department of Chemistry, Stanford University, Stanford, CA 94305; [e]Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel POB 653; and [f]Chan Zuckerberg Biohub, San Francisco, CA 94158

Edited by David Baker, University of Washington, Seattle, WA, and approved February 23, 2018 (received for review September 22, 2017)

Transcription factors (TFs) are primary regulators of gene expression in cells, where they bind specific genomic target sites to control transcription. Quantitative measurements of TF–DNA binding energies can improve the accuracy of predictions of TF occupancy and downstream gene expression in vivo and shed light on how transcriptional networks are rewired throughout evolution. Here, we present a sequencing-based TF binding assay and analysis pipeline (BET-seq, for Binding Energy Topography by sequencing) capable of providing quantitative estimates of binding energies for more than one million DNA sequences in parallel at high energetic resolution. Using this platform, we measured the binding energies associated with all possible combinations of 10 nucleotides flanking the known consensus DNA target interacting with two model yeast TFs, Pho4 and Cbf1. A large fraction of these flanking mutations change overall binding energies by an amount equal to or greater than consensus site mutations, suggesting that current definitions of TF binding sites may be too restrictive. By systematically comparing estimates of binding energies output by deep neural networks (NNs) and biophysical models trained on these data, we establish that dinucleotide (DN) specificities are sufficient to explain essentially all variance in observed binding behavior, with Cbf1 binding exhibiting significantly more nonadditivity than Pho4. NN-derived binding energies agree with orthogonal biochemical measurements and reveal that dynamically occupied sites in vivo are both energetically and mutationally distant from the highest affinity sites.

protein–DNA binding | transcription factor binding | transcription factor specificity | microfluidics | transcriptional regulation

Gene expression is extensively regulated by transcription factors (TFs) that bind genomic sequences to activate or repress transcription of target genes (1). The strength of binding between a TF and a given DNA sequence at equilibrium depends on the change in Gibbs free energy (ΔG) of the interaction (2–5). Thermodynamic models that explicitly incorporate quantitative binding energies more accurately predict occupancies, rates of transcription, and levels of gene expression in vivo (4, 6–11). In addition, binding energy measurements for TF–DNA interactions can provide insights into the evolution of regulatory networks. Unlike coding sequence variants that manifest at the protein level to influence fitness, noncoding TF target site variants affect phenotype by modulating the binding energies of these interactions to affect gene expression (12–14). Understanding how TFs identify their cognate DNA target sites in vivo and how these interactions change during evolution therefore requires the ability to accurately estimate binding energies for a wide range of sequences.

Most high-throughput efforts to develop accurate models of TF binding specificity have focused on mutations within known TF target sites that dramatically change binding energies. However, even subtle changes in binding energies can have dramatic effects on both occupancy and transcription (15–18). Sequences surprisingly distal from a known consensus motif can affect affinities and levels of transcription (19–22), and genomic variants in regulatory regions outside of known transcription factor binding sites (TFBSs) may be subject to nonneutral evolutionary pressures (23). Therefore, understanding the fundamental mechanisms that regulate transcription requires measurement of binding energies at sufficient resolution to resolve even small effects.

Despite the utility of comprehensive binding energy measurements, existing methods often lack the energetic resolution and scale required to yield such datasets. Currently, three dominant technologies are used to query DNA specificities: methods based on systematic evolution of ligands by exponential enrichment (SELEX) (24–29), protein binding microarrays (PBMs) (30, 31), and mechanically induced trapping of molecular interactions

Significance

Transcription factors (TFs) are key proteins that bind DNA targets to coordinate gene expression in cells. Understanding how TFs recognize their DNA targets is essential for predicting how variations in regulatory sequence disrupt transcription to cause disease. Here, we develop a high-throughput assay and analysis pipeline capable of measuring binding energies for over one million sequences with high resolution and apply it toward understanding how nucleotides flanking DNA targets affect binding energies for two model yeast TFs. Through systematic comparisons between models trained on these data, we establish that considering dinucleotide (DN) interactions is sufficient to accurately predict binding and further show that sites used by TFs in vivo are both energetically and mutationally distant from the highest affinity sequence.

Author contributions: D.D.L., T.C.S., A.K.A., and P.M.F. designed research; D.D.L., T.C.S., A.K.A., and A.M.K. performed research; D.D.L., T.C.S., A.K.A., S.A.L., and Y.O. contributed new reagents/analytic tools; D.D.L. and T.C.S. analyzed data; and D.D.L., T.C.S., and P.M.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, https://www.ncbi.nlm.nih.gov/geo (accession no. GSE111936). All processed data have been uploaded to Figshare (https://figshare.com/articles/_/5728467) and Github (https://github.com/FordyceLab/BET-seq).

[1]D.D.L. and T.C.S. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: pfordyce@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1715888115/-/DCSupplemental.

Published online March 27, 2018.

Downloaded at Palestinian Territory, occupied on December 23, 2021

www.manaraa.com

(MITOMI) (32, 33). SELEX-based methods require repeated enrichment and amplification cycles followed by low depth sequencing of the enriched TF-bound material. Thus, these methods are optimized to identify the highest affinity substrates from extremely large random populations but fail to detect weakly bound sequences or measure their affinities. Although PBMs quantify the binding of TFs to DNA microarrays using a fluorometric readout with a broad dynamic range, the precise relationship between measured fluorescence intensities and binding energies is unclear because PBMs require wash steps that disrupt binding equilibrium. In addition, PBM arrays usually contain many replicates for each sequence, limiting the number of distinct sequences probed to ∼50,000. The MITOMI platform, based on mechanical trapping by microfluidic valves, enables high-resolution measurements of concentration-dependent binding to yield absolute affinities, but throughput is limited to several hundred sequences. Recent iterations of MITOMI-based assays have addressed throughput limitations by using massively parallel sequencing to increase sequence space coverage but at the cost of resolving binding energies (29, 34). High-throughput sequencing–fluorescent ligand interaction profiling (HiTS-FLIP) couples massively parallel sequencing with the ability to perform concentration-dependent binding measurements; however, adoption of this technology has been limited by the requirement for extensively customized sequencing hardware (35). Taken together, these TF–DNA binding assays can sample vast sequence spaces, but it remains challenging to simultaneously measure binding energies at the scale and resolution necessary to derive complete landscapes.

The most popular and widely used models represent TF specificities as a position weight matrix (PWM), in which each nucleotide at each position contributes additively and independently to overall binding energies (36). These mononucleotide (MN) models are easily implemented, visualized, and interpreted and provide useful approximations of binding specificity for the majority of studied TFs (9, 37–39). However, PWM-based models fail to capture nonadditivity between nucleotides, which can lead to inaccurate predictions, particularly for low-affinity sites (32, 40). This approximation can be refined by including contributions of higher order sequence features, such as dinucleotides (DNs) or longer k-mers (41–50). Recently developed models predict binding based on local DNA shape using structural parameters (e.g., minor groove width, propeller twist, helical twist, and roll) (51–55). However, these biophysical variables are determined by primary sequence, rendering the relationship between the two somewhat degenerate. Deep neural network (NN)-based models can learn complex patterns from large datasets across many applications, including predicting the function of noncoding genomic sequences (56). Training NN models on large sets of binding data therefore has the potential to yield accurate, high-resolution estimates of binding at a per-sequence level, revealing local topography of binding energy landscapes.

To address the need for technologies capable of high-throughput thermodynamic measurements, we developed BET-seq (Binding Energy Topography by sequencing), an integrated sequencing assay and analysis pipeline that yields relative and absolute binding energies ($\Delta\Delta G$ and ΔG, respectively) for >1 million sequences in parallel, even for relatively small energetic differences. Using Monte Carlo simulations to mimic the effects of stochastic sampling noise on energetic resolution, we establish guidelines for the sequencing depth required to resolve accurate binding energies for libraries of different sizes and expected energy ranges. We then deployed this assay to measure comprehensive and quantitative binding energy landscapes for >1 million mutations surrounding the known consensus motif for two model yeast TFs (Pho4 and Cbf1). Deep NN models that incorporate all possible higher order, nonadditive contributions were then trained on these large datasets to yield high-resolution estimates of binding energy for each sequence. Comparisons to orthogonal biochemical affinity measurements established that NN predictions are highly quantitative, accurately predicting measured binding energies over a range of 3 kcal/mol. A surprisingly large number of flanking sequences have effects on binding energies as great or greater than mutations in the core, suggesting that current definitions of TFBSs are too restrictive and may limit accurate predictions of TF occupancy in vivo. Comparisons between NN predictions and predictions from a series of biophysically motivated models reveal that DN specificity preferences explain nearly all observed binding behavior, with Cbf1 exhibiting significantly more nonadditivity than Pho4. Strikingly, most dynamically occupied target loci for both Pho4 or Cbf1 are mutationally distant from the energetically optimal flanking sequence, providing evidence of evolutionary molecular selection for near-neutral effects on binding energies. These data demonstrate the utility of our high-throughput approach to measure binding energies and model determinants of substrate specificity required to understand biological behaviors. Furthermore, this assay and analysis pipeline may be extended to a wide variety of TFs, improving predictive models of TF–DNA affinities across species.

Results

A Microfluidic Approach Using High-Throughput Sequencing (HTS) to Derive Comprehensive Binding Affinity Landscapes. We sought to develop an assay that significantly extends the scale at which TF–DNA interactions can be probed while maintaining the ability to quantitatively measure binding energies at high resolution. TF–DNA interactions can be considered a two-state system, such that the affinity of a given interaction can be determined by the equilibrium partitioning of sequences into bound and unbound states:

$$\Delta G = -RT \ln\left(\frac{[TF \cdot DNA_{bound}]}{[TF_{unbound}][DNA_{unbound}]}\right)$$

.

Several groups have established that molecular counting of individual DNA molecules via HTS can reliably measure bound and input concentrations for each species (10, 40, 57). These assays generally use electromobility shift assays (EMSAs) to isolate bound material. However, TF–DNA complexes are not at chemical equilibrium during the electrophoresis step, and complexes with particularly fast dissociation rates may be underrepresented within the bound fraction, leading to a systematic underestimation of weak affinity interactions (58). To address this, we used a microfluidic device incorporating pneumatic valves with fast (∼100 ms) actuation times to mechanically "trap" DNA associated with TF proteins at equilibrium (29, 32–34) (Fig. 1*A*). This device requires small amounts of DNA substrate and expressed protein, eliminating the need for cell-based protein production. Antibody-patterned surfaces within the device capture monomeric enhanced GFP (meGFP)-tagged TFs produced via in vitro transcription/translation before washing, purifying TFs in situ. After TF capture, libraries of DNA sequences are introduced and allowed to interact with surface-immobilized TFs until equilibrium is reached. Mechanical valves then sequester TF-bound DNA sequences, making it possible to wash out unbound material without loss of weak interactions (32, 33, 59) (Fig. 1*B*). Bound DNA species can then be eluted from the device and quantified using HTS (Fig. 1*C*).

As a first application of BET-seq, we focused on two model TFs from *Saccharomyces cerevisiae*, Pho4 and Cbf1. Although Pho4 and Cbf1 bind the same CACGTG variant of the six-nucleotide enhancer-box (E-box) motif both in vitro and in vivo (16, 18, 28, 60–62), they bind largely nonoverlapping sets of genomic loci and regulate distinct target genes (19, 63). To comprehensively probe how flanking nucleotides affect Pho4

Downloaded at Palestinian Territory, occupied on December 23, 2021

www.manaraa.com

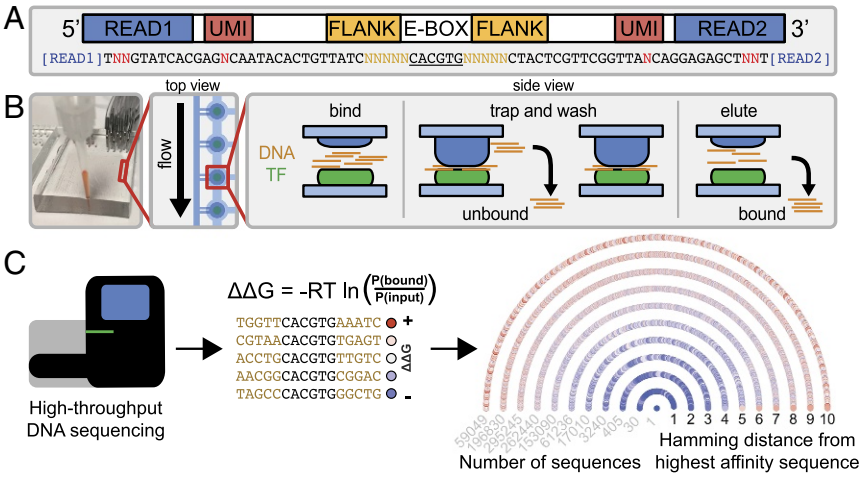

Fig. 1. DNA library design and assay overview. (*A*) Schematic of flanking sequence library design indicating Illumina sequencing adapters (blue), unique molecular identifiers (UMIs, red), variable flanking regions (orange), and E-box consensus. (*B*) Photograph of MITOMI device and schematic showing device operation. (*C*) Schematic showing downstream sample analysis. Counting of individual molecules within bound and input fractions allows calculation of relative binding energies for each sequence (*Left* and *Middle*) to yield a comprehensive thermodynamic binding affinity landscape (Pho4 example). Each color-coded point (blue = $-\Delta\Delta G$; red = $+\Delta\Delta G$) represents a sequence, grouped by Hamming distance from the highest affinity sequence and alphabetically ordered in clockwise polar coordinates.

and Cbf1 binding affinities, we designed a library of 1,048,576 sequences in which the core E-box sequence was flanked by all possible random combinations of five nucleotides upstream and downstream, embedded within a constant sequence shown to exhibit negligible binding (33) (Fig. 1*A*). Constant sites at the 5′ and 3′ ends allowed simultaneous PCR amplification and incorporation of Illumina adapters. UMIs included within each sequence allowed accurate counting of library species even in the presence of PCR bias (64). Each UMI barcode was segmented and interspersed along the library sequence to prevent formation of an additional CACGTG consensus site. After sequencing, relative binding affinities ($\Delta\Delta$Gs) were calculated for all sequences by considering relative enrichment of individual DNA species, thereby generating a comprehensive binding affinity landscape (example shown in Fig. 1*C*).

Assay Simulations Determine Sequencing Depth Requirements for Binding Affinity Measurements. Accurately estimating concentrations of DNA in TF-bound and input samples via sequencing requires that measured read counts reflect true abundances. However, read counts can be distorted by stochastic sampling error, particularly for low read count numbers (65, 66). To understand how stochastic sampling error depends on read depth, library size, and the expected range of binding energies across library sequences, we considered a previously published experiment that quantified interactions between the *Escherichia coli* LacI repressor and a library of 1,024 binding site variants via deep sequencing (40). Each sequence was sampled with roughly 10^3 reads per species, yielding $\Delta\Delta$G measurements with negligible sampling noise. To understand how read depth affects recovery of accurate $\Delta\Delta$G measurements, we down-sampled these data to simulate lower sequencing depths of 10^2-10^6 reads, split evenly between bound and input fractions (*ca.* 0.05–5,000 reads per sequence). We then assessed the accuracy of recovered $\Delta\Delta$G values at these lower sequencing depths by calculating the squared Pearson's correlation coefficient (r^2) between $\Delta\Delta$G values for each species calculated from down-sampled data and published values calculated from the full dataset. To minimize the effect of a few high accuracy values dominating the correlation statistic, each r^2 was normalized by the fraction of observed species (Fig. 2*A*). For this 1,024-species library with binding energies that span ~3 kcal/mol, $\sim 2 \times 10^5$ total reads (~100 reads per sequence) were sufficient to recover highly accurate $\Delta\Delta$G values for every sequence.

Measuring accurate $\Delta\Delta$Gs for a 1,048,576 species library represents a 1,000-fold increase in scale from these prior experiments. To understand more generally the determinants of $\Delta\Delta$G measurement accuracy, we generated a simulated test set of true relative binding energies and implemented Monte Carlo simulations to mimic stochastic sampling during HTS. For given library sizes, sequencing depths, and binding energy ranges, we again calculated the correlation coefficient between calculated $\Delta\Delta$G values and true values. As expected, accuracy improves and library coverage expands with increasing sequencing depth (Fig. 2*B* and *SI Appendix*, Figs. S1 and S2). As the range of expected binding energies increases, accuracy improves but the fraction of sequences observed from the input library decreases. Nearly all existing motif discovery libraries used in SELEX-seq, MITOMI-seq, and SMiLE-seq experiments probe on the order of 10^{18}–10^{24} species with read depths of several thousand total reads. These simulations establish that such sparse sequencing will sample only the highest affinity sequences, representing an infinitesimal fraction of the input library.

To delineate conditions under which unbound concentrations can be approximated by sequencing the input library, reducing assay cost, we modeled the distribution of "apparent" per-sequence $\Delta\Delta$Gs for a given population of sequences under competitive binding conditions in which we explicitly consider effects of ligand depletion. These simulations consider the total number of library sequences, respective concentrations of the TF and the DNA library, and expected range and distribution of $\Delta\Delta$G values and return predicted equilibrium concentrations of each species within the bound and unbound fractions (*SI Appendix*, Fig. S3). To make these simulations computationally feasible, we modeled 100 species uniformly distributed across the $\Delta\Delta$G range with a single high concentration "dummy" substrate to represent the majority of species. As the $\Delta\Delta$G range and number of species increases, species become depleted from the unbound fraction (*SI Appendix*, Fig. S3), causing $\Delta\Delta$G values estimated from sequencing input to systematically underestimate true $\Delta\Delta$G values for high affinity interactions (*SI Appendix*, Fig. S4). However, under the conditions used here (30 nM and 1 μM concentrations for TF and DNA libraries, respectively, with 1,048,576 species and a range

Downloaded at Palestinian Territory, occupied on December 23, 2021
www.manaraa.com

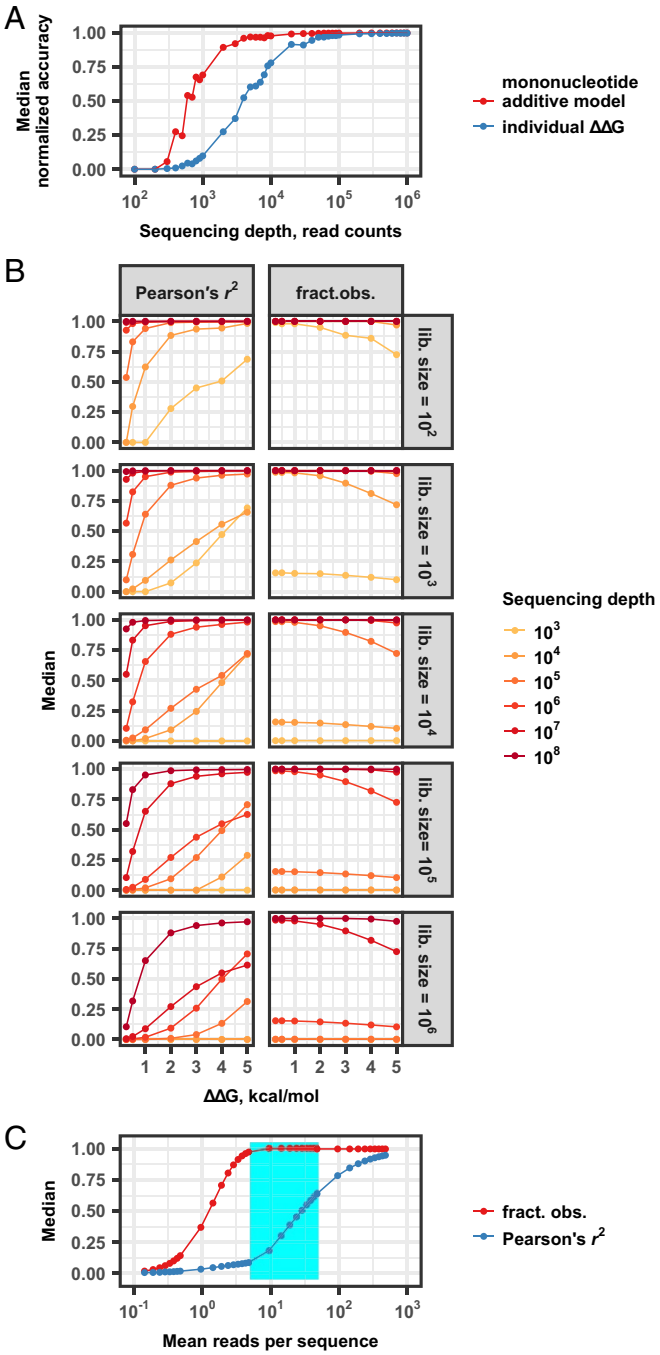

Fig. 2. Probing the relationship between assay accuracy, read depth, library size, and energy range. (*A*) Normalized median accuracy (Pearson's $r^2 \times f$, where f is the fraction of observed species) comparing results for down-sampled data with "true" values as a function of read depth for MN model coefficients (red) (40) and individual ΔΔG values (blue). (*B*) Median squared Pearson's correlation coefficients between recovered and true values (r^2, *Left*) and median fraction of observed species (*Right*) as a function of binding affinity range for various library sizes (rows) sequenced to different depths. (*C*) Median squared Pearson's correlation coefficient (r^2) between recovered and true values (blue) and fraction of observed species (red) as a function of mean reads per sequence for 10 replicate simulations; cyan rectangle denotes flank library assay conditions presented here.

of ΔΔGs < 4kcal/mol), this approximation is justified. Calculated assay accuracies ($r^2 \times fraction\ of\ library\ observed$) for simulated HTS of these distributions reveal two regimes with decreased accuracy (*SI Appendix*, Fig. S5): Large libraries with small ΔΔG ranges are subject to sequencing-associated counting error, while libraries with large ΔΔG ranges are subject to sampling error and effects of ligand depletion.

Previous observations of concentration-dependent Pho4 and Cbf1 binding to E-box motifs with mutations in the first flanking nucleotide revealed differences in affinities spanning ~1 kcal/mol (32). While these observations may not reflect the full energetic range of binding for the larger library queried here, they provide a reasonable initial estimate given that the most proximal flanking nucleotide likely has the largest impact on binding. To guide sequencing assays, we examined in detail simulations sampling a 1,048,576-member library with this energy range at mean read depths per species of 10^{-1}–10^3 (10^3–10^8 total reads) (Fig. 2*C*). Although 95% of sequences can be recovered from as few as 4–5 reads per species, high read depths of ~10^2 counts per species (10^8 total reads per TF) are required to yield individual ΔΔG measurements with accuracies of ~80% and errors of ~0.2 kcal/mol.

Modeling Specificity from Noisy Individual Measurements Improves Assay Resolution. Very high-depth sequencing may be cost-prohibitive for studies involving many TFs or when considering large DNA libraries. In those scenarios, modeling can be used to infer determinants of binding specificity while minimizing stochastic sampling noise. To illustrate the power of this approach, we again considered the published LacI repressor dataset (40). Although ~10^2 reads per sequence were required for accurate ΔΔG estimates, 10^1 to 10^2-fold fewer reads per sequence allowed generation of additive MN PWM models with similar predictive power to those generated from the entire dataset (Fig. 2*A*). However, while PWMs predict high affinity binding, they fail to explain variance among lower affinity target sites with high sequence diversity (32, 40).

A NN trained on millions of noisy per-sequence measurements can capture all measurable higher order complexity, yielding a high-resolution model capable of accurately predicting binding over a wide range of energies. However, this increased predictive power comes at the cost of interpretability. To improve the accuracy of our energetic estimates while preserving the ability to gain mechanistic insights, we applied an integrated measurement and modeling approach (Fig. 3*A*). First, we collected millions of sequencing-based estimates of per-sequence ΔΔGs. Next, we trained a NN model on these sequencing data to obtain high-resolution energetic predictions for each substrate that capture the effects of all higher order nonadditive interactions among nucleotides. Finally, we parsed and quantified the biophysical mechanisms responsible for observed TF–DNA binding by systematically comparing correlations between predictions made by the NN model and biophysically motivated linear models (MN, nearest neighbor DN, and all DN models). This integrated scheme yielded a binding energy landscape of unprecedented scale and energetic resolution and allowed dissection of the biophysical mechanisms responsible for Pho4 and Cbf1 specificity.

High-Throughput, Comprehensive Estimates of Absolute Binding Affinities for Pho4 and Cbf1. We used the BET-seq assay and DNA library described above to acquire four replicate measurements of Pho4 and three of Cbf1 at sequencing depths ranging from ~5 to 50 million reads allocated to either bound or input samples (*SI Appendix*, Table S1). For each experiment, ΔΔGs were calculated for each sequence from the measured ratio of bound to input read counts (*SI Appendix*, Fig. S6). As predicted, measured per-sequence ΔΔGs between two experiments at low read depth (*ca.* 6–8 million limiting counts) show no correlation; at higher read depths (~24 million limiting counts), this correlation increases to $r^2 = 0.67$ (*SI Appendix*, Fig. S7 and Table S2). To further improve resolution, we trained a NN regression

Downloaded at Palestinian Territory, occupied on December 23, 2021
www.manaraa.com

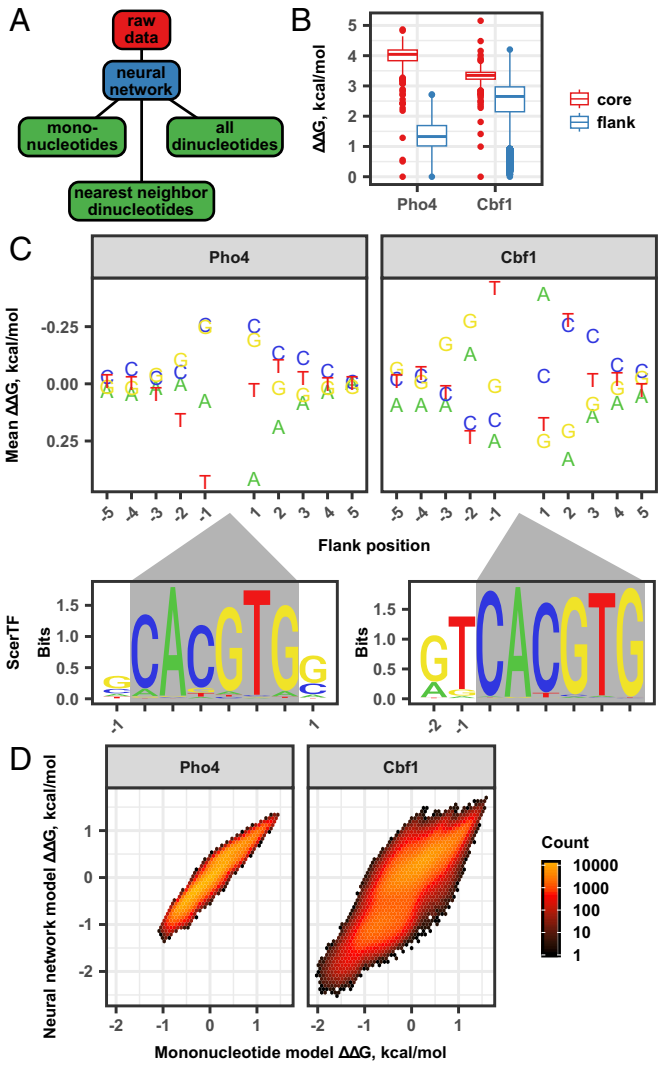

Fig. 3. Modeling and interpretation of binding specificity based on MN features. (*A*) Data analysis diagram shows NN modeling from raw data followed by model interpretation. (*B*) Magnitude of energetic changes (ΔΔGs) for core (red) (32) and flanking (blue) mutations for Pho4 and Cbf1. (*C*) Pho4 and Cbf1 mean MN ΔΔG values as a function of flanking sequence position (*Top*), compared with the ScerTF database (67–70) sequence logos (*Bottom*). Gray boxes show position of core consensus CACGTG. (*D*) Density scatter plots comparing NN model estimates and MN additive model predictions based on those estimates.

model that predicted the measured ΔΔG for each flanking sequence. Accuracy of the NN against observed training data and an unobserved validation dataset was recorded throughout training; training was stopped once accuracy against the validation dataset failed to improve, protecting against overfitting to the training data (*SI Appendix*, Fig. S8). Predictions from NN models trained on the two high-depth Pho4 replicates showed excellent correlation (r^2 = 0.94), validating the ability to apply such models to derive accurate, reproducible estimates of binding energies. For all analyses moving forward, we therefore use per-sequence ΔΔG estimates output from the NN trained on a composite dataset of all replicates (*SI Appendix*, Fig. S9).

Absolute binding energies and dissociation constants (ΔG and K_d, respectively) allow direct comparison between different TFs and across experimental platforms and further enable quantitative predictions of TF occupancy in vivo under known cellular conditions. However, sequencing-based measurements of ΔGs from sparse datasets can underestimate the true affinity range due to systematic undersampling of bound reads for low-affinity sequences. In addition, the NN is trained only on relative binding affinities (ΔΔGs) and therefore cannot return estimates of absolute energies (ΔGs). NN-derived ΔΔG estimates can be projected onto an absolute scale by calibrating to a set of high-resolution biochemical measurements of ΔGs with a linear scaling factor and offset. To generate a set of high-confidence ΔGs, we measured concentration-dependent binding for surface-immobilized Pho4 and Cbf1 TFs interacting with all single-nucleotide variants of AGACA_TCGAG, a medium affinity reference flanking sequence (where the underscore indicates the CACGTG core motif), via traditional fluorometric MITOMI (*SI Appendix*, Figs. S10 and S11). For each sequence, observed binding was globally fit to a single-site binding model, yielding both K_ds and ΔGs (*SI Appendix*, Table S3). All NN values were then scaled by fit parameters returned from a linear regression between NN predictions and experimental measurements for these sequences. Median K_d values for all flanking library sequences were 100 and 63 nM for Pho4 and Cbf1, respectively, in agreement with prior work (32). Strikingly, flanking sequence variation can modulate K_d values by over two orders of magnitude, ranging between 11–1,036 nM and 1–866 nM, respectively. In some cases, the magnitude of these effects exceeds that of mutations within the CACGTG core consensus (Fig. 3*B* and *SI Appendix*, Fig. S12), demonstrating the importance of flanking sequences to specificity.

Pho4 and Cbf1 Flanking Preferences Extend Far Beyond the Known Consensus Sequence. To understand the biophysical features that contribute to the predictive performance of the NN model, we generated PWMs (71), which estimate the mean energetic contribution of each nucleotide at each position, from the full set of scaled NN-predicted ΔΔG values (Fig. 3*C*). While the assumption of additivity fails to explain all specificity, these models offer a close approximation (11, 44) and PWMs are easily visualized and interpreted. These MN model results confirm that positions proximal to the E-box core motif exhibit the largest mean effect on binding, in agreement with PWMs generated by orthogonal techniques (67–69) (Fig. 3*C*). However, nucleotides up to four and five positions from the consensus contribute to specificity for Pho4 and Cbf1, respectively, significantly farther than previously reported.

To quantitatively assess the degree to which MN features dictate binding behavior, we determined the proportion of NN-derived per-sequence ΔΔG variance explained by a simple PWM (Fig. 3*D*). If MN models capture all determinants of observed specificity, PWM predictions should explain all of the variance in NN-derived ΔΔG values; conversely, discrepancies may indicate the presence of higher order interactions. PWMs explained a majority of the variance in NN predictions (r^2 = 0.92 and r^2 = 0.70 for Pho4 and Cbf1, respectively) (*SI Appendix*, Table S4), consistent with the prevailing sentiment that PWMs provide good approximations of specificity (9, 11). Intriguingly, PWMs explain a significantly smaller proportion of observed Cbf1 measurement variance, suggesting that Cbf1 recognition may rely on higher order determinants of specificity.

To evaluate BET-seq assay reproducibility, we generated PWMs from each of the Pho4 and Cbf1 technical replicates (*SI Appendix*, Fig. S13). Linear model coefficients for MNs at each position were strongly correlated between replicates of a TF (Pho4 r^2 = 0.95–0.97 and Cbf1 r^2 = 0.79–0.87) and uncorrelated between TFs (*SI Appendix*, Table S5 and Fig. S14); a meGFP negative control protein exhibited no sequence specificity (*SI Appendix*, Fig. S15). Using the fraction of unexplained variance($1 - r^2$) as a precision metric, the expected error range in NN-derived mean MN ΔΔG values for Pho4 is 0.02–0.04 kcal/mol and that of Cbf1 is 0.09–0.16 kcal/mol,

Downloaded at Palestinian Territory, occupied on December 23, 2021
www.manaraa.com

highlighting the robustness of binding specificity models derived from the assay and data presented here.

DN Models Reveal that Flanking Nucleotides Exhibit Significant Nonadditivity for Cbf1. The remaining unexplained variance observed between NN-derived values and PWM-predicted values (~8% and ~30% for Pho4 and Cbf1, respectively) could indicate higher order nonadditive interactions governing specificity or could simply represent experimental noise (41) (*SI Appendix*, Table S4). To probe for higher order interactions, we fit two DN models to the NN-derived scaled ΔΔG values: a nearest neighbor model that considers only contributions from adjacent DNs and a complex model that considers contributions from all DN combinations, including nonadjacent pairs (46).

Comparisons between nearest neighbor DN model-predicted and NN-derived binding energies showed increased correlation for both Pho4 and Cbf1, with r^2 values of 0.98 and 0.94 (*SI Appendix*, Table S4 and Fig. 3 *A* and *B*). These improvements, corresponding to ~5% and 24% increases in explanatory power over MN models, are consistent with the potential for physically interacting nucleotides to affect binding energies through local structural distortions. Considering all possible DN features accounts for nearly all of the remaining variance in NN-derived binding energies (improvements of ~1% and 5% for Pho4 and Cbf1, respectively). These findings highlight the differential degree to which nonadditivity defines binding even among structurally related TFs, which ultimately determines the predictive power and accuracy of widely used PWMs.

To visualize and interpret binding energy contributions of DNs alone, we calculated the mean residual ΔΔG from the linear regression against PWM-predicted ΔΔG values for all possible DNs within and across flanking sequences (Fig. 4*C*). Nucleotide interactions with measured ΔΔGs lower than expected based on considering the linear combination of individual MNs exhibit cooperativity; conversely, interactions that exhibit negative cooperativity increase measured ΔΔGs more than expected. The largest magnitude nonadditivity is observed for DNs immediately upstream or downstream of the E-box (N4/N5 or N6/N7 pairs), with absolute energetic differences among combinations spanning ~0.5 kcal/mol and epistatic interactions occurring primarily within flanks rather than between them. Inter- and intraflank DNs exhibited palindromic arrangements near the core motif, consistent with the expectation of binding site symmetry for homodimeric TFs like Pho4 and Cbf1 (62, 72). For Cbf1, TT and TG DNs upstream of the motif (and the corresponding downstream palindromes) exhibit large positive and negative nonadditivity, respectively. Although the overall magnitude of nonadditivity is significantly smaller for Pho4, a GG DN downstream of CACGTG shows strong synergistic effects. Interestingly, the Pho4 crystal structure reveals direct contacts between the Arg2 and His5 residues and this GG DN, providing a potential structural basis for this observation (62).

Incorporating Weight Constraints into DN Models Confirms that Cbf1 Interactions Are Significantly More Epistatic. Models that incorporate additional free parameters should always increase explanatory power. While MN models attempt to describe all 1,048,576 observed measurements using only 40 free parameters (4 nucleotides per position across 10 positions), the nearest neighbor DN model adds another 128 free parameters (16 pairs across 8 positions), and the all DNs model includes 720 free parameters [all combinations of nucleotide identities (4^2 = 16) and positional pairs ($\binom{10}{2}$ = 45)]. In most cases, DN coefficients are near zero (Fig. 4*C*), meaning they contribute little explanatory power. To identify the minimal set of features that define sequence specificity in an unbiased fashion, we used least absolute shrinkage and selection operator (LASSO) regression to develop parsimonious linear models with weight constraints (73). Nonzero coefficients in the model are penalized, leading to inclusion of only the most explanatory variables with respect to reduction in squared error (*SI Appendix*, Fig. S16). The regression explores a range of penalization stringencies to distinguish important sequence features based on differential coefficient minimization rates.

From the selected Cbf1 features, four nearest neighbor DNs exhibited large initial coefficient magnitudes and persisted throughout most of the penalization regime [two pairs of palindromic DNs spanning the core motif: (NNNAT_NNNNN, NNNNN_ATNNN) and (NNNGT_NNNNN, NNNNN_ACNNN)]. Strikingly, these DN model coefficients are up to 3-fold larger than that of the most significant MN feature (*SI Appendix*, Figs. S17 and S18), highlighting the importance of DNs to Cbf1 binding specificity. Among selected DN features for both Pho4 and Cbf1, nearest neighbor pairs exhibited the largest coefficient magnitudes (*SI Appendix*, Fig. S17).

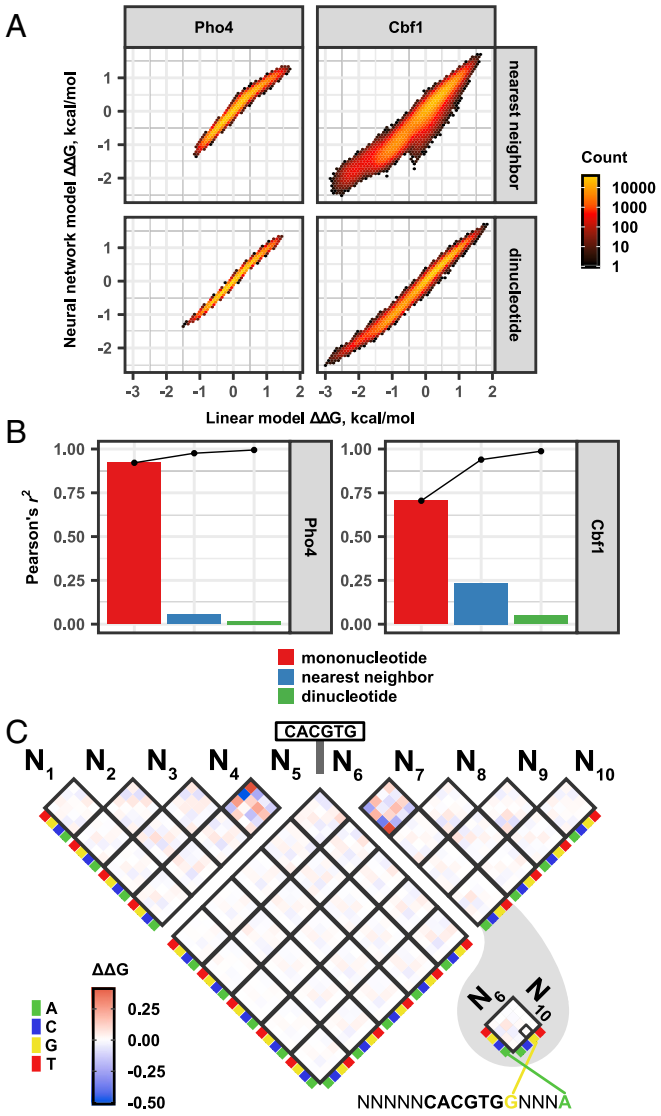

Fig. 4. NN model interpretation using DN features. (*A*) Density scatter plots of NN model estimates vs. DN (nearest neighbor) additive model predictions based on those estimates. (*B*) Fraction of variance in Pho4 and Cbf1 NN model estimates explained by MN and DN models. (*C*) Heatmap of mean residual energetic contributions when MN effects are removed for Cbf1.

Downloaded at Palestinian Territory, occupied on December 23, 2021
BIOPHYSICS AND COMPUTATIONAL BIOLOGY
www.manaraa.com

Orthogonal in Vitro Biochemical Measurements Confirm Results Obtained Via HTS. To confirm that NN model predictions provide accurate per-sequence estimates of true binding energies, we quantitatively compared titration-based ΔΔG values with unprocessed measurements and NN predictions. Using traditional fluorometric MITOMI, we determined ΔΔGs for Pho4 and Cbf1 binding to single-site variants of the ACAGA_TCGAG flanking sequence (*SI Appendix*, Table S3 and Figs. S10 and S11). In addition, we compared NN predictions with previously reported ΔΔG measurements of CACGTG flanking site mutations (18, 32). Consistent with Monte Carlo simulations, ΔΔG values calculated directly from raw sequencing data showed essentially no correlation to direct measurements, with r^2 values ranging between 0.07–0.16 and 0–0.24 for Pho4 and Cbf1, respectively (*SI Appendix*, Fig. S19). NN-predicted values showed remarkable agreement, with r^2 values ranging between 0.76–0.94 and 0.61–0.69 for Pho4 and Cbf1. In all cases, predictions agreed with observations within ∼1 kcal/mol. These results establish that the Pho4 and Cbf1 NN models presented here yield accurate measurements of binding energies for >1 million TF–DNA interactions with similar resolution to "gold-standard" biochemical measurements.

High-Resolution in Vitro Affinity Measurements Can Be Used to Identify Biophysical Mechanisms Underlying in Vivo Behavior. The role of transcriptional activators in vivo is not simply to bind DNA but to bind specific genomic loci and regulate transcription of downstream target genes. The high resolution of these comprehensive binding energy measurements makes it possible to quantitatively estimate the degree to which measured binding affinities explain differences in measured TF occupancies, rates of downstream transcription, and ultimate levels of induction.

First, we compared NN-modeled ΔG values with measured rates of transcription and fold change induction for engineered promoters containing CACGTG Pho4 consensus sites with different flanking sequences driving the expression of fluorescent reporter genes (17, 18). As reported previously, rates of transcription and induction scaled with measured ΔG values (*SI Appendix*, Fig. S20). Next, we compared NN-derived binding energies with reported levels of TF occupancy in vivo at CACGTG consensus sites in the *S. cerevisiae* genome for both Pho4 and Cbf1 (63). Large magnitude TF enrichment induced by phosphate starvation was observed at loci with measured K_d values of around 100 nM or lower (Fig. 5*A*). While TF enrichment roughly correlated with binding energy, very high affinity sequences showed strikingly low enrichment. The observed nonlinearities may indicate the degree to which other regulatory mechanisms, such as cooperation and competition among TFs or changes in DNA accessibility due to nucleosome positioning, contribute to reported TF enrichment (63). Alternatively, these nonlinearities may reveal the need for higher resolution in vivo measurements to test the degree to which binding energies alone dictate occupancies.

Previous analyses of TF binding energies used landscape visualization strategies to identify energy-dependent patterns in the data (74). To better understand the relationship between binding energies and in vivo occupancies, we similarly visualized the binding energy landscapes for both Pho4 and Cbf1 as a function of sequence space (Fig. 5*B* and *SI Appendix*, Fig. S21). The highest affinity sequence for each TF was placed at the center of a series of concentric rings, each of which includes all sequences at a given Hamming distance from this sequence. Within each ring, points representing each sequence are arranged alphabetically, with the color of each point reporting the measured ΔΔG for that sequence. As expected, the landscape forms a somewhat rugged funnel, with binding energies increasing with mutational distance from the highest affinity site (*SI Appendix*, Fig. S22). Next, we projected flanking site occupancies from ChIP-seq

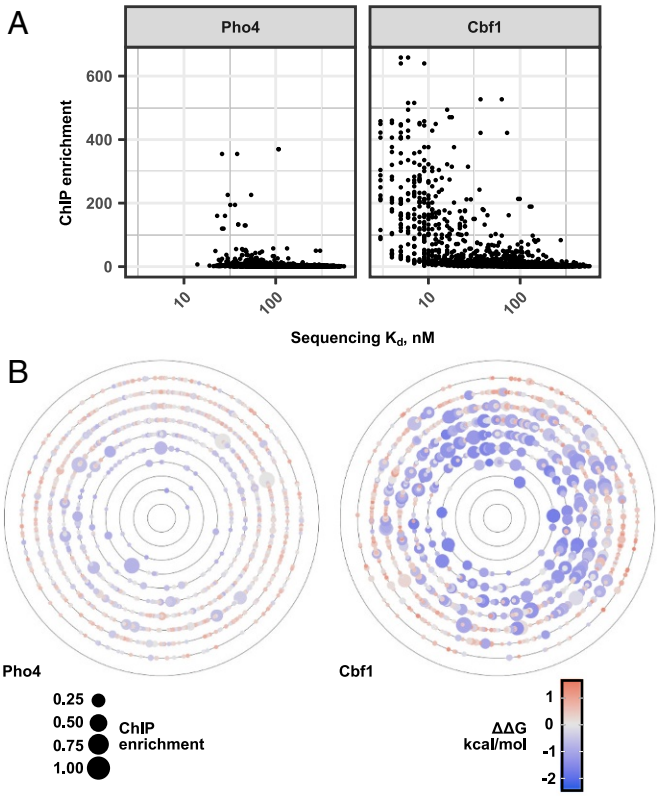

Fig. 5. Pho4 and Cbf1 binding energies and in vivo activity. (*A*) Pho4 and Cbf1 affinities (K_d, in nM) compared with in vivo ChIP-seq enrichment (63). (*B*) Functional-energetic landscapes of ChIP enrichment at dynamically regulated loci, relative to the measured highest affinity sequence in Hamming distance space.

experiments (63) onto these binding affinity landscapes to yield a composite functional-energetic landscape (Fig. 5*B*). For both Pho4 and Cbf1, the majority of enriched genomic loci are greater than four mutational steps away from the global minimum, corresponding to mean increases in binding energy of approximately 0.8 and 1.5 kcal/mol, respectively (*SI Appendix*, Fig. S23). These quantitative comparisons between measured affinities and in vivo occupancies establish that even relatively small differences in ΔΔG are associated with differential TF enrichment.

Discussion

TFs play a central role in regulating gene expression throughout development and allowing organisms to adapt to changing environmental conditions. The ability to quantitatively predict levels of TF occupancy in vivo from DNA sequence would therefore be transformative for our understanding of cellular function. TF binding at a given locus depends on multiple factors, including accessibility of a particular site (75, 76), effects of cooperation and competition with other TFs and nucleosomes (77, 78), the presence of nucleotide modifications (79, 80), and the nuclear concentration of a TF at a given time (81, 82). For accessible, unmodified target sites, the probability of TF occupancy at a given locus includes an exponential dependence on the corresponding TF–DNA binding energy (83); therefore, accurate occupancy predictions require the ability to resolve even small (∼1–2 kcal/mol) changes in binding energies. Toward this goal, we developed an assay to provide comprehensive and quantitative measurements of near-neutral changes in binding energies caused by mutations in the flanking sequences surrounding TF consensus sites. By training a NN on noisy estimates of

Downloaded at Palestinian Territory, occupied on December 23, 2021
www.manaraa.com

binding energies for millions of sequences, we obtained a model that incorporates all higher order complex interactions required for accurate binding energy estimates for each sequence. In future work, we anticipate that high-resolution in vitro binding energy landscapes can be combined with genomic [e.g., methylation state (84) and chromatin accessibility data (85, 86)] and mechanistic data (e.g., quantifying cooperation and competition between other TFs and nucleosomes) to yield comprehensive, predictive models of TF binding in vivo.

Many mutations in flanking nucleotides outside of "core" consensus motifs can change binding energies by an amount equal to or greater than mutations within the core. For example, the difference in binding energy between a *TCCCC***CACGTG***CCCCA* sequence and a *AATTT***CACGTG***AAAAG* sequence is ~2.6 kcal/mol, equivalent to mutating the core motif from **CACGTG** to **CGTGTG**. The bold sequence indicates the consensus TF motif 'CACGTG' that is identical for all sequences within the library, the italics indicate specific upstream and downstream flanking sequences, and the underlined letters highlight a change in the consensus sequence whose effects were previously measured in ref. 32. However, current representations of TFBSs would predict a change in binding energy for only the core mutation. This discrepancy may explain mysteries regarding ChIP data in which some genomic loci are occupied despite an apparent lack of a consensus site while other accessible regions containing consensus sites remain unoccupied. In addition, many current efforts to infer the presence of bound TFs first analyze DNase-seq or ATAC-seq data to identify regions of accessible DNA and then scan these regions for putative bound TFs by searching for sequence similarities to known TFBSs. Failing to consider flanking sequence effects could return a significant number of both false-positives and false-negatives.

In practice, measuring complete binding energy landscapes remains rare, with most assay development focused toward discovery of the highest affinity sequences. The binding energy landscapes presented here provide a unique opportunity to explore the mechanisms that drive evolution of transcriptional regulatory networks. High-affinity, but submaximal, TF binding sites may be evolutionarily favorable due to the potential for greater dynamic transcriptional control (87). Consistent with this, we find that the most highly occupied sites in vivo are mutationally distant from the highest affinity flanking sequences, potentially indicating the existence of an evolutionary buffer used to avoid sequence proximity to a suboptimal binding extreme. In addition, elevated levels of nonadditivity are thought to produce more rugged energetic landscapes compared with those created by additive binding interactions (88). Given that nonadditive DN interactions play a larger role in determining Cbf1 specificity, we speculate that Cbf1 binding sites can traverse fewer nondeleterious evolutionary pathways than Pho4, ultimately rendering Pho4 binding sites more evolutionarily plastic.

Systematic comparisons between per-sequence estimates of binding energies output by a NN and a series of linear models revealed the mechanistic features that drive specificity and quantified their contributions to observed binding energies. These results have relevance to recent debates surrounding the relative utility of DNA sequence-based models (PWMs) and DNA shape-based models representing TF specificity. Both models parameterize DNA binding preferences by a set of four values at each position [nucleotides (A, C, G, and T) for PWMs (36) and structural features (minor groove width, propeller twist, helical twist, and roll) for shape-based models (51–55)]. While these models can extract mechanistic determinants of specificity from sparse data, higher order information is lost in the process. Here, we demonstrate that models based on nearest neighbor DN preferences fully explain observed binding behavior, consistent with biophysical observations that local DNA structure is largely determined by base stacking interactions and interbase pair hydrogen bonds in the major groove between adjacent base pairs (55, 89, 90). Such nearest neighbor DN models require only a modest increase in the number of required free parameters relative to MN models. While the NN's capacity to incorporate higher order complexity ultimately proved unnecessary for accurately modeling Pho4 and Cbf1 binding specificities, high-resolution predictions output by the NN were essential to quantify the degree to which simpler models could explain observed behavior. The high resolution of these measurements further allows direct quantification of the degree to which thermodynamic models based on binding energies can predict behavior in vivo. The simulation-guided assay design and experimental assay presented here should allow a broader diversity of labs to make comprehensive and high-resolution measurements of binding energy landscapes. While BET-seq was deployed here for a specific use case (measurement of near-neutral effects over a small energy range), these simulations can guide choice of sequencing depths to resolve absolute binding energies across a variety of applications and platforms (9, 11, 29, 91, 92), including target site discovery efforts. The assay further offers the resolution of traditional MITOMI or HiTS-FLIP fluorescence-based assays while requiring significantly less equipment and infrastructure. Traditional MITOMI fluorescence assays require a DNA microarray printer and either a high-cost fluorescence scanner or fully automated microscope capable of imaging a slide with a microfluidic device attached; HiTS-FLIP assays require access to a customized Illumina GAIIx sequencing platform. A sequencing readout eliminates these requirements, allowing any laboratory with access to educational or commercial deep-sequencing services to measure energies at this scale and resolution. Moreover, the microfluidic valving is significantly simpler than for traditional MITOMI assays, reducing the pneumatics infrastructure required. Finally, BET-seq provides unique opportunities in future work to probe additional control mechanisms that influence TF binding in vivo. Introduction of synthesized DNA libraries containing modified bases involved in epigenetic regulation (e.g., 5-methylcytosine, 5-hydroxymethylcytosine) could allow systematic investigation of how these modifications affect TF specificities. In addition, BET-seq should be compatible with DNA libraries assembled into nucleosomal arrays in vitro, facilitating direct and quantitative investigation of how competition between TFs and nucleosomes dictates occupancies and how site-specific histone modifications influence this competition (93–95). The simulations presented here can guide the development of sequencing-based assays to measure binding energies for additional systems, including both protein–RNA and protein–protein interactions. In future work, BET-seq can complement initial SELEX-seq and PBM efforts to probe TF target specificity by providing high-resolution, quantitative mapping of the topography of these binding energy landscapes.

Materials and Methods

NN Binding Models. NN input was defined as a flattened 4×10 one-hot encoded matrix; NN output was the predicted $\Delta\Delta G$ value for the species of interest. The network consisted of three hidden layers of size 500, 500, and 250 units, respectively. All weights were initialized with Xavier initialization (96), and all layers used batch normalization (97) and ReLU activation. The entire dataset was randomly divided into training (60%), validation (10%), and test (30%) datasets; networks were trained using stochastic gradient descent until the validation set root-mean-squared error failed to decrease for three consecutive epochs. At this point, learning rate (initialized at 10^{-3}) was decreased in 10-fold increments, and training continued until error failed to improve for a further two epochs (*SI Appendix*, Fig. S8).

Linear Binding Models. Linear binding models were trained on scaled binding energy estimates output by the NN. The MN model includes sequence

Downloaded at Palestinian Territory, occupied on December 23, 2021

www.manaraa.com

features consisting of nucleotide identities at each flanking position; nearest neighbor DN included all MN features plus all possible combinations of adjacent nucleotide pairs. The full DN model adds all nonadjacent (gapped) DN combinations. All linear binding models were trained using the same 60% of the data as the NN; reported accuracies are calculated with respect to the held-out 40% of the data.

Material Availability. Detailed methods are available in *SI Appendix*, raw data are available from the Gene Expression Omnibus (accession number GSE111936), processed and intermediate files are available from Figshare (DOI 10.6084/m9.figshare.5728467), and code used for analysis and figure generation is available on GitHub (https://github.com/FordyceLab/BET-seq).

ACKNOWLEDGMENTS. We thank Justin Kinney, Hua Tang, Anshul Kundaje, Daniel Herschlag, and Rhiju Das for helpful discussions. T.C.S. and A.K.A. acknowledge NSF Graduate Research Fellowships; A.K.A. acknowledges the Stanford ChEM-H (Chemistry, Engineering, and Medicine for Human Health) Predoctoral Training Program. This work was supported by NIH/National Institute of General Medical Sciences Grant R00GM09984804. P.M.F. is a Chan Zuckerberg Biohub Investigator and acknowledges Sloan Research Foundation and McCormick and Gabilan faculty fellowships.

1. Latchman DS (1990) Eukaryotic transcription factors. *Biochem J* 270:281–289.
2. Kim HD, O'Shea EK (2008) A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol* 15:1192–1198.
3. Kim HD, Shay T, O'Shea EK, Regev A (2009) Transcriptional regulatory circuits: Predicting numbers from alphabets. *Science* 325:429–432.
4. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* 451:535–540.
5. Raveh-Sadka T, Levo M, Segal E (2009) Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res* 19:1480–1496.
6. Gertz J, Siggia ED, Cohen BA (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature* 457:215–218.
7. Foat BC, Morozov AV, Bussemaker HJ (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 22:e141–e149.
8. Riley TR, Lazarovici A, Mann RS, Bussemaker HJ (2015) Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. *eLife* 4:e06397.
9. Zhao Y, Stormo GD (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol* 29:480–483.
10. Zhao Y, Granas D, Stormo GD (2009) Inferring binding energies from selected binding sites. *PLoS Comput Biol* 5:e1000590.
11. Weirauch MT, et al. (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol* 31:126–134.
12. Mustonen V, Kinney J, Callan CG, Lässig M (2008) Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. *Proc Natl Acad Sci USA* 105:12376–12381.
13. Haldane A, Manhart M, Morozov AV (2014) Biophysical fitness landscapes for transcription factor binding sites. *PLoS Comput Biol* 10:e1003683.
14. Crocker J, et al. (2015) Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* 160:191–203.
15. Bintu L, Buchler NE, Garcia HG, Gerland U (2005) Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev* 15:116–124.
16. Lam FH, Steger DJ, O'Shea EK (2008) Chromatin decouples promoter threshold from dynamic range. *Nature* 453:246–250.
17. Aow JSZ, et al. (2013) Differential binding of the related transcription factors Pho4 and Cbf1 can tune the sensitivity of promoters to different levels of an induction signal. *Nucleic Acids Res* 41:4877–4887.
18. Rajkumar AS, Dénervaud N, Maerkl SJ (2013) Mapping the fine structure of a eukaryotic promoter input-output function. *Nat Genet* 45:1207–1215.
19. Gordân R, et al. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep* 3:1093–1104.
20. Levo M, et al. (2015) Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res* 25:1018–1029.
21. Afek A, Schipper JL, Horton J, Gordân R, Lukatsky DB (2014) Protein-DNA binding in the absence of specific base-pair recognition. *Proc Natl Acad Sci USA* 111:17140–17145.
22. Farley EK, Olson KM, Zhang W, Rokhsar DS, Levine MS (2016) Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc Natl Acad Sci USA* 113:6508–6513.
23. Afek A, Cohen H, Barber-Zucker S, Gordân R, Lukatsky DB (2015) Nonconsensus protein binding to repetitive DNA sequence elements significantly affects eukaryotic genomes. *PLoS Comput Biol* 11:e1004429.
24. Jolma A, et al. (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* 20:861–873.
25. Slattery M, et al. (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* 147:1270–1282.
26. Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249:505–510.
27. Ellington AD, Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature* 346:818–822.
28. Zykovich A, Korf I, Segal DJ (2009) Bind-n-Seq: High-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic Acids Res* 37:e151.
29. Chen D, et al. (2016) SELMAP–SELEX affinity landscape MAPping of transcription factor binding sites using integrated microfluidics. *Sci Rep* 6:33351.
30. Mukherjee S, et al. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* 36:1331–1339.
31. Berger MF, et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 24:1429–1435.
32. Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315:233–237.
33. Fordyce PM, et al. (2010) De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat Biotechnol* 28:970–975.
34. Isakova A, et al. (2017) SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat Methods* 14:316–322.
35. Nutiu R, et al. (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat Biotechnol* 29:659–664.
36. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res* 10:2997–3011.
37. Stormo GD, Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci* 23:109–113.
38. Stormo GD, Hartzell GW (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci USA* 86:1183–1187.
39. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15:563–577.
40. Zuo Z, Stormo GD (2014) High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. *Genetics* 198:1329–1343.
41. Bulyk ML, Johnson PLF, Church GM (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res* 30:1255–1261.
42. Mordelet F, Horton J, Hartemink AJ, Engelhardt BE, Gordân R (2013) Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics* 29:i117–i125.
43. Mathelier A, Wasserman WW (2013) The next generation of transcription factor binding site prediction. *PLoS Comput Biol* 9:e1003214.
44. Zhao Y, Ruan S, Pandey M, Stormo GD (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* 191:781–790.
45. Tomovic A, Oakeley EJ (2007) Position dependencies in transcription factor binding sites. *Bioinformatics* 23:933–941.
46. Siddharthan R (2010) Dinucleotide weight matrices for predicting transcription factor binding sites: Generalizing the position weight matrix. *PLoS One* 5:e9722.
47. Badis G, et al. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324:1720–1723.
48. Annala M, Laurila K, Lähdesmäki H, Nykter M (2011) A linear model for transcription factor binding affinity prediction in protein binding microarrays. *PLoS One* 6:e20059.
49. Zhao X, Huang H, Speed TP (2005) Finding short DNA motifs using permuted Markov models. *J Comput Biol* 12:894–906.
50. Sharon E, Lubliner S, Segal E (2008) A feature-based approach to modeling protein-DNA interactions. *PLoS Comput Biol* 4:e1000154.
51. Rohs R, et al. (2009) The role of DNA shape in protein-DNA recognition. *Nature* 461:1248–1253.
52. Abe N, et al. (2015) Deconvolving the recognition of DNA shape from sequence. *Cell* 161:307–318.
53. Chiu TP, et al. (2016) DNAshapeR: An R/bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* 32:1211–1213.
54. Yang L, et al. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol Syst Biol* 13:910.
55. Zhou T, et al. (2013) DNAshape: A method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* 41:W56–W62.
56. Quang D, Xie X (2016) DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 44:e107.
57. Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res* 13:2381–2390.
58. Hellman LM, Fried MG (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* 2:1849–1861.
59. Fordyce PM, et al. (2012) Basic leucine zipper transcription factor Hac1 binds DNA in two distinct modes as revealed by microfluidic analyses. *Proc Natl Acad Sci USA* 109:E3084–E3093.
60. Jones S (2004) An overview of the basic helix-loop-helix proteins. *Genome Biol* 5:226.
61. Fisher F, Goding CR (1992) Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. *EMBO J* 11:4103–4109.
62. Shimizu T, et al. (1997) Crystal structure of PHO4 bHLH domain-DNA complex: Flanking base recognition. *EMBO J* 16:4689–4697.

Downloaded at Palestinian Territory, occupied on December 23, 2021
www.manaraa.com

63. Zhou X, O'Shea EK (2011) Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Mol Cell* 42:826–836.
64. Kivioja T, et al. (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* 9:72–74.
65. Fu GK, et al. (2014) Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proc Natl Acad Sci USA* 111:1891–1896.
66. Fu GK, Hu J, Wang PH, Fodor SPA (2011) Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc Natl Acad Sci USA* 108:9026–9031.
67. Spivak AT, Stormo GD (2012) ScerTF: A comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic Acids Res* 40:D162–D168.
68. Morozov AV, Siggia ED (2007) Connecting protein structure with predictions of regulatory sites. *Proc Natl Acad Sci USA* 104:7068–7073.
69. MacIsaac KD, et al. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* 7:113.
70. Wagih O (2017) ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* 3:3645–3647.
71. Stormo GD, Schneider TD, Gold L (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res* 14:6661–6679.
72. Mellor J, et al. (1990) CPF1, a yeast protein which functions in centromeres and promoters. *EMBO J* 9:4017–4026.
73. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 58:267–288.
74. Carlson CD, et al. (2010) Specificity landscapes of DNA binding molecules elucidate biological function. *Proc Natl Acad Sci USA* 107:4544–4549.
75. Thurman RE, et al. (2012) The accessible chromatin landscape of the human genome. *Nature* 489:75–82.
76. Degner JF, et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482:390–394.
77. Segal E, et al. (2006) A genomic code for nucleosome positioning. *Nature* 442:772–778.
78. Gebhardt JCM, et al. (2013) Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. *Nat Methods* 10:421–426.
79. Khund-Sayeed S, et al. (2016) 5-Hydroxymethylcytosine in E-box motifs ACAT|GTG and ACAC|GTG increases DNA-binding of the B-HLH transcription factor TCF4. *Integr Biol* 8:936–945.
80. Yin Y, et al. (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* 356:eaaj2239.
81. Hao N, O'Shea EK (2011) Signal-dependent dynamics of transcription factor translocation controls gene expression. *Nat Struct Mol Biol* 19:31–39.
82. Tay S, et al. (2010) Single-cell NF-κB dynamics reveal digital activation and analogue information processing. *Nature* 466:267–271.
83. Bintu L, et al. (2005) Transcriptional regulation by the numbers: Models. *Curr Opin Genet Dev* 15:116–124.
84. Frommer M, et al. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA* 89:1827–1831.
85. Boyle AP, et al. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132:311–322.
86. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10:1213–1218.
87. Crocker J, Noon EPB, Stern DL (2016) The soft touch: Low-affinity transcription factor binding sites in development and evolution. *Curr Top Dev Biol* 117:455–469.
88. Aguilar-Rodríguez J, Payne JL, Wagner A (2017) A thousand empirical adaptive landscapes and their navigability. *Nat Ecol Evol* 1:45.
89. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci USA* 95:11163–11168.
90. Yang L, et al. (2014) TFBSshape: A motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res* 42:D148–D155.
91. Jolma A, et al. (2013) DNA-binding specificities of human transcription factors. *Cell* 152:327–339.
92. Tuğrul M, Paixão T, Barton NH, Tkačik G (2015) Dynamics of transcription factor binding site evolution. *PLoS Genet* 11:e1005639.
93. Simon MD, et al. (2007) The site-specific installation of methyl-lysine analogs into recombinant histones. *Cell* 128:1003–1012.
94. Yang A, et al. (2016) A chemical biology route to site-specific authentic protein modifications. *Science* 354:623–626.
95. McGinty RK, Kim J, Chatterjee C, Roeder RG, Muir TW (2008) Chemically ubiquitylated histone H2B stimulates hDot1L-mediated intranucleosomal methylation. *Nature* 453:812–816.
96. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research: Workshop & Conference Proceedings* [Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), Sardinia, Italy], Vol 9, pp 249–256.
97. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*, Journal of Machine Learning Research: Workshop & Conference Proceedings (Proceedings of the 32nd International Conference on Machine Learning, Lille, France), Vol 37, pp 448–456.

Downloaded at Palestinian Territory, occupied on December 23, 2021

www.manaraa.com