

Supplementary Material

Contents

Contents	1
Supplementary Methods.....	2
Strain Construction.....	2
Native Tagging for <i>in vivo</i> ChIP-Seq.....	2
Inducible Tagging for <i>in vivo</i> ChIP-Seq.	2
Inducible Tagging for <i>in vitro</i> Experiments.....	3
ChIP-Seq	3
Standardized Culture Conditions and (Optional) TF Induction.....	3
Crosslinking, DNA Shearing, and Extraction.	3
Immunoprecipitation.	4
DNA Purification.....	4
Library Preparation.	4
In Vitro ChIP-Seq.....	5
Generation of Sheared gDNA.	6
<i>In Vitro</i> Binding of Purified Proteins.....	6
Comparison to DAP-Seq.	6
ChIP-Seq Analysis Pipeline.....	6
Read Quality Control.	7
Read Trimming.	7
Alignment of Reads.	7
Calculating Coverage from Alignments.	7
Identifying Statistically Enriched Regions.....	7
Filtering for Expected Peak Shift.	8
Filtering of Artifacts.....	8
Filtering Watermarks.....	9
Verifying the Correct Transcription Factor.	9
TFs That Could Not Be Mapped.....	9
Mathematical Model of TF-DNA Binding and ChIP Coverage.....	10
Thermodynamic Model of Binding to a Single Site	10
Relationship to Hill Function	12
Equilibrium Constant Relation to Free Energy	13
Probability of TF Binding to a Sequence with Multiple Sites.....	13
Probability a TF is Genome-Bound vs In Solution.....	15
Motif (PFM) vs. Coverage Analysis	16

1	Motif Comparison.....	17
2	BoltzNet Stability Analysis	17
3	Comparison of BoltzNet with Other Tools.....	20
4	TFs that Failed to Generate Models	24
5	Library-ChIP	24
6	Library-ChIP Library Construction	24
7	Growing Cells and Preparing for Immunoprecipitation	25
8	Immunoprecipitation	25
9	Purification of Pre- and Post-IP DNA	25
10	Library-ChIP Library Prep and Pooling.....	25
11	Analysis.....	26
12	Biolayer Interferometry.....	26
13	Generation of Double-Stranded Biotinylated DNA	26
14	Preparing Solutions and Assay Plate	27
15	Running BLI.....	27
16	Analyzing BLI	27
17	Supplementary Figures.....	29
18	Supplementary Tables.....	93
19	Supplementary References	95

21 **Supplementary Methods**

22 ***Strain Construction***

23 We used a total of 4 tagging methods for TFs presented here, 3 of these were used for *in vivo* ChIP-Seq
24 and Library-PCR, these included two methods for tagging TFs at their genomic locus for native expression
25 levels, and one method for tagging TFs on an inducible plasmid to help achieve adequate TF expression.
26 Additionally, we used one tagging method to create strains that allowed protein purification for *in vitro*
27 ChIP-Seq and BLI.

28 **Native Tagging for *in vivo* ChIP-Seq.**

29 TFs tagged at their native locus were generated with either P1 transduction [1] or the FRUIT method [2].
30 Across these methods, we used 3 different tags (Figure S1). We tagged 133 TFs with a SPA tag and 64 TFs
31 with a 3xFLAG tag using P1 transduction by initially introduced into *E. coli* strain DY330 using
32 recombineering as described previously [3] and transferred to MG1655 by P1 transduction before
33 removing the KanR marker with FLP recombinase [4]. Additionally, 134 TFs with a V5-tag were made with
34 FRUIT. This yielded 314 TFs with at least one of these tags.

35 **Inducible Tagging for *in vivo* ChIP-Seq.**

36 We generated 311 pBAD24-3xFLAG (Figure S2) inducible strains for *in vivo* ChIP-Seq by performing PCR on
37 genomic DNA using primers that targeted the TF of interest (Supplementary Data 5). The forward primer
38 included an overhang with homology 5' to the pBAD insert site as well as a start codon followed by the

first 15-18 bases of the TF sequence after the start codon. The reverse primer included an overhang with homology 3' of the pBAD insert site followed by the last 15-18 bases of the TF, omitting the stop codon. Homologous overhangs matched a KpnI insertion site on pBAD24. Plasmid was linearized with KpnI from NEB (#R0142) and purified with Qiagen's PCR Purification Kit (Qiagen #28104). PCR fragments and linearized vector were ligated with NEB HiFi Assembly Mix (NEB #E2621) and ligation was transformed by heat shock back into *E. coli* K-12 MG1655. Colonies were restreaked and sent for Sanger sequencing for clone validation. This generated 311 TFs tagged with this method.

Inducible Tagging for *in vitro* Experiments.

We generated tagged strains for AllR, GlnG, Nac, PdhR, and UlaR using Lucigen's Expresso Rhamnose Cloning System (#49012-2) to create inducible 6xHis-tagged TF strains that can produce TFs easily purified for *in vitro* assays. Briefly, we followed kit instructions to clone each TF into the pRham vector included with the kit, where the 6xHis tag was added on the C-terminal end. Clones were selected using kanamycin and verified with Sanger sequencing, valid strains were preserved in 50% glycerol stocks.

ChIP-Seq

We have established a ChIP-Seq method that we have now applied to a range of bacteria (Figure S3) [5-9]. The method involves the following steps: (1) growing cells and inducing TFs if necessary; (2) crosslinking, DNA shearing, and DNA extraction; (3) immunoprecipitation; (4) purification of immunoprecipitated DNA; and (5) creation of NGS libraries for sequencing from purified DNA. These steps are described in more detail below.

Standardized Culture Conditions and (Optional) TF Induction.

For *E. coli* experiments reported in this paper, we used a standard culture condition for all experiments to prevent potential media-dependent differences in binding. We used M9 minimal media supplemented with 0.4% glycerol as a carbon source and 100µg/mL ampicillin for inducible strains. Cultures were inoculated from glycerol stocks into 5mL media and grown at 30°C with shaking at 250rpm for a day then subcultured into 50mL media in vented flasks until OD ~0.5. For inducible ChIP-Seq, we added arabinose to a final concentration of 0.2% and continued growth for 1hr to allow TF induction before crosslinking.

Crosslinking, DNA Shearing, and Extraction.

Cells were crosslinked to fix all TFs bound to DNA by adding formaldehyde to a final concentration of 1% and shaking at 100rpm, RT for 30min. Crosslinking was ceased by adding glycine to a final concentration of 250mM and shaking for 15min at 100rpm, RT. Crosslinked cultures were prepared for shearing by pelleting in a 4°C centrifuge at 3220g for 10min and washed twice in 50mL PBS. A mixture of Buffer 1 (20mM HEPES Potassium Salt pH 7.9, 50mM KCl, 0.5mM DTT, 10% glycerol) + Protease Inhibitor (PI) (Sigma Aldrich #04693116001) was used to prepare cells for shearing via sonication while protecting TFs from degradation. This was made by dissolving 1 PI tablet in 1.5mL DI water and adding 8.5mL Buffer 1. After the second wash, cells were resuspended in 600µL Buffer 1+PI mixture, transferred to Covaris-compatible tubes (Covaris Cat. #520056), and left on ice until sonication.

We used a Covaris S2 with an optimized sonication protocol to shear DNA. This allowed us to obtain DNA fragments with a tight length distribution centered around 250bp, which proved critical in our analysis of the data. Cells were sheared for 12min with amplitude 20%, intensity=5, and 200 cycles/burst. After sonication, lysate was transferred to 1.5mL centrifuge tubes and spun at 13000g, 4°C for 10 min. DNA was

1 extracted by transferring supernatant to fresh 1.5mL tubes. Salt adjustment was performed by adding
2 Tris-HCl pH8, NaCl, and NP-40 to final concentrations of 10mM, 150mM, and 0.1%, respectively.

3 **Immunoprecipitation.**

4 We used ChIP-grade antibodies for each epitope tag, ensuring specific pulldown of our TF of interest and
5 its bound DNAs, while minimizing nonspecific pulldown. For TFs tagged with SPA or 3xFLAG, we used Anti-
6 FLAG antibody (Sigma-Aldrich #F1804), and for V5-tagged TFs, we used Anti-V5 antibody (Sigma-Aldrich
7 #V8012). 5 μ L (5 μ g) of antibody was added to the salt-adjusted supernatant and samples were rocked at
8 4°C overnight for primary antibody binding.

9 We used Protein G Agarose Beads (Thermo Scientific #20399) to pull down antibody-bound TFs. These
10 interact specifically with the antibodies added above and allowed us to wash away the rest of the lysate,
11 keeping only the TF of interest and its bound DNAs. We washed 50 μ L of beads by adding 1mL IPP150
12 (10mM Tris-HCl pH 8, 150mM NaCl, 0.1% NP-40) buffer. We pelleted the beads by spinning at 2000g at RT
13 for 2min and discarding buffer. Lysate with primary antibody was added to the bead pellet and rocked at
14 4°C for 30min, followed by another 90min of rocking at RT for primary & secondary antibody binding. We
15 pelleted the beads at 2000g for 2 min, discarded the lysate, and started washes. These were carried out
16 by adding 1mL of buffer, rocking tubes for 2min, spinning for 2min at 2000g, and discarding the
17 supernatant; 5 washes were done with IPP150 as buffer, followed by 2 washes with 1x TE.

18 **DNA Purification.**

19 To purify immunoprecipitated DNA, we eluted TF-DNA complexes from beads by adding 150 μ L EBB (50mM
20 Tris-HCl pH 8, 10mM EDTA, 1% SDS) to the bead pellet and incubated at 65°C for 15 minutes. Tubes were
21 spun at 2000g for 5 min, and the elution was transferred to a fresh 1.5mL tube. We performed a second
22 elution using 100 μ L TE + 1% SDS, incubated at 65°C for 5 min, and spun at 2000g for 5 min. The two
23 elutions were pooled for a total volume of 250 μ L. We added 12.5 μ L Proteinase K (Sigma-Aldrich #P4850)
24 to the pooled elution and incubated at 37°C for 1hr to digest proteins, releasing DNA from TF-complexes,
25 followed by incubation overnight at 65°C to deactivate the proteinase. ChIP DNA was purified using
26 Qiagen's PCR Purification Kit (Qiagen #28104) using standard procedures, with a final DNA elution volume
27 of 35 μ L EB.

28 **Library Preparation.**

29 We converted purified ChIP DNA to NGS libraries for sequencing with New England Biolabs' NEBNext Ultra
30 II Library Prep Kit for Illumina (E7645). Library prep consists of the following steps, where we modify
31 reaction volumes as described below: (1) end preparation; (2) adapter ligation; (3) adapter cleanup; (4)
32 library PCR; (5) size-selection; and (6) library quality control and pooling for NGS. These steps are
33 described in more detail below.

34 End Preparation. The first stage blunted fragment ends, phosphorylated the 5' nucleotides, and added a
35 dA tail on the 3' nucleotides to prepare the ends of DNA for adapter ligation. This was achieved by mixing
36 33.3 μ L ChIP DNA with 2 μ L End Prep Enzyme Mix and 4.7 μ L End Prep Reaction Buffer and incubating on a
37 thermocycler at 20°C for 30min, and 65°C for 30min.

38 Adapter Ligation. We ligated adapters with sequences for targeting in PCR by diluting NEBNext Adaptor
39 for Illumina 1:25 in 10mM Tris-HCl pH 8 + 10mM NaCl (as suggested for low-input samples) and adding
40 1.67 μ L of diluted adapter to the end prep reaction along with 20.67 μ L of premixed ligation master mix
41 and ligation enhancer. This reaction was incubated at 20°C for 15min with the thermocycler lid off. Then

we added 2μL USER enzyme and incubated for another 15min at 37°C with the thermocycler lid on and set > 47°C to complete adapter ligation.

Cleanup of Adapter Ligation. Samples were cleaned up to remove residual adapters that could prevent efficient library PCR using AMPure XP beads (Fisher Scientific #NC9933872). Briefly, we added 60μL room temperature beads to the adapter ligation reaction (~0.9X bead ratio) and incubated for 5min allowing adapter-ligated DNA to bind the beads, while shorter fragments remained free. Samples are moved to a magnetic rack to separate the DNA-bound beads. The supernatant was discarded, and beads were washed twice with 200μL 80% ethanol. After the second wash, tubes were removed from the magnetic rack, residual ethanol was allowed to dry with the tube tops open (without over-drying beads which can reduce DNA recovery), and beads were resuspended in 11μL 0.1X TE. We allowed this to incubate for 2min which separated adapter-ligated DNA from the beads before returning the tubes to the magnetic rack. After beads were separated from the eluted DNA, we transferred 9μL of the DNA-containing supernatant to PCR fresh tubes.

Library PCR. We performed PCR on the cleaned-up adapter-ligated fragments to amplify DNA while adding Illumina NGS adapters and unique barcodes that were used to identify each sample on an NGS flowcell. This was done by adding 15μL NEBNext Ultra II Q5 Master Mix to the 9μL DNA fragments. For samples sequenced with NextSeq500, we used NEBNext Multiplex Oligos for Illumina (E7335, E7500, E7710, and E7730) for multiplexing. We added 3μL of one of the 48 i7 primers to each sample, and 3μL of the universal i5 primer to all samples. For samples sequenced with NextSeq2000, we used unique dual indexes (E6440, E6442, E6444, E6446, E6448). These primer sets have i5 and i7 primers premixed, so we added 6μL of a primer pair to each sample. Reactions were mixed and thermocycled with an initial denaturation step of 30sec at 98°C, followed by 14 cycles of 10sec denaturation at 98°C then 75sec annealing/extension at 65°C. After the last cycle, a final elongation was performed at 65°C for 5min, followed by cooling to 4°C until samples were retrieved.

Size-Selection. After library-PCR, we performed size-selection to retrieve adapter-linked DNA fragments with a tight size distribution centered around 350bp. This was done by increasing the PCR volume to 100μL with water and adding 65μL room temp AMPure XP beads for a 0.65X bead ratio. We separated beads and unbound DNAs by incubation on a magnetic rack and transferred supernatant (which contains the library fragments of interest) to fresh PCR tubes. Left-side selection was performed by adding 110μL of beads to the supernatant for an approximate bead ratio of 1.8X. Beads were washed according to the same procedure for cleanup after adapter-ligation, namely magnetically separating the beads and washing in 80% ethanol twice. Library DNA was eluted in 33μL 0.1X TE.

Library QC, Pooling, and Sequencing. Libraries were analyzed with on-chip electrophoresis to verify size distributions fit the expected profile and to obtain library molarities for pooling before NGS. We used Agilent's Bioanalyzer 2100 with High-Sensitivity DNA chips (Agilent #5067-4626). We typically diluted libraries 1:10 in water to ensure concentrations were within Bioanalyzer's measurable range. To create pools for NGS, each library was normalized to 4nM concentration, and equal volumes of each library were added to the pool. In cases where library concentration was < 4nM, we added an equimolar volume of library to the pool. Libraries were sequenced on either a NextSeq500 or NextSeq2000 with 75bp single-end reads. In a few cases, 75bp paired-end sequencing was performed.

In Vitro ChIP-Seq

For five TFs, we performed an *in vitro* variant of ChIP-Seq. Our *in vitro* ChIP-Seq protocol differs from *in vivo* ChIP-Seq by utilizing pre-sheared gDNA and using purified epitope-tagged TFs to bind gDNA *in vitro*. The overall workflow is shown graphically in Figure S4, and gDNA shearing and binding reactions are described in more detail below, along with a brief discussion of the relationship between our *in vitro* ChIP-Seq protocol and DAP-Seq.

Generation of Sheared gDNA.

We generated sheared gDNA for *in vitro* ChIP-Seq by extracting gDNA from WT *E. coli* and shearing to the 250 bp size targeted for *in vivo* ChIP-Seq. We started by growing WT *E. coli* in 2mL LB from glycerol stocks at 37°C with shaking at 250rpm in vented tubes. Genomic DNA was extracted using the GenElute Bacterial Genomic DNA Kit from Millipore-Sigma (#NA2110). gDNA was sheared by increasing gDNA volume to 500μL with 1x TE buffer, and sonicating on a Covaris S2 for 12min with amplitude 20%, intensity=5, and 200 cycles/burst. Fragments with the correct size were retrieved by mixing the sample from Covaris with 750μL of AMPure XP beads in a 1.5mL tube (1.5x bead ratio) and incubating for 10 minutes. Tubes were transferred to a magnetic rack for 5 minutes, isolating beads bound to target gDNA fragments. Beads were washed twice with 200μL 80% ethanol, and air dried for up to 5 minutes, ensuring not to over-dry the beads. gDNA fragments were eluted by resuspending beads in 150μL 0.1x TE buffer, incubating for 5 minutes off the magnetic rack, followed by 5-minute incubation on the rack. After beads separated, supernatant containing sheared gDNA was transferred to a new tube and ran on Bioanalyzer to confirm gDNA concentration and fragment size.

***In Vitro* Binding of Purified Proteins.**

We combined gDNA and different TF concentrations to assess the relationship between TF concentration and binding. We created a Protease Inhibitor-buffered IPP150 and by dissolving 1 tablet of Protease Inhibitor (Sigma Aldrich #04693116001) in 10mL IPP150 (10mM Tris-HCl pH 8, 150mM NaCl, 0.1% NP-40). We combined 500ng gDNA and either 50ng or 500ng purified 6xHis-tagged TF in the PI-buffered IPP150, with a final reaction volume of 500μL. This was mixed and allowed to incubate for 1hr at RT for protein binding. After an hour, we added 5μL (5μg) anti-6xHis antibody (Invitrogen cat #MA1-21315) and allowed tubes to rock at 4°C overnight. All subsequent steps from immunoprecipitation to analysis were carried out identically to *in vivo* ChIP-Seq.

Comparison to DAP-Seq.

During the completion of this work, a similar *in vitro* binding assay called DAP-Seq was reported [10, 11]. Our *in vitro* ChIP-Seq protocol differs from DAP-Seq in two primary ways. First, we purify protein as described above while DAP-Seq uses an *in vitro* expression and purification protocol. Moreover, a variant of DAP-Seq called biotin-DAP-Seq introduces biotinylated lysines during translation for use in IP. Second, DAP-seq begins by performing library preparation on pre-fragmented gDNA prior to IP, whereas our *in vitro* ChIP-Seq protocol performs library preparation after IP.

ChIP-Seq Analysis Pipeline

We developed a comprehensive pipeline for analyzing our ChIP-Seq data (Figure S5A). Applied to *E. coli*, the high-level steps the pipeline performs are as follows: (1) quality control on raw reads; (2) trimming adapters and low-quality bases from reads; (3) aligning the reads to the *E. coli* reference genome; (4) generating read coverage matrices; (5) identifying statistically enriched regions of coverage; (6) filtering likely artefactually enriched regions; (7) verifying the correct TF as a ChIP-Seq-specific QC. Each of these steps are applied uniformly to all samples and are described in more detail below.

Read Quality Control.

All samples began with an initial quality control check to ensure raw read quality with FastQC (version 0.11.5) [12], which provided several metrics assessing the reads and underlying base calls should any troubleshooting be needed. One metric is the distribution of base quality scores, which are a measure of the probability that a base call is incorrect:

$$\text{Quality score} = -10 \log p \quad (1)$$

where p is the probability of an incorrect base call. From equation 1, we can see that higher quality scores indicate a lower probability of an incorrect call. From FastQC, our median quality score over the first 5 bases was typically around 32, with a sharp increase to 36 or higher for the remainder of the read. For reference, a quality score of 30 indicates a 0.001 chance that the base call was incorrect, and thus scores of 30 or higher are typically considered high-quality bases.

Read Trimming.

Next, we removed Illumina adapters and low-quality base calls from the ends of our reads. We used Cutadapt (version 2.6) [13] to clip adapters and base calls with a quality score below 30. However, as discussed above, all base calls typically scored above this threshold, and thus the primary function of this step was generating reads containing only *E. coli* sequences inside of the adapters.

Alignment of Reads.

Reads were aligned to the *E. coli* reference genome. Alignment was done with Bowtie2 (version 2.3.4.3) [14] with default settings, and bam files were returned. These were converted to sam files with samtools (version 1.5) [15] for downstream applications. We also converted the alignments to tdf format, which is compatible with most genome viewers, using IGVTools (version 2.3.72) [16]. We also collected statistics on the alignments using Picard's (version 2.10.0) CollectAlignmentSummaryMetrics and CollectWgsMetrics [17]. These included the number & fraction of reads that were aligned, the number of bases that were/were not covered by reads, how many bases were covered by 1, 5, 10, etc. reads, and more. Results were imported to our MySQL database.

Calculating Coverage from Alignments.

Next, we calculated the number of reads covering each base. We used our previously published peak calling tool, SPAT [18], to parse aligned reads into an $N \times 4$ matrix (N = genome size), with columns for the genomic position, number of reads covering the position aligned to the forward strand, reverse strand, or either strand.

Identifying Statistically Enriched Regions.

We used the coverage matrices to identify enriched regions that may contain one or more TF binding sites. An enriched region is a contiguous sequence of DNA, at least 150bp long, that has statistically enriched coverage at each position compared to a background model. SPAT discovered enriched regions by fitting total genome-wide coverage to a lognormal distribution and identifying regions that with coverage in the top 0.1 percentile of this distribution for the whole sequence. We also required the maximum coverage in the region to be at least 4.75-fold enriched compared to the genomic average. This threshold was determined empirically by ROC analysis, where we used a set of RegulonDB [19] known binding sites as a set of positive regions and randomly selected genomic coordinates as negative regions. A custom python script flagged regions that did not pass the enrichment threshold to be omitted from downstream analyses.

1 Filtering for Expected Peak Shift.

2 ChIP-Seq produces a strand-specific signature of enrichment that can be used to identify true binding
3 peaks [8]. When ChIP-DNA fragments are sequenced, reads are generated starting at the edges of
4 fragments and moving inward, with the TF binding site(s) occurring somewhere in between. Reads will
5 thus align to either side of the binding site in a strand-specific manner, where reads aligning to the forward
6 strand will be shifted upstream of the actual binding site and reads aligning to the reverse strand will be
7 shifted downstream. The expected distance between the forward and reverse strand profiles can be
8 determined by the combination of DNA fragment size and NGS read length and measured by finding the
9 distance that maximizes the cross-correlation between stranded enrichment. For our configuration (350
10 bp fragments after library preparation and 75 bp reads) the expected shift is 45 bases, thus enriched
11 regions with shift < 45 were flagged and omitted from downstream analyses.

12 Filtering of Artifacts.

13 The pipeline filters out two broad categories of IP artifacts: (1) regions enriched in IP controls, and (2)
14 regions enriched across experiments for large numbers of different TFs.

15 Filtering For Enrichment in Controls. For each promoter/tag configuration we performed multiple control
16 experiments. For native tagging, we performed experiments with WT *E. coli* both with and without anti-
17 tag antibody. For inducible tagging, we performed experiments with an empty vector with and without
18 anti-tag antibody. For any given ChIP-Seq experiment, we filter out regions enriched in at least one control
19 corresponding to that experiment type unless the region in the experiment was sufficiently enriched over
20 the region in all control experiments. For native expression samples using the anti-FLAG antibody, all peaks
21 <1.5-fold enriched over controls were filtered, for native expression samples using anti-V5 antibody, peaks
22 <2.2-fold enriched were filtered, and for inducible experiments, peaks <2-fold enriched were filtered.
23 Thresholds were determined through an ROC analysis relative to confirmed binding sites in RegulonDB.

24 Filtering for Regions Enriched in Multiple TFs. By performing ChIP-Seq on many TFs, we were able to
25 identify regions that were enriched across many or all TFs (Supplementary Data 4). In some cases, these
26 regions were not enriched in controls and were therefore not detected by that filter (Figure S7). Of note,
27 several of these regions are frequently reported in published ChIP-Seq experiments. An examination of
28 these artifacts suggested that they can broadly be categorized as SPA tag-specific and tag-independent.

29 SPA tag specific artifacts consisted of 12 regions enriched in nearly all experiments using SPA-tags. An
30 examination of these regions revealed that the 12 regions included all previously reported binding sites
31 for the TF NagC, and all twelve regions were discovered in our ChIP-Seq of NagC. This suggested that these
32 artifacts were somehow associated with binding and pull-down of NagC. We confirmed this in two ways.
33 First, we performed experiments with SPA tags in strains deleted for NagC and the 12 regions did not
34 appear enriched. Second, it is known that binding by NagC is prevented by Glc-Nac [20]. Experiments
35 performed using SPA tags in WT *E. coli* with the addition of N-Acetyl-D-glucosamine (Glc-Nac) also
36 removed these 12 regions. Based on these results, if a SPA experiment had enriched regions overlapping
37 at least 5/12 of the canonical regions, all overlapping regions were flagged for filtering. This threshold was
38 determined empirically by comparing the proportion of known sites flagged or not, and the proportion
39 random sites flagged or not at different thresholds.

40 Tag-independent artifacts were classified as strong or weak depending on the number of different TFs for
41 which the region was enriched. Strong artifacts were enriched for at least 50 TFs and for at least half of
42 these experiments, the artifact was at least half as strong as the most enriched region. These regions were
43 removed by flagging any enriched region that overlapped an identified artifact. Weak artifacts were

enriched in more than 25 TFs but less than 50. Enriched regions overlapping these artifacts were filtered if the region was less than 7x enriched relative to the average across all other TFs.

Filtering Watermarks

Inducible ChIP-Seq uses a high-copy plasmid, causing genes on the plasmid to be present in high number compared to the single-genomic copy of other genes. As a result, plasmid genes will have background coverage 100s-1000s of fold-enriched compared to background, a signature we call the “watermark” (Figure S8). In addition, because AraC is present in the plasmid, this TF also displays a watermark.

We filter enriched regions that overlap these expected watermarks. After removing regions overlapping those genes, we applied a custom detection script that aimed to recover enriched regions that could represent autobinding, while avoiding the artefactual enrichment of the plasmid. We used the same approach to detect enriched regions, but only look in a 500 bp window around the edges of the tagged-TF that would have artefactual inducible enrichment. Positive matches are added back to the set of passing regions.

Verifying the Correct Transcription Factor.

We developed two custom QC scripts to determine that the correct TF was tagged in a ChIP-Seq experiment. One script is used on every experiment where we search the NGS reads for sequences matching our epitope tags to determine which genes bore a tag. For inducible ChIP-Seq, we performed the additional QC step that aimed to identify genes that were present on the inducible plasmid. In both cases, we require these stages to demonstrate that only one TF was unambiguously tagged.

TF Verification Via Epitope Tag. We searched reads in every sample for sequences from the beginning and end of our epitope tags to verify that the correct TF was tagged (Figure S9). Because we encoded the tag alongside the TF, some reads span the junctions around the epitope and read into gDNA. We searched reads for the first or last 25bp of the tag sequence, and for reads that matched we retrieved the side of the read that would have come from the genome (taking the 5’ side of a forward-aligned read that found the first 25bp of the tag, or a reverse-aligned read that found the last 25bp; or taking the 3’ side of a forward-aligned read with the last 25bp of the tag, or a reverse-aligned read finding the first 25bp). We required the retrieved read fragment to be at least 8bp to find its genomic locus, aligned fragments back to the reference genome with Bowtie2. We identified which genes (if any) the aligned reads overlapped using Bedtools2 (version 2.30.0) intersect [21]. Successful experiments are marked by clear enrichment of the TF of interest, where nearly all identified reads align to the correct TF (or potentially the downstream gene for reads found from the end of the tag).

TF Verification Via Watermark Detection. We developed a custom script that aimed to identify if the correct TF displays a watermark (see above) (Figure S10). Potentially watermarked regions were those with at least 9x average genomic fold-enrichment for a consecutive 200bp. We filtered these hits, considering only those regions that covered at least 80% of a gene, and at most 80% of the enrichment was genic. Because AraC was present on the pBAD plasmid to enable inducibility, it should always be watermarked. For this case, we permitted up to 1/3 of the enriched region to be intergenic, as the araBAD promoter upstream of AraC was also present on the plasmid and thus part of its watermark. Finally, a passing inducible experiment was one where at most one TF and AraC are watermarked.

TFs That Could Not Be Mapped

We sought to map all 318 proteins in *E. coli* that we curated to be potential transcription factors (Supplementary Data 1). For 179 TFs we were not able to generate replicate ChIP-Seq experiments that both passed quality control filters. In two cases we were not able to generate either a natively tagged strain or an inducible strain. In 64 (35%) cases we were only able to generate on chip experiment that passed quality control filters (22 of these had only one native promoter experiments and 42 had one successful inducible experiment). In all 64 cases, additional attempts were made to generate replicate. In 113 cases (65%), we were not able to generate any ChIP-Seq experiments from any strain that passed our quality control filters.

We can only speculate as to the reasons for most of these failures. In some cases, we expect that our single consistent growth condition did not support binding by the TF, even when induced (for example, in the case of allosteric TFs, the appropriate allosteric binding condition might not have been satisfied). Particularly in the 35% of cases where we were able to generate a single successful replicate, we expect that further optimization of the growth condition could (in future work) yield a replicate. In a minority of cases, inducible strains did not grow well and we expect that the TF was not well expressed by the native strain either.

The 179 TFs for which we were not able to generate data from *in vivo* ChIP represent an important gap to fill. However, for 5 TFs (AllR, GlnG, Nac, PdhR, and UlaR) we also generated *in vitro* ChIP-Seq data. As described below, we have now performed stability analysis in which we built models from subselections of experiments. One crucial result from this analysis is that data from *in vitro* ChIP-Seq provides an equally accurate measure of binding energy as *in vivo* experiments, at least for the 5 TFs examined. We thus expect that a future strategy of mapping the remaining 179 TFs with *in vitro* ChIP-Seq would yield a complete picture of TF binding energy for *E. coli*.

Mathematical Model of TF-DNA Binding and ChIP Coverage

Thermodynamic Model of Binding to a Single Site

To analyze TF-DNA binding, we seek to model the sequence specific binding energy of a TF to DNA. In the simplest case, we consider a system consisting of one sequence of interest and N_{NS} non-specific sites. We then consider N_{TF} transcription factors each of which can be either be bound to the sequence of interest with energy ϵ_{seq} or occupy a non-specific site with energy ϵ_{NS} . By convention, more negative energies are more favorable. The nature of the non-specific sites depends on context, as we describe in more detail below. We further assume that occupancy of these states is in thermal equilibrium over the timescales of interest for our analysis, motivated by the fact that speed of binding reactions is very fast relative to the timescale of a typical ChIP experiment.

Given the assumption of thermal equilibrium, the probability that the sequence of interest will be bound by a TF can be calculated via the Boltzmann distribution [22]. This is accomplished by considering two states for the system: (1) an “unbound” state where all N_{TF} molecules occupy “non-specific” sites, and (2) a “bound” state where $N_{TF} - 1$ TF molecules occupy non-specific sites, and one TF molecule is bound to the sequence of interest. Each state is associated with a multiplicity, Z , which is the number of configurations by which that state can be realized. For example, for the unbound state, $Z_{unbound}$ is the number of ways in which N_{TF} TFs can be arranged in N_{NS} non-specific sites.

$$Z_{unbound} = Z(N_{NS}, N_{TF}) = \frac{N_{NS}!}{N_{TF}! (N_{NS} - N_{TF})!} \quad (2)$$

Similarly, $Z_{bound}=Z(N_{NS}, N_{TF}-1)$, and describes the number of ways in which $N_{TF} - 1$ TFs can be arranged in N_{NS} non-specific sites. For $N_{NS} \gg N_{TF}$, this can be approximated as:

$$Z(N_{NS}, N_{TF}) \sim \frac{N_{NS}^{N_{TF}}}{N_{TF}!} \quad (3)$$

The energy of each configuration can be calculated by summing the energies of the N_{TF} TFs in that configuration. For example, for any unbound configuration:

$$\varepsilon_{unbound} = N_{TF} \varepsilon_{NS} \quad (4)$$

and for any bound configuration:

$$\varepsilon_{bound} = (N_{TF} - 1) \varepsilon_{NS} + \varepsilon_{seq} \quad (5)$$

For each configuration, we then calculate the Boltzmann factor (also called statistical weight), defined as the exponential of the negative energy relative to $k_b T$ for that configuration where k_b is the Boltzmann constant and T is temperature:

$$\text{Boltzmann factor} = e^{-\varepsilon/k_b T} \quad (6)$$

For the following, we assume that all energies are given relative to $k_b T$ so that this term will be omitted, and energies be unitless. The probability that the system will be in the bound state can then be calculated as the sum of the Boltzmann factors for all bound configurations divided by the sum of the Boltzmann factors for all configurations divided by the sum of the Boltzmann factors for all configurations:

$$P_{bound} = \frac{Z_{bound} e^{\varepsilon_{bound}}}{Z_{bound} e^{\varepsilon_{bound}} + Z_{unbound} e^{\varepsilon_{unbound}}} \quad (7)$$

(where, as noted above, we assume energies are in units of $k_b T$). The denominator of equation 6 is the partition function of the system. As shown previously [22, 23], for the simple system of one sequence and N_{NS} non-specific sites, equation 7 can be simplified to:

$$P_{bound} = \frac{1}{1 + \frac{N_{NS}}{N_{TF}} e^{\Delta\varepsilon}} \quad (8)$$

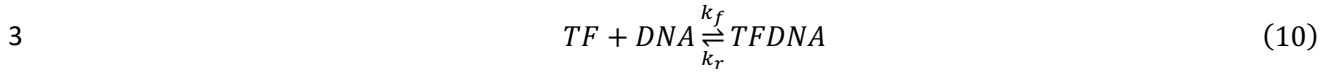
Where $\Delta\varepsilon$ is defined as the difference in energy between a single TF bound to the specific sequence and the TF occupying a non-specific site:

$$\Delta\varepsilon = \varepsilon_{seq} - \varepsilon_{NS} \quad (9)$$

$\Delta\varepsilon$ can be interpreted as the energy released when a TF transitions from a non-specific site to the specific sequence, or conversely the energy that must be provided to dislodge a TF from the specific sequence to a non-specific site. From equation 8, more negative $\Delta\varepsilon$ increases P_{bound} and thus reflects a sequence with higher affinity to the TF relative to non-specific sites.

1 Relationship to Hill Function

2 Equation 8 can also be related a unimolecular model of TF binding to a DNA sequence:



4 Here, k_f is the rate of complex association (TF + DNA binding to form TF-DNA), and k_r is the rate of complex
5 dissociation (TF-DNA unbinding into TF + DNA). We can model dynamics of TF-DNA concentration using
6 the law of mass action:

$$7 \quad \frac{d[TFDNA]}{dt} = k_f[TF][DNA] - k_r[TFDNA] \quad (11)$$

8 At thermal equilibrium, the association & dissociation reactions are balanced, and thus equation 11 is 0,
9 allowing us to rewrite it as:

$$10 \quad k_f[TF]_{eq}[DNA]_{eq} = k_r[TFDNA]_{eq} \quad (12)$$

11 Separating the rate-constants and equilibrium concentrations to different sides allows us to recover the
12 association/dissociation constants, K_A & K_D :

$$13 \quad \frac{[TF]_{eq}[DNA]_{eq}}{[TFDNA]_{eq}} = \frac{k_r}{k_f} = K_D = \frac{1}{K_A} \text{ (mol)} \quad (13)$$

14 where K_D for has units of moles. To determine the probability of a site being bound at equilibrium, we
15 simply divide the amount of TF-DNA complex by all DNA:

$$16 \quad P_{bound} = \frac{[TFDNA]_{eq}}{[DNA]_{eq} + [TFDNA]_{eq}} \quad (14)$$

17 If we multiply top and bottom by $1/[TFDNA]_{eq}$, we get:

$$18 \quad P_{bound} = \frac{1}{\frac{[DNA]_{eq}}{[TFDNA]_{eq}} + 1} \quad (15)$$

19 Multiplying top and bottom by $[TF]_{eq}$ gives:

$$20 \quad P_{bound} = \frac{[TF]_{eq}}{\frac{[TF]_{eq}[DNA]_{eq}}{[TFDNA]_{eq}} + [TF]_{eq}} \quad (16)$$

21 Where we can recognize the dissociation constant in the denominator, and we recover the familiar Hill
22 function form for TF binding at equilibrium:

$$23 \quad P_{bound} = \frac{[TF]_{eq}}{K_D + [TF]_{eq}} \quad (17)$$

Equilibrium Constant Relation to Free Energy

K_D (and K_A) can in turn be related to the Gibbs free energy of the reaction. From thermodynamics, we know that the energy change of the reaction of Equation 10 can be determined by:

$$\Delta G = \Delta G^0 + N_A k_b T \ln \left(\frac{[TFDNA]/[TFDNA]_0}{[TF]/[TF]_0 [DNA]/[DNA]_0} \right) \text{ (J/mol)} \quad (18)$$

Where ΔG^0 is the Gibbs free energy of associated with the reaction under standard conditions in which all reactants are present a 1M, k_b is the Boltzmann constant, and N_A is Avogadro's number or the number of molecules in 1 mole. The terms $[X]/[X]_0$ represent the ratios of the concentration of species X relative to the standard condition of 1M ($[X]_0$ is thus 1M for all X). When $\Delta G^0 = 0$, the system is at equilibrium, and thus:

$$\Delta G^0 = -N_A k_b T \ln \left(\frac{[TFDNA]_{eq}}{[TF]_{eq} [DNA]_{eq}} \frac{[TF]_0 [DNA]_0}{[TFDNA]_0} \right) = -N_A k_b T \ln \left(\frac{[TFDNA]_{eq}}{[TF]_{eq} [DNA]_{eq}} \mathbb{C} \right) \text{ (J/mol)} \quad (19)$$

where $\mathbb{C} = \frac{[TF]_0 [DNA]_0}{[TFDNA]_0} = 1 \text{ mol}$. We can recognize the quantity inside the log as K_A , and rearrange to solve for it:

$$K_A = \frac{1}{\mathbb{C}} e^{\frac{-\Delta G^0}{N_A k_b T}} = e^{\frac{-\Delta G^0}{N_A k_b T}} \text{ (mol}^{-1}\text{)} \quad (20)$$

or equivalently:

$$K_D = \mathbb{C} e^{\frac{\Delta G^0}{N_A k_b T}} = \mathbb{C} e^{\frac{\Delta G^0}{N_A k_b T}} \text{ (mol)} \quad (21)$$

In equations 20 & 21, ΔG^0 is the change in energy (joules) per 1 mol, and N_A is the number molecules per 1 mol. From this, we can recognize that:

$$\frac{\Delta G^0}{N_A k_b T} = \Delta \varepsilon \quad (22)$$

which is the change in energy of a molecule in units of $k_b T$ (and it thus unitless). We can thus rewrite equations 20 and 21 as:

$$K_D = \mathbb{C} e^{\Delta \varepsilon} = e^{\Delta \varepsilon} \text{ (mol)} \quad (23)$$

and:

$$K_A = \mathbb{C} e^{-\Delta \varepsilon} = e^{-\Delta \varepsilon} \text{ (mol)} \quad (24)$$

Providing an interpretation of the equilibrium constants in terms of the change in energy (in units of $k_b T$) when a single TF molecule binds DNA.

Probability of TF Binding to a Sequence with Multiple Sites

A more realistic model of TF-DNA binding assumes that each of N_{region} positions in a DNA sequence (for example, a 101mer) can be considered as an individual binding site i , each with their own energy ε_i . As

before, there are N_{TF} proteins available to bind any site within this region, and N_{NS} nonspecific sites that can be bound each with energy ϵ_{NS} . We further make the simplifying assumption that only 1 TF can bind the sequence at any time but can do so at any of the N_{region} positions. Determining the probability of binding anywhere in this sequence requires deriving the appropriate Boltzmann factors and multiplicities for configurations where the sequence is bound by a TF or not. For configurations where the sequence is not bound, the state weight is:

$$\frac{N_{NS}!}{N_{TF}! (N_{NS} - N_{TF})!} e^{-N_{TF} \epsilon_{NS}} \quad (25)$$

For bound configurations, we sum the Boltzmann factors of a TF binding to each position in the region:

$$\frac{N_{NS}!}{(N_{TF} - 1)! (N_{NS} - (N_{TF} - 1))!} \sum_{i=1}^{N_{region}} e^{-\epsilon_i - (N_{TF}-1) \epsilon_{NS}} \quad (26)$$

Then, the probability of a TF binding the sequence is:

$$P_{bound} = \frac{\frac{N_{NS}!}{(N_{TF} - 1)! (N_{NS} - (N_{TF} - 1))!} \sum_{i=1}^{N_{region}} e^{-\epsilon_i - (N_{TF}-1) \epsilon_{NS}}}{\frac{N_{NS}!}{(N_{TF} - 1)! (N_{NS} - (N_{TF} - 1))!} \sum_{i=1}^{N_{region}} e^{-\epsilon_i - (N_{TF}-1) \epsilon_{NS}} + \frac{N_{NS}!}{N_{TF}! (N_{NS} - N_{TF})!} e^{-N_{TF} \epsilon_{NS}}} \quad (27)$$

Using the assumption that $N_{NS} \gg N_{TF}$, we can simplify:

$$P_{bound} = \frac{\frac{N_{NS}^{(N_{TF}-1)}}{(N_{TF} - 1)!} \sum_{i=1}^{N_{region}} e^{-\epsilon_i - (N_{TF}-1) \epsilon_{NS}}}{\frac{N_{NS}^{(N_{TF}-1)}}{(N_{TF} - 1)!} \sum_{i=1}^{N_{region}} e^{-\epsilon_i - (N_{TF}-1) \epsilon_{NS}} + \frac{N_{NS}^{N_{TF}}}{N_{TF}!} e^{-N_{TF} \epsilon_{NS}}} \quad (28)$$

We then multiply both the numerator and denominator by $(N_{TF}-1)!/N_{NS}^{N_{TF}-1}$:

$$P_{bound} = \frac{\sum_{i=1}^{N_{region}} e^{-\epsilon_i - (N_{TF}-1) \epsilon_{NS}}}{\sum_{i=1}^{N_{region}} e^{-\epsilon_i - (N_{TF}-1) \epsilon_{NS}} + \frac{N_{NS}}{N_{TF}!} e^{-N_{TF} \epsilon_{NS}}} \quad (29)$$

and divide the sum $\sum_{i=1}^{N_{region}} e^{-\epsilon_i - (N_{TF}-1) \epsilon_{NS}}$ from the top and bottom:

$$P_{bound} = \frac{1}{1 + \frac{N_{NS}}{N_{TF}!} \frac{e^{-N_{TF} \epsilon_{NS}}}{\sum_{i=1}^{N_{region}} e^{-\epsilon_i - (N_{TF}-1) \epsilon_{NS}}}} \quad (30)$$

which is equivalent to:

$$P_{bound} = \frac{1}{1 + \frac{N_{NS}}{N_{TF}} \frac{1}{\sum_{i=1}^{N_{region}} e^{N_{TF}\epsilon_{NS}} e^{-\epsilon_i - (N_{TF}-1)\epsilon_{NS}}}} \quad (31)$$

which we can simplify as:

$$P_{bound} = \frac{1}{1 + \frac{N_{NS}}{N_{TF}} \frac{1}{\sum_{i=1}^{N_{region}} e^{-\epsilon_i + \epsilon_{NS}}}} \quad (32)$$

or

$$P_{bound} = \frac{1}{1 + \frac{N_{NS}}{N_{TF}} \frac{1}{\sum_{i=1}^{N_{region}} e^{\Delta\epsilon_i}}} \quad (33)$$

where $\Delta\epsilon_i = \epsilon_i - \epsilon_{NS}$ as in equation 9. From this, we can see that the probability of a single TF binding to a sequence with multiple binding sites is a function of the number of protein molecules (N_{TF}), and the sum of energy differences at each site (difference between binding the site vs a nonspecific site).

If we then equate these equations to equation 8 above and consider a single sequence with $\Delta\epsilon_{eff}$, then we have:

$$e^{\Delta\epsilon_{eff}} = \frac{1}{\sum_{i=1}^{N_{region}} e^{\Delta\epsilon_i}} \quad (34)$$

Which implies an effective energy of binding to the whole sequence of length N_{region} is related to the sum of energetic contributions to each individual site in the region:

$$\Delta\epsilon_{eff} = -\log \left(\sum_{i=1}^{N_{region}} e^{-\Delta\epsilon_i} \right) \quad (35)$$

Probability a TF is Genome-Bound vs In Solution

To estimate the probability of a TF being bound to genomic DNA relative to solution, we define two states: (1) a single TF is bound to one of N_{genome} sites with energy ϵ_{genome} ; and (2) a single TF is in one of $N_{solution}$ sites with energy $\epsilon_{solution}$. By convention, binding energies are taken with solution as the reference, thus $\epsilon_{solution}$ is 0. We seek to define the probability a TF is bound vs in solution, and start by defining the appropriate Boltzmann factors, for the case where TF is in solution:

$$\frac{N_{solution}!}{N_{solution}! (N_{solution} - 1)!} \quad (36)$$

And the case where TF is bound to the genome:

$$\frac{N_{genome}!}{N_{genome}! (N_{genome} - 1)!} e^{-\epsilon_{genome}} \quad (37)$$

The probability that of being genome-bound is the ratio of its Boltzmann factor to the sum of all Boltzmann factors:

$$p_{bound} = \frac{\frac{N_{genome}!}{N_{genome}!(N_{genome} - 1)!} e^{-\epsilon_{genome}}}{\frac{N_{genome}!}{N_{genome}!(N_{genome} - 1)!} e^{-\epsilon_{genome}} + \frac{N_{solution}!}{N_{solution}!(N_{solution} - 1)!}} \quad (38)$$

With the assumption that $N_{solution} \gg N_{genome}$ we can simplify:

$$p_{bound} = \frac{N_{genome} e^{-\epsilon_{genome}}}{N_{genome} e^{-\epsilon_{genome}} + N_{solution}} \quad (39)$$

Divide the top and bottom by $N_{genome} e^{-\epsilon_{genome}}$:

$$p_{bound} = \frac{1}{1 + \frac{N_{solution}}{N_{genome}} e^{\epsilon_{genome}}} \quad (40)$$

We see that the probability a single TF is bound to the genome is a function of the genome binding energy relative to solution, and the ratio of solution sites to genomic sites. For our analyses, we calculate the value for a range of ratios of solution sites to genome sites. To select one value of this ratio for discussion, we used a previously published measurement suggesting that genomic DNA accounts for roughly 1% of the volume of an *E. coli* cell [24]. Given that binding to DNA occurs on a roughly one-dimensional manifold while the solution space is a three dimensional volume, we take the cube of this percent and invert it to arrive at a rough estimate of 10^6 solution sites for every genomic base.

Motif (PFM) vs. Coverage Analysis

Position frequency matrixes (PFMs) are the most common means of representing binding sites. PFMs model the frequency of bases at each position in a collection of sites, but do not directly model the energy contribution of each base at each position. PFMs thus depend on the sites used to count frequencies, and consequently the frequency of a base need not be positively associated with its energy contribution. For example, most TFs display a small number of high affinity binding sites (Figure 1H, Figure 2, Figure 3), and as discussed in the main text these high affinity sites are typically distinguished by accessory bases outside a common conserved core binding sequence. Although such accessory bases can be associated with a strong favorable binding energy, this contribution can be masked in a PFM because the base is infrequent – present only in the infrequent strong sites. In such cases, the frequency of the base would underestimate its affinity contribution.

When applied to novel sequences, PFMs provide scores of statistical significance (for example, the log-likelihood of a sequence being generated by the PFM relative to a background model). In effect, PFMs are typically used to solve a binary classification problem (motif or not motif). The prediction of coverage and/or binding affinity, by contrast, is a regression problem.

Because PFMs do not model binding energy, they do not predict ChIP-Seq coverage (a function of binding energy). We demonstrate this empirically in (Figure S27, Figure S28, Figure S29, Figure S30, Figure S31, and Figure S32). For this analysis we used our reported binding sites to generate motifs for all 113 TFs

with more than one binding region using MEME, MEME-ChIP, and XSTREME [25]. For each motif, we use FIMO to scan the genome for motif occurrences. For each occurrence, we then identified the maximum coverage in a window of sequence spanning 250bp on either side of the FIMO hit in each experiment performed for the corresponding TF. We normalized these coverages to the maximum coverage in each experiment across all motif occurrences. For all TFs, we plotted the normalized coverage against the FIMO score of motif occurrences. We created figures for the 5 TFs calibrated with BLI (AlIR, GlnG, Nac, PdhR, UlaR) showing FIMO scores generated by each of the three MEME tools (MEME – Figure S27; MEME-ChIP – Figure S29; XSTREME – Figure S31), with each ChIP-Seq sample plotted individually. For TFs apart from the 5 calibrated with BLI, we created figures for FIMO scores generated by each of the three MEME tools (MEME – Figure S28; MEME-ChIP – Figure S30; XSTREME – Figure S32) with all ChIP-Seq samples on one plot. For all TFs and all three tools, the FIMO score was not predictive of coverage.

Different MEME-based approaches differ only slightly on how motif sequences are identified [25]. MEME is fundamentally a maximum likelihood approach that seeks models that maximize the likelihood ratio of a sequence belonging to a probabilistic model of a motif as compared to a background model. XTREME performs additional analyses on discovered motifs (e.g., enrichment analysis, similarity to known motifs, motif clustering). MEME-ChIP expects input sequences to be centered on expected motifs, and performs sampling of these input sequences, both to increase the efficiency of analyzing specifically ChIP-Seq data. But in all cases, the fundamental binding model remains a model of the probability of a base at each position of a motif.

Motif Comparison

To align PFMs, we maximized the pairwise information content from comparing all possible PFM alignments (including the forward- and reverse-complement of one motif). To compare PFMs from BoltzNet and RegulonDB v12.0 [19], we converted PFMs into information matrices using `transform_matrix` in `logomaker` [26]. Information at each base was normalized to 1, and only bases with a total information of at least 0.6 in at least one of the matrices were compared. Alignments were scored by taking the inner product of the information vectors at each aligned position and summing those inner products across aligned positions. In cases where alignments included overhangs, edges were padded with 0.0001. The alignment maximizing the sum of information inner products was returned as the best alignment.

BoltzNet Stability Analysis

To assess the stability of BoltzNet models, for the five TFs calibrated with BLI (AlIR, GlnG, Nac, PdhR, UlaR) we analyzed the impact on accuracy of weight matrix length, neural network width, subsampled called peaks, and subsampled experiments. By performing this analysis for the five calibrated TFs, accuracy could be independently assessed by comparing predicted binding energy to actual binding energy for the sequences measured in Figure 5. In addition, for all models constructed, the MSE on the training set coverage prediction and genome coverage prediction was calculated.

Neural Network Width. To assess the impact on the number of nodes in the middle layer of the neural network component, we attempted to build models with 2, 3, 5, 10, 15, 25, 50, 100, 140, 200, 300, 500, and 2000 nodes. Not all cases resulted in models for all TFs. In these cases, a sufficiently accurate model could not be found during pretraining. Models were largely insensitive to neural network width. Widths from 5 to 2000 nodes in the middle layer displayed virtually identical performance (Figure S34, Figure S35, Figure S36, Figure S37, and Figure S38). These results are consistent with the design of BoltzNet. The role of the neural network is to learn a mapping from the single affinity score for a sequence to the coverage

of that sequence in one or more ChIP-Seq experiments. Based on the biophysical model of ChIP-Seq, we expect this mapping to be smooth and well approximated by one sigmoid (Hill) function for each output experiment.

A concern with a model with too few parameters is that it will not have enough expressive power to fit the desired function. Our results indicate that for BoltzNet as few as 5 nodes (as well as one node for each output) in the middle layer provides sufficient expressive power to learn the expected coverage mapping for all TFs. In many cases, as few as two middle nodes was sufficient, although in such cases the mapping became noticeably piecewise linear.

A concern with a model that has too many parameters is that it will simply “memorize” a mapping between each observed input and output. Our results indicate that even for 2000 nodes in the middle layer, this did not occur. As indicated in the plots of Genome Coverage vs Affinity Score, the models all learned a mapping that was smoother than the underlying actual genome coverage (grey points). An overfit model would learn a mapping that jumped from one grey point to the next. Our results indicate that for 2000 nodes or less (we did not test a larger number of nodes), the regularization applied was sufficient to ensure that the neural network layer was not overfit.

Most importantly, the number nodes in the neural network layers did not impact the energy weight matrix learned. This is also expected from the design of BoltzNet. The parameters of the neural network can only transform the total affinity score, and thus can only impact the contribution of all parameters of the weight matrix equally. The neural network cannot modify the contributions of individual bases to the affinity score, and thus cannot affect the internal structure of the weight matrix. The design of BoltzNet thus insulates the structure of the weight matrix – which maps sequence to binding affinity – from the structure of the neural network – which maps affinity to coverages.

Weight Matrix Length. To assess the impact of weight matrix length, we attempted to build models with matrix lengths of 10, 15, 20, 25, 30, 40, 50, 60, 75, 100, 125, and 200 bp (Figure S39, Figure S40, Figure S41, Figure S42, and Figure S43). Not all matrix lengths resulted in models for all TFs. In some cases, a sufficiently accurate model could not be found during pretraining.

Models were robust to range of weight matrix lengths. In all cases where a model could be built, the same relative positions and bases with significant energy contributions were identified. However, very short (<15bp) or very long (>40bp) matrices resulted in degraded accuracy. The highest accuracy was achieved with matrices with margins <~10bp around the core sequence and accessory bases. Much shorter matrices failed to capture all bases with substantial affinity contributions, lowering accuracy. And in some cases, models could not be built for matrices of length 10 or 15. Much longer matrices beyond a threshold weight matrix length (typically 40bp) also decreased model accuracy, and in several cases accurate models with longer matrices could not be built. For example, no models passing pretraining accuracy filters with matrices longer than 40bp could be built for PdhR.

Longer matrices beyond a threshold length (typically 40bp) also decreased model accuracy, and in several cases accurate models with longer matrices could not be built. For longer matrices, regularization ensured that most positions outside the main motif typically had smaller energy contributions. However, positions with larger contributions were not uncommon. In some cases, such positions appeared to reflect degenerate or partial copies of the main binding site weights. These might have been learned from training set data that included closely spaced clusters of binding sites (as discussed in the main text, this was common). While it was not uncommon for such long sequences to accurately predict coverage on the

training set and genome, poor performance on the independent BLI measurements indicates that matrices learned were overparameterized and overfit and thus did not generalize to independent predictions. This result highlights the common issue of overparameterized machine learning models. A further side effect of weight matrices longer than 60bp is that we could not use these to score the sequences used in BLI as these sequences were restricted to 61bp.

These results are broadly consistent with the design of BoltzNet. The role of the weight matrix is to model the binding affinity of a TF to a single binding site. We selected a weight matrix of 25 bp to be consistent with the distribution of known binding sites for *E. coli* in RegulonDB (Figure S16A). All but three of the 97 experimentally verified high confidence RegulonDB binding sites are less than 25bp in length. Moreover, in the case the three that have longer motifs, an inspection suggests that it contains more than one copy of the site (Figure S16B, C, &D). This interpretation is supported by the motifs learned by BoltzNet for this TFs which are all shorter than 25 bp and are similar to one copy of the repeated sequences in the known motifs.

While longer sequences often have multiple binding sites, the affinity of a sequence containing multiple independent binding sites is biophysically accounted for by adding the *exponentiated energy of each binding site* within a sequence (Methods - Equations 15-17). This requires that multiple independent binding sites be scored independently by the weight matrix and exponentiated *before summation*, as occurs in the architecture of BoltzNet. Longer weight matrices containing multiple binding sites break this model. And they would be expected to generalize less well on independent sequences with a different pattern of clustered binding than the training set. This is consistent with our results.

Subsampling Called Peaks. To assess the importance of the number of called peaks in the input training data, we built models by subsampling called peak positions. For each subsample, only the top N peaks were provided as positive samples. For all TFs except Nac, we attempted to build models with only the top 50, 20, 10, 5, 4, 3, 2, and 1 peaks. For Nac, we attempted to build models with only the top 300, 200, 100, 50, 20, 10, 5, 4, 3, 2, and 1 peaks. For all TFs, we also rebuilt models with all called peaks as a reference. The number of negative samples (5000) in each model remained unchanged. Removed peaks could be included in the negative set, in which case they would be assigned a coverage of zero as with all other negative peaks.

The impact of subsampling the number of peaks varied by TF (Figure S44, Figure S45, Figure S46, Figure S47, and Figure S48). For the majority of cases, accurate models could be built with as few as the five strongest called peaks (PdhR, GlnG, AilR, UlaR), and in some cases the three strongest peaks (PdhR, GlnG). And moderately accurate models could be built with as few as two strong peaks. In contrast, while Nac models were robust to removal of more than 50% of the 504 called peaks, models for Nac with 100 or fewer subsampled peaks could not be generated. Nac has the most called binding sites for any TF we have mapped (Figure 1). And as discussed in the main text, the affinity of these called sites exists on a continuum with a high and highly variable background affinity to the genome as a whole (Figure 6). This broad binding profile is consistent with the preference of Nac to bind to a short string of largely 'A's and 'T's. It is also consistent with a larger number of training samples required for learning a model.

Subsampling Experiments. To assess the impact of different experiments, we built models on subsets of available experiments. The subsets were selected to span the different types of experiments available for each TF (e.g. in vitro, native promoter, inducible promoter) and included multiple subsets of only two *in vivo* experiments as would be the case for the majority of the 124 models we report. We also generated

models on subsets of experiments that were used to train ProBound (see Comparison of BoltzNet with Other Tools below).

Models for PdhR, AlIR, GlnG, and Nac were essentially insensitive to the selection of experiment subsets, with equivalent accuracy on the prediction of binding energy as compared to BLI measurements (we note that MSE cannot be compared between models built on different experiments as the magnitudes of the peak coverages in different experiments alters the magnitude of the MSE). Only UlaR displayed a sensitivity to the selection of experiments. Maximum accuracy was achieved with a combination of *in vitro* and *in vivo* experiments ($R^2=0.8$) while subsets of experiments displayed more moderate accuracy ($0.58 < R^2 < 0.67$). Importantly, models could be built for all five TFs from subsampled pairs of *in vivo* experiments, whether native or inducible promoters, and performed as well as models from the full set of experiments for four of the five TFs, and with acceptable accuracy for UlaR.

The difference in behavior of UlaR may be attributed to its binding profile. UlaR possesses two strong (and previously confirmed) binding sites identified in all experiment, and a large set of weak binding sites. The weak binding sites could be clearly observed with *in vitro* experiments with 10x protein but had much lower coverage in *in vivo* experiments and in *in vitro* experiments with 1x protein. The highest accuracy was obtained for models that included *in vitro* 10x experiments, suggesting that model accuracy benefitted when the signal from weaker sites could be observed.

Comparison of BoltzNet with Other Tools

We sought to compare BoltzNet with six different tools: ProBound, BPNet, Catchitt, DeepBind, DeepLift and FactorNet. Not all tools were available, and different tools provided different capabilities (Table S2). Of the six tools, two could not be compared to BoltzNet: DeepBind provides code for using pre-trained models but does not support building new models, DeepLift is an interpretation engine for extracting patterns from neural networks, (a variant of DeepLift is employed in BPNet) but is not itself a standalone tool for analyzing genomic data. In addition, we were not able to compile the code base for FactorNet. The code no longer appears to be supported and we received no reply when we contacted the authors. We were thus able to compare BoltzNet to three state-of-the-art tools for predicting TF binding: ProBound, BPNet, and Catchitt.

BoltzNet provides three key predictions: (1) a predicted model of binding energy for single site (the weight matrix), that can be used to generate (2) a prediction of the relative affinity of any input sequence, and (3) a prediction of the ChIP-Seq coverage associated with a sequence. For our analysis, we trained all three tools on the five TFs calibrated by BLI (PdhR, AlIR, GlnG, Nac, UlaR), and attempted to compare these BoltzNet predictions with those of the tool.

Each of the three tools had individual requirements for what data can be used as inputs for training and prediction, and what predictions and outputs are provided. Consequently, the three tools investigated did not all allow for the same comparisons.

BoltzNet outperformed all other other tools, as described in detail in the following:

ProBound.

ProBound learns biophysical parameters of TF binding from SELEX experiments and can also be applied to ChIP-Seq data[27]. ProBound is based on a custom maximum-likelihood framework that models both the biophysical process of TF-DNA binding as well as the biophysical process of experimental binding

enrichment. ProBound was designed for eukaryotes, and has not previously been applied to bacterial ChIP-Seq data, to our knowledge.

Training. For model training, ProBound uses as input a set of sequences each associated with two numbers: the count of that read after enrichment the count of that read prior to enrichment. For our analysis, the count after enrichment was derived from ChIP-Seq data with a tagged TF, and the count prior to enrichment derived from control ChIP-Seq data (no tagged TF), and input sequences were acquired by taking the 200bp downstream of the 5' aligned base, as was done in the original ProBound paper.

In our hands, we were not able to train ProBound on data for more than 5 ChIP-seq replicates. Thus, we trained both BoltzNet and ProBound on the same set of 5 experiments for each TF to allow for direct comparison.

The output of training is a position specific affinity matrix with a predicted $\Delta\Delta G$ energy for every base at every position of the binding sequence for the target TF. This is analogous to the BoltzNet weight matrix.

Predictions. ProBound provides tools (ProBoundTools version 1.0) that enable a user to apply a learned model to predict the TF-DNA binding affinity of any sequence. We applied these tools to predict the TF-Binding $\Delta\Delta G$ in 100 bp windows every 50 bp over the entire *E. coli* genome. This is analogous to the affinity score of BoltzNet. However, ProBound does not provide a mapping of predicted coverage for every predicted $\Delta\Delta G$ score. Our biophysical model of ChIP-Seq predicts this mapping will be smooth and roughly sigmoidal. To allow for comparison to BoltzNet, we fit the relationship between ProBound $\Delta\Delta G$ and actual ChIP-Seq coverage to a Hill function. We also predicted the binding affinity of every 61bp sequence whose TF-binding energy we experimentally measured using BLI and purified TF and DNA (as Figure 5).

Results. We were able to perform the following comparisons between BoltzNet and ProBound (Figure S59): (1) the predicted binding energy matrix for BoltzNet compared to the binding energy matrix for ProBound, (2) coverage vs. affinity score for BoltzNet as compared to coverage vs. $\Delta\Delta G$ for ProBound respectively, and (3) the accuracy in predicting the binding energy of sequences whose binding energies were measured with BLI.

The ProBound binding matrices learned fell into two categories. ProBound matrices for PdhR, Nac, and UlaR captured positive binding energy contributions for the expected core binding sequences and relative positions, and these bases and positions matched closely the positive bases and positions learned by BoltzNet. However, bases with negative weights were often more uniform in the matrices generated by ProBound vs BoltzNet. For GlnG and AlIR, by contrast, the ProBound matrix differed more significantly from the BoltzNet matrix. With respect to predicting coverage on the genome, ProBound models differed by TF. PdhR, AlIR, and UlaR displayed sigmoidal trends. Nac as well, although with much higher variability. GlnG, by contrast, could not readily be fit with a Hill function.

ProBound performed much less well in the prediction of binding energies on independent sequences measured with BLI. In nearly all cases, ProBound performed poorly on this test ($0 < R^2 \leq 0.39$). The one exception was PdhR where ProBound predicted more accurately ($R^2=0.76$), but still less accurately than the BoltzNet model with the same five training experiments ($R^2=0.93$). Notably, the accuracy on independent sequences was not always correlated with predictions on the genome. AlIR, for example, displayed the lowest variance relationship between coverage and $\Delta\Delta G$ on the genome, but the lowest accuracy on the independent test.

A number of observations can also be derived from the results. First, ProBound prediction on independent sequences did poorly when the underlying affinity matrix failed to capture the expected bases (AIIIR and GlnG). Second, even when the matrix did capture the expected positive bases (UlaR, Nac) or when the model was able to predict genome coverage, independent testing was not accurate suggesting a lack of generalization. Third, the ProBound prediction of affinity of novel sequences appeared to differentiate between expected peaks and non-specific sequences, with ProBound failing to capture differences in affinity between non-specific sequences.

BPNet.

BPNet is a deep convolutional neural network designed to predict genome coverage from sequence [28]. It does not explicitly generate a model of TF-DNA binding affinity. Instead, additional tools have been developed to interpret the behavior of the model: TF_MoDISco[29] is used to predict a PFM binding model, and SHAP [30] is used to assign each nucleotide an importance value reflecting how much it contributed for the model prediction at each position of a sequence. Two different code repositories exist. We used the most recent code base (<https://github.com/kundajelab/bpnet-refactor>).

Training. BPNet is trained on coverage from ChIP-Seq or ChIP-exo data as well as a list of filtered peaks locations. Importantly, BPNet (version 0.4.0) can only be trained on a single experiment at a time. We thus trained individual models for each dataset for each TF. BPNet is also designed for eukaryotes. Training is typically performed on a subset of chromosomes and tested and validated on the rest of the chromosomes. This approach assumes a large number of binding sites such that any split of chromosomes has a sufficient and balanced number of sites in each subset. *E. coli* data contain a smaller number of sites. Moreover, we validate our models on independent sequences using BLI. We thus trained BPNet on the full *E. coli* genome and all called peaks for each experiment. We then validated on independent sequences as described below.

Results. As BPNet does not explicitly generate a model of TF-DNA binding affinity, we cannot make a direct comparison with the BoltzNet weight matrix. Instead, we compare the PFM derived from the BoltzNet weight matrix with the top PFM returned by BPNet TF-MoDISco (Figure S60). We generate one top PFM for each individual BPNet model. Not all models resulted in TF-MoDISco motifs. Experiments with fewer peaks, even if they generated models, often failed to generate motifs. For models with PFMs, there was often excellent agreement with the PFMs from BoltzNet.

We next tested the ability of BPNet to model binding affinity. BPNet uses SHAP to provide a score of importance for each base at each position of a sequence. There is no explicit relationship between this score and binding energy at any given position. Moreover, there is no documented approach for deriving an affinity measure of a sequence using the SHAP scores at individual bases. To derive a measure of affinity for an arbitrary sequence, we therefore summed the SHAP scores (to derive a SHAP Sum) across the sequence where the SHAP score at each position was taken as the element-wise multiplication of the SHAP 'hyp_scores' with the one-hot encoding of the input sequence.

We then performed two comparisons. First, we sought to test the ability of BPNet to model binding affinity on the training set. As described above, BoltzNet learns a mapping from affinity score and coverage for every input experiment. While BPNet does not generate such an explicit mapping, SHAP scores are derived by backpropagating the prediction of coverage through the network and thus it might be expected that SHAP scores would be related to predicted coverage. To allow for comparison, we fit the relationships between the SHAP Sum coverage for all called peaks in each experiment specific model (Figure S60). In the vast majority of cases, coverage could be modeled as a sigmoidal function of SHAP Sum.

Second, we sought to assess the ability of BpNet to predict binding affinity on sequences never seen during training. To do so, we ran predictions using all BpNet models on the novel designed sequences from Figure 5. The length of the test sequences were shorter than the default input length for BpNet. To perform predictions, we selected a random genomic sequence with the correct input length, and which did not contain any peaks. We then replaced the same middle portion of this sequence with each test sequence. SHAP Sum scores were then calculated on just the portion of the input sequence containing the test sequence. Finally, we plotted the SHAP Sum against the BLI measured energy for each BpNet model.

BpNet performed worse than BoltzNet in all comparison for all TFs (Figure S61). In the case of some models for some TFs, SHAP Sum performed well in differentiating specific binding (colored points) vs non-specific point (gray) points in most cases. However, the SHAP Sum score poorly differentiated between different specific binding sites in the majority of cases. Moreover, the SHAP Sum score did not accurately predict the energies of non-specific sites. Overall accuracy in predicting sequence binding energy was strongly associated with the ability of TF-MoDISco to generate PFMs. Models without PFMs, associated with experiments with fewer peaks, performed less well in predicting energy. This was particularly prominent with models for GlnG. Moreover, within the set of models with PFMs for all TFs, accuracy in predicting sequencing binding energy often differed between models with highly similar PFMs.

Catchitt.

Unlike BoltzNet (and ProBound and BpNet), Catchitt is not a tool for binding motif/weight matrix discovery [31]. For training, Catchitt (version 0.1.4) requires a pre-existing PFM model of binding for a TF. It then combines this model with the location of ChIP-Seq binding sites, chromosome accessibility data (typically from DNase-Seq or ATAC-Seq), and optional DNA methylation data. Catchitt was designed for eukaryotic prediction as part of the 2017 “ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge” in which this information was available[31, 32]. The trained model can then be used to predict the binding probability across the genome.

Catchitt can thus be seen as a tool that uses a machine learning framework to combine PFMs with additional data to predict the probability of binding. While it is not directly comparable to BoltzNet (which discovers a binding weight matrix directly from ChIP-Seq without the need for additional data), it provides another test of the ability of PFMs to predict binding probability.

Training. Catchitt does not utilize ChIP-Seq data directly. In order to apply Catchitt to our data, we used MEME to generate PFM models from called peaks derived from our ChIP-Seq data for our five calibrated TFs. We further provided the full set of called peaks for each TF to Catchitt for training. Catchitt also requires chromosome accessibility data which is not required by BoltzNet. While chromosome accessibility is not strictly applicable to prokaryotes, genome accessibility more broadly, especially as regulated by nucleoid associated proteins (NAPs), is an analogous characteristic. To provide a measure of genome accessibility, we utilized control ChIP-Seq data where the coverage is expected to vary with overall DNA accessibility. The output of training was a model that was used to make predictions of the probability of binding across the *E. coli* genome.

Results. A comparison of the PFMs derived from the BoltzNet binding weight matrix with the PFMs used with Catchitt shows overall high agreement (Figure S62A). This is expected given the MEME input, as MEME motifs are already in agreement with PFMs derived from the BoltzNet binding weight matrices.

In contrast, the Catchitt probability of binding was not predictive of binding coverage (Figure S62B). These results are consistent with our results that demonstrate that PFM matrices are not accurate predictors of

coverage and binding affinity for our ChIP-Seq data. We note that accuracy in these tests may have been impacted by lack of DNase-Seq or ATAC-Seq accessibility data which may be important for Catchitt predictions. These data are not required by BoltzNet. More generally, the largest difference between the two tools is Catchitt was designed for the ENCODE-DREAM challenge which employed an extensive set of features derived from chromatin accessibility, binding motifs, gene expression, sequence and annotation to train classifiers for TF-DNA binding[32]. By contrast, BoltzNet learns TF-DNA binding de novo directly from ChIP-Seq data only.

TFs that Failed to Generate Models

Applied to *in vivo* ChIP-Seq data for 139 *E. coli* TFs, we have generated BoltzNet models for 124 that passed criteria for accuracy in predicting coverage and for specificity on the whole genome (Supplementary Data 2). BoltzNet was unable to build models for 15 TFs (Table S1). An inspection of these TFs fails to identify a single common feature that explains why accurate models could not be generated. We can only speculate about explanations for individual TFs. In particular, 4 of the 15 TFs have only a single strong peak (YedW) or one strong peak and one very weak peak (MazE, YfiE, YjgJ), with no other intermediate peaks. A fifth TF, sgrR, has only two very strong peaks with not other intermediate peaks. The meta-analysis of models built with subsamples of called peaks (Figure S44, Figure S45, Figure S46, Figure S47, and Figure S48). suggests that while BoltzNet can build models with as few as 2-3 called peaks, accuracy improves when peaks with intermediate affinities are included. It is thus possible that a single strong peak did not provide sufficient information for BoltzNet to build an accurate model in these five cases. In another case, StpA is a known nucleoid associated protein (NAP), and consistent with pattern for many NAPs the peaks for StpA are broader and more smeared out than is typical for most TFs. This might explain the failure of StpA, although the example of Nac (Figure 3) demonstrates that BoltzNet can build models for TFs that bind with broader peaks.

Library-ChIP

We employed Library-ChIP following a previously published method [33] (Figure S33), which allowed us to perform ChIP on designed binding sites in high-throughput. The protocol consists of the following steps: (1) library construction and transformation; (2) growing cells and preparing for IP; (3) immunoprecipitation of post-IP sample; (4) purification of pre- and post-IP DNA; (5) library prep and pooling; and (6) Library-ChIP analysis. These steps are detailed below.

Library-ChIP Library Construction

We cloned designed binding sites into a single copy pAMD plasmid (Figure S63) to create a plasmid library where each plasmid had a single designed site. These could be transformed into our tagged TF strains to assay binding via Library-ChIP.

We ordered an oligo-pool from IDT that contained over 100 novel binding sites (Supplementary Data 6) designed with varied specificities toward PdhR, AlIR, and GlnG, along with 3 genomic reference sites for each TF (2 in the case of GlnG). Each site was designed as a 61-mer, and 20 bp were added to each end with homology to the pAMD insertion site, creating 101-mer oligos. The oligo-pool was received as ssDNA and a universal primer pair we designed was used to create dsDNA with PCR (Supplementary Data 5—oPool dsDNA oligos). We mixed 1ng oPool DNA with 12.5μL KAPA HiFi HotStart ReadyMix (Roche Cat. #07958935001) and 0.75μL of each primer at 10μM, then filled the reaction to 25μL with sterile water. We performed PCR with a 3min initial denaturation at 95°C, followed by 5 PCR cycles with a 98°C denaturation for 20 seconds, 15 second annealing at 62°C (based on primer T_m), and a 15 second extension

at 72°C. After the final PCR cycle, we did a 1-minute extension at 72°C before cooling to 4°C. Reactions were cleaned up with Qiagen's PCR Purification Kit (Qiagen #28104).

To facilitate cloning, pAMD was linearized with BamHI and HindIII restriction enzymes from NEB (Cat #'s R0136 & R0104, respectively) and then ligated with the dsDNA pool using NEBuilder HiFi DNA Assembly Kit (#E2621). Ligation was transformed into *E. coli* EPI300 received from the Wade Lab via electroporation and selected with carbenicillin. 12 colonies were selected for Sanger sequencing to confirm that different sequences from the pool had been inserted (Supplementary Data 5 – oPool Test Primers). To ensure that each designed sequence is present in the pool, we aimed to get ~10X the number of colonies as there were binding sites in the pool, in this case, we targeted ~1200 colonies. These were scraped from plates and grown in 50mL LB at 37°C with shaking at 250rpm and carbenicillin selection. pAMD replication was induced in the EPI300 cells with arabinose and plasmid was extracted with Qiagen's Miniprep Kit (#27104). The resulting plasmid pool was electroporated into a natively V5-tagged AllR strain, an inducible GlnG strain, and native SPA-tagged and inducible PdhR strains for Library-ChIP.

Growing Cells and Preparing for Immunoprecipitation

Library-ChIP started identically to our ChIP-Seq protocol until the addition of appropriate primary antibody. Briefly, we grew cells from glycerol stocks and subcultured into 50mL fresh media, culturing until log-phase at which point arabinose was added for induction if necessary (note: arabinose does not induce pAMD replication in our tagged-TF strains – thus pAMD remains single-copy under TF induction). We crosslinked cells and stopped crosslinking with glycine, then pelleted and washed cells twice in PBS. After the second wash, we resuspended cells in Buffer 1+PI, and sheared via sonication with Covaris S2. At this point, we pelleted the lysate and reserved ~10% of the DNA-containing supernatant as a reference to quantify the abundance of designed sequences before IP.

Immunoprecipitation

The remaining 90% of lysate continued with immunoprecipitation exactly as previously described. Briefly, the lysate was salt-adjusted to 10mM Tris-HCl pH8, 150mM NaCl, and 0.1% NP-40, 5μL (5μg) of appropriate primary antibody was added, and samples were rocked at 4°C overnight. The following day, we performed secondary antibody binding with Protein G Agarose beads, utilizing the same 7-wash setup used in ChIP-Seq. Following the 7th wash, DNA was eluted once with EBB and once with TE + 1% SDS, and the two elutions were pooled.

Purification of Pre- and Post-IP DNA

Proteins were digested from pre- and post-IP DNA samples by adding Proteinase K at a 1:20 ratio, incubating for 1hr at 37°C, followed by proteinase deactivation at 65°C overnight. DNAs were purified from this reaction using Qiagen's PCR Purification Kit.

Library-ChIP Library Prep and Pooling

To prepare Library-ChIP DNA for NGS, we employed a 2-step PCR process following Illumina's 16S Metagenomic Library Preparation (Illumina Ref #15044223). The first PCR used primers targeted the flanking regions that were common across all binding sites in the pool (Supplementary Data 5 – oPool Universal PCR Primers). The primers also contained overhangs that could be targeted by Illumina multiplexing primers. Briefly, we added 2.5μL of sample DNA (normalized to 5ng/μL if the concentration was higher), along with 5μL each of 1μM Forward & Reverse primer, and 12.5μL KAPA HiFi HotStart ReadyMix. PCR used 25 cycles, and the reaction was cleaned up with Qiagen's PCR Purification Kit (#28104).

We performed a second PCR to barcode fragments into NGS-compatible libraries using primers from Illumina's Nextera XT Index Kit v2 (Illumina Cat. #FC-131-2001). This time, we used 25 μ L KAPA HiFi HotStart ReadyMix, along with 5 μ L of cleaned up DNA from the first PCR, 10 μ L sterile water, and 5 μ L each of Nextera Kit Index 1 & Index 2 primers, ensuring each sample received a unique pair of primers. This PCR utilized 8 cycles and libraries were cleaned up with AMPure XP beads using single-side selection to remove any adapter-dimers.

We used the Bioanalyzer 2100 to confirm fragment length distributions and obtain library molarities as we did for ChIP-Seq. However, because Library-ChIP libraries are much less complex than ChIP-Seq (Library-ChIP libraries are a collection of designed binding sites with high similarity while ChIP-Seq surveys a genome on the scale of megabases), care should be taken to ensure that Library-ChIP libraries do not dominate the overall composition of a pool, as this can lead to sequencing artifacts. Some applications may require spiking in exogenous DNA, though this was not needed in our case. Libraries were sequenced on a NextSeq2000 with 150bp paired-end reads to ensure that reads reached from the adapters across the designed sequence.

Analysis

We quantified the abundance of designed sequences in pre- and post-IP samples to determine the enrichment of each designed site in Library-ChIP. First, we created a database of designed 61mers for each of the 3 TFs. We developed a custom script which searched Library-ChIP reads for a 20 bp anchor sequence present at the end of the insertion site, and common among all sequences in the pool. Upon a successful match, we took the next 37 bases after the anchor, which was sufficient to cover all variant nucleotides in our pool. We recorded the sequence of each 37-mer, and their abundance in the raw reads of pre- and post-IP samples. Then, we searched for the 37-mers in each of the designed variants and marked any identified matches. In some cases, a sequence could be present in more than one 37-mer, in which case those counts were merged. We filtered out any sequences present in fewer than 0.1% of the raw reads, and then normalized the abundance of each sequence within the sample by the average number of sequence counts. Finally, the ratio of normalized post-IP:pre-IP counts is presented as the Library-ChIP enrichment.

Biolayer Interferometry

We used an optimized version of our previously published BLI protocol [34, 35]. This assay allowed us to derive binding energies for our TF-DNA pairs by fitting to the 1:1 binding model used in the Octet Analysis software, which is a particular application of the unimolecular TF-DNA binding reaction in equation 10. The steps involved in BLI are as follows: (1) annealing dsDNA; (2) preparing and plating buffered TF and DNA solutions; (3) running the BLI assay; and (4) analyzing the data. These are described in more detail below.

Generation of Double-Stranded Biotinylated DNA

We ordered biotinylated-forward and corresponding reverse oligo to measure TF-DNA binding kinetics in BLI (Supplementary Data 7). Oligos were ordered from IDT and resuspended to 100 μ M in 0.1X TE. We annealed oligos by combining the forward and reverse strands to 10 μ M in annealing buffer (10mM Tris-HCl, 50mM NaCl) in 1.5mL tubes and incubated at 95°C for 5min. Tubes were removed and allowed to cool to room temperature for annealing.

Preparing Solutions and Assay Plate

We used a common binding buffer (20mM Tris-HCl, 0.1mM EDTA, 10mM MgCl₂, 1mM DTT, 120mM KCl, 5% glycerol, and 0.05% Tween-20) for all assays. PdhR was previously reported to be functional in this buffer [36], and the other TFs tested empirically worked well in it. We used the biotinylated-dsDNA as the load molecule, becoming immobilized on the streptavidinized biosensor probe. Load solutions were made by mixing biotinylated-dsDNA at a 1:1 ratio with biotin for a final concentration of 10nM in binding buffer based on suggestions in [37] that spiking in biotin can mitigate potential overloading of BLI probes that confounds analysis.

We used different TF concentrations to measure affinity to DNAs that were or were not designed to have affinity toward the TF. For specific TF-DNA pairs, we determined the maximum TF concentration empirically by finding the concentration where the maximum and ½ the maximum TF concentration equilibrate near the same value. This was found to be 50nM for AIIIR, GlnG, Nac, UlaR, and 100nM for PdhR. Subsequent experiments assayed a 2-fold serial dilution of TF concentrations below the maximum. For nonspecific TF-DNA pairs, more TF was needed to observe a signal, and we uniformly used 100nM, 250nM, 500nM, 750nM, and 1µM concentrations in those cases.

We prepared a soak plate for probe preincubation by adding 200µL binding buffer to each well of a Greiner 96W Plate, PP, Flat Bottom, Chimney Style, Black (#655209) that would contain an assay probe. The assay plate was prepared by adding 80µL of solutions to a Greiner 384W Plate, PP, Flat Bottom, Black (#781209). Each probe required 3 wells in the assay plate: (1) buffer for baselining and dissociation; (2) loading well with biotin and biotinylated-dsDNA; (3) analyte well with or without TF. For each TF-DNA pair, we tested 6 TF concentrations using 8 probes. One probe was used as a “reference sensor”, which sees buffer wells for the entire assay and allows subtraction of any background signal. Another probe was used as a “reference sample”, which has DNA loaded onto it but sees no TF and allows subtraction of the effect of DNA loading. The remaining 6 probes assayed the response of different TF concentrations to the same concentration of DNA.

Running BLI

Each BLI assay measured the binding of a TF-DNA pair over 5 steps: (1) initial baseline; (2) DNA loading; (3) second baseline; (4) TF association; and (5) TF dissociation. All steps were performed on an Octet Red 384 at 30°C with shaking at 1000rpm using SA-tips (Sartorius #18-5020). Data were acquired with Sartorius Octet BLI Discovery 13.0. Before a group of assays began, probes were allowed to soak in binding buffer for a 10-minute pre-incubation. Assays began with a 60 second baseline step, where probes read from buffer wells. Next, probes were moved into wells containing biotinylated DNAs, which were loaded onto the probe’s surface for 2 minutes. A second baseline was taken in buffer for 60 seconds. Probes were then moved into TF-containing wells for 10 minutes, allowing TF to associate with DNA. This step was optimized such that enough time was given for the 2 highest TF concentrations to reach equilibrium. Finally, probes were moved back to buffer wells with no TF for 15 minutes, and TF-DNA complex dissociation was recorded. This step was also optimized such that enough time was given to observe dissociation of all TF concentrations.

Analyzing BLI

Our BLI analysis followed that of the 1:1 Binding Model used in the Octet Analysis software, which we implemented in MATLAB to determine binding energies for our designed sites. This model describes the BLI assay as a unimolecular binding reaction. When solved analytically using the assay parameters as initial

conditions for association and dissociation, this model gives rise to one equation per phase that allows simultaneous fitting of association and dissociation, which we use to derive binding energies.

In association, we have:

$$y_A(t) = \frac{R_{max}}{(\frac{K_D}{[TF]_0} + 1)} (1 - e^{-(k_f[TF]_0 + k_r)t}) \quad (41)$$

Where R_{max} is the maximum response in BLI across all traces for a given TF, k_f is the rate of TF and DNA binding, k_r is the rate of TF-DNA unbinding, K_D is the dissociation constant (equation 13), and $[TF]_0$ is the initial concentration of TF for a given trace.

In dissociation, we have:

$$y_D(t) = Y_A e^{-k_r(t-t_A)} \quad (42)$$

Where t_A is the time at the end of association, and Y_A is the signal at the end of association, given by:

$$Y_A(t) = \frac{R_{max}}{(\frac{K_D}{[TF]_0} + 1)} (1 - e^{-(k_f[TF]_0 + k_r)t_A}) \quad (43)$$

Taken together, we have a pair of equations that completely describes the association and dissociation phases of our BLI assay.

To fit our data to these equations, we started by processing raw BLI traces. Reference sensor and sample traces were subtracted from all traces where TF was present, which removed background noise and the effect of DNA loading from the signal. The beginning of the association phase was zeroed to the average value from the last second of the pre-association baseline. The beginning of the dissociation phase was aligned to the average value from the last second of the association phase. Processed traces were fit to equations 41 and 42. Resulting kinetic constants were converted to equilibrium constants through equation 13, and to binding energies through equations 22 and 23 (Supplementary Data 8).

1

3xFLAG atgGCCTACAGC...CAACGAAAGAATggcggtggcGACTACAAAGAC...TGAGCTCCCAAC

SPA atgGCCTACAGC..CAACGAAAGAATtccatggaaAAGAGAAGATGG...GATATTCATATG

V5 atgGCCTACAGC..CAACGAAAGAATggcgggtggcgacGGTAAACCTATT...GATTCAACATAG

TF Sequence Minus
Stop Codon

Linker

Tag Sequence

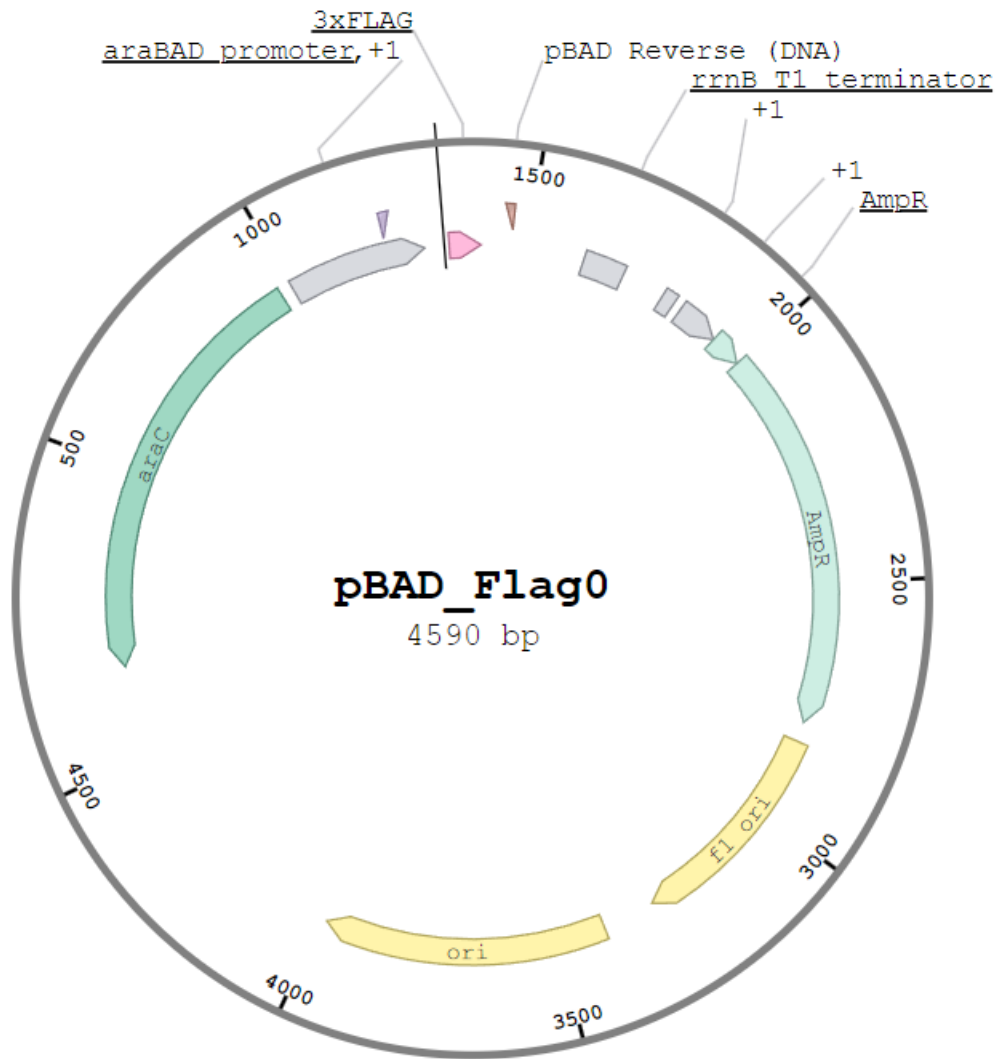
2

Figure S1: Examples of TF Tags. Tagged TF strains were generated by cloning tag sequences onto the C terminal of a TF's sequence, omitting its stop codon (green). Tags contained a short linker sequence (black) between the TF and tag sequence (orange).

5

6

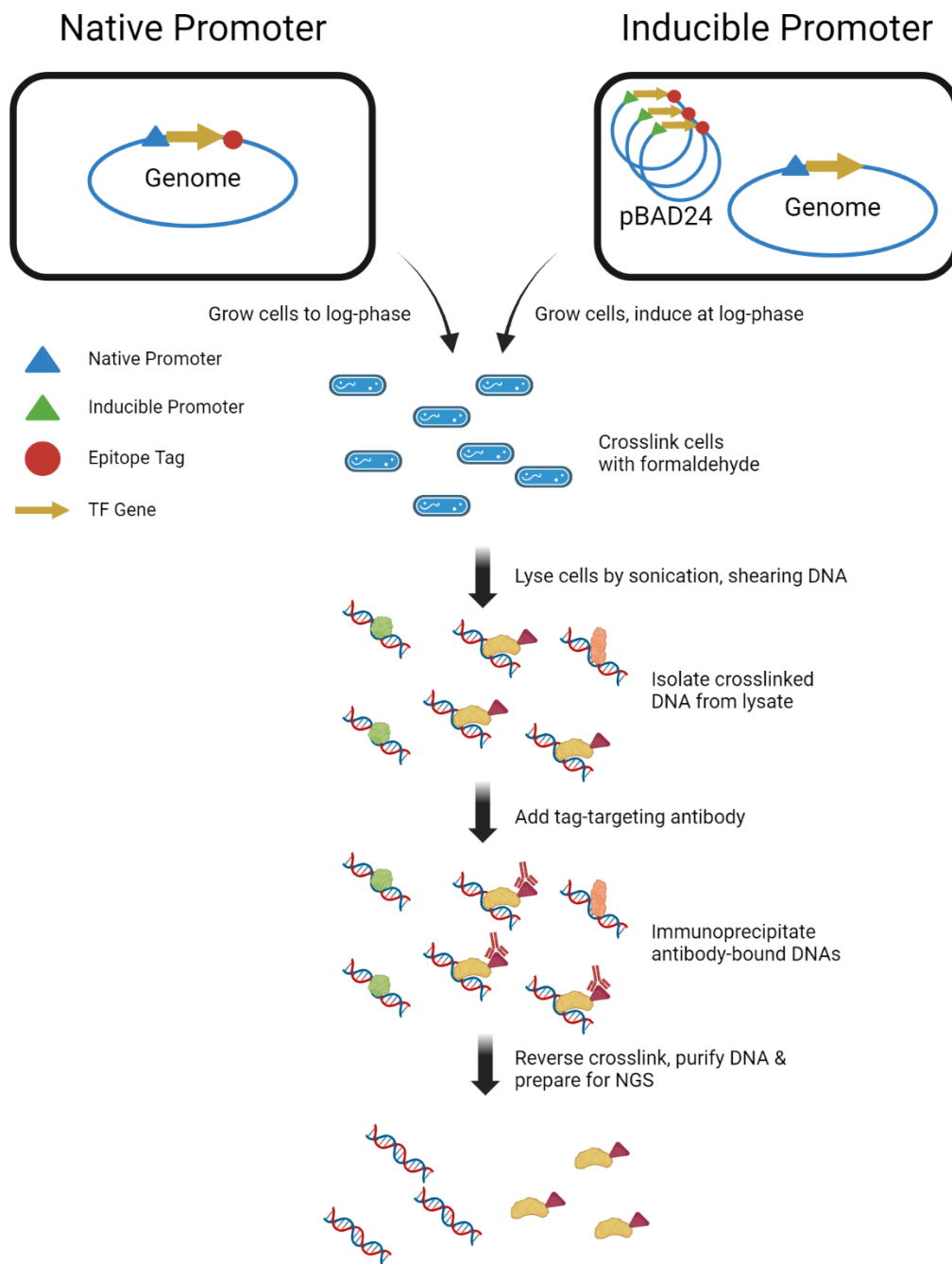
1



2

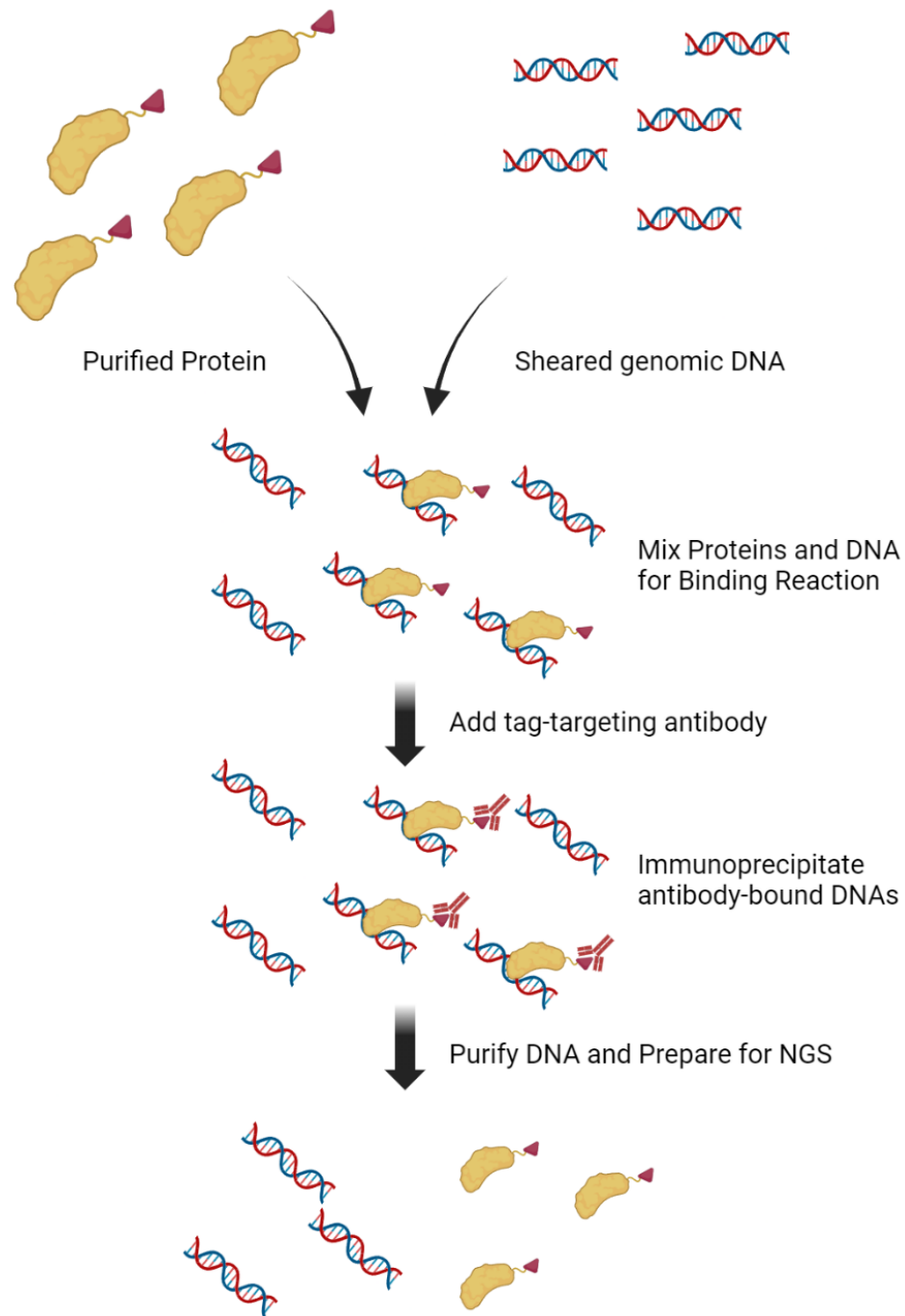
3 **Figure S2: pBAD24 Plasmid Map.** Our inducible ChIP-Seq experiments are facilitated by cloning different TFs into the pBAD24 plasmid
 4 between the araBAD promoter and a 3xFLAG epitope tag. Successful clones can be selected for resistance to ampicillin due to the
 5 plasmid's AmpR gene. Tagged-TF expression can then be induced using arabinose which causes AraC to drive transcription from
 6 the araBAD promoter.

1



2

3 **Figure S3: ChIP-Seq Experiment Schematic.** We started by growing cells to logarithmic growth, at which point we may have added
4 treatments (e.g. arabinose) as necessary. We crosslinked cells to fix TFs to any DNA sequences they were bound to, and lysed the
5 cells via sonication, shearing the DNA at the same time. We isolated DNA fragments and added a tag-specific antibody to allow
6 immunoprecipitation of fragments bound by the TF of interest. These fragments are purified and turned into NGS libraries for
7 Illumina sequencing. Created in BioRender. Lally, P. (2025) <https://BioRender.com/o72a005>



1

2 **Figure S4: *In Vitro* ChIP-Seq Experiment Schematic.** We started by combining a purified tagged-TF with pre-sheared genomic
 3 DNA. Next, we added a tag-specific antibody to immunoprecipitate DNA fragments bound by the TF. These fragments are purified
 4 and turned into NGS libraries for Illumina sequencing. Created in BioRender. Lally, P. (2025) <https://BioRender.com/o72a005>

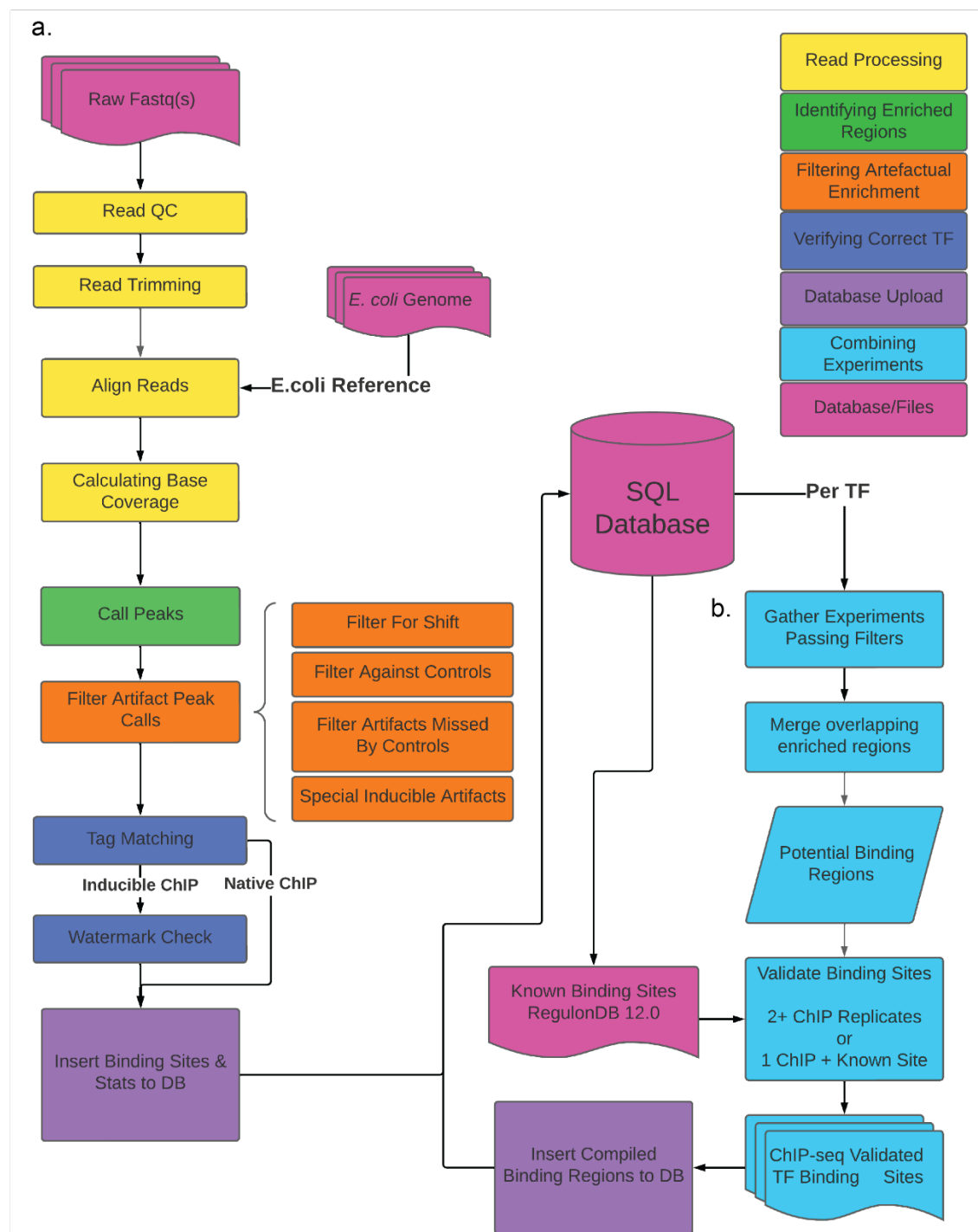
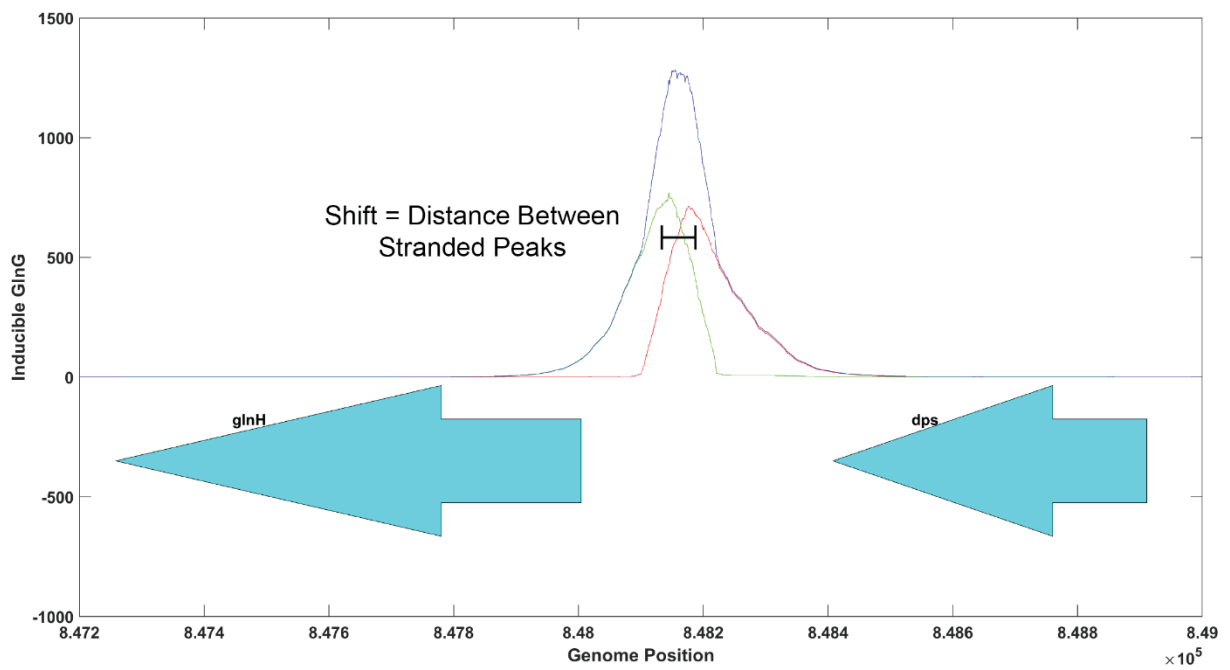


Figure S5: ChIP-Seq Analysis Workflow. For each experiment we uniformly processed our NGS reads to call high-confidence enriched regions. **(a)** We began by assessing read quality and trimming low-quality bases and Illumina adapters from reads. We aligned reads to the *E. coli* genome and calculated the number of reads aligning to each strand at each base. We identified enriched regions and filtered likely artefactual enrichment. Finally, we performed custom QC to verify that the correct TF was tagged. All analysis results are imported to our MySQL Database. **(b)** After all TF ChIP samples have been analyzed, we combine enriched regions across a TF's replicates to identify high-confidence binding regions.



1

2

Figure S6: Peak Shift Example. Enriched regions that are a result of TFs binding point sources will appear with a characteristic shift between forward (green) and reverse (red) strand coverage as a consequence of sequencing bound fragments from the outside-in.

3

4

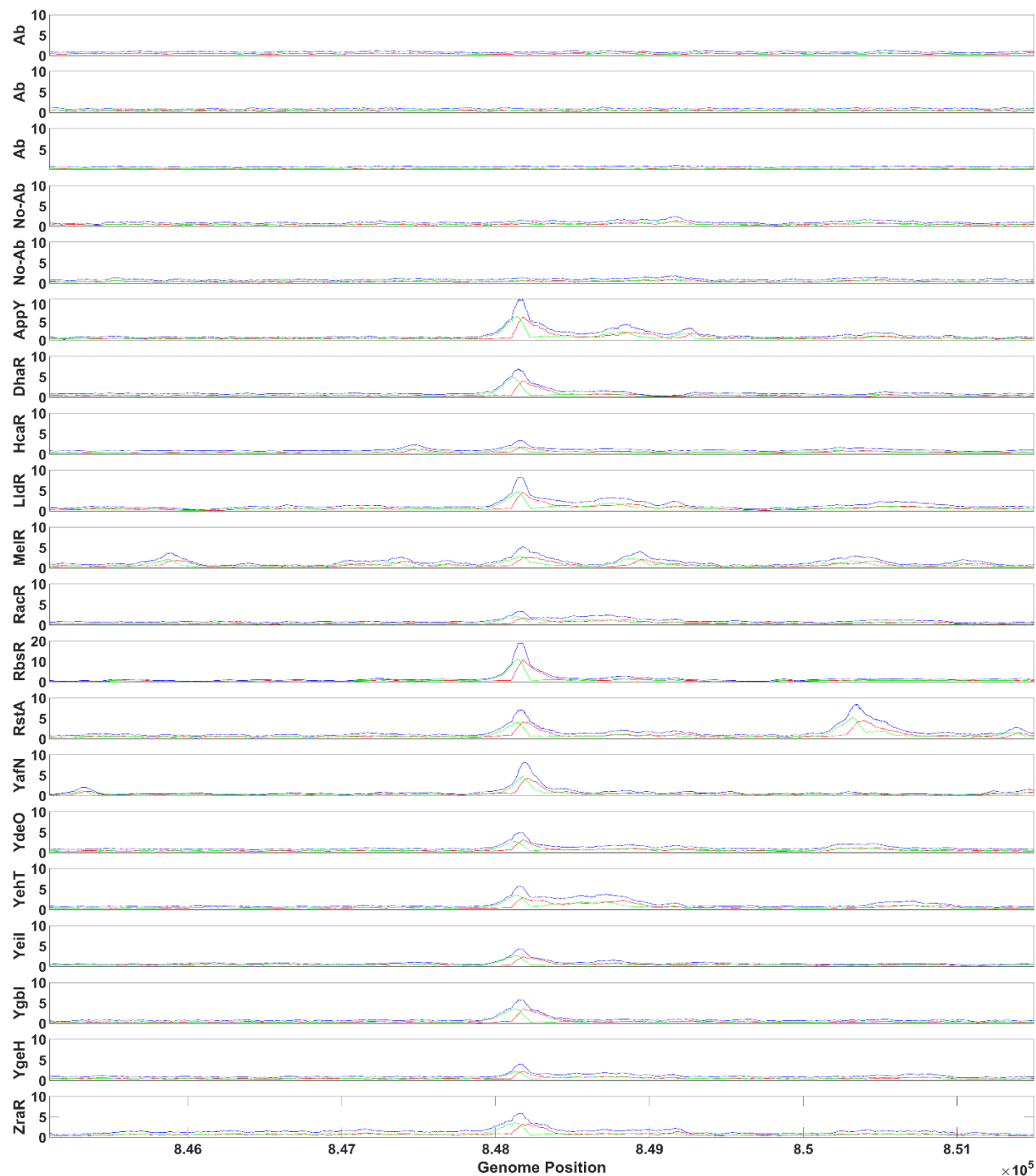
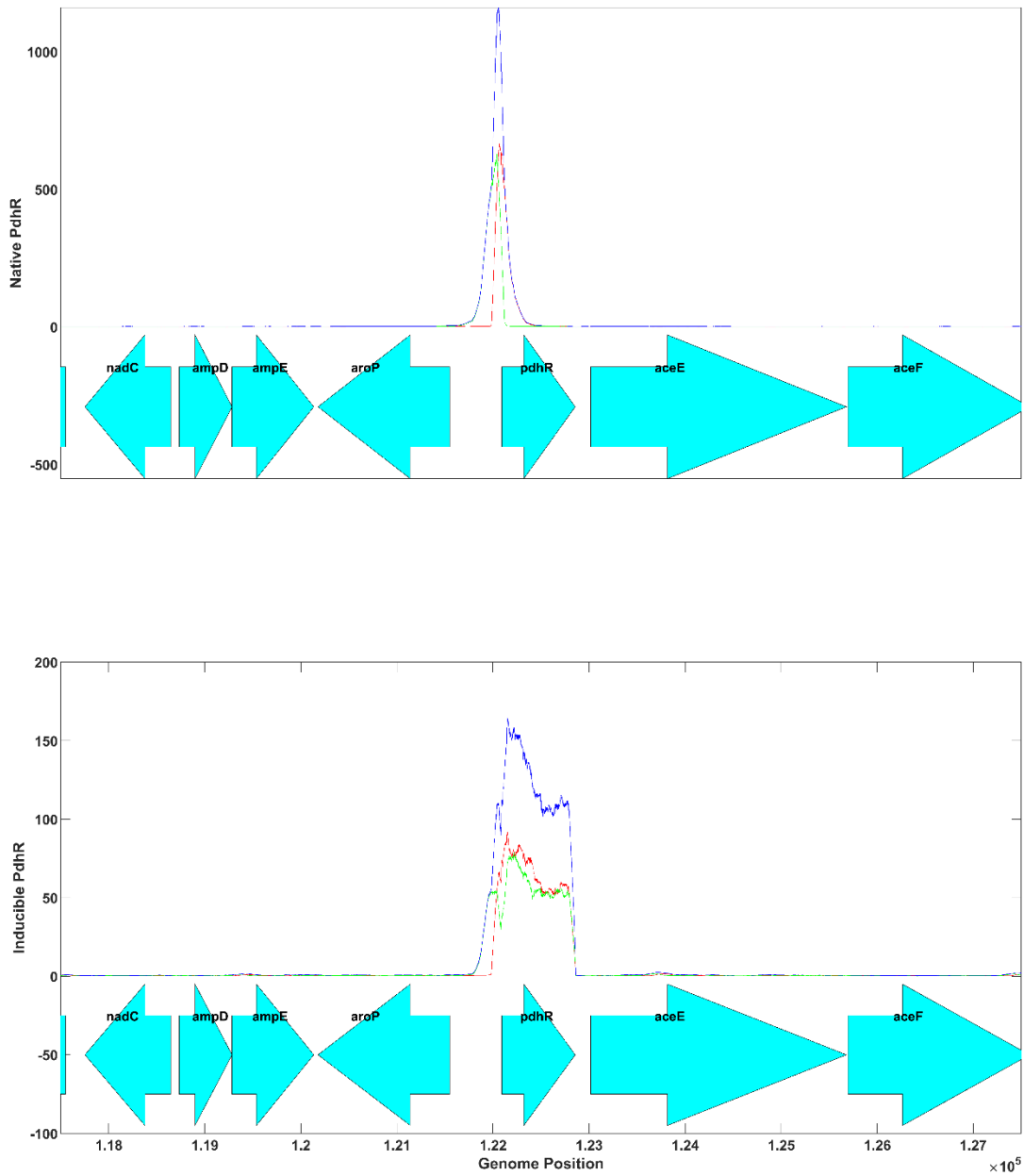


Figure S7: ChIP-Seq Artifacts are not Always Apparent in Controls. We have identified several genomic regions that appear to be artifacts on account of their ubiquity across TF ChIP-Seq datasets yet are not revealed in control experiments. The first 5 rows show different control experiments we performed, where this artifact region is not found. The following 15 tracks show a sample of different TFs where this region nearly always appears enriched and is therefore artefactual.



1

2 **Figure S8: High-Copy Plasmids Create Watermarks.** Coverage plot over PdhR's genomic locus showing a natively tagged
3 experiment (top), and inducible experiment (bottom). Because of the high copy pBAD24 plasmid, we see significant enrichment
4 across the entire PdhR sequence that is not present in the native experiment and can tarnish the appearance of the autobinding
5 region seen clearly in the native experiment.

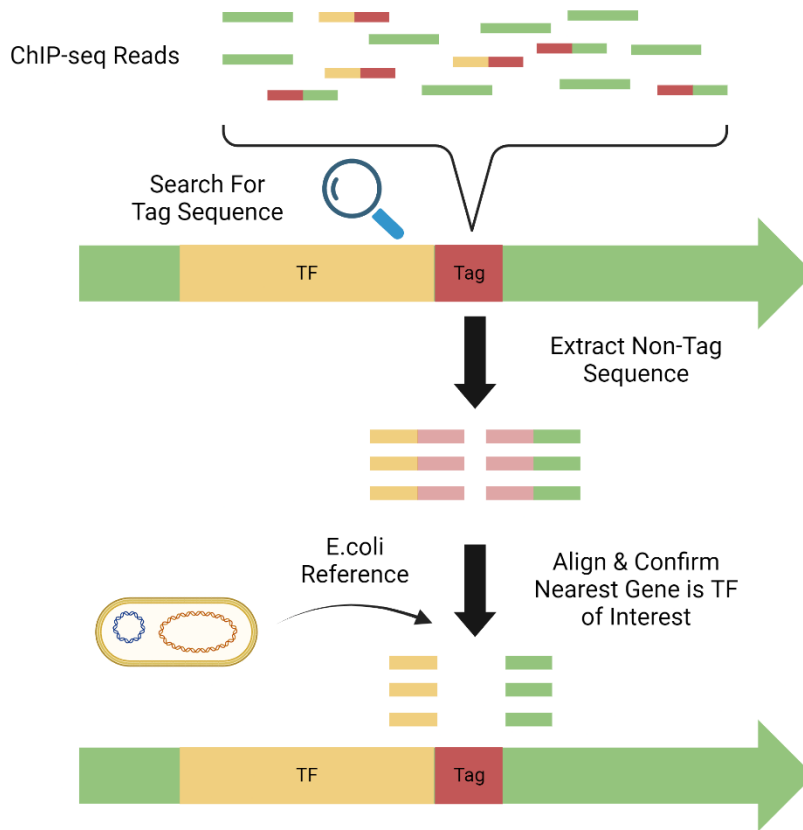
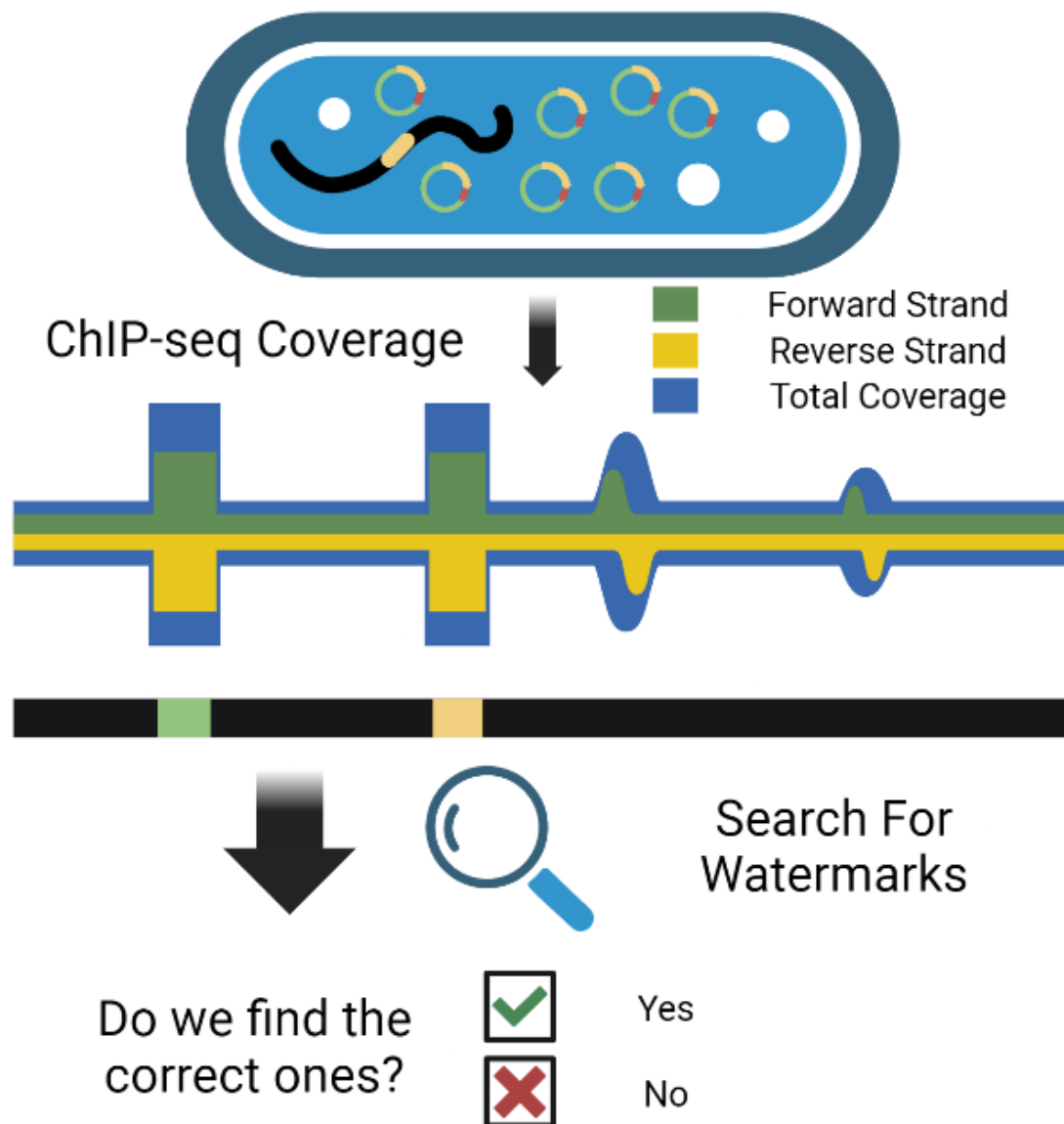


Figure S9: Tag Matching Schematic. We search raw ChIP-Seq reads for sequences matching the epitope tag cloned onto the TF's C terminus. The fragment of the read that does not match the epitope tag is aligned to the *E. coli* genome and any genes that the fragment overlaps are identified. For successful ChIP-Seq, we expect to see nearly all matched reads aligning at the end of the TF of interest. Created in BioRender. Lally, P. (2025) <https://BioRender.com/r38v839>

1



2

3 **Figure S10: Watermark Detection Schematic.** We perform QC on inducible ChIP-Seq experiments by searching for signatures of
 4 "watermarks". These appear visually as smears of coverage over the coordinates of genes on the inducible plasmid and are unlike
 5 ChIP-Seq peaks in two ways: 1) watermarks lack the peak shift characteristic of a real binding sites, and 2) coverage is enriched
 6 consistently across the length of the gene, often with much higher enrichment than ChIP-Seq peaks. Because our plasmids contain
 7 AraC (for induction) and the TF of interest, we consider a successful ChIP-Seq experiment one where at most these two watermarks
 8 are detected. Created in BioRender. Lally, P. (2025) <https://BioRender.com/w44d280>
 9

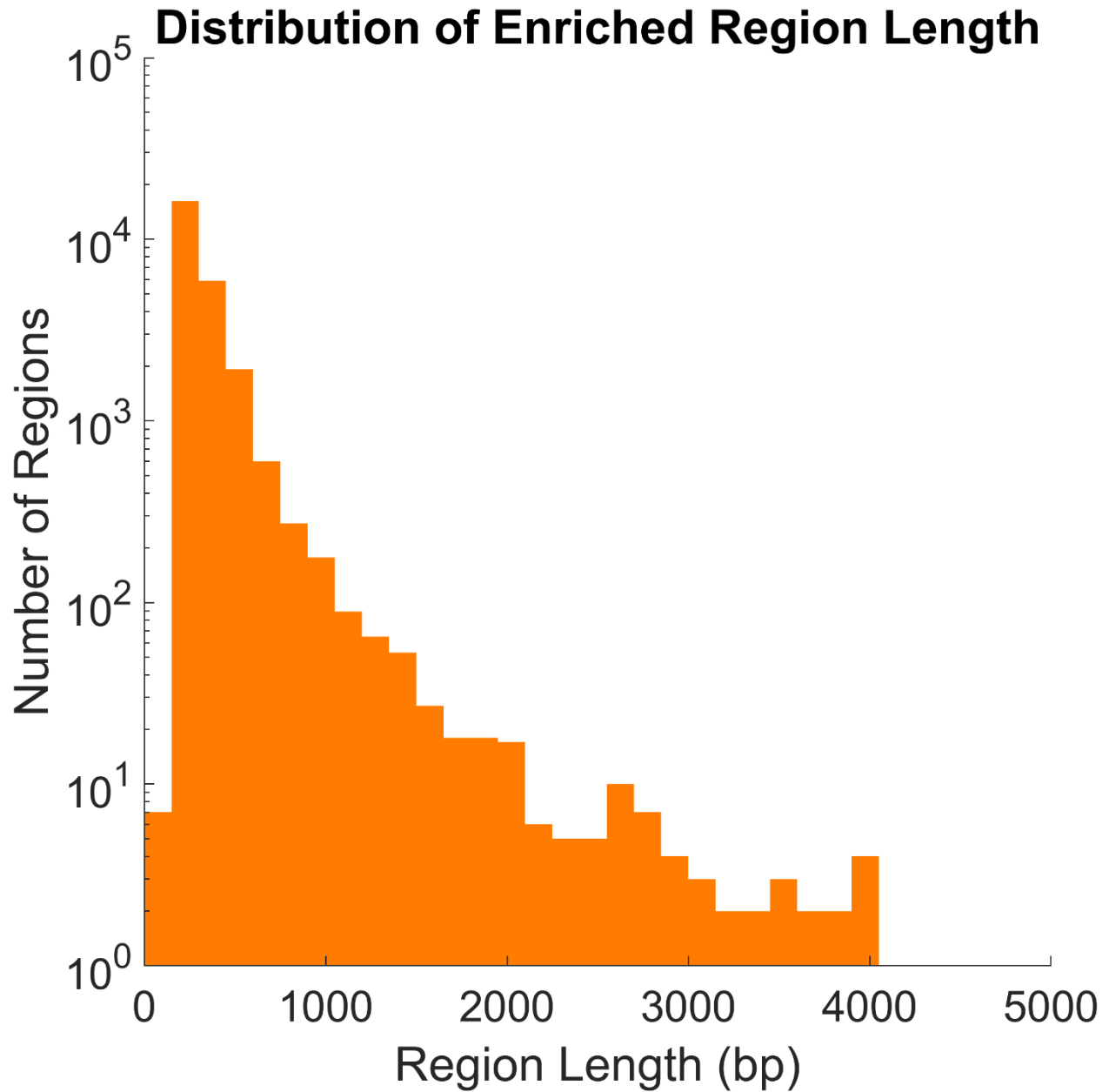


Figure S12: Distribution of Enriched Region Length. Histogram of enriched region lengths, with frequencies shown on the y-axis in log-scale. With the exception of a few special cases near the TF of interest in inducible ChIP-Seq, we required enriched regions to be at least 150 bp in length. The vast majority of regions were 500 bp or shorter (~90%).

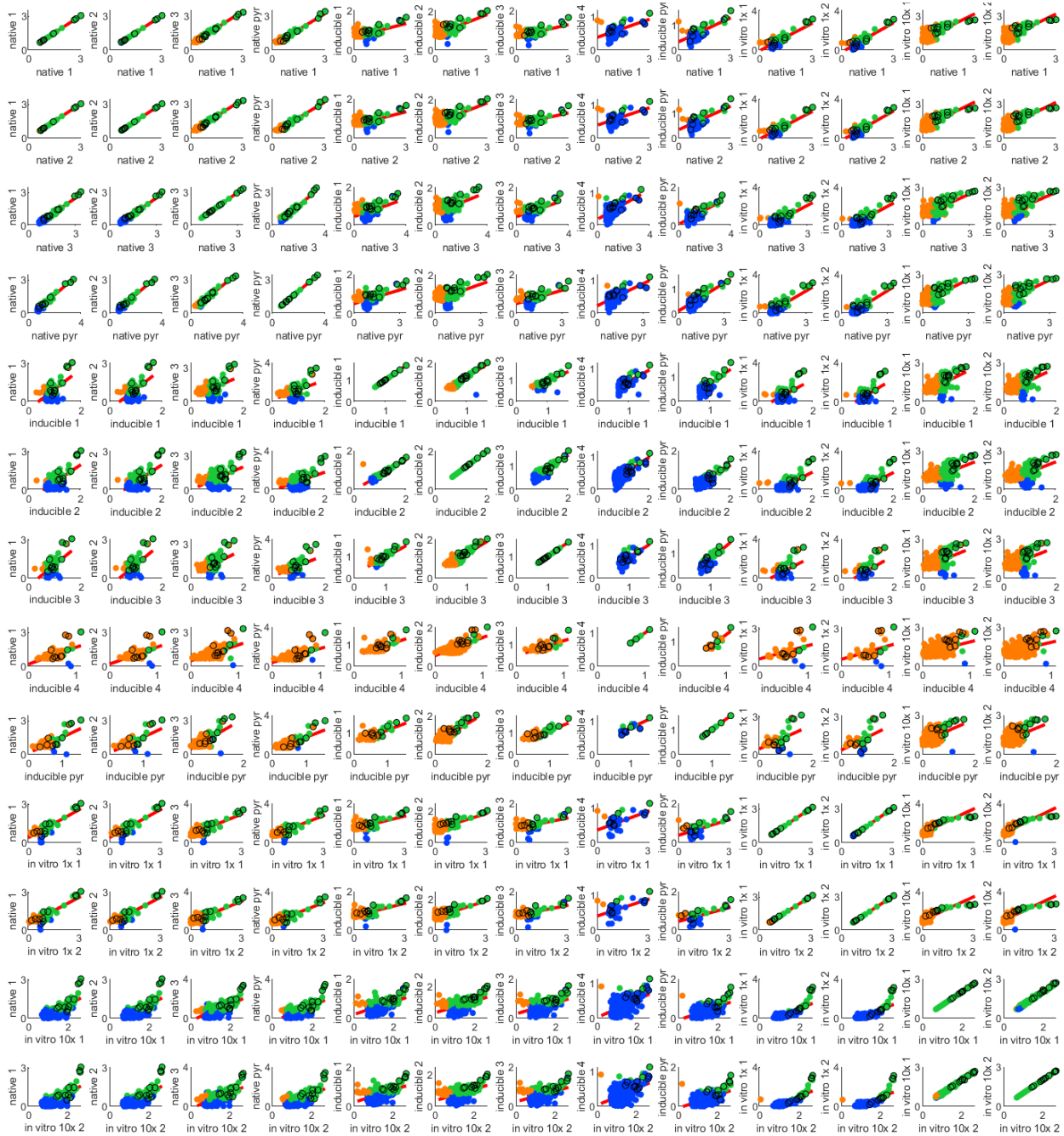


Figure S13: Enrichment Scatterplots of All PdhR Replicate Comparisons. Across 13 experiments performed *in vivo* (native or inducible) and *in vitro*, we see close agreement even across expression modes (median $R^2 = 0.65$). Data point colors follow that of Figure 1E where peaks found in both experiments are green; those found only in the x-axis experiment are blue; those only in the y-axis experiment are orange; and previously reported binding sites are shown with a black border. Between technical replicates of the same mode, we see exceptional agreement (median native $R^2 = 0.98$, median inducible $R^2 = 0.76$, median *in vitro* $R^2 = 0.88$). We also see the effect of saturation as a function of TF levels, where binding sites found in low TF-level (e.g., native) experiments have enrichment closer to the strongest binding site, and weaker sites not found with low-TF are revealed. This can be seen clearly with *in vitro* ChIP-Seq, where experiments with 10x TF find all of the sites from the 1x experiments with a tight range of enrichment, as well as a large number of sites that were lowly-enriched (and not identified) in the 1x experiments.

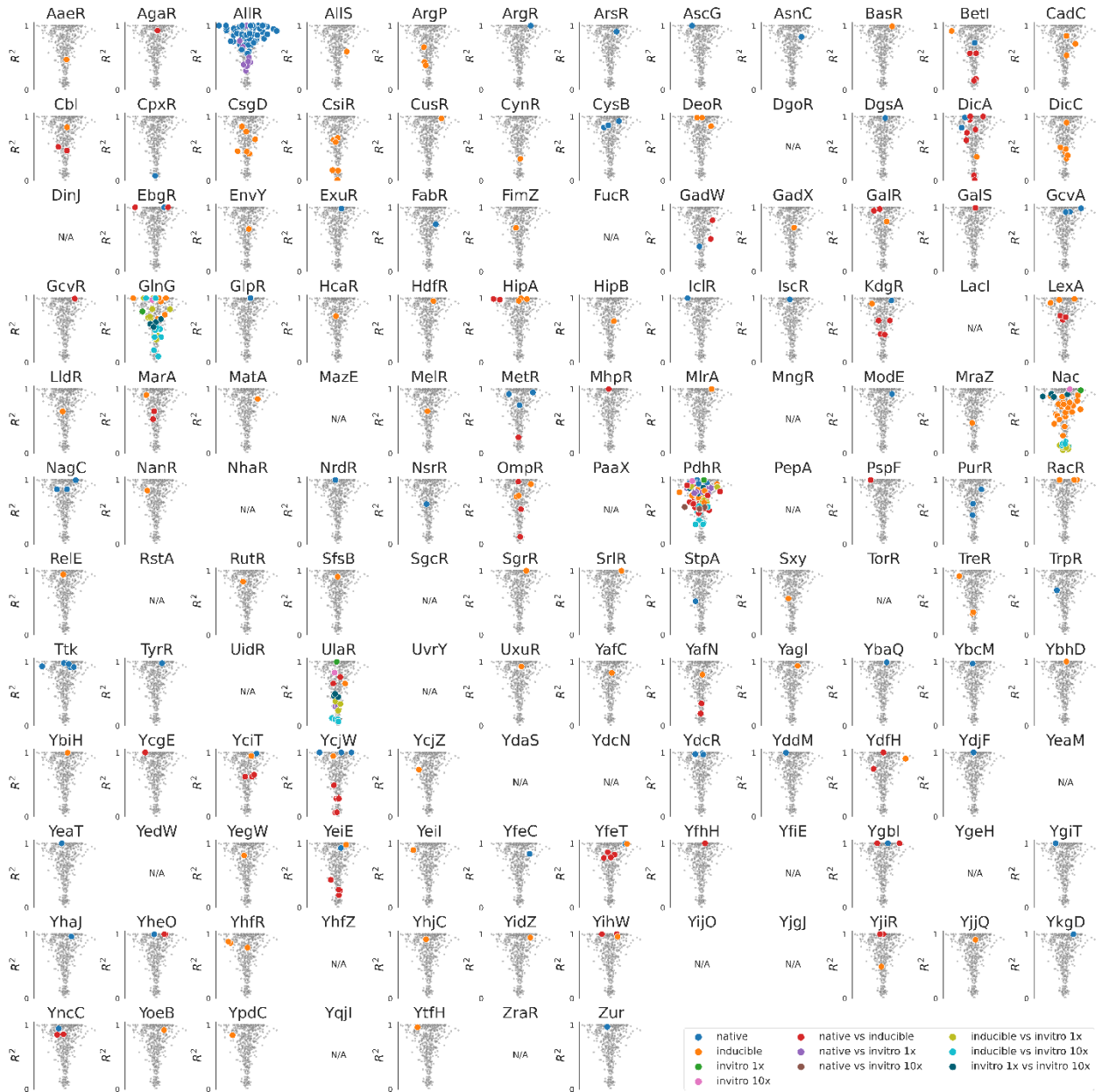


Figure S14: Correlation Analysis of All TFs. Each plot shows R^2 values of all ChIP-Seq replicates (gray) with values corresponding to the titular TF displayed in color depending on the pair of experiments used (blue – native vs native; orange – inducible vs inducible; green – *in vitro* 1x vs *in vitro* 1x; pink – *in vitro* 10x vs *in vitro* 10x; red – native vs inducible; purple – native vs *in vitro* 1x; brown – native vs *in vitro* 10x; yellow – inducible vs *in vitro* 1x; cyan – inducible vs *in vitro* 10x; teal – *in vitro* 1x vs *in vitro* 10x). N/A present in subplots for TFs where only one peak was in common for all pairs of experiments. Correlation of like-expression modes tend to have higher correlation than pairs of different expression modes.

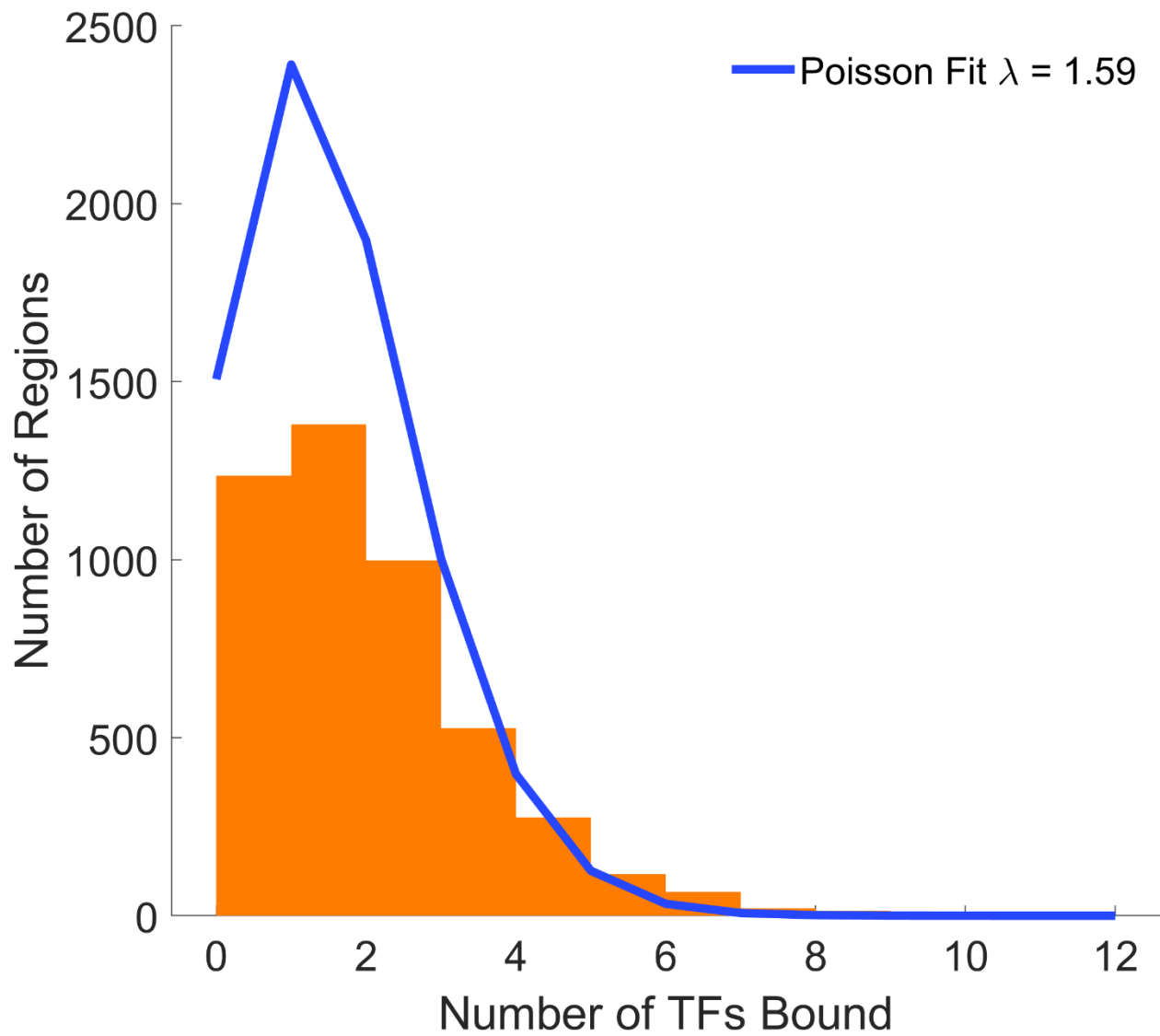


Figure S15: Distribution of Number of TFs Binding All Non-Overlapping 1kb Regions. Most regions only have at most 2 TFs binding within them while a handful of regions have several TFs binding (as many as 11).

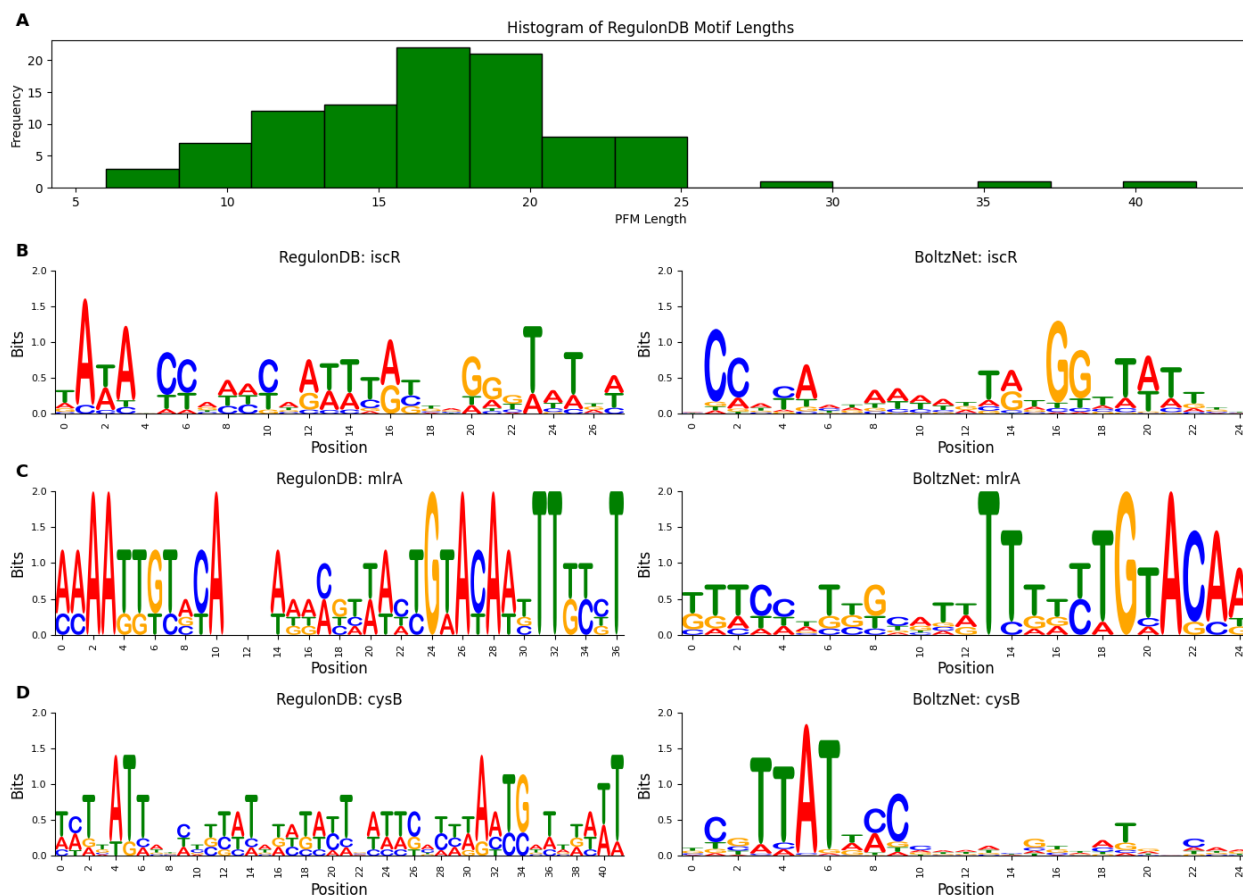


Figure S16: Selection of 25 bp for BoltzNet weight matrix. (A) Distribution of RegulonDB known motif length. A matrix of length 25 was selected for all models reported in this manuscript as the majority of known motifs for *E. coli* in RegulonDB are less than 25 bp in length. (B-D) Out of 97 motifs in RegulonDB, the motifs for only 3 TFs are longer than 25 bp, and in these three cases the RegulonDB motifs appear to include repeats of a core motifs (left column). In all three cases, the Boltznet models for these three TFs learn a shorter motif consistent with this interpretation (right column).

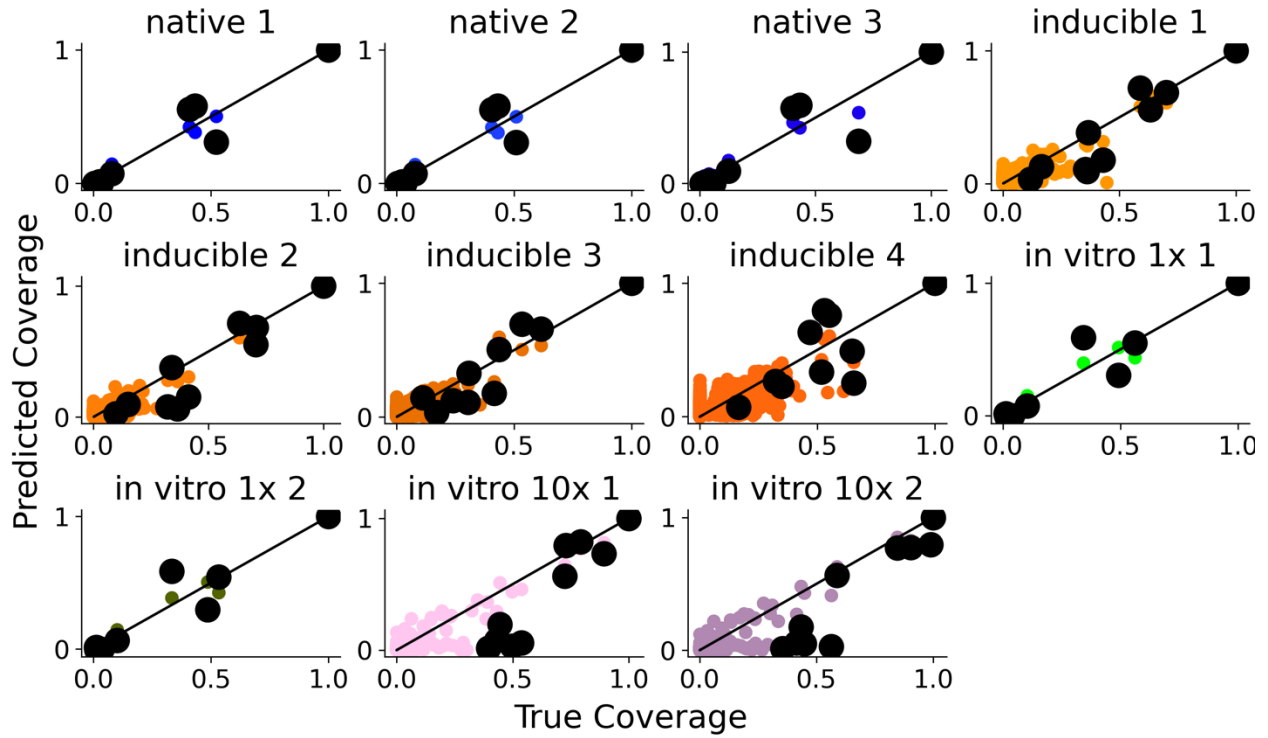


Figure S17: Prediction of Normalized Coverage and Cross-Validation on Training Set for All Experiments for PdhR. Black circles show leave-one-out cross-validation predictions. All coverage is normalized between zero (baseline) and 1 (strongest called peak). Native= native promoter replicate number, inducible= inducible promoter replicate number, *in vitro* #x = *in vitro* ChIP-Seq replicate numbers at two different protein concentration.

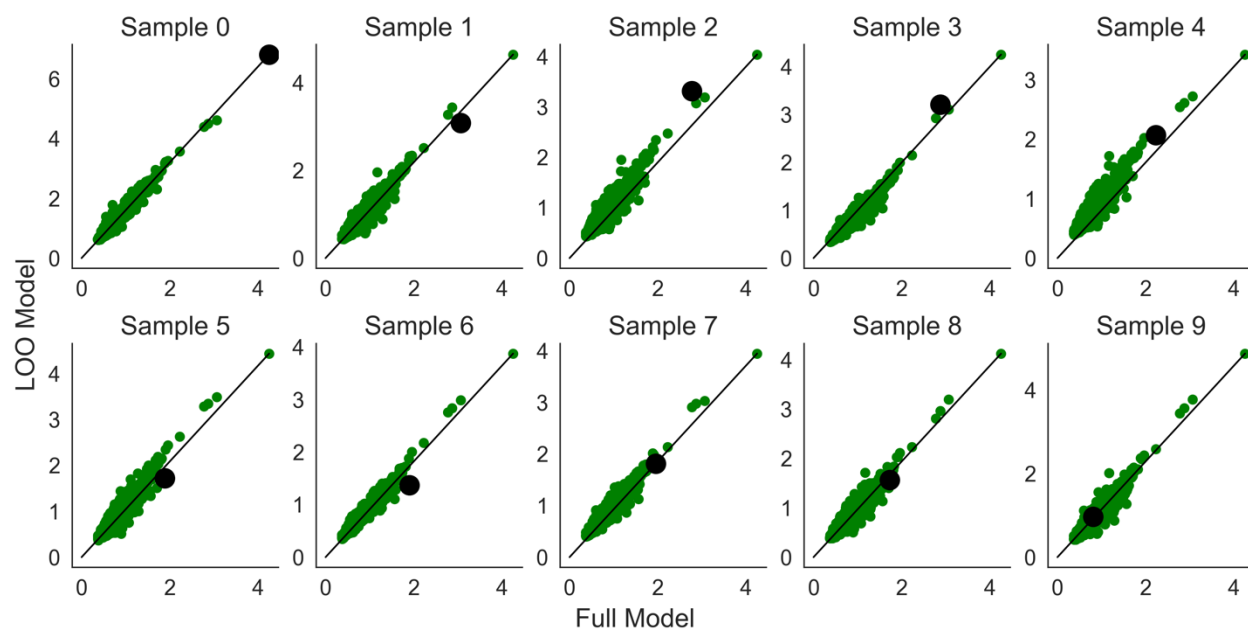


Figure S18: Leave-One-Out Cross-Validation of Affinity Scores for PdhR. For each of 10 samples starting with the strongest binding region, the sample was left out and an LOO model trained. For each LOO model, we plot the affinity score predictions LOO model against the predictions of the model built on the full dataset. The dark circle is the LOO sample.



Figure S19: Alignments of motifs for all models with known RegulonDB motifs. PFM for BoltzNet models were generated using a threshold of 0.6 applied to all called peaks (see Methods in Main document).

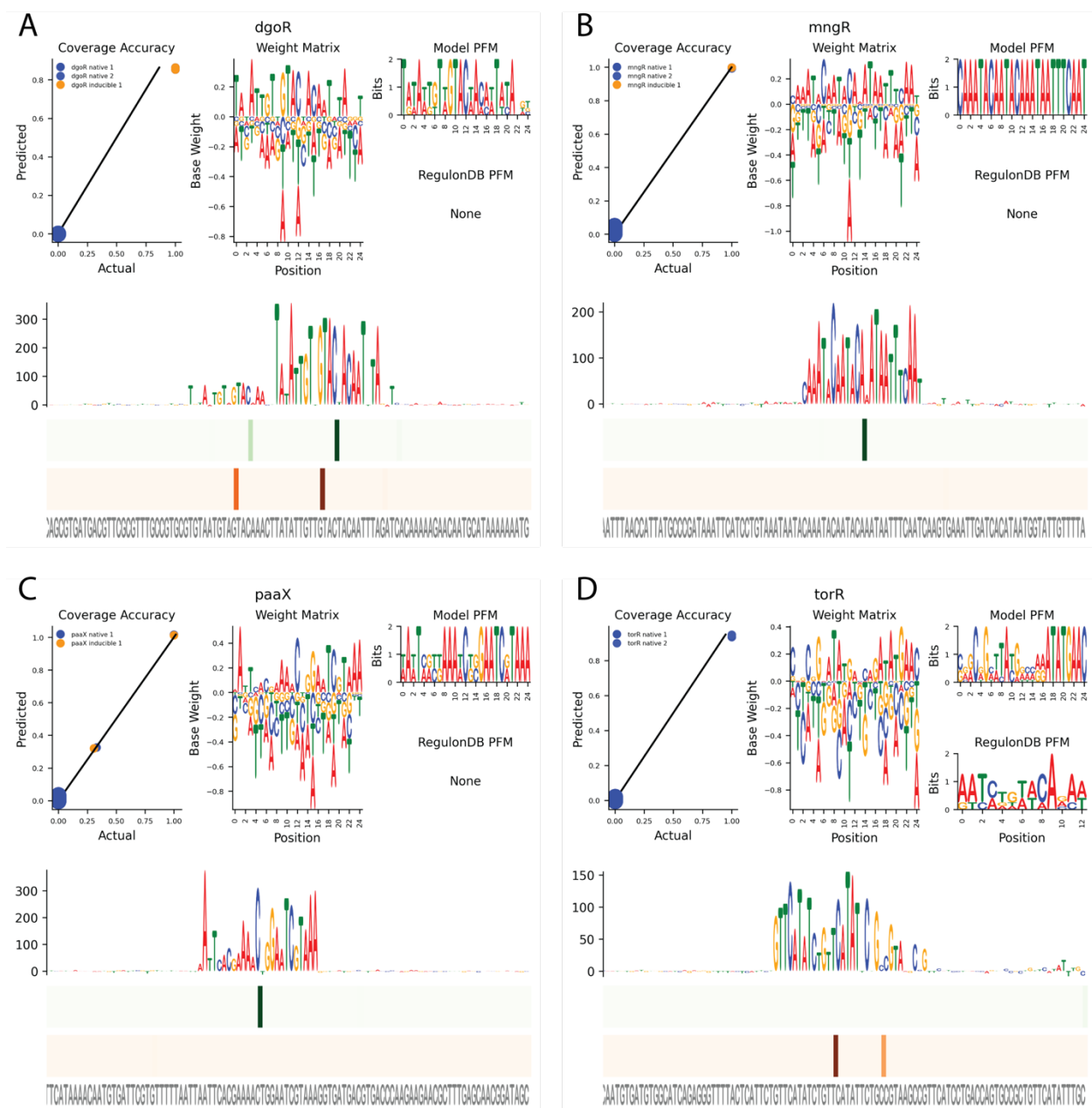


Figure S20: Examples of Models Built for TFs with Only a Single Called Binding Region. Four panels for TFs (A) dgoR, (B) mngR, (C) paaX, and (D) torR. Top row of each panel displays a scatter plot of peak coverage prediction accuracy (left) and the weight matrix (right). Points may not be visible due to nearly perfect overlap. The bottom half of each panel shows nucleotide level resolution predictions for the binding site for each TF. Sequence logo shows base contribution score. Cyan shading shows known binding site regions. Heatmaps shows predicted affinity at each position in positive (green) and negative (red) orientation. Sequence for binding site region show in gray.

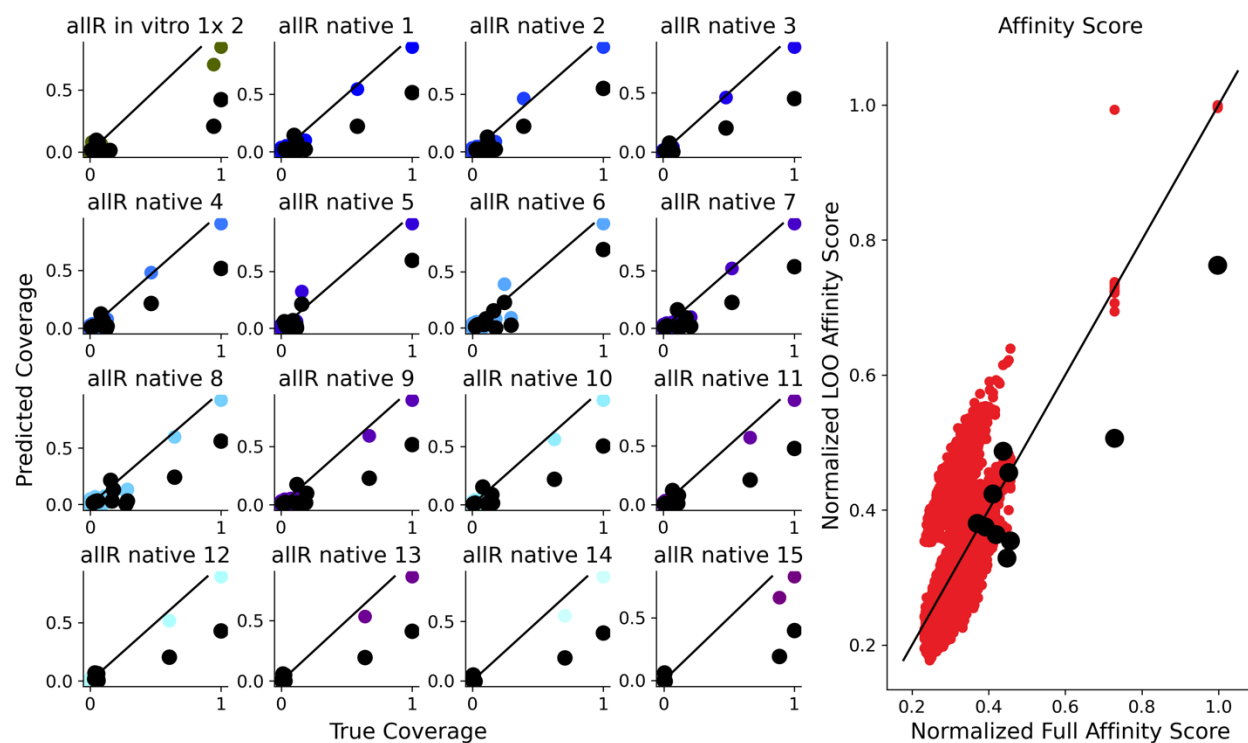


Figure S21: Leave-One-Out Cross-Validation of Coverage prediction and Affinity Scores for AIIR. For each of 10 samples starting with the strongest binding region, the sample was left out and an LOO model trained. In the left panes, we plot the predicted coverage against the true coverage for each LOO model (dark points) and the model trained on all peaks (colored points) in all experiments. In the right panel, we plot the affinity score predictions of each LOO model (dark points) against the predictions of the model built on the full dataset (red points) normalized across experiments so that all data can be visualized in one plot (as in Figure 2B).

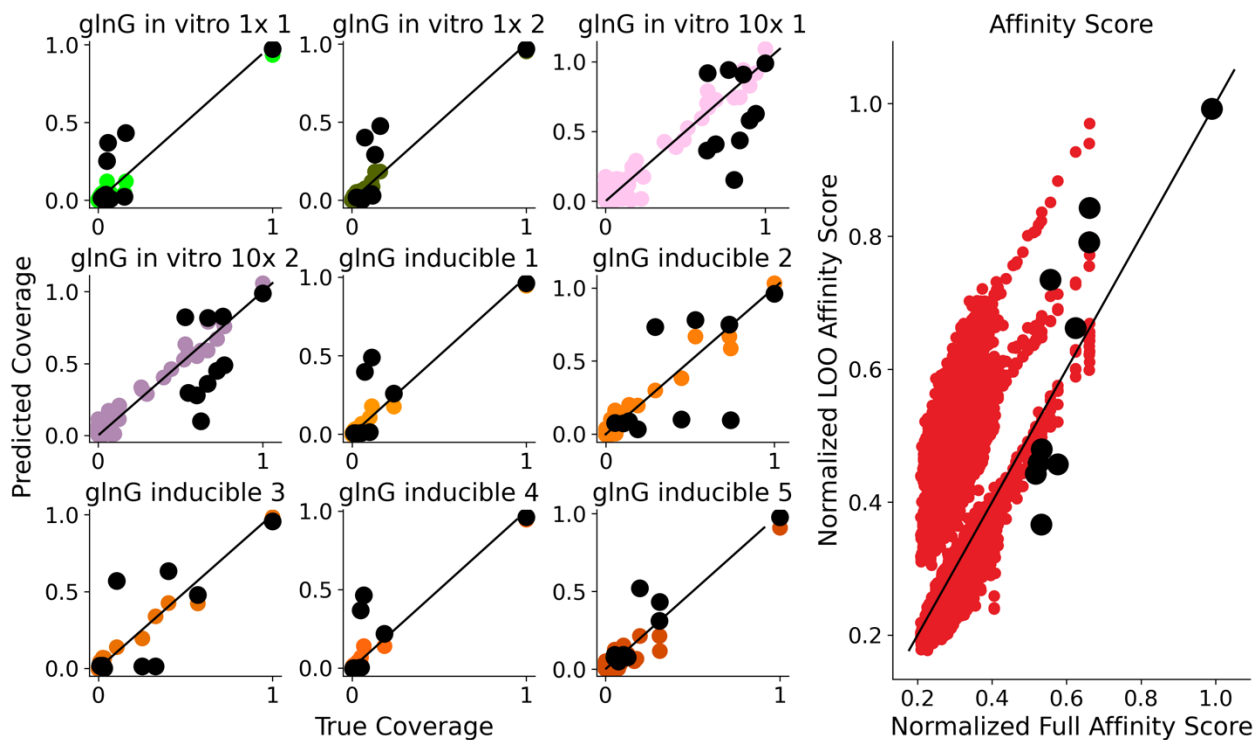


Figure S22: Leave-One-Out Cross-Validation of Coverage prediction and Affinity Scores for GlnG. For each of 10 samples starting with the strongest binding region, the sample was left out and an LOO model trained. In the left panes, we plot the predicted coverage against the true coverage for each LOO model (dark points) and the model trained on all peaks (colored points) in all experiments. In the right panel, we plot the affinity score predictions of each LOO model (dark points) against the predictions of the model built on the full dataset (red points) normalized across experiments so that all data can be visualized in one plot (as in Figure 2B).

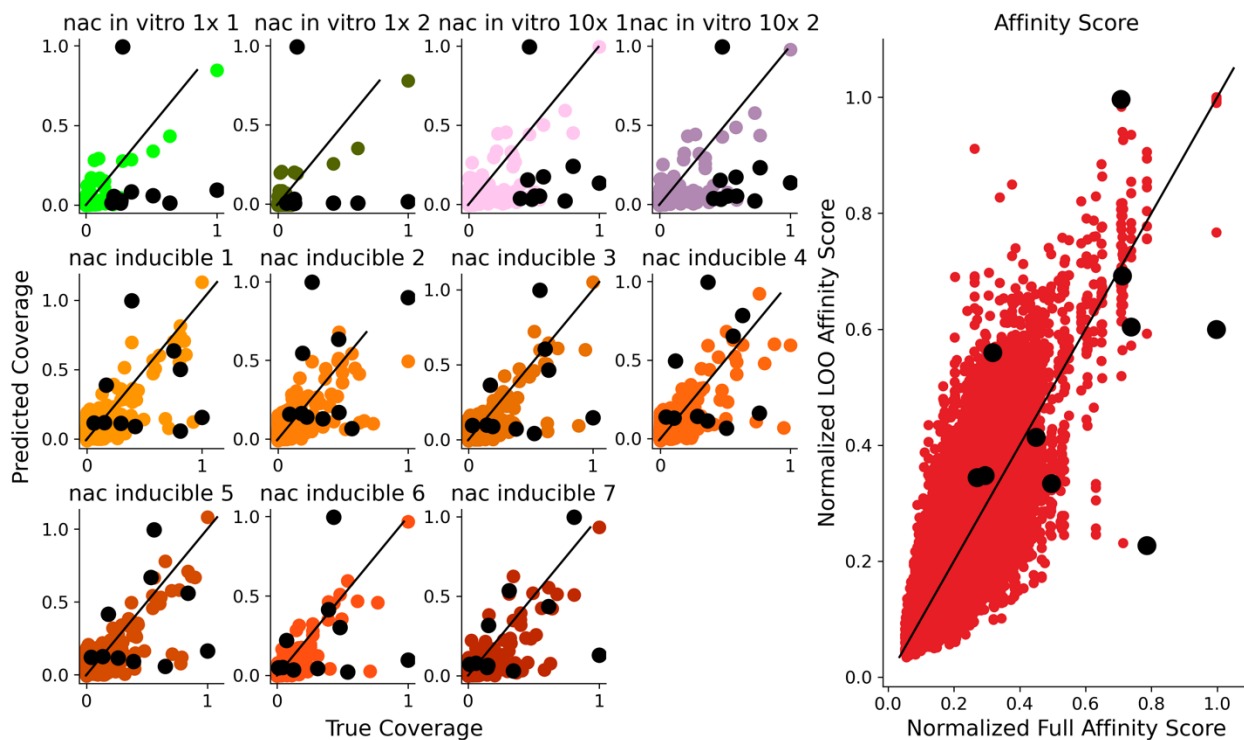


Figure S23: Leave-One-Out Cross-Validation of Coverage prediction and Affinity Scores for Nac. For each of 10 samples starting with the strongest binding region, the sample was left out and an LOO model trained. In the left panes, we plot the predicted coverage against the true coverage for each LOO model (dark points) and the model trained on all peaks (colored points) in all experiments. In the right panel, we plot the affinity score predictions of each LOO model (dark points) against the predictions of the model built on the full dataset (red points) normalized across experiments so that all data can be visualized in one plot (as in Figure 2B).

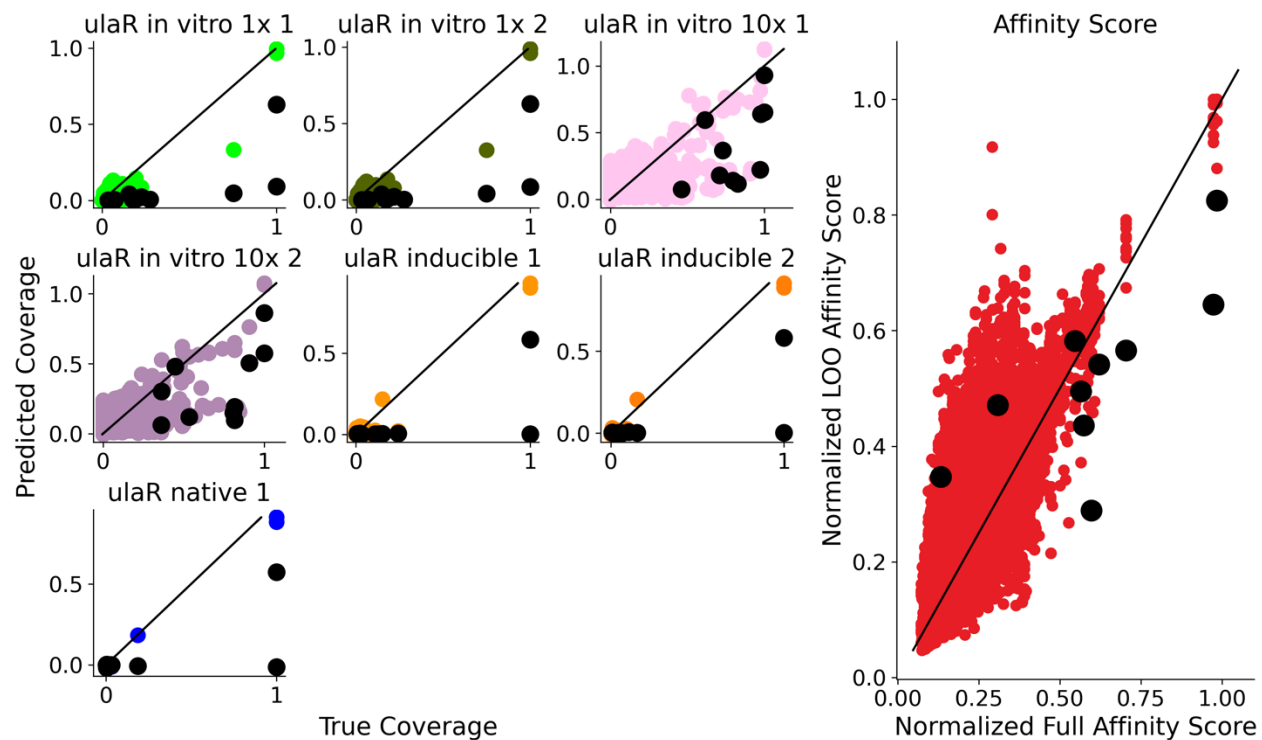


Figure S24: Leave-One-Out Cross-Validation of Coverage prediction and Affinity Scores for UlaR. For each of 10 samples starting with the strongest binding region, the sample was left out and an LOO model trained. In the left panes, we plot the predicted coverage against the true coverage for each LOO model (dark points) and the model trained on all peaks (colored points) in all experiments. In the right panel, we plot the affinity score predictions of each LOO model (dark points) against the predictions of the model built on the full dataset (red points) normalized across experiments so that all data can be visualized in one plot (as in Figure 2B).

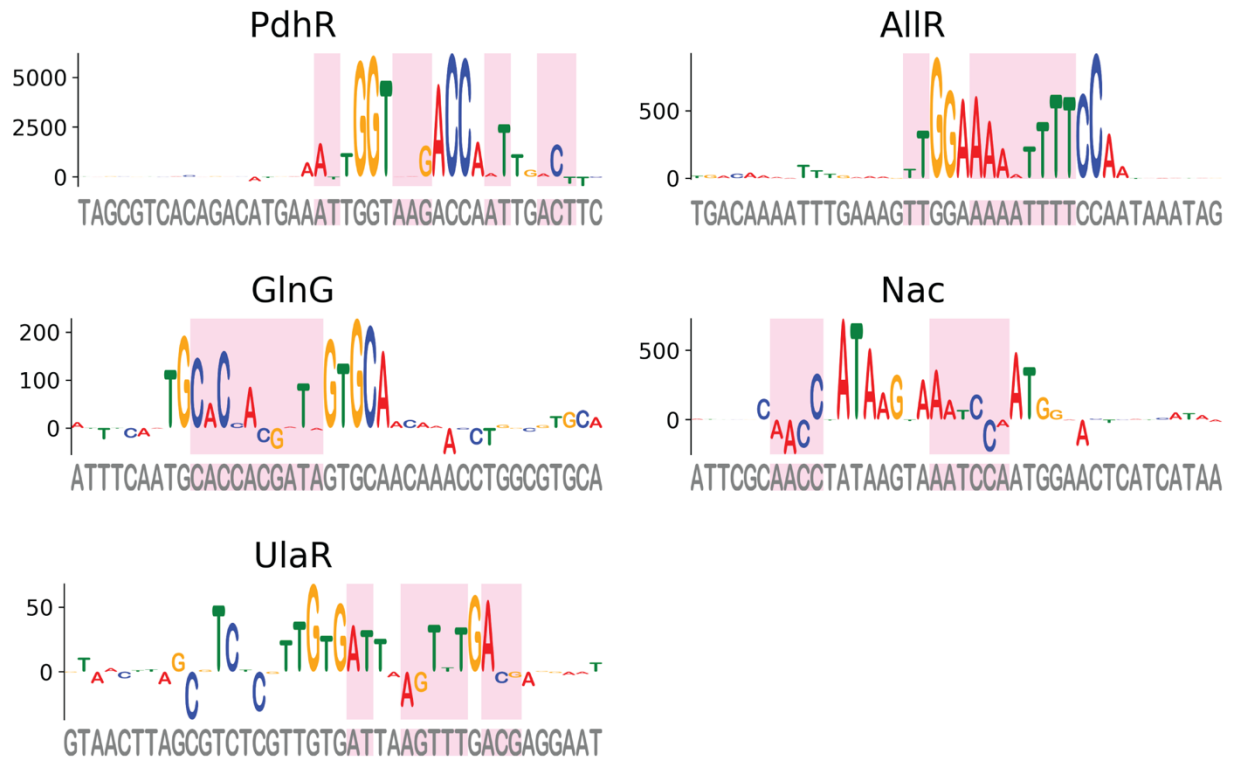


Figure S25: Novel Site Design. Novel binding sites were designed for 5 TFs. For each TF, the top singleton binding site was selected as a reference (sequence logo) and all combination of accessory bases (pink shading) we generated and used as input to the corresponding BoltzNet model.

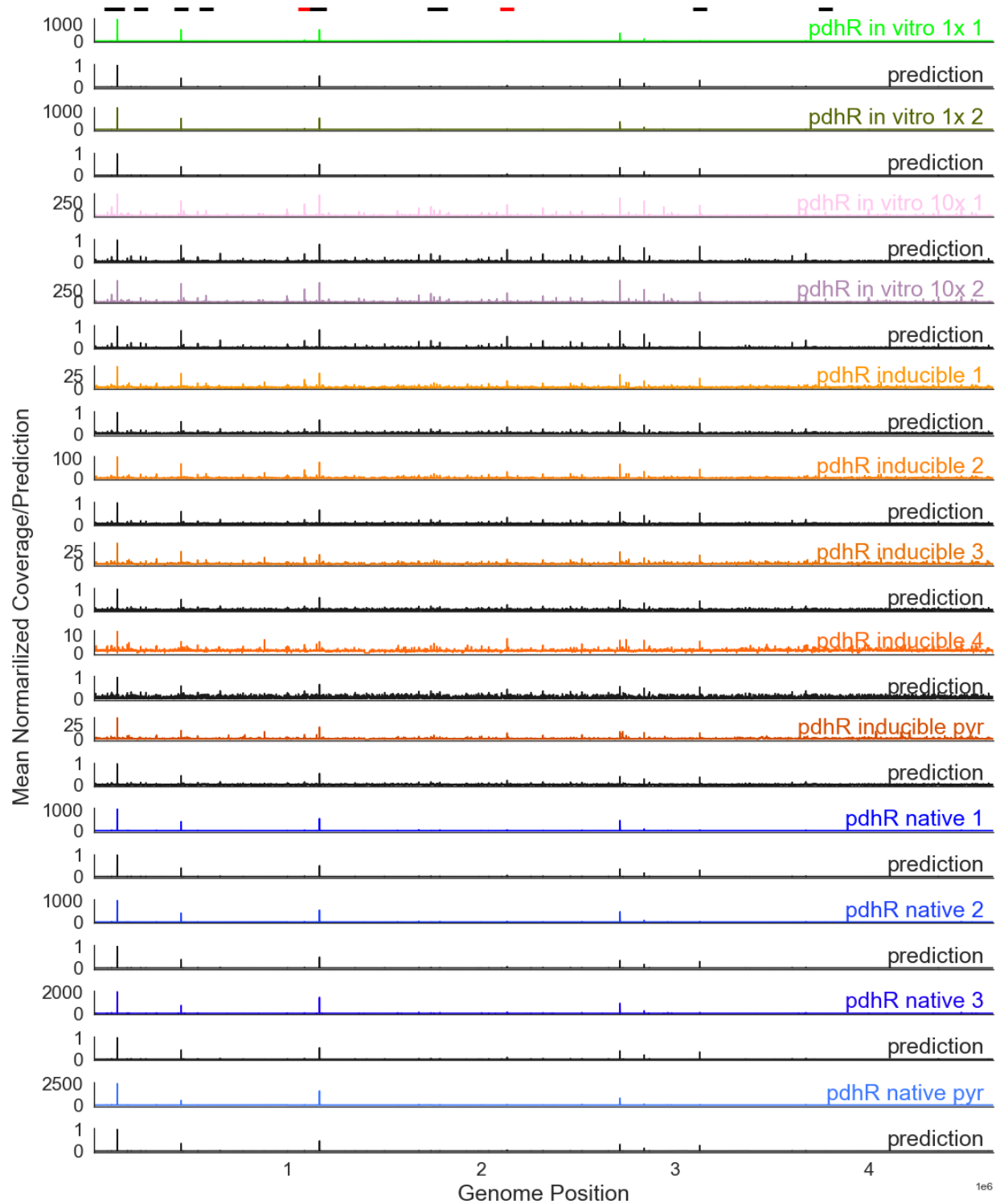
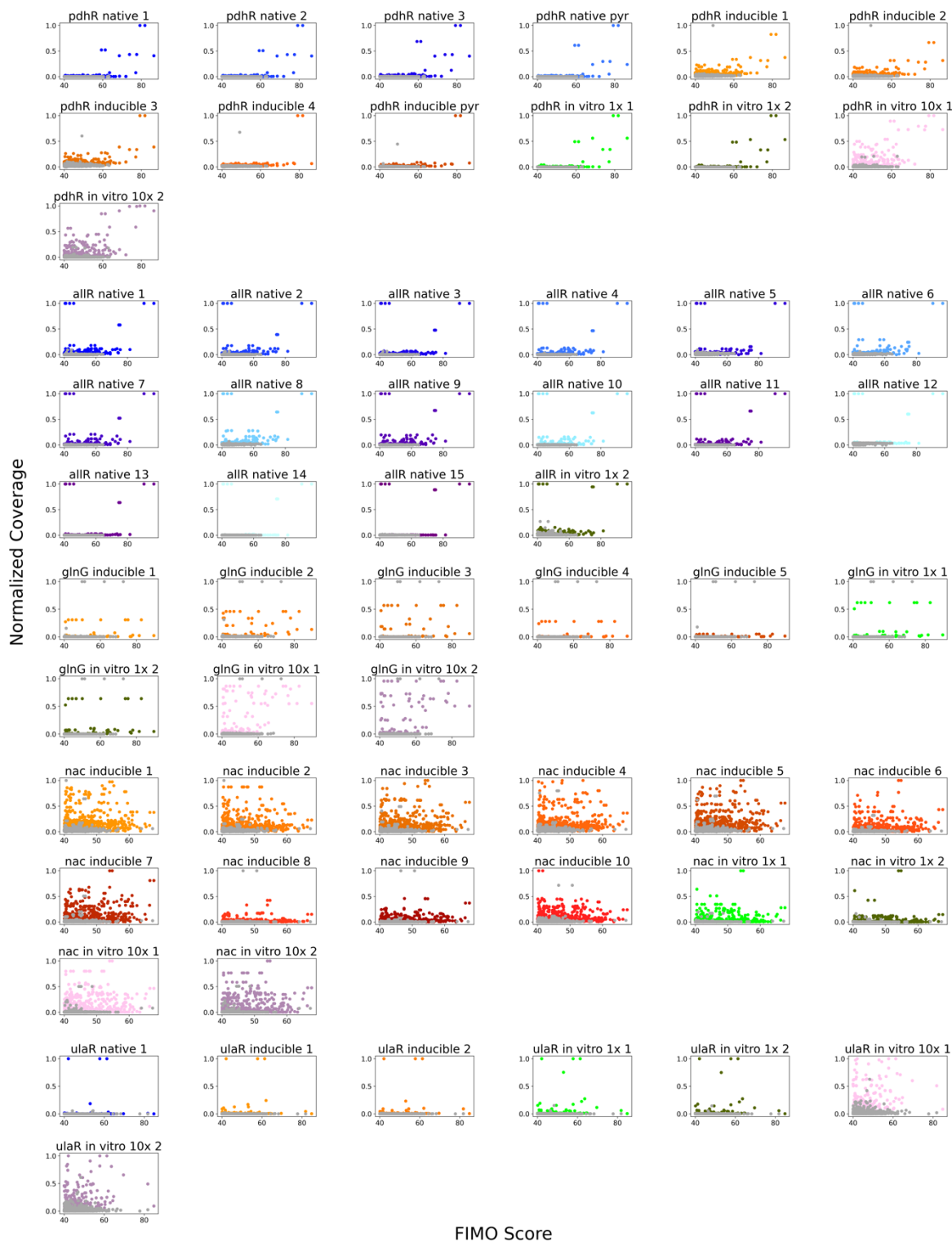


Figure S26: Genome Wide Plot of Coverage and Coverage Predictions for All PdhR Experiments. Each pair of tracks show actual coverage (blue) and model predicted coverage (black) for one experiment. Actual coverage in units of fold enrichment over mean coverage. Predicted coverage from zero to one. Positions of known sites (black rectangles) and Figure 2 details regions (red rectangles) shown above top track.



1
2 **Figure S27: MEME PFMs are not Predictive of ChIP-Seq Coverage on 5 Calibrated TFs.** We used MEME to generate PFMs directly
3 from ChIP-Seq data and scored all genomic sequences against these motifs with FIMO (where higher scores indicate a better match
4 to the PFM). Plots show the FIMO score of each motif match on the x-axis and the maximum coverage in a window of sequence
5 spanning 250bp on either side of the FIMO hit normalized to the strongest binding site on the y-axis. If the hit overlaps a called
6 peak the point is colored according to experiment. Hits that do not overlap a called peak are colored gray.

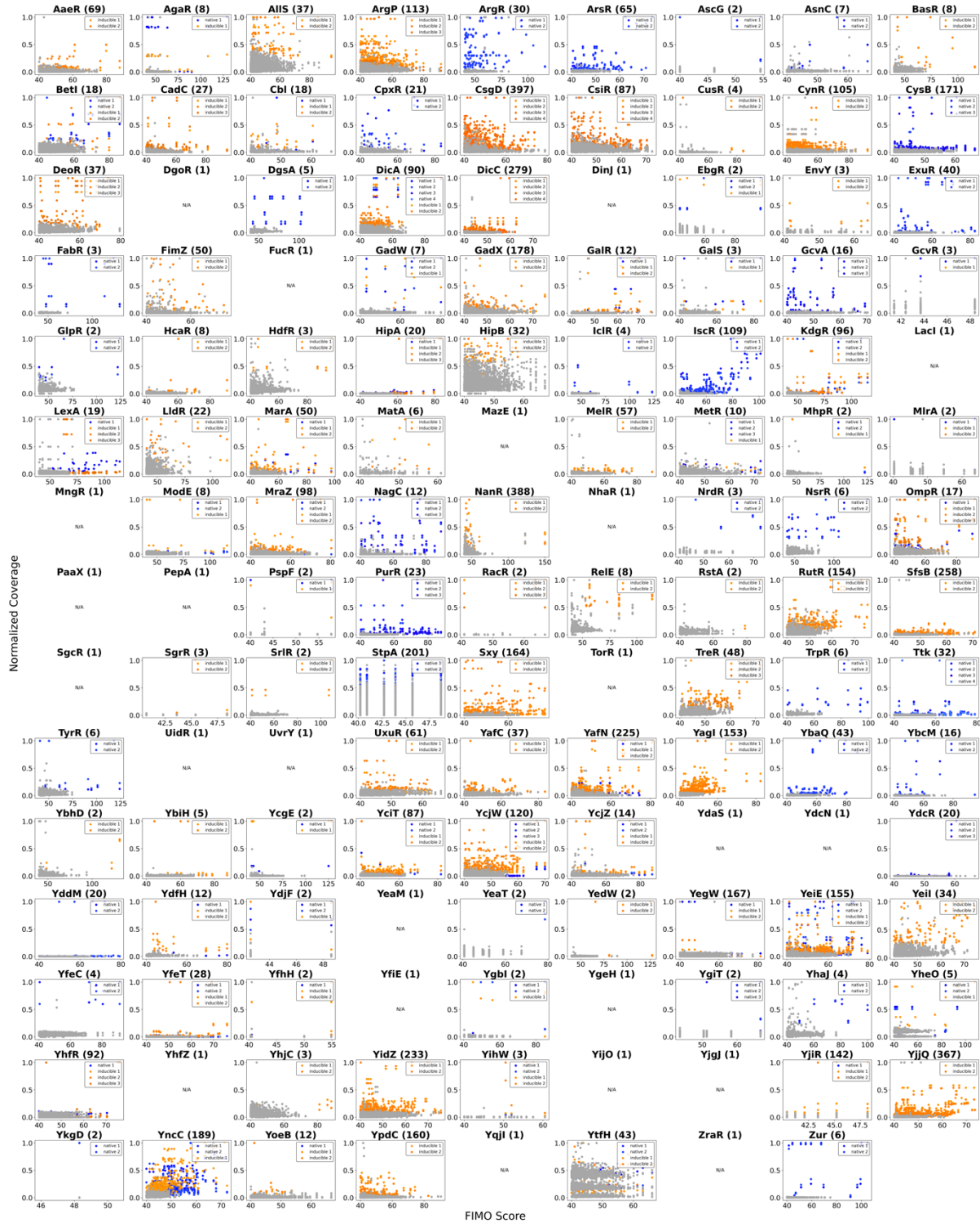
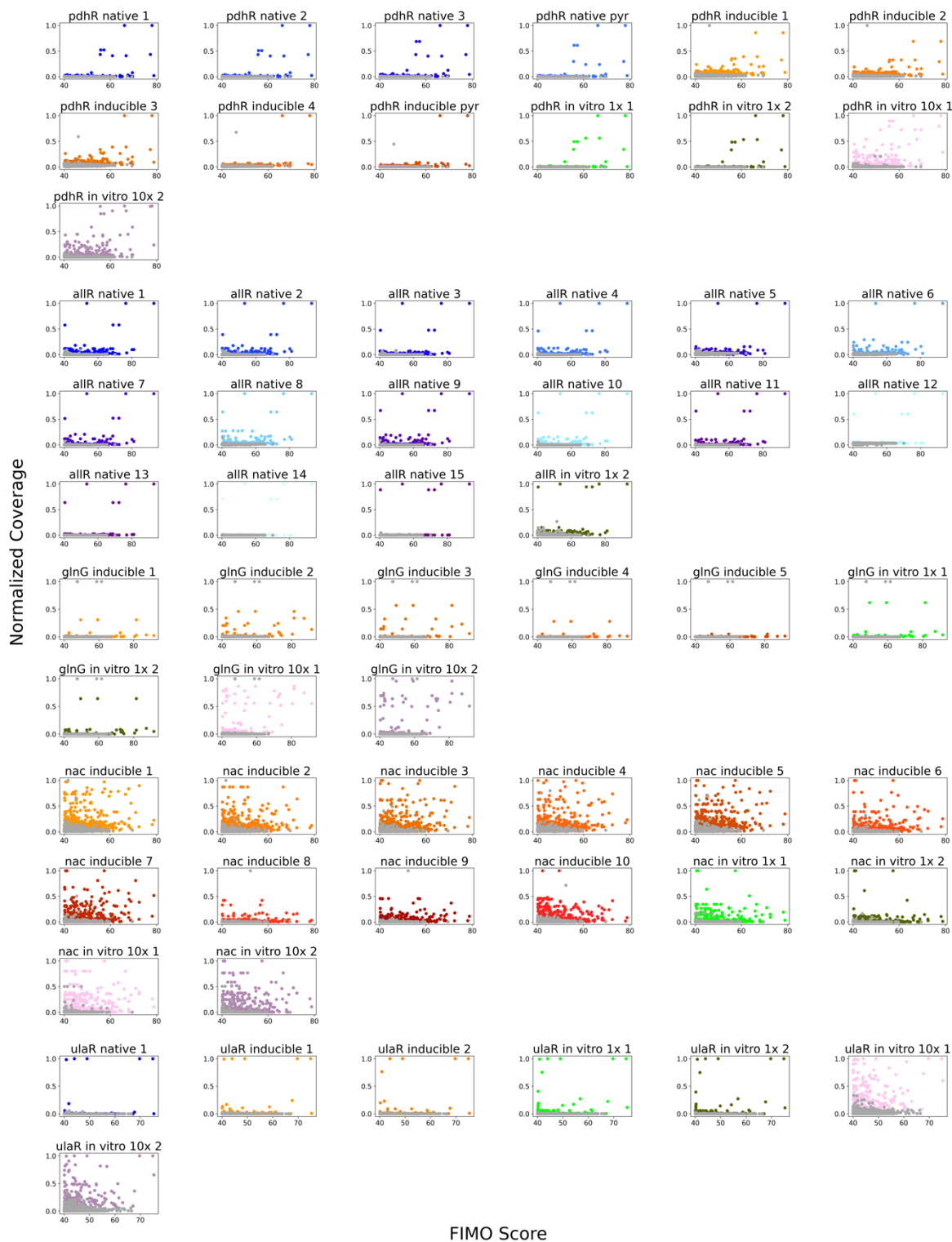
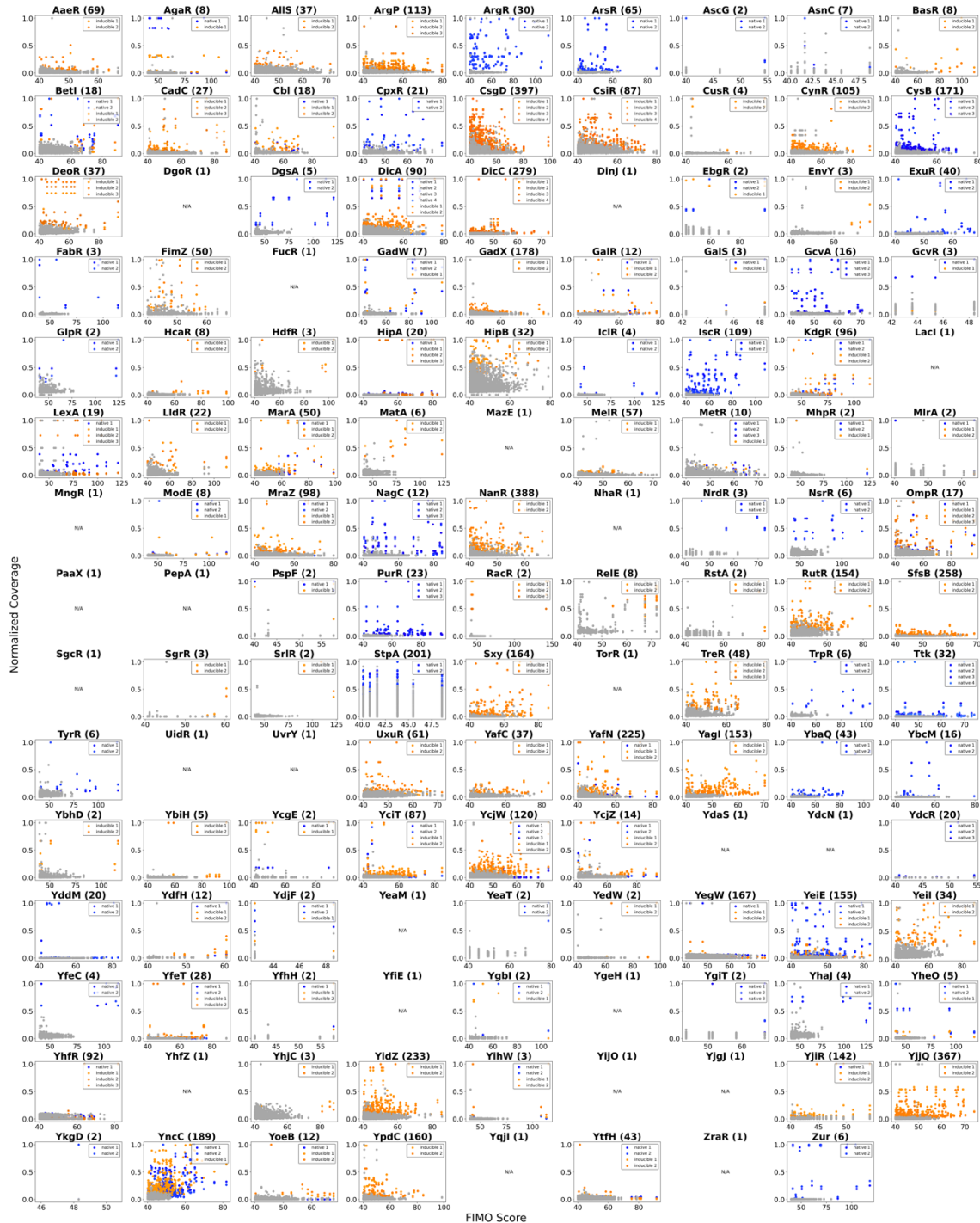


Figure S28: MEME PFMs are not Predictive of ChIP-Seq Coverage on all TFs with ChIP-Seq. We used MEME to generate PFMs for all TFs with replicated ChIP-Seq data and scored all genomic sequences against these motifs with FIMO (where higher scores indicate a better match to the PFM). Plots show the FIMO score of each motif match on the x-axis and the maximum coverage in a window of sequence spanning 250bp on either side of the FIMO hit normalized to the strongest binding site on the y-axis. Each point represents a FIMO hit. If the hit overlaps a called peak the point is colored according to experiment. Hits that do not overlap a called peak are colored gray.



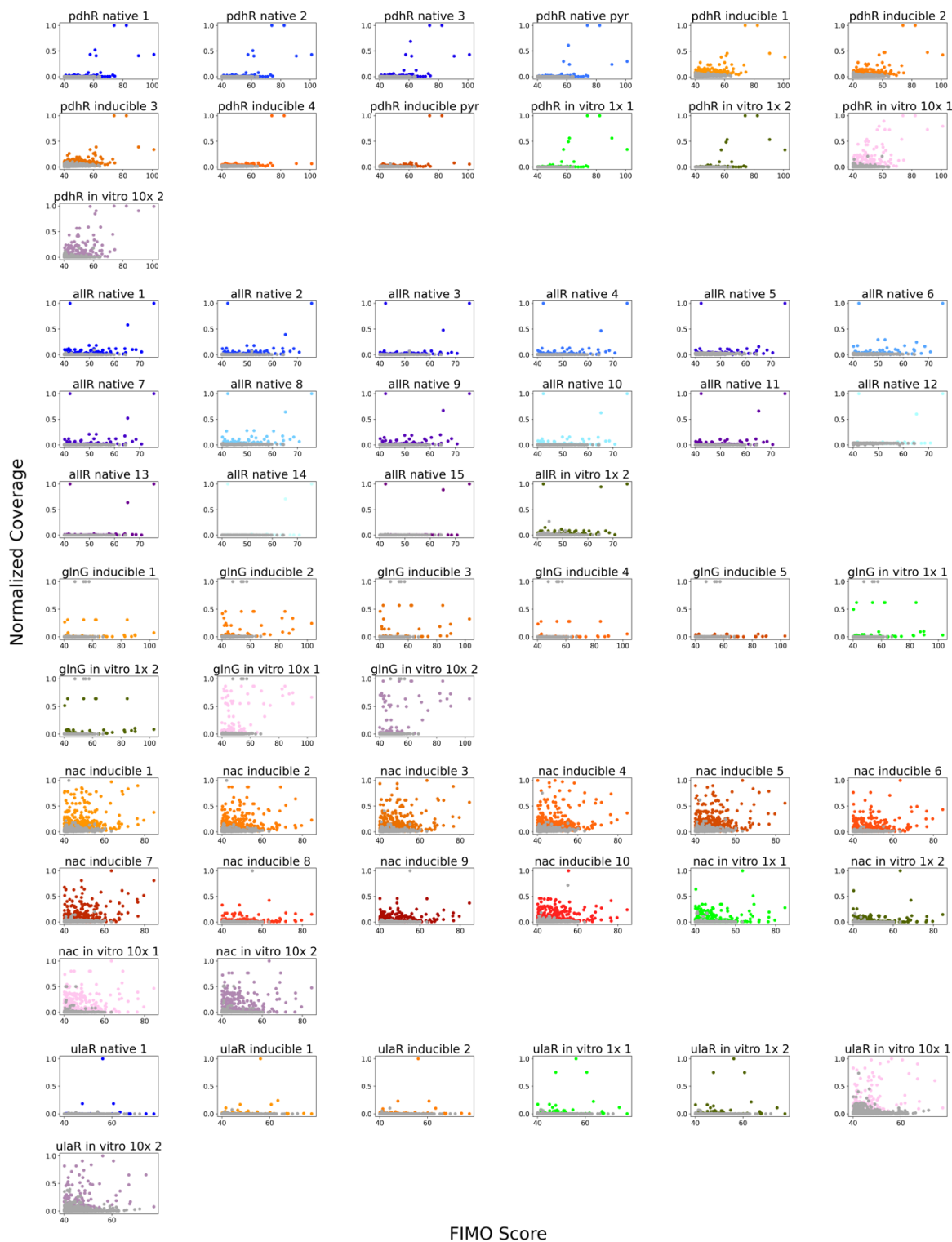
1
2 **Figure S29: MEME-ChIP PFMs are not Predictive of ChIP-Seq Coverage on 5 Calibrated TFs.** We used MEME-ChIP to generate
3 PFMs directly from ChIP-Seq data and scored all genomic sequences against these motifs with FIMO (where higher scores indicate
4 a better match to the PFM). Plots show the FIMO score of each motif match on the x-axis and the maximum coverage in a window
5 of sequence spanning 250bp on either side of the FIMO hit normalized to the strongest binding site on the y-axis. If the hit overlaps
6 a called peak the point is colored according to experiment. Hits that do not overlap a called peak are colored gray.

1



2

3 **Figure S30: MEME-ChIP PFMs are not Predictive of ChIP-Seq Coverage on all TFs with ChIP-Seq.** We used MEME-ChIP to generate
 4 PFMs for all TFs with replicated ChIP-Seq data and scored all genomic sequences against these motifs with FIMO (where higher
 5 scores indicate a better match to the PFM). Plots show the FIMO score of each motif match on the x-axis and the maximum coverage
 6 in a window of sequence spanning 250bp on either side of the FIMO hit normalized to the strongest binding site on the y-axis. Each
 7 point represents a FIMO hit. If the hit overlaps a called peak the point is colored according to experiment. Hits that do not overlap
 8 a called peak are colored gray.



1
2 **Figure S31: XSTREME PFMs are not Predictive of ChIP-Seq Coverage on 5 Calibrated TFs.** We used XSTREME to generate PFMs
3 directly from ChIP-Seq data and scored all genomic sequences against these motifs with FIMO (where higher scores indicate a better
4 match to the PFM). Plots show the FIMO score of each motif match on the x-axis and the maximum coverage in a window of
5 sequence spanning 250bp on either side of the FIMO hit normalized to the strongest binding site on the y-axis. If the hit overlaps
6 a called peak the point is colored according to experiment. Hits that do not overlap a called peak are colored gray.

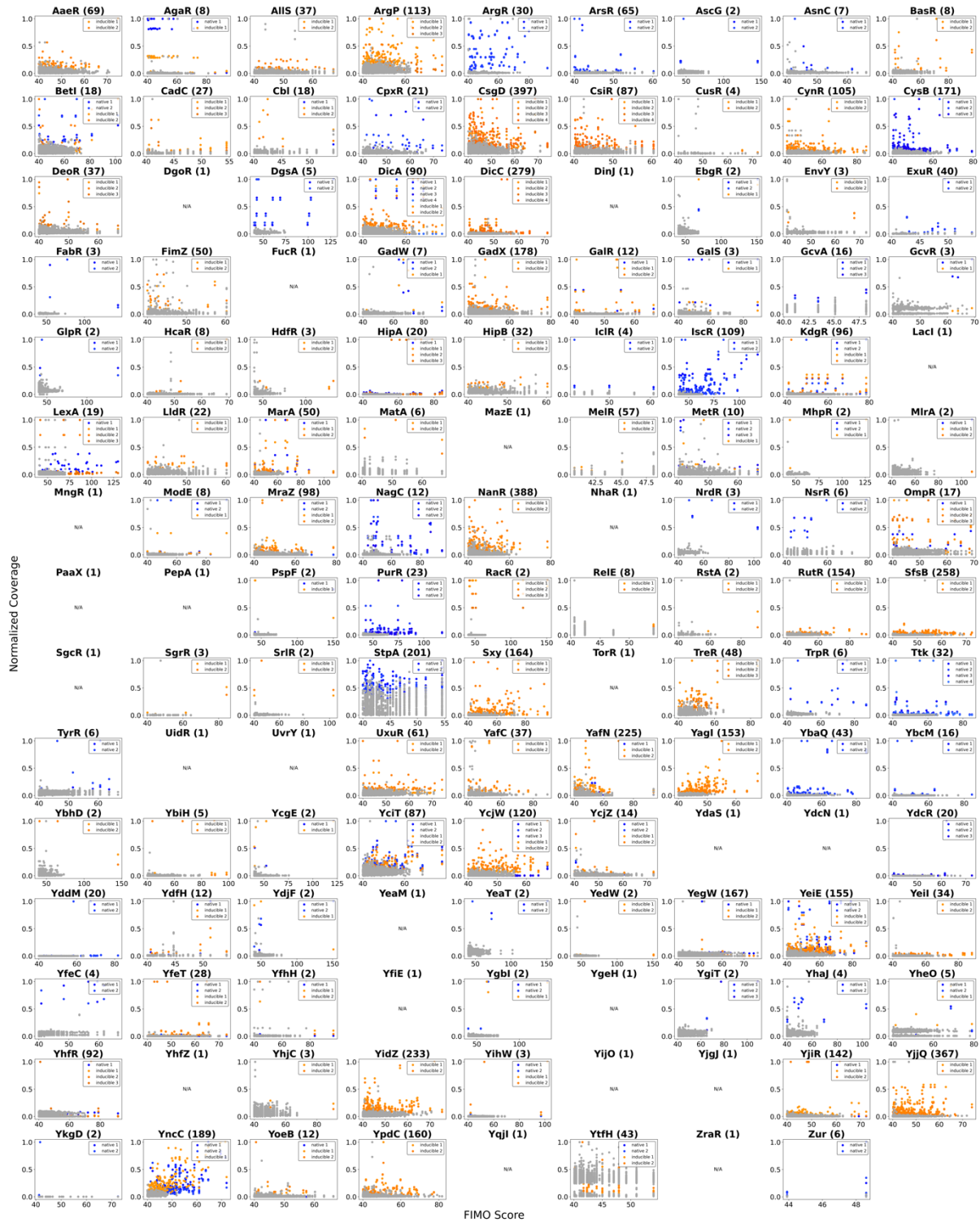


Figure S32: XSTREME PFMs are not Predictive of ChIP-Seq Coverage on all TFs with ChIP-Seq. We used XSTREME to generate PFMs for all TFs with replicated ChIP-Seq data and scored all genomic sequences against these motifs with FIMO (where higher scores indicate a better match to the PFM). Plots show the FIMO score of each motif match on the x-axis and the maximum coverage in a window of sequence spanning 250bp on either side of the FIMO hit normalized to the strongest binding site on the y-axis. Each point represents a FIMO hit. If the hit overlaps a called peak the point is colored according to experiment. Hits that do not overlap a called peak are colored gray.

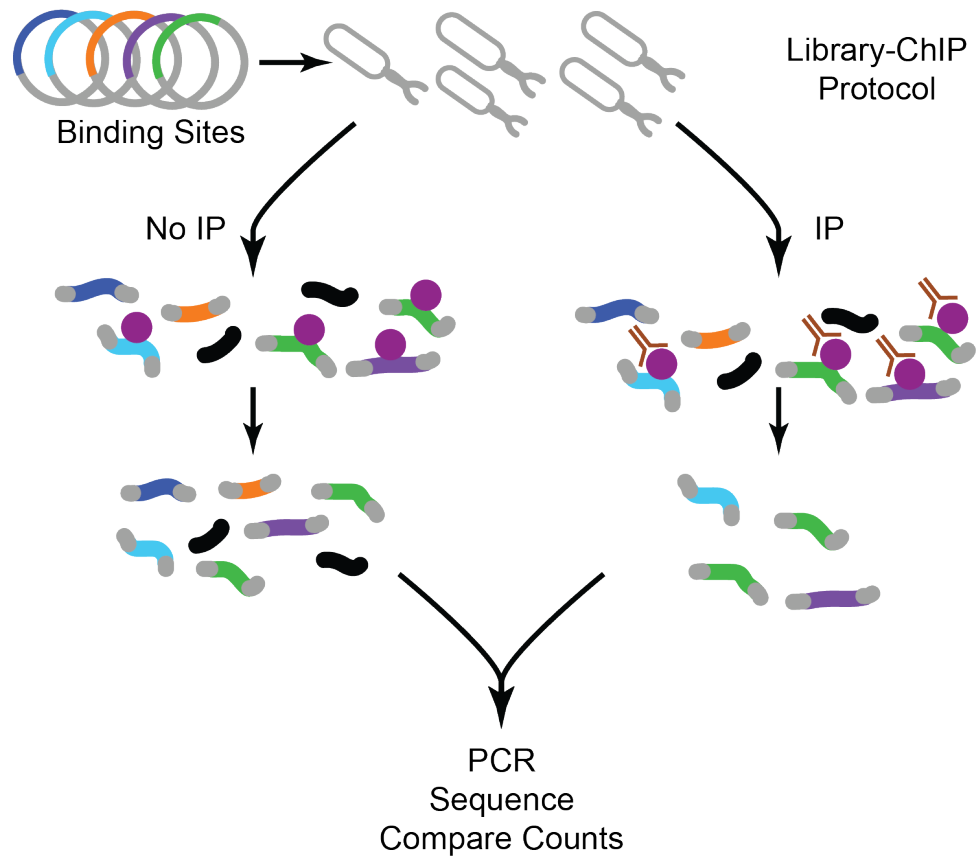


Figure S33: Library-ChIP schematic. A library of plasmids each containing 1 designed (or genomic reference) binding site is transformed into *E. coli* (top). After the crosslinking and DNA-shearing steps of ChIP, a small amount of DNA-containing lysate is reserved as a no-IP control (left). The remaining fraction of lysate undergoes IP as in a normal ChIP-Seq experiment (right). Control and IP samples then undergo two rounds of PCR to add Illumina adapters and barcodes, and NGS is performed. We then compare the abundance of each sequence in IP and control samples to determine the Library-ChIP enrichment.

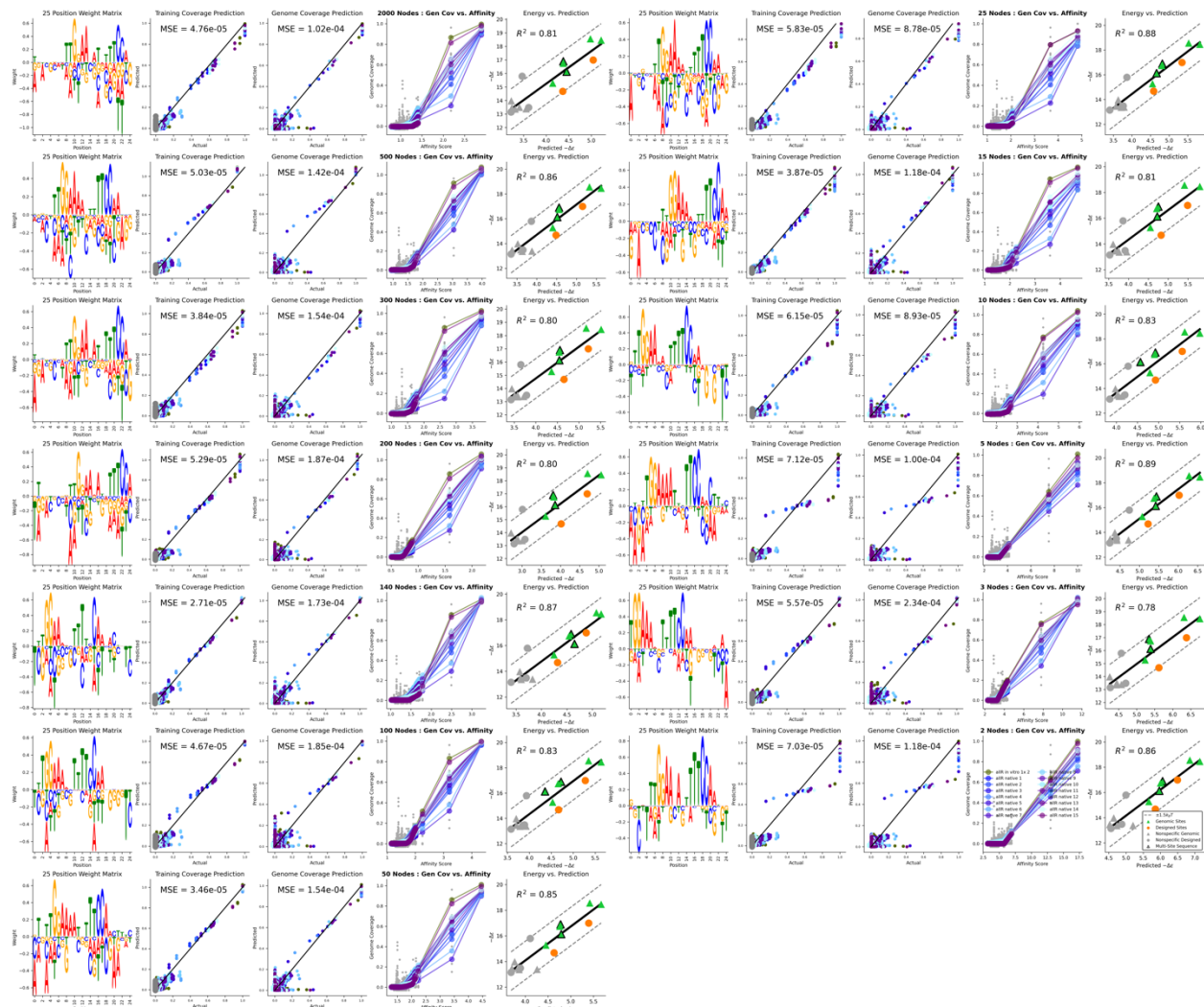


Figure S34: Analysis of impact of neural network width for AIIR. Each repeated set of 5 panels corresponds to a model built with a different number of nodes in the middle layer of the BoltzNet fully connected neural network component as indicated in the title of the figures in the last column. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

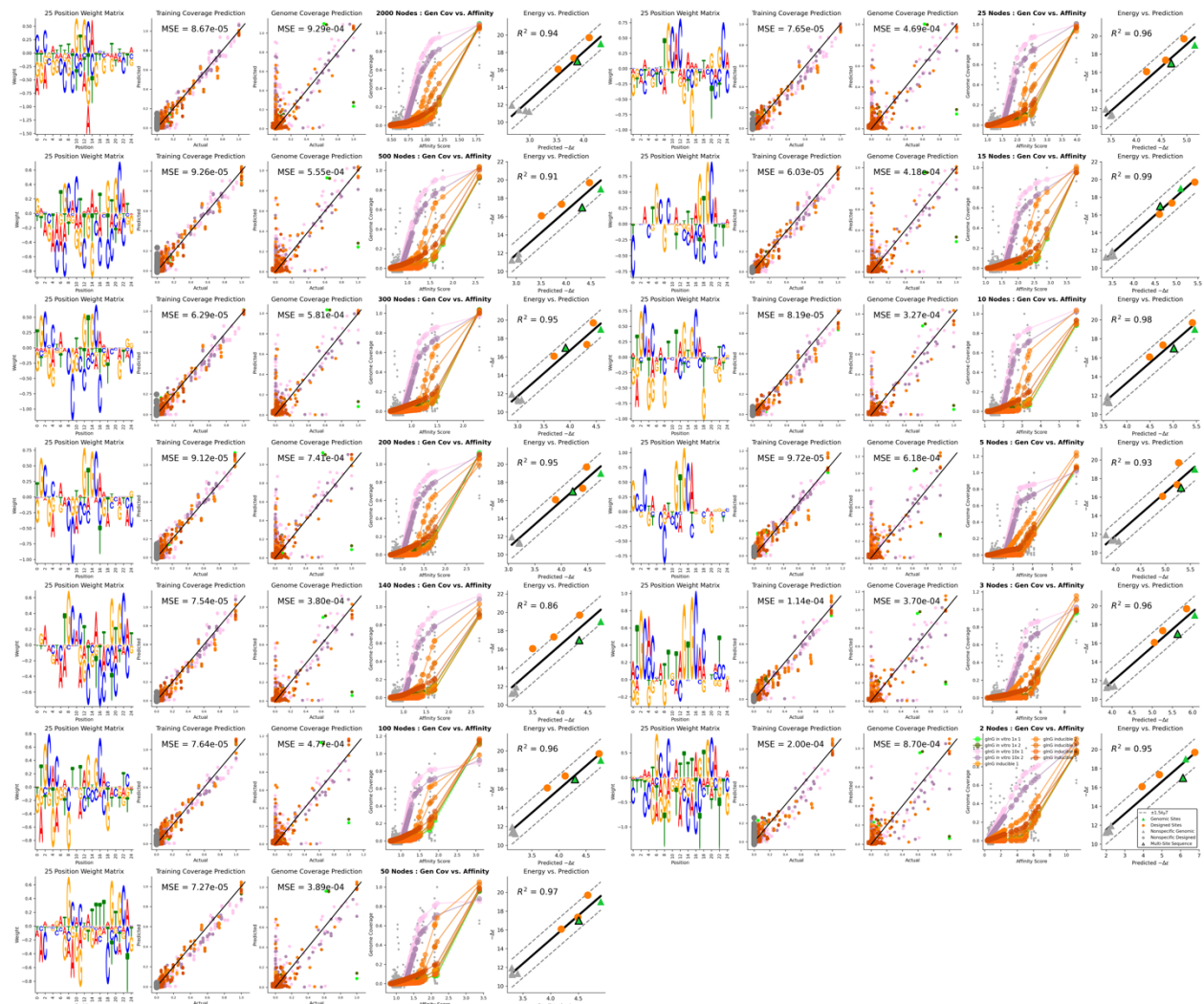


Figure S35: Analysis of impact of neural network width for GlnG. Each repeated set of 5 panels corresponds to a model built with a different number of nodes in the middle layer of the BoltzNet fully connected neural network component as indicated in the title of the figures in the last column. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

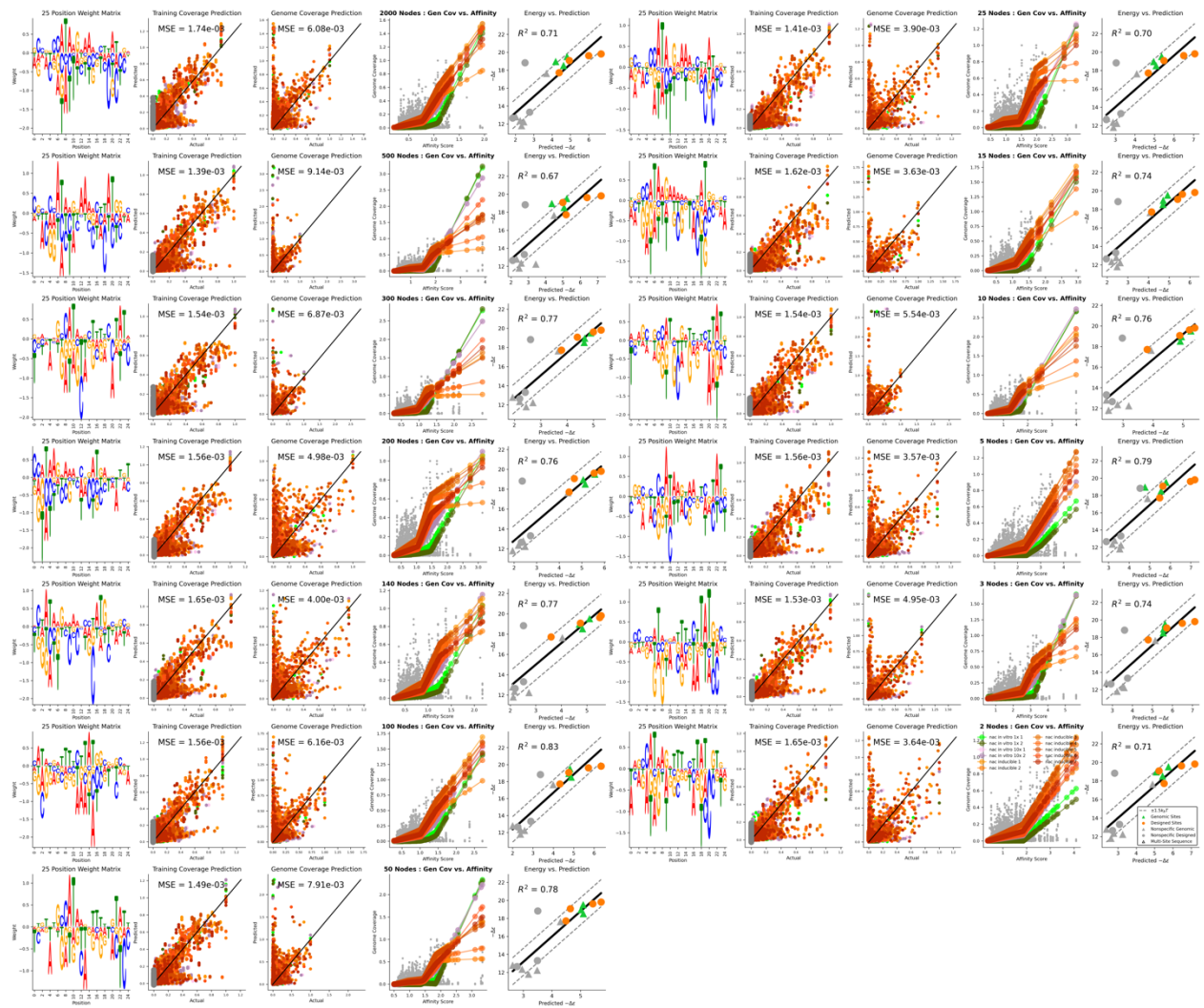


Figure S36: Analysis of impact of neural network width for Nac. Each repeated set of 5 panels corresponds to a model built with a different number of nodes in the middle layer of the BoltzNet fully connected neural network component as indicated in the title of the figures in the last column. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

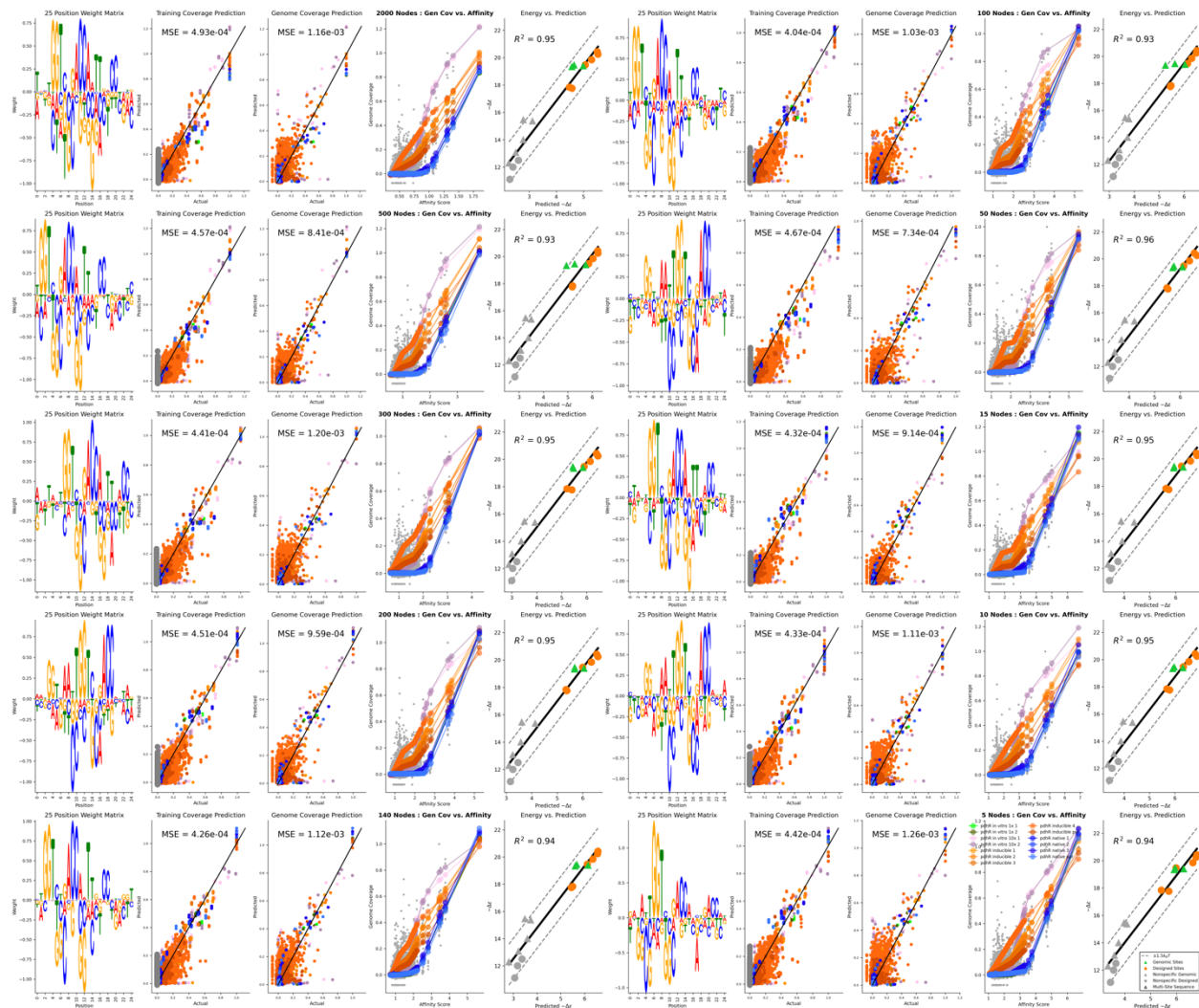


Figure S37: Analysis of impact of neural network width for PdhR. Each repeated set of 5 panels corresponds to a model built with a different number of nodes in the middle layer of the BoltzNet fully connected neural network component as indicated in the title of the figures in the last column. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

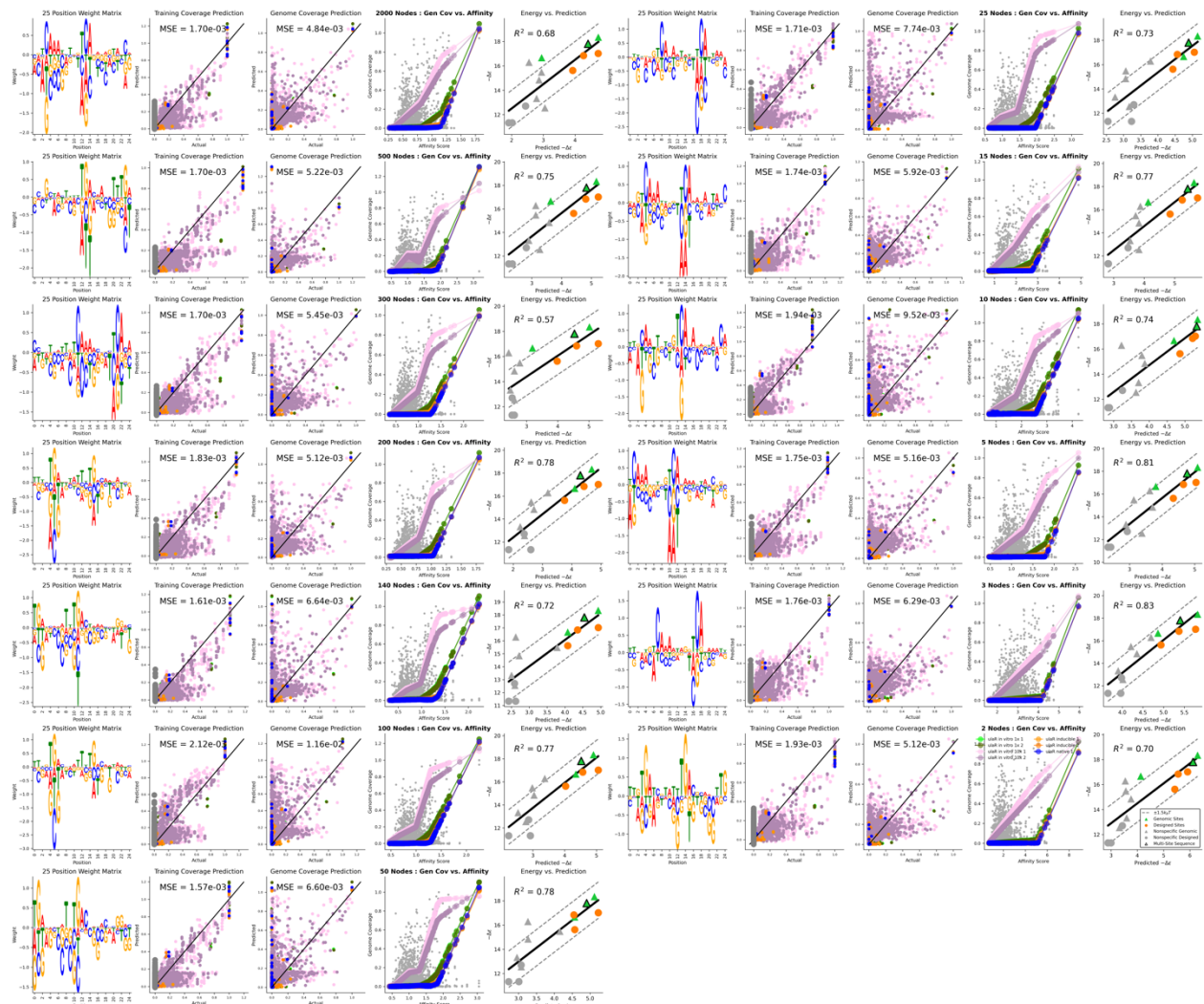


Figure S38: Analysis of impact of neural network width for UlaR. Each repeated set of 5 panels corresponds to a model built with a different number of nodes in the middle layer of the BoltzNet fully connected neural network component as indicated in the bolded title. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

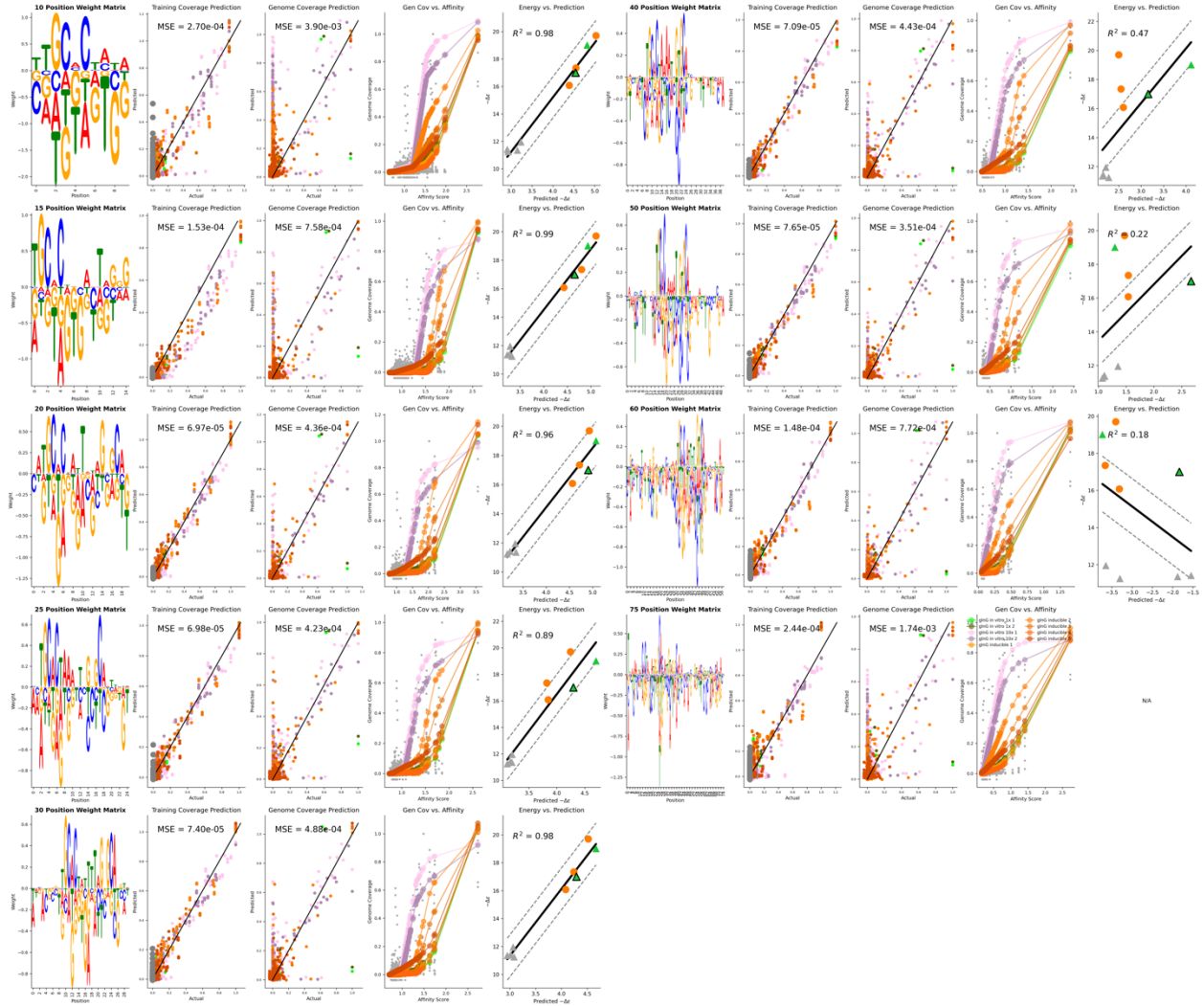


Figure S40: Analysis of impact of matrix width for GlnG. Each repeated set of 5 panels corresponds to a model built with a different weight matrix width as indicated in the bolded title. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

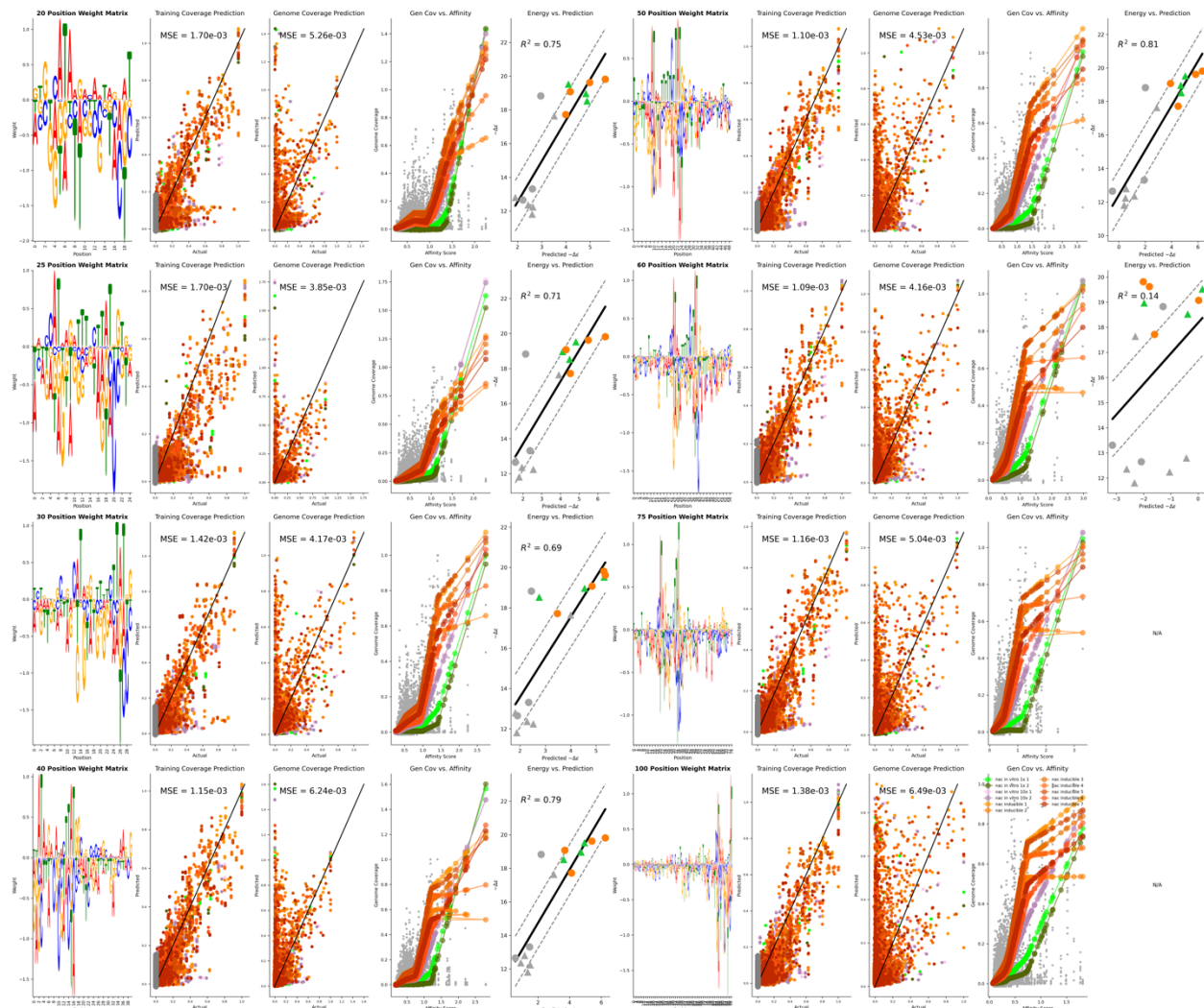


Figure S41: Analysis of impact of matrix width for Nac. Each repeated set of 5 panels corresponds to a model built with a different weight matrix width as indicated in the bolded title. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

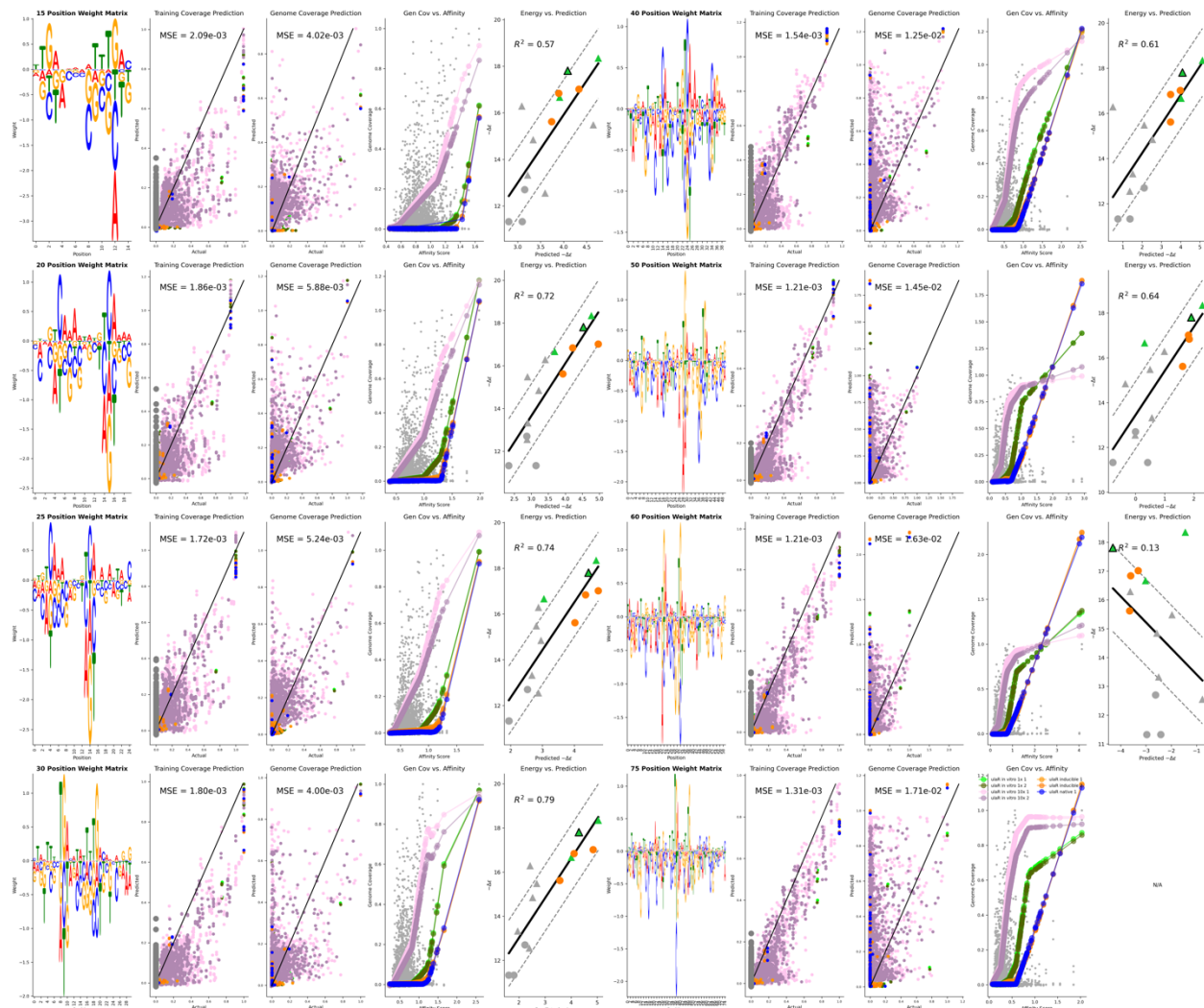


Figure S43: Analysis of impact of matrix width for UlaR. Each repeated set of 5 panels corresponds to a model built with a different weight matrix width as indicated in the bolded title. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

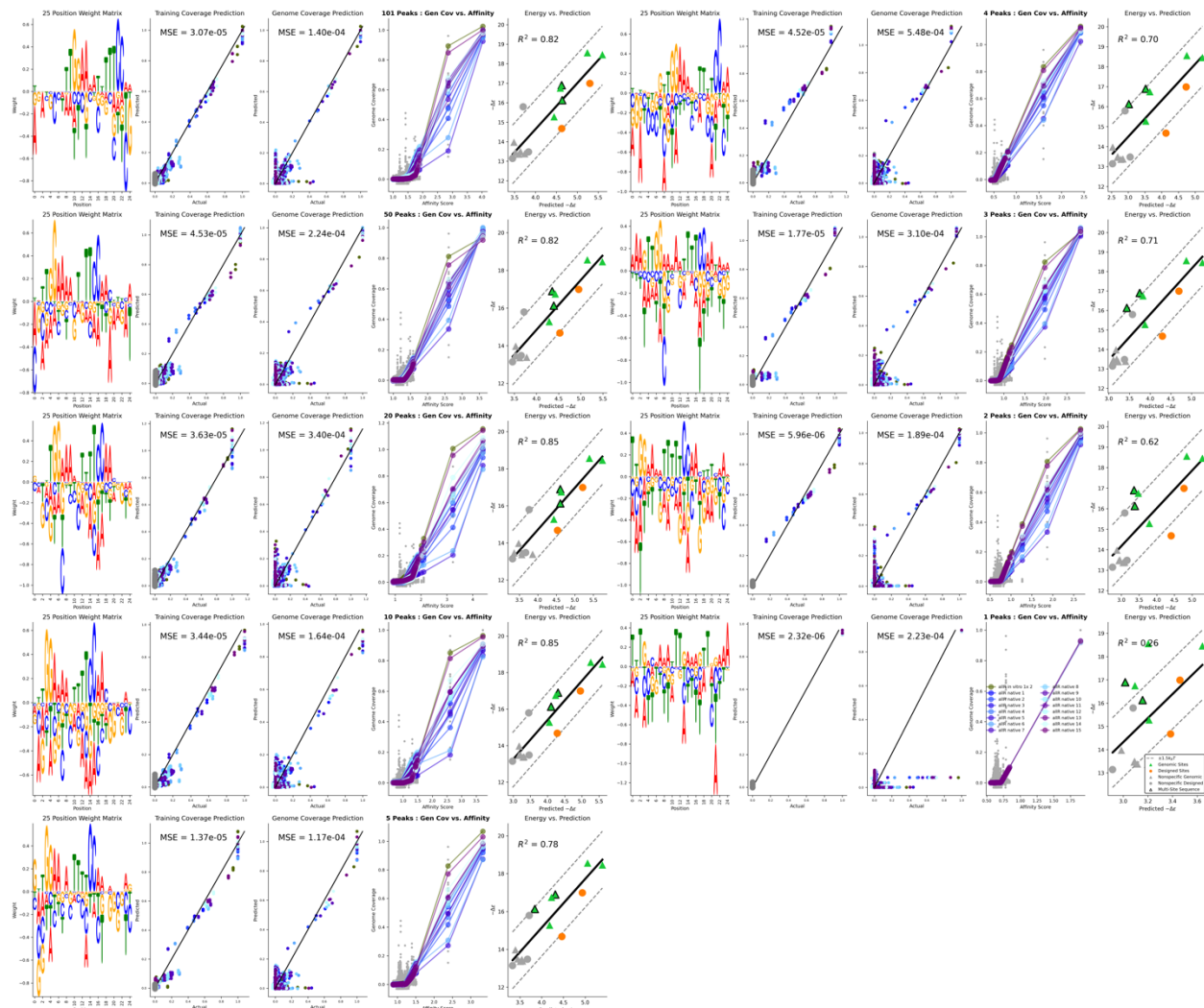


Figure S44: Analysis of impact of sub-selection called peaks for AIRR. Each repeated set of 5 panels corresponds to a model built with a different number of the strongest called peaks as indicated in the bolded title. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

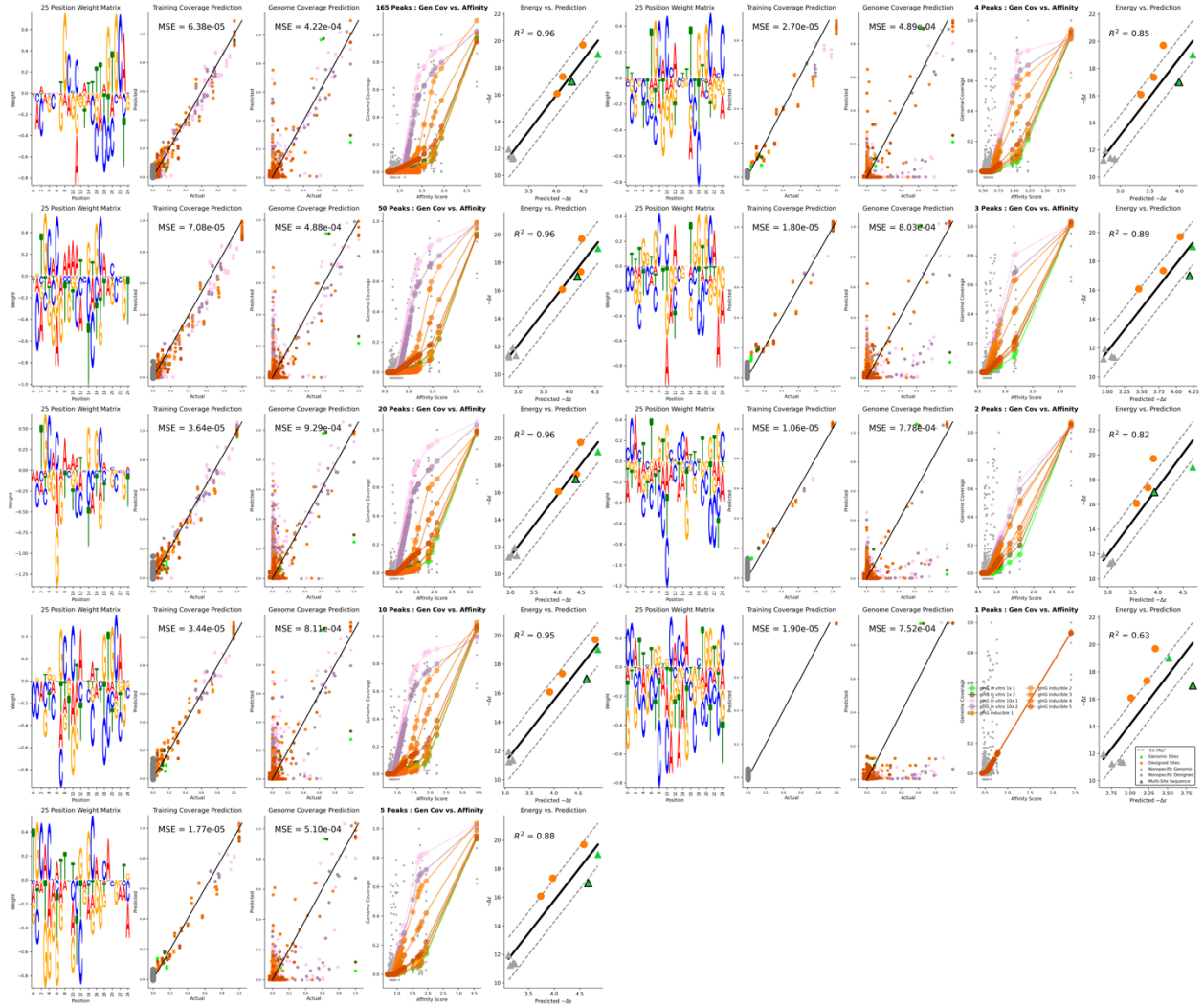


Figure S45: Analysis of impact of sub-selection called peaks for GlnG. Each repeated set of 5 panels corresponds to a model built with a different number of the strongest called peaks as indicated in the bolded title. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

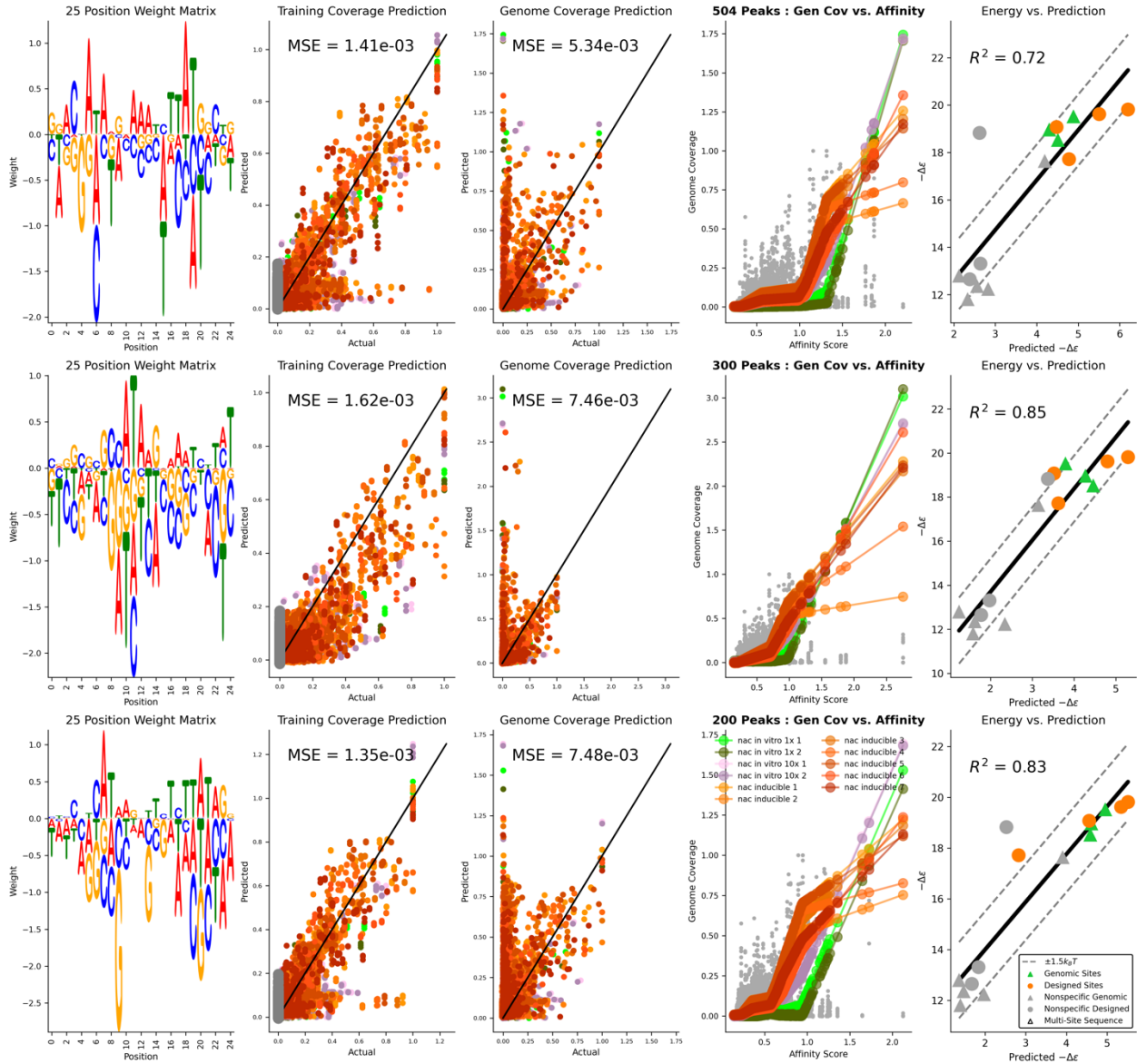


Figure S46: Analysis of impact of sub-selection called peaks for Nac. Each repeated set of 5 panels corresponds to a model built with a different number of the strongest called peaks as indicated in the bolded title. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

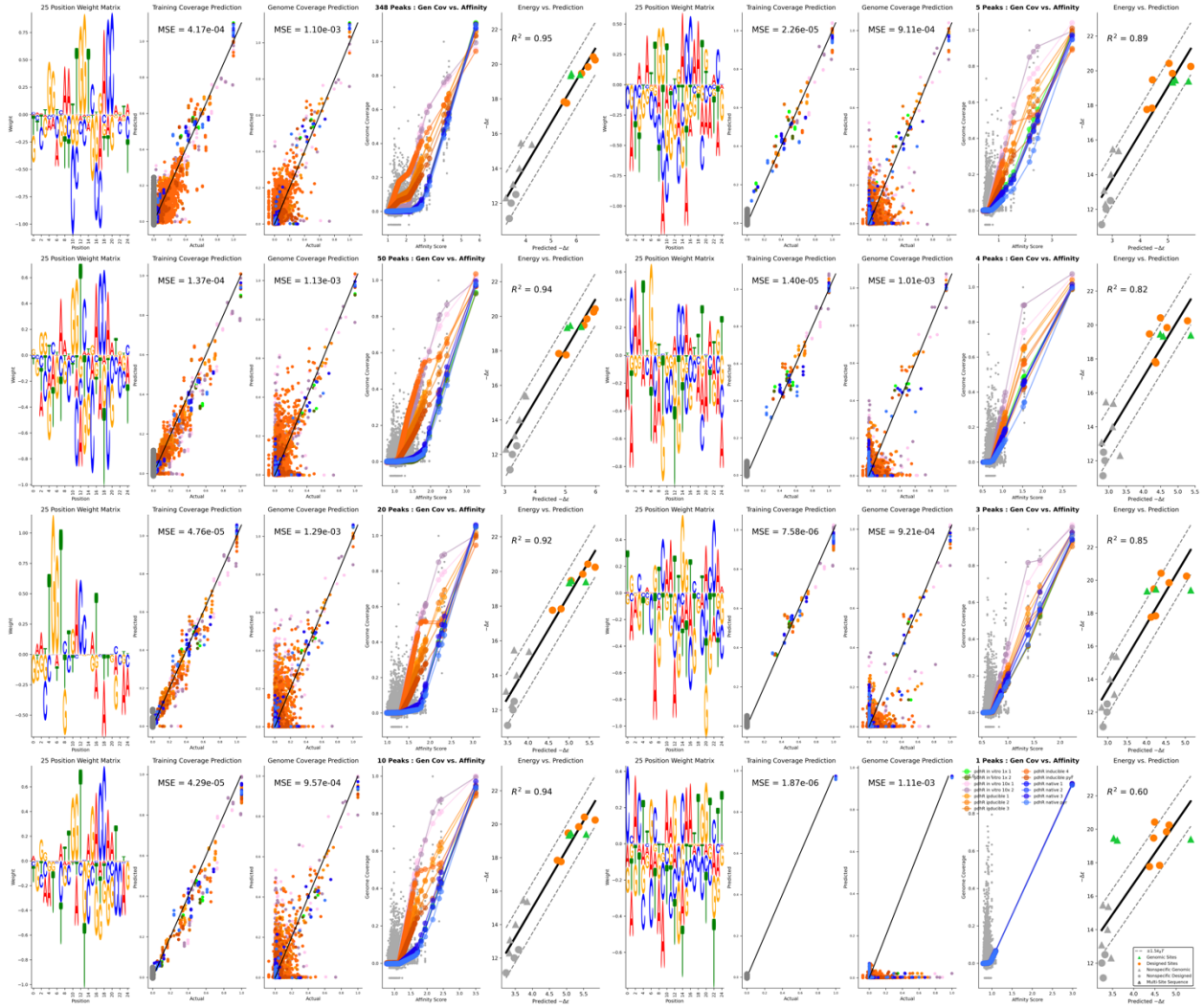


Figure S47: Analysis of impact of sub-selection called peaks for PdhR. Each repeated set of 5 panels corresponds to a model built with a different number of the strongest called peaks as indicated in the bolded title. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

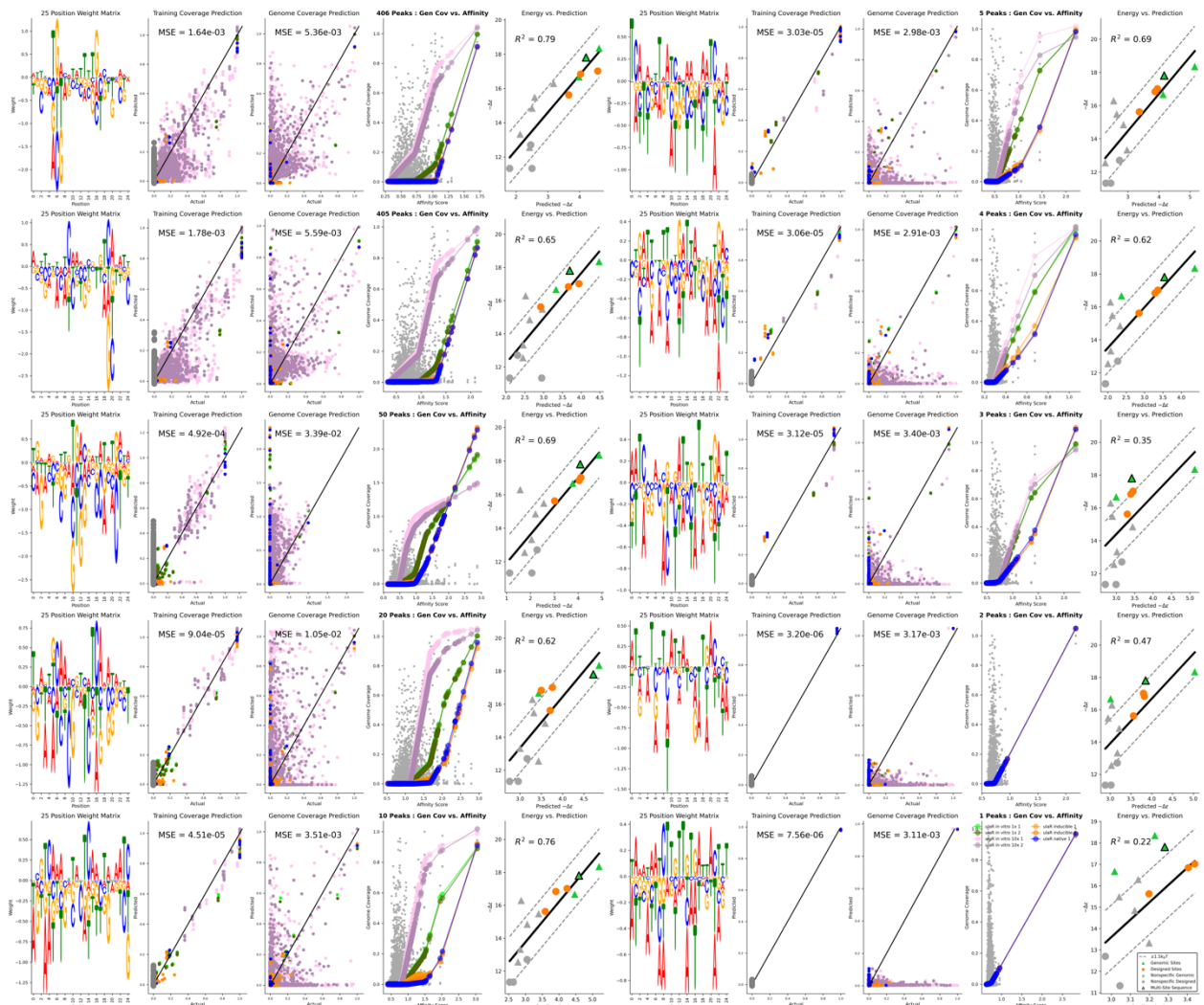


Figure S48: Analysis of impact of sub-selection called peaks for UlaR. Each repeated set of 5 panels corresponds to a model built with a different number of the strongest called peaks as indicated in the bolded title. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in the last plot of coverage vs. affinity.

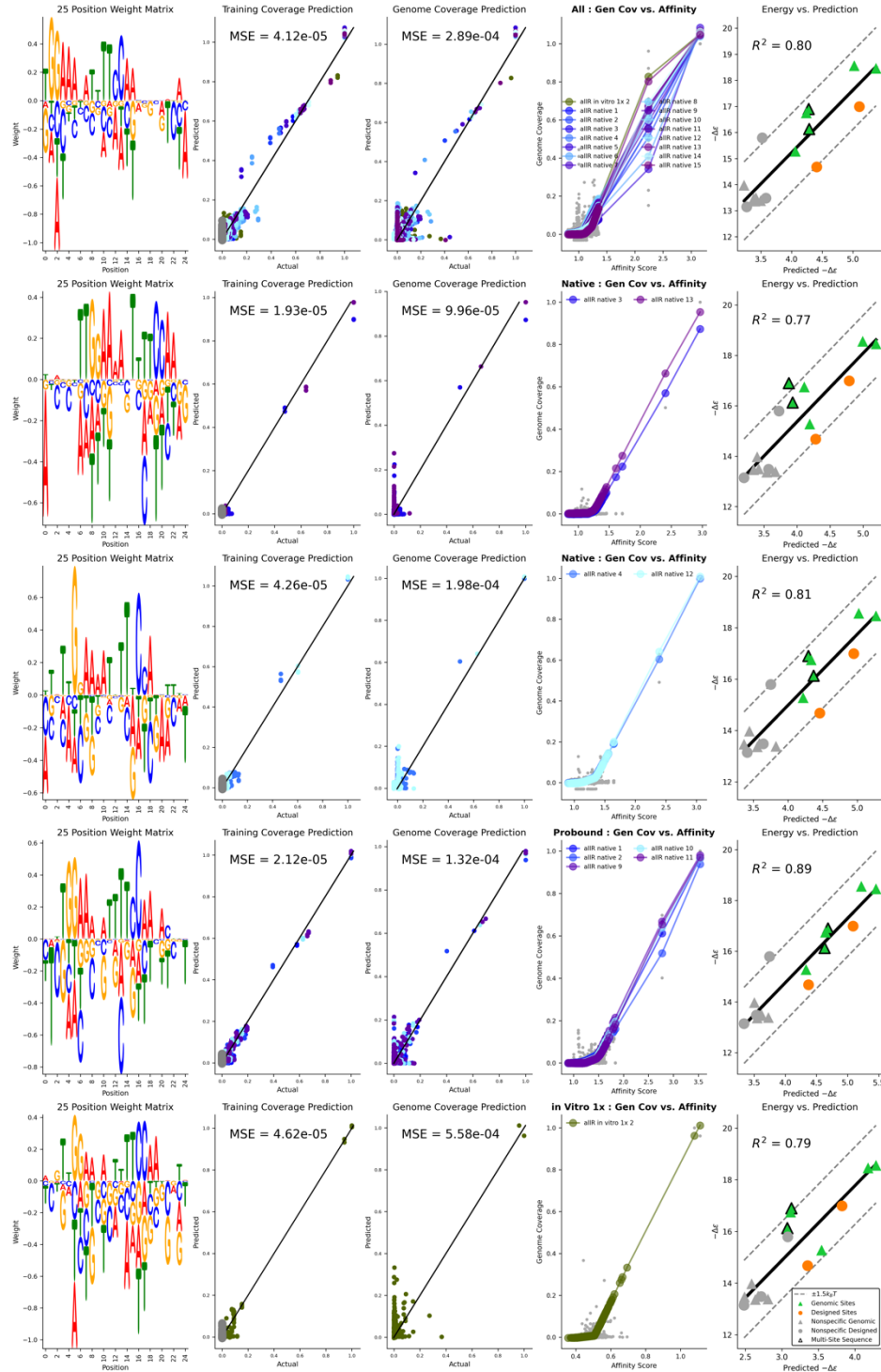


Figure S49: Analysis of impact of sub-selection called experiments for AIIR. Each repeated set of 5 panels corresponds to a model built with a different sub-selection of experiments as in bolded title and experiment legends. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in of coverage vs. affinity plots.

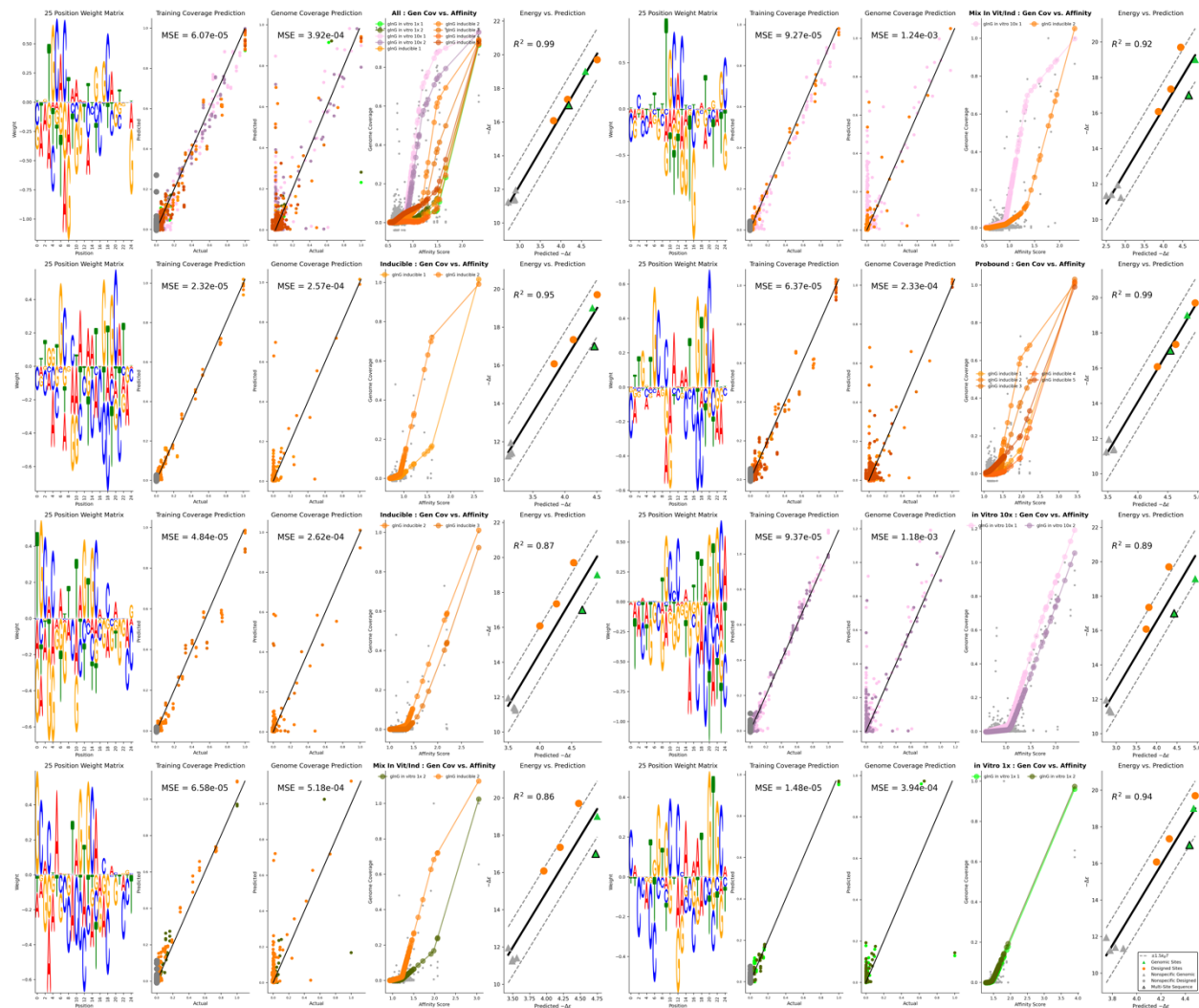


Figure S50: Analysis of impact of sub-selection called experiments for GlnG. Each repeated set of 5 panels corresponds to a model built with a different sub-selection of experiments as in bolded title and experiment legends. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in of coverage vs. affinity plots.

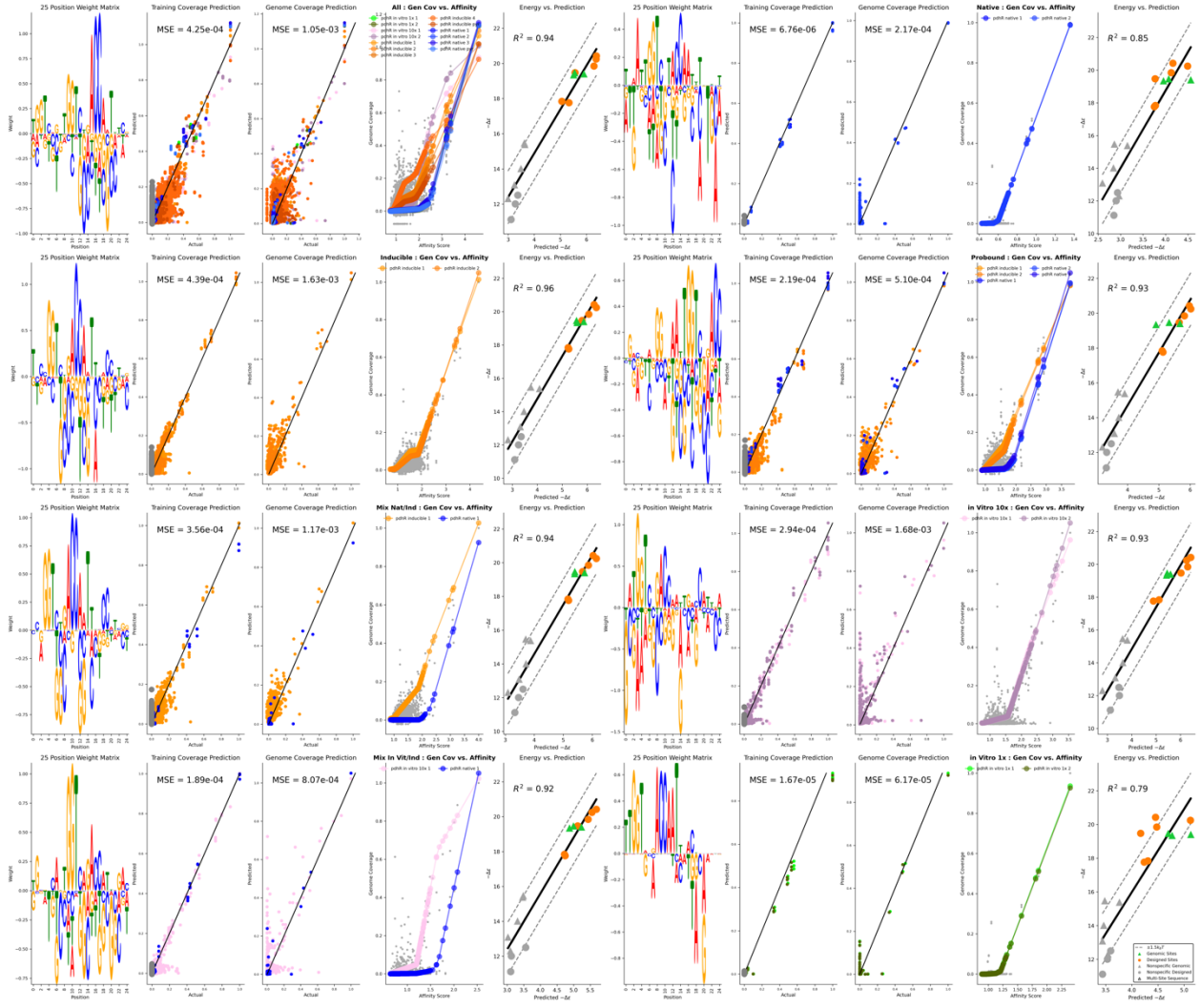


Figure S52: Analysis of impact of sub-selection called experiments for PdhR. Each repeated set of 5 panels corresponds to a model built with a different sub-selection of experiments as in bolded title and experiment legends. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in of coverage vs. affinity plots.

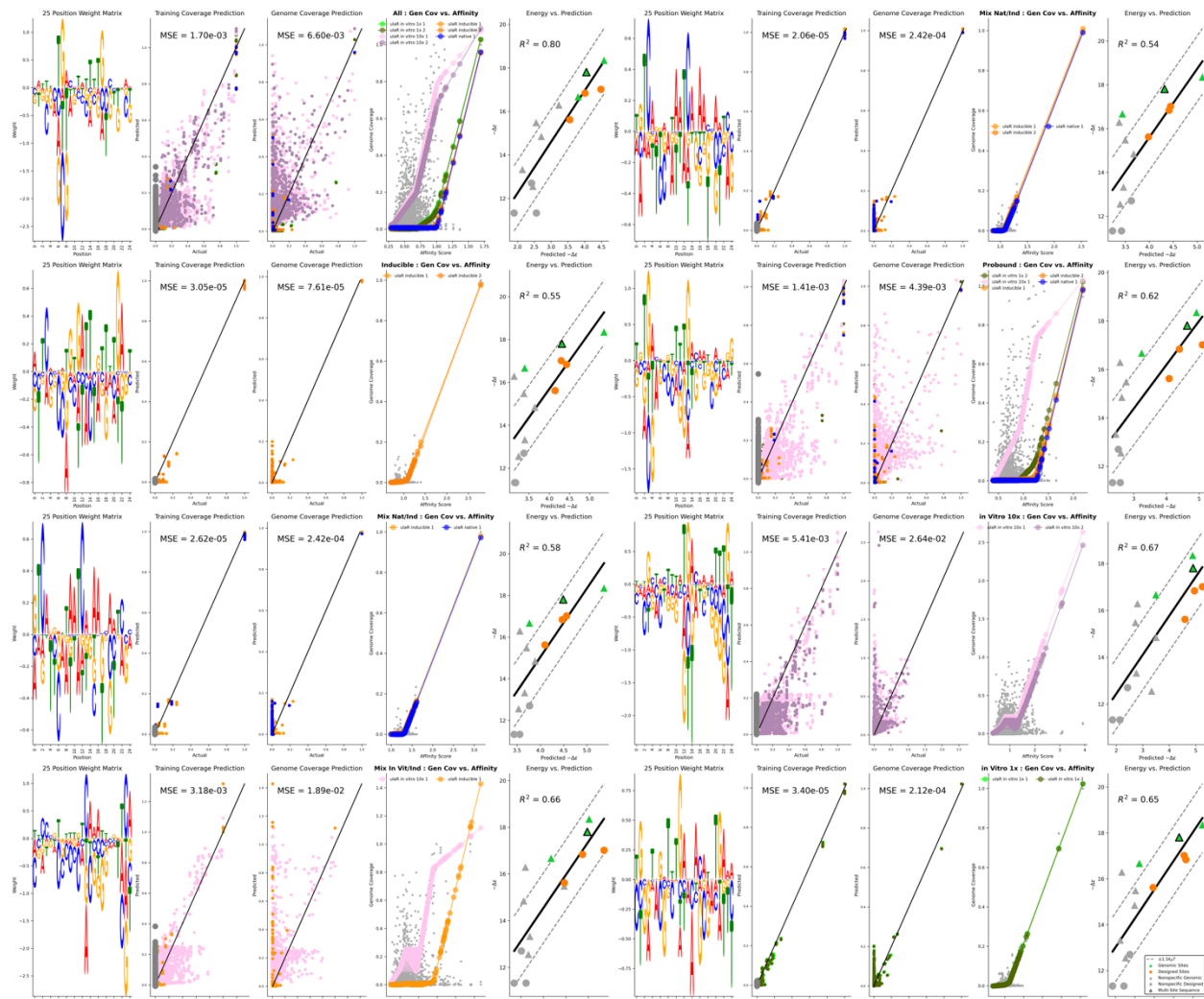
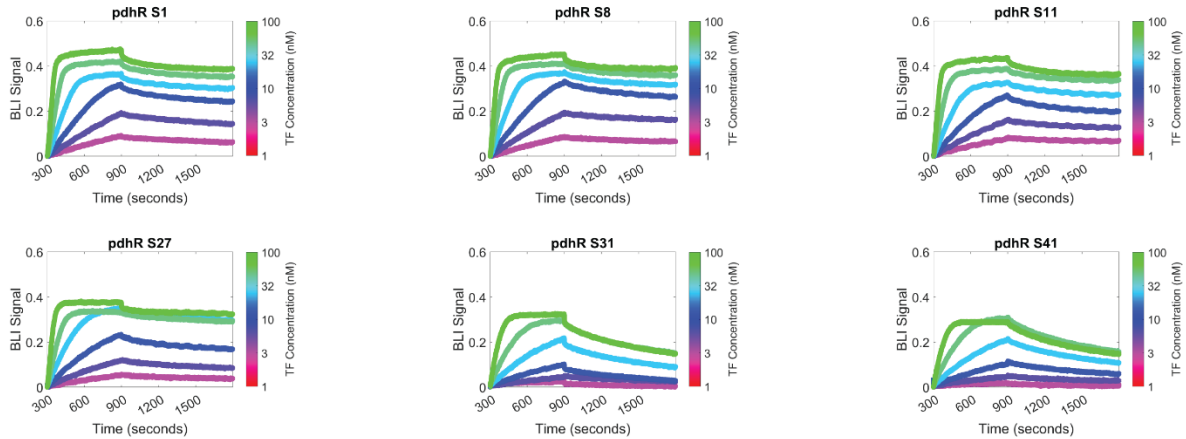
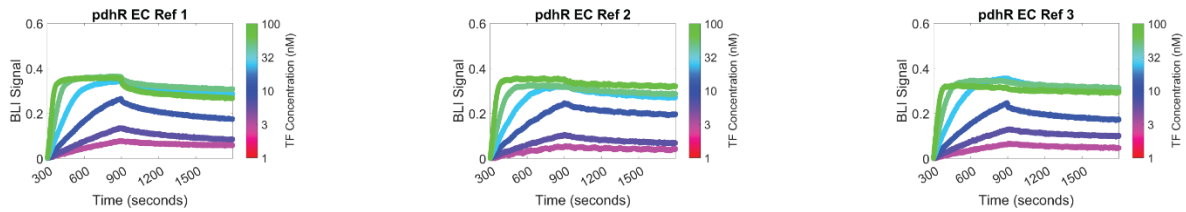


Figure S53: Analysis of impact of sub-selection called experiments for UlaR. Each repeated set of 5 panels corresponds to a model built with a different sub-selection of experiments as in bolded title and experiment legends. For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage. All experiments are color coded as indicated in the legend in of coverage vs. affinity plots.

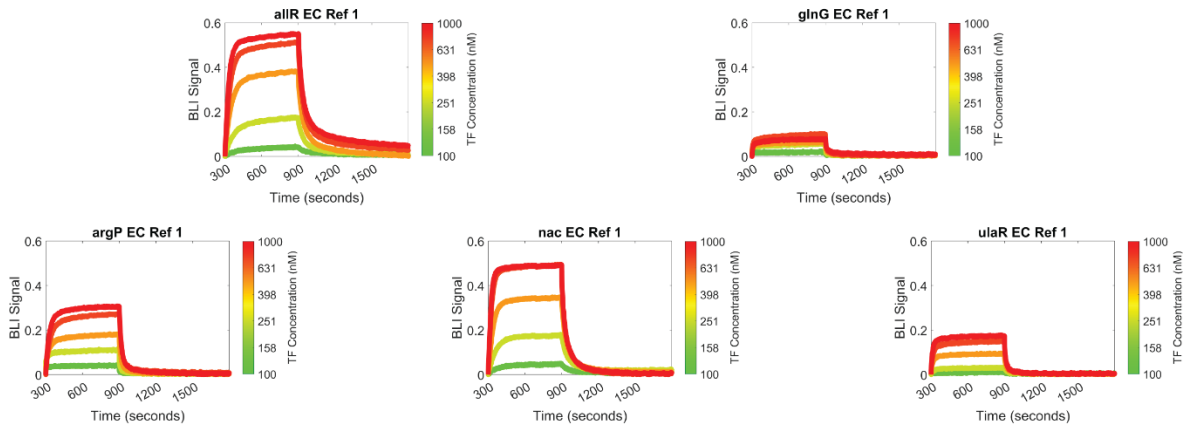
PdhR Designed Sequences



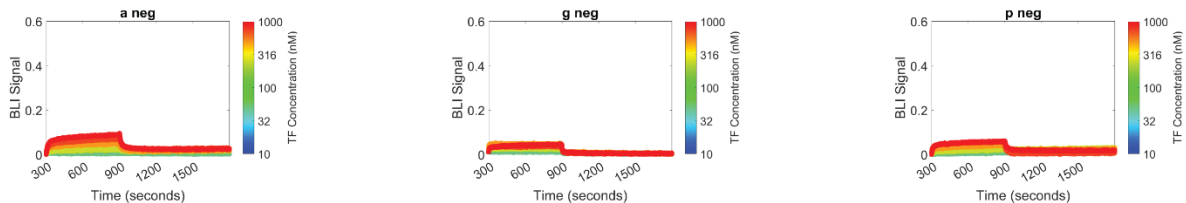
PdhR Genomic Sequences



Non-Specific Genomic Sequences



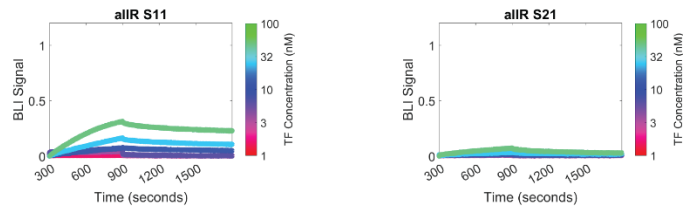
Scrambled Sequences



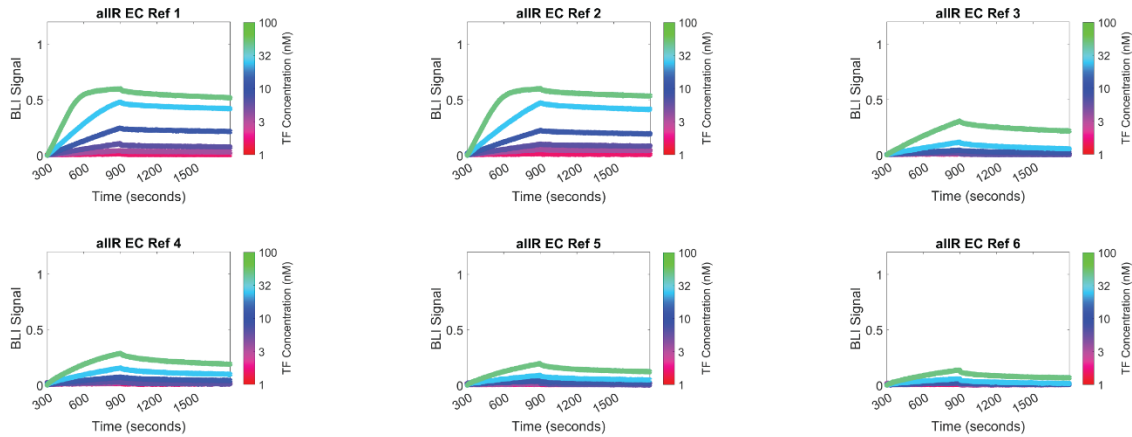
1

2 **Figure S54: Plots of BLI data for PdhR.** BLI binding curves of PdhR against sequences designed from BoltzNet (top two rows), PdhR
3 genomic sites (third row), genomic sites of other TFs (fourth and fifth rows), or scrambled versions of genomic sites for AllR (a neg),
4 GlnG (g neg), or PdhR (p neg) (bottom row). Colors of curves indicate TF concentration used – note that higher concentrations were
5 required for non-specific DNAs to achieve a measurable signal.

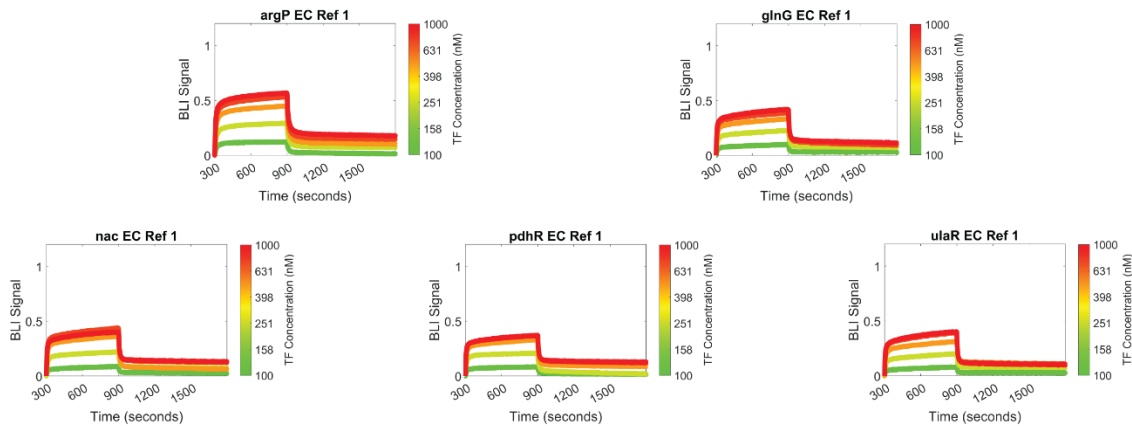
AlIR Designed Sequences



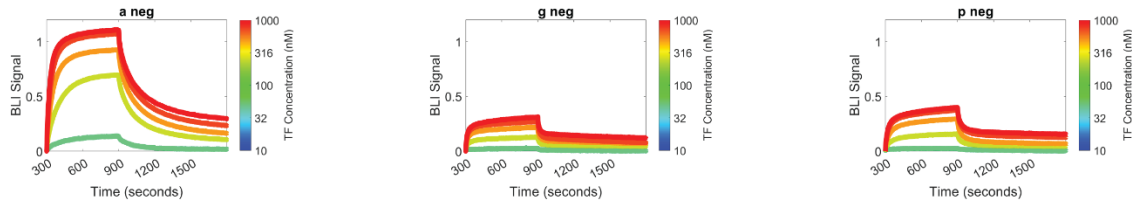
AlIR Genomic Sequences



Non-Specific Genomic Sequences



Scrambled Sequences



1

2

3

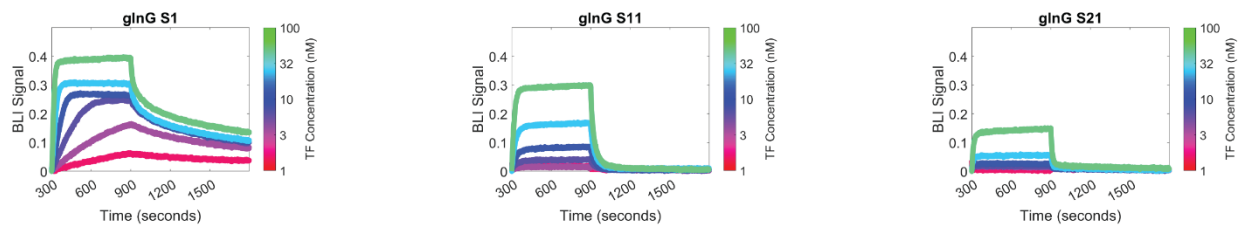
4

5

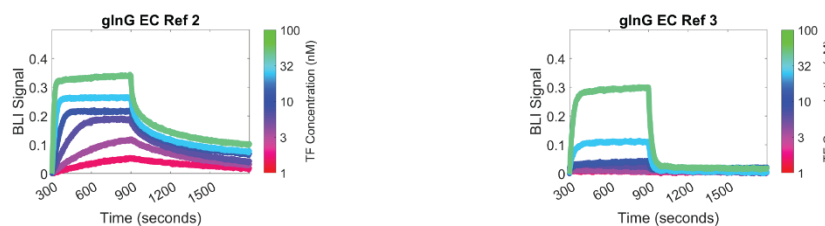
Figure S55: Plots of BLI data for AlIR. BLI binding curves of AlIR against sequences designed from BoltzNet (top row), AlIR genomic sites (second and third row), genomic sites of other TFs (fourth and fifth rows), or scrambled versions of genomic sites for AlIR (a neg), GlnG (g neg), or PdhR (p neg) (bottom row). Colors of curves indicate TF concentration used – note that higher concentrations were required for non-specific DNAs to achieve a measurable signal.

1

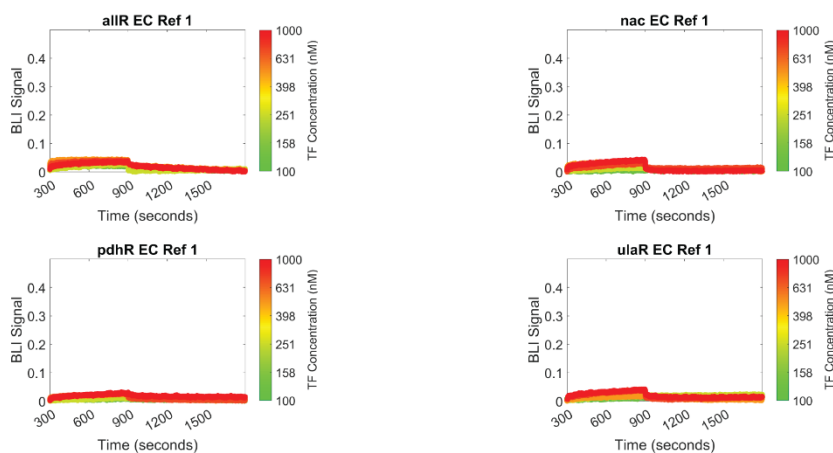
GlnG Designed Sequences



GlnG Genomic Sequences



Non-Specific Genomic Sequences

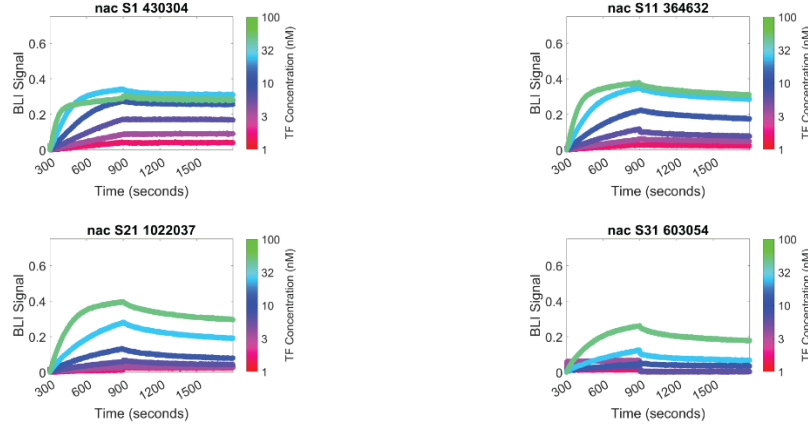


2

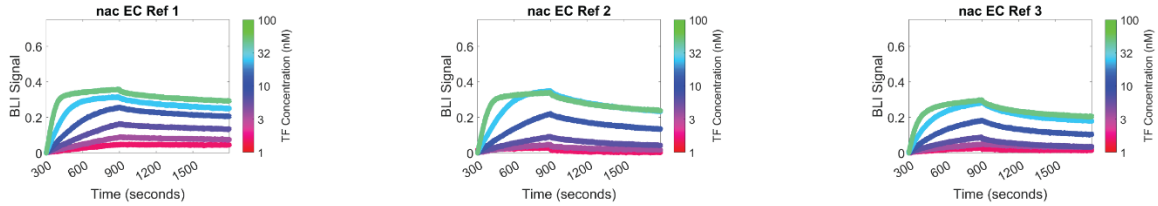
3 **Figure S56: Plots of BLI data for GlnG.** BLI binding curves of GlnG against sequences designed from BoltzNet (top row), GlnG
 4 genomic sites (second row), genomic sites of other TFs (third and fourth rows). Colors of curves indicate TF concentration used –
 5 note that higher concentrations were required for non-specific DNAs to achieve a measurable signal.

6

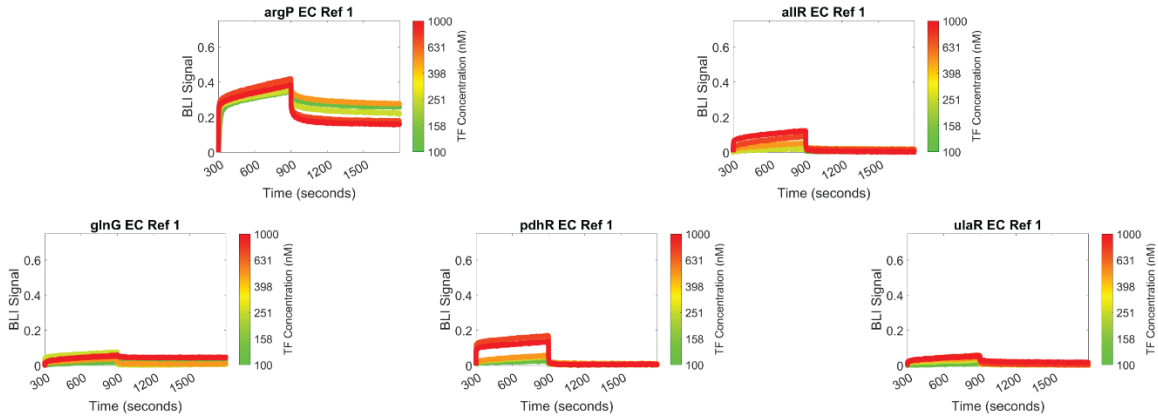
Nac Designed Sequences



Nac Genomic Sequences



Non-Specific Genomic Sequences



Scrambled Sequences

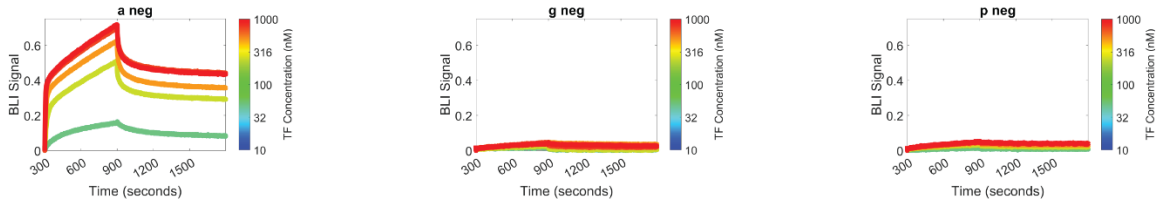
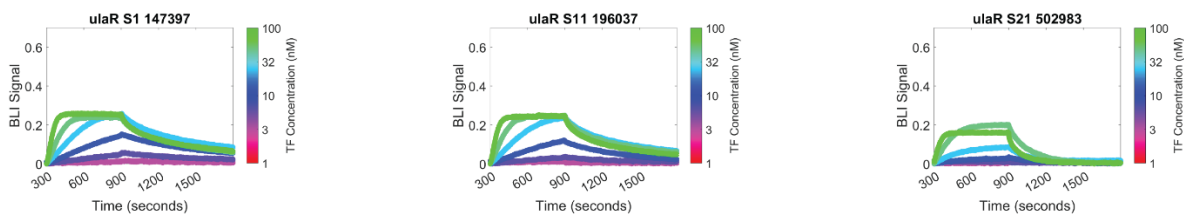
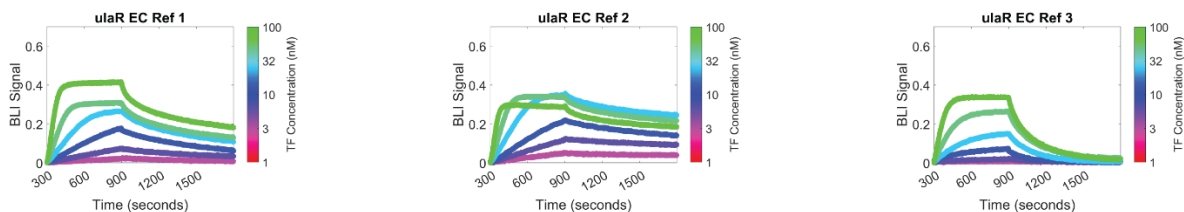


Figure S57: Plots of BLI data for Nac. BLI binding curves of Nac against sequences designed from BoltzNet (top two rows), Nac genomic sites (third row), genomic sites of other TFs (fourth and fifth rows), or scrambled versions of genomic sites for AllR (a neg), GlnG (g neg), or PdhR (p neg) (bottom row). Colors of curves indicate TF concentration used – note that higher concentrations were required for non-specific DNAs to achieve a measurable signal.

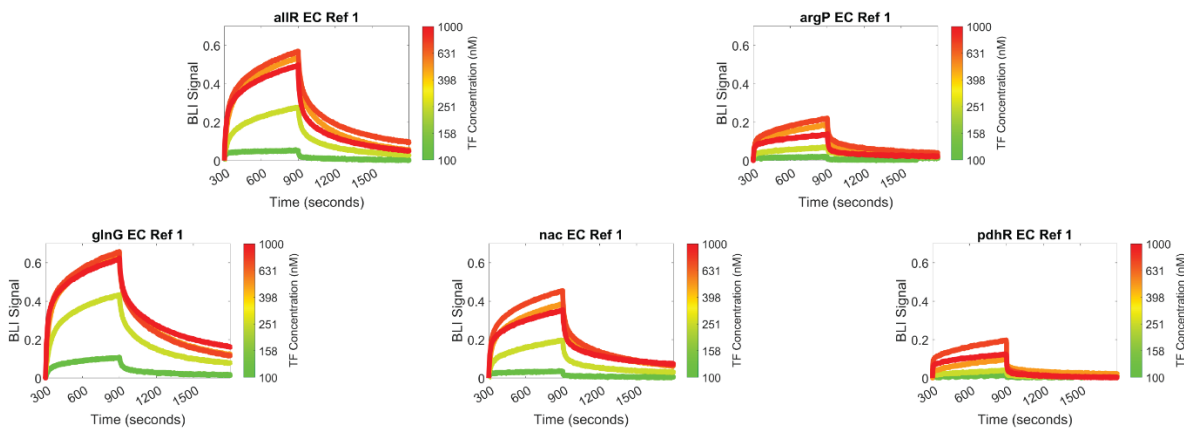
UlaR Designed Sequences



UlaR Genomic Sequences



Non-Specific Genomic Sequences



Scrambled Sequences

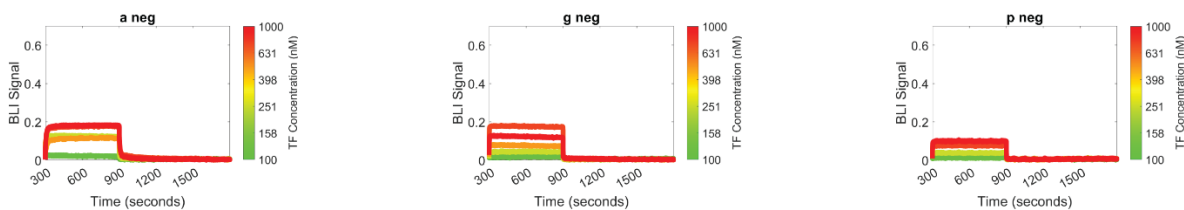


Figure S58: Plots of BLI data for UlaR. BLI binding curves of UlaR against sequences designed from BoltzNet (top row), UlaR genomic sites (second row), genomic sites of other TFs (third and fourth and rows), or scrambled versions of genomic sites for AllR (a neg), GlnG (g neg), or PdhR (p neg) (bottom row). Colors of curves indicate TF concentration used – note that higher concentrations were required for non-specific DNAs to achieve a measurable signal.

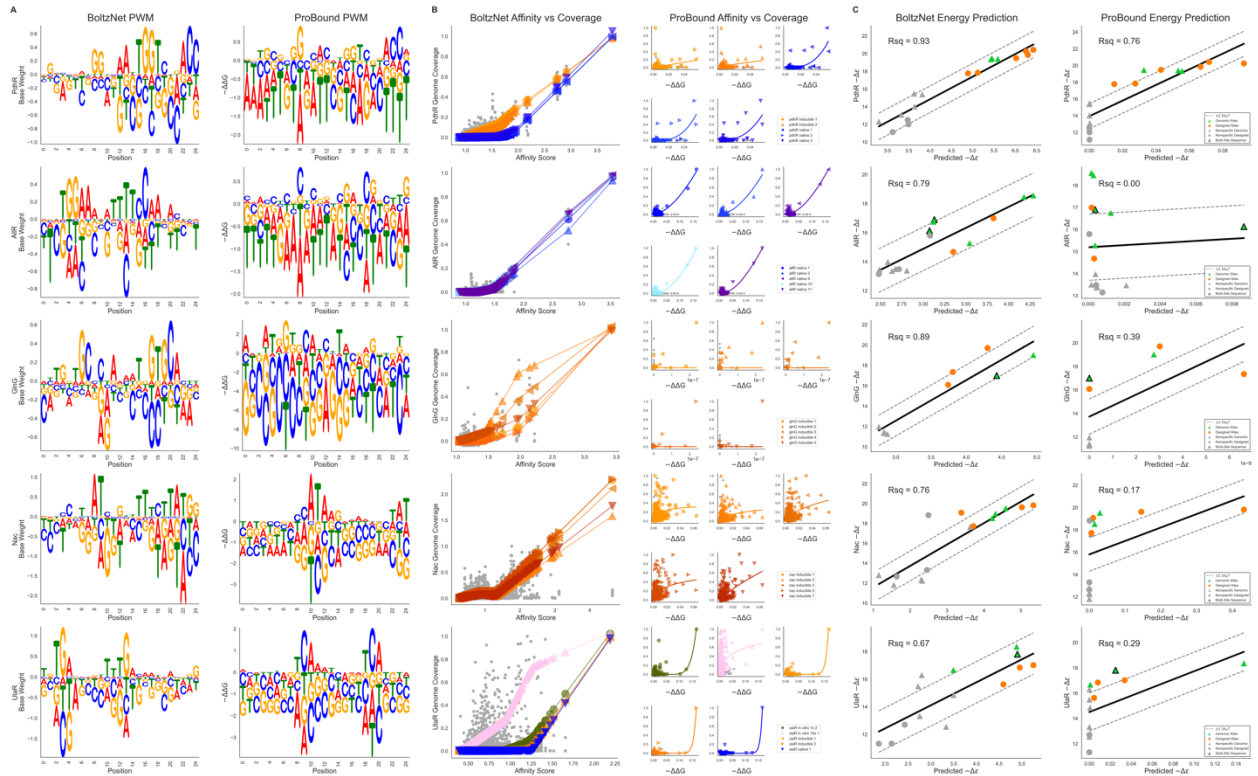


Figure S59: Comparison of BoltzNet to ProBound. Comparison of binding weight matrices, genome affinity prediction, and novel sequence prediction for 5 TFs calibrated by BLI. Each row corresponds to a TF. **(A)** The first two columns compare the binding energy weight matrices from BoltzNet (left) with the binding energy weight matrices from ProBound (right). **(B)** The second two columns compare the prediction of coverage as a function of sequence binding affinity for BoltzNet (left) and Probound (right). Colored lines represent predicted coverage from BoltzNet or fit to a Hill function for ProBound, while gray points and colored points show actual genome coverage for BoltzNet (left) and ProBound (right) respectively. **(C)** The final pair of columns compares predicted binding energy for sequences from Figure 5 vs the measured binding energy from BLI. Dotted lines represents margins of ± 1 kBT relative to the linear regression line (solid line). Data points as labeled in the far right plots.

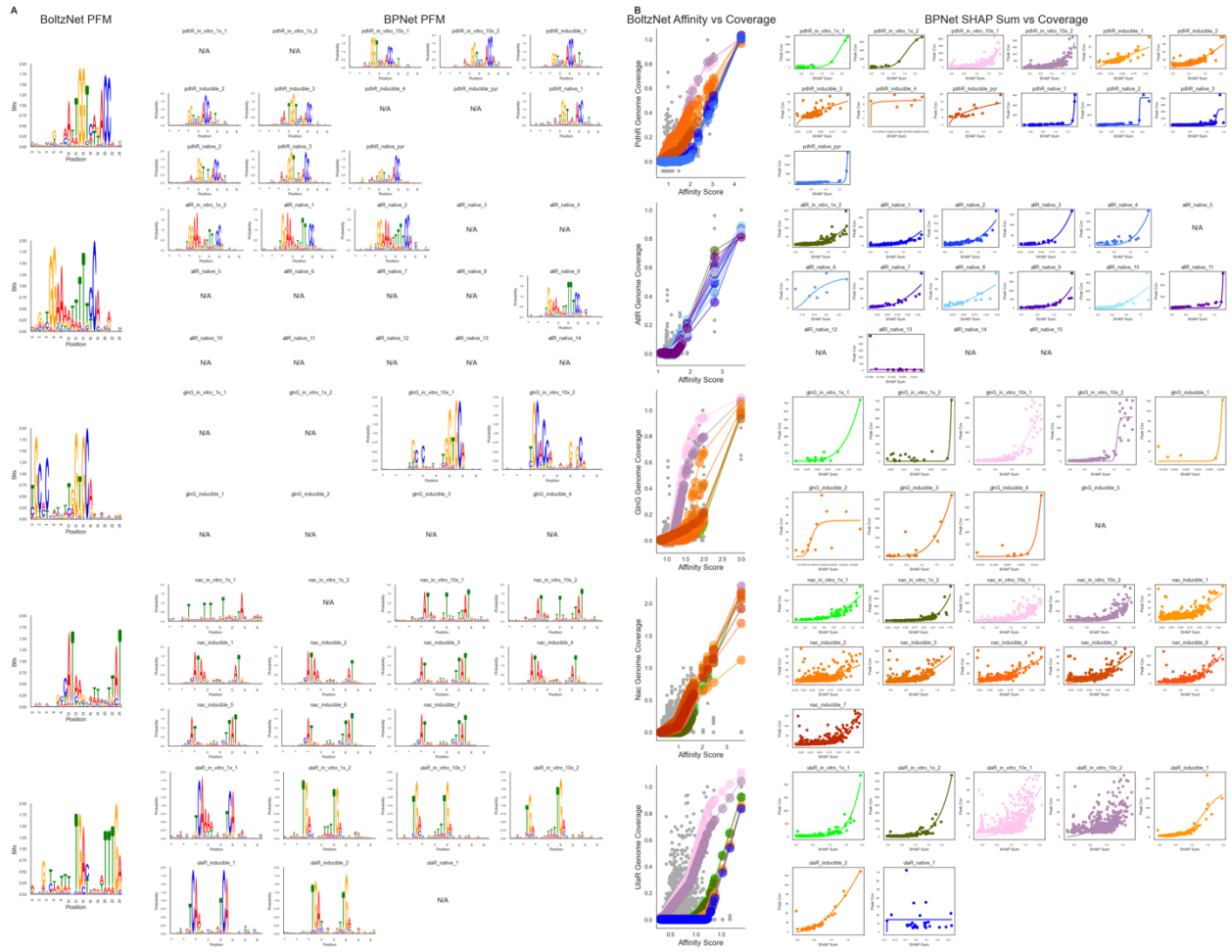


Figure S60: Comparison of BoltzNet to BPNet – PFM and Genome Coverage. Comparison of frequency weight matrices and genome affinity predictions for 5 TFs calibrated by BLI. Each row corresponds to a TF. **(A)** The first two columns compare the position frequency matrices (PFM) derived from BoltzNet weight matrices (left) with the PFMs from BPNet using TF-Modisco (right). BPNet does not directly provide a binding weight matrix. BPNet cannot be trained on more than one experiment, so PFMs are shown for each individually trained model. **(B)** The second two columns compare the prediction of coverage as a function of sequence binding affinity for BoltzNet and the sum of the SHAP scores for BPNET. (Right) In the BoltzNet figure, colored lines represent predicted coverage from BoltzNet outputs while gray points and colored points show actual genome coverage. (Left) In the BPNet figures, we plot the sum of SHAP scores versus genome coverage. As described in Supplementary Material, BPNet does not provide an affinity score. SHAP can be used to derive an importance score at each position of sequence. We sum this score over the entire sequence to generate a measure of affinity. The lines represent hill function fits to the plots.

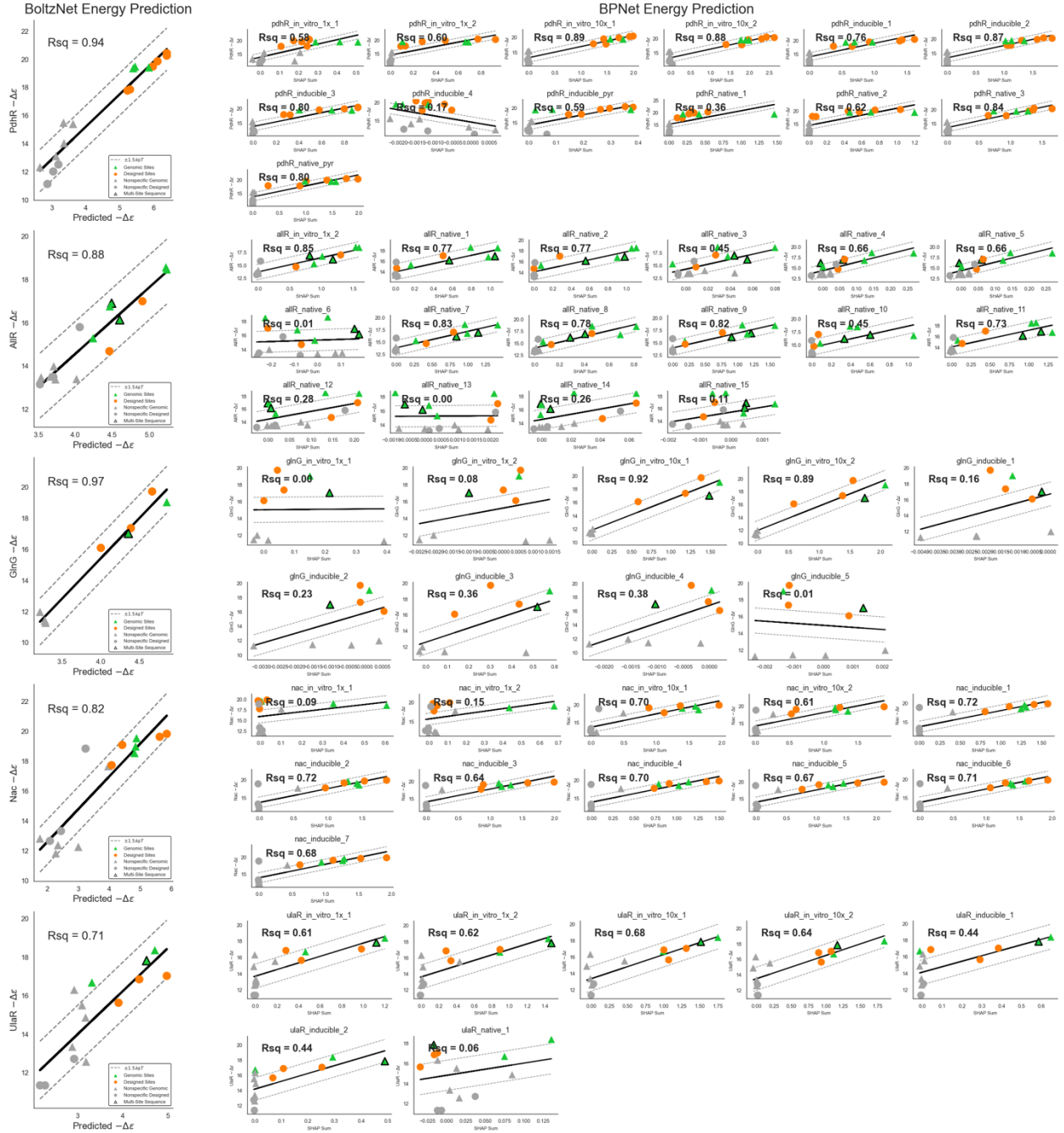


Figure S61: Comparison of BoltzNet to BPNet – Novel Sequence Energy Prediction. Prediction of predicted energy (BoltzNet) or SHAP sum scores (BPNet) versus BLI measurements of energies for the sequences of Figure 5. Each row represents a TF. (Left) Scatter plot of the BoltzNet predicted energy vs the BLI measured energy. (Right) Scatter plot of the sum of SHAP scores for a sequence vs the BLI measured energy. BPNet cannot be trained on more than one experiment at a time, so we show plots for each individually trained model. Dotted gray lines represents margins of ± 1.5 kT relative to the linear regression line (solid black line). Data points for each plot follows the legend for the BoltzNet plot for that TF.

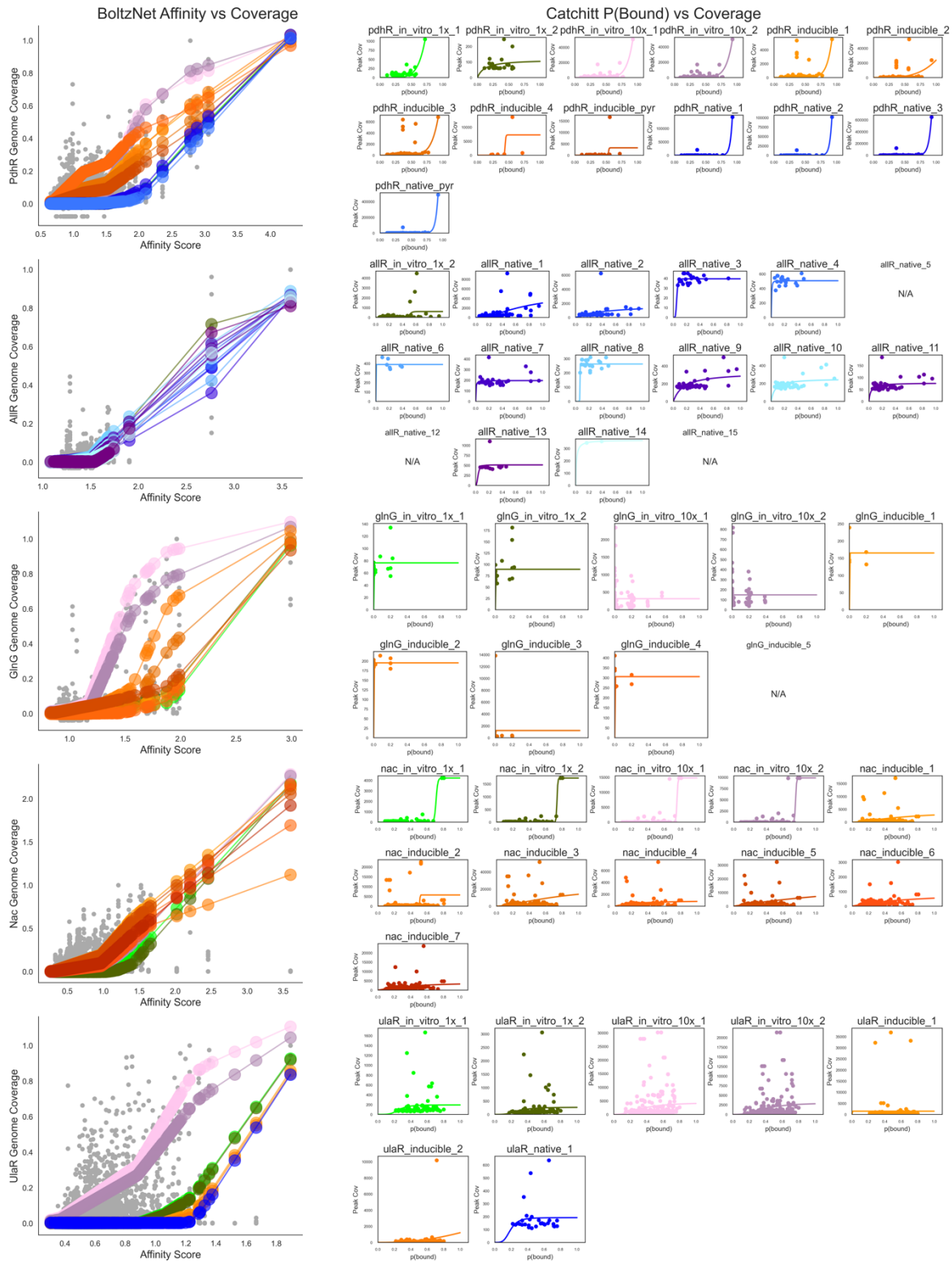


Figure S62: Comparison of BoltzNet to Catchitt. Comparison of coverage vs affinity score (BoltzNet) or P(bound) (Catchitt). Colored lines represent predicted coverage from BoltzNet or fit to a Hill function for Catchitt, while gray points and colored points show actual genome coverage for BoltzNet (left) and ProBound (right) respectively.

[illegible]

Figure S63: pAMD Plasmid Map. Library-ChIP pool construction begins with inserting designed binding sites into the pAMD-BA plasmid. This vector is selected with chloramphenicol resistance (Cat gene), and contains two origins of replication – the standard oriV (single-copy), and ori2, which can make the plasmid high-copy under induction of a TrfA gene present in certain *E. coli* strains.

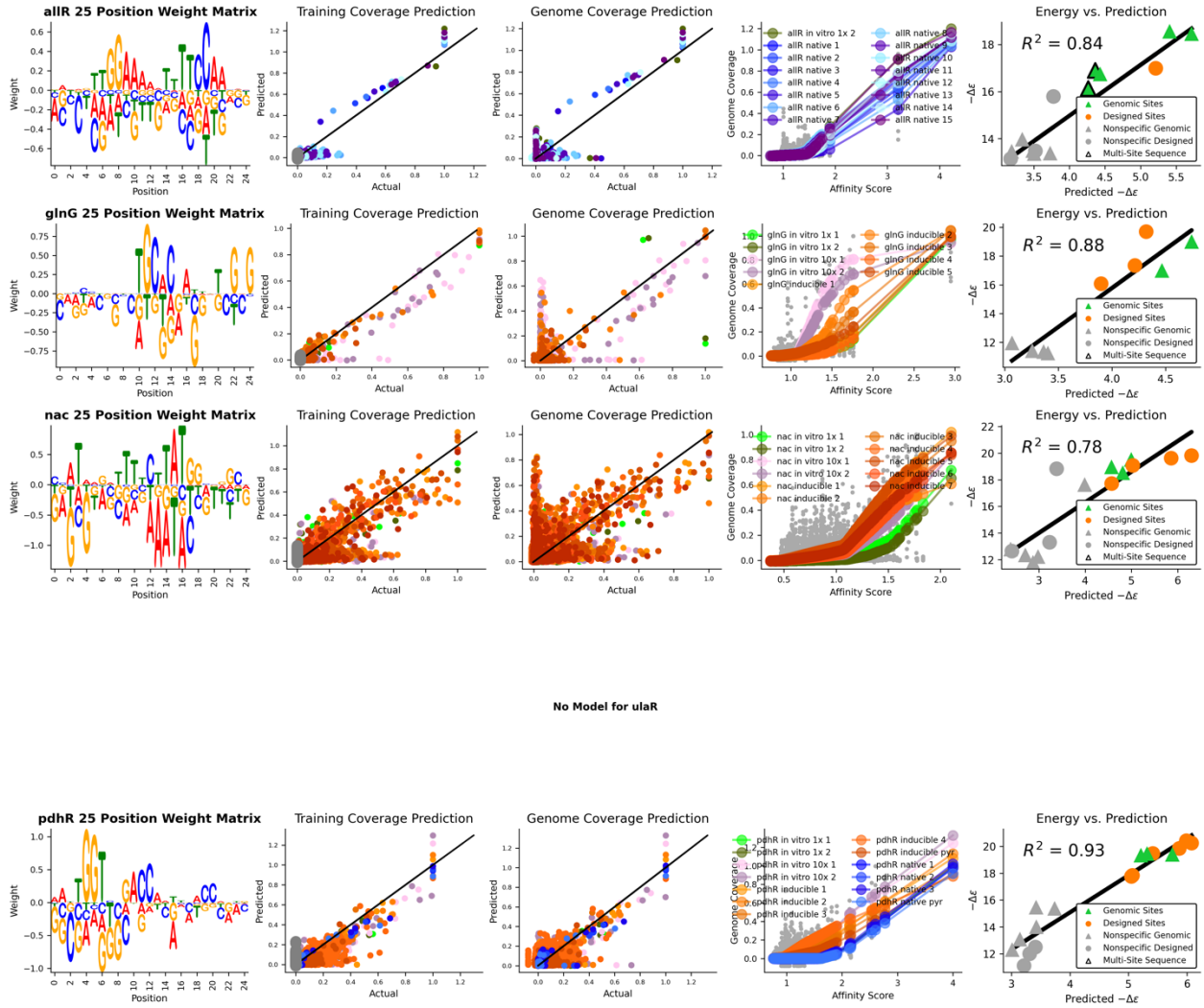


Figure S64: Analysis of Models Built Without Data Augmentation. We built models for PdhR, AllR, GlnG, Nac, and UlaR without augmentation with circular shift permutation (see Methods). For each model we plot the following from left to right: the model weight matrix, the accuracy on the training set, the accuracy of prediction on the full genome, the accuracy of learned model of coverage vs. BoltzNet affinity score, and predicted affinity score vs. experimentally determined binding energy for sequences from Figure 5. In the plots of coverage vs BoltzNet affinity score (last column) the colored points and lines are the learned models, and the gray points are the true coverage.

1 Supplementary Tables

2 Table S1: Table of TFs that failed to generate BoltzNet Models.

TF	Known Peaks	Experiments	Called Peaks	Max Fold Enrichment
allS	0	2	37	34.6
csgD	0	4	397	128.45
csiR	0	4	87	76.49
gadX	0	2	178	59.16
gcvR	0	2	3	11.13
hipA	0	4	20	7276.35
hipB	0	2	32	25.99
mazE	0	2	1	199.94
nanR	0	2	388	201.83
sgrR	0	2	3	112.897
stpA	0	2	201	28.75
yafC	0	2	37	35.67
yedW	0	2	2	61.81
yfiE	0	2	1	25.09
yjgJ	0	2	1	333.42

3

1 Table S2: Tools Investigated for Comparison to BoltzNet.

	BoltzNet	ProBound[27]	BPNet[28]	Catchitt[31]	DeepBind[38]	DeepLift[39]	FactorNet[40]
Software Available							(3)
Trainable on ChIP-Seq Data					(1)	(2)	
Support Replicates							
Predict Genome Coverage							
Predict Binding Energy/ P(bound)			(4)				
Predict Motif (PFM)							
Predict Binding Weight Matrix							
Predict Novel Sequence Energy							

2 (1) DeepBind provides code for using pre-trained models but does not support building new models.

3 (2) DeepLift is an interpretation engine for extracting patterns from neural networks but is not itself a standalone tool
4 for analyzing genomic data.

5 (3) We were not able to compile the code for FactorNet. The code no longer appears to be supported.

6 (4) SHAP scores are not explicit binding energies.

Supplementary References

1. Zeghouf, M., et al., *Sequential Peptide Affinity (SPA) system for the identification of mammalian and bacterial protein complexes*. J Proteome Res, 2004. **3**(3): p. 463-8.
2. Stringer, A.M., et al., *FRUIT, a scar-free system for targeted chromosomal mutagenesis, epitope tagging, and promoter replacement in Escherichia coli and Salmonella enterica*. PLoS One, 2012. **7**(9): p. e44841.
3. Butland, G., et al., *Interaction network containing conserved and essential protein complexes in Escherichia coli*. Nature, 2005. **433**(7025): p. 531-537.
4. Cherepanov, P.P. and W. Wackernagel, *Gene disruption in Escherichia coli: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant*. Gene, 1995. **158**(1): p. 9-14.
5. Aquino, P., et al., *Coordinated regulation of acid resistance in Escherichia coli*, in *BMC Systems Biology*. 2017, BMC Systems Biology. p. 1-15.
6. Galagan, J., A. Lyubetskaya, and A. Gomes, *ChIP-Seq and the complexity of bacterial transcriptional regulation*. Curr Top Microbiol Immunol, 2013. **363**: p. 43-68.
7. Galagan, J.E., et al., *The Mycobacterium tuberculosis regulatory network and hypoxia*. Nature, 2014. **499**: p. 178-183.
8. Jaini, S., et al., *Transcription Factor Binding Site Mapping Using ChIP-Seq*. 2014.
9. Grazon, C., et al., *A progesterone biosensor derived from microbial screening*.
10. O'Malley, R.C., et al., *Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape*. Cell, 2016. **165**(5): p. 1280-1292.
11. Baumgart, L.A., et al., *Persistence and plasticity in bacterial gene regulation*.
12. Andrews, S. *FastQC - A quality control analysis tool for high throughput sequencing data*. [Github Repository]; Available from: <https://github.com/s-andrews/FastQC>.
13. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads*. 2011, 2011. **17**(1): p. 3.
14. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2*. Nat Methods, 2012. **9**(4): p. 357-9.
15. Danecek, P., et al., *Twelve years of SAMtools and BCFtools*. Gigascience, 2021. **10**(2).
16. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration*. Brief Bioinform, 2013. **14**(2): p. 178-92.
17. *Picard toolkit*. 2019, Broad Institute: Broad Institute, GitHub repository.
18. Peterson, M.W., *Network reconstruction by ChIP-seq in Mycobacterium tuberculosis and Neurospora crassa*. 2013.
19. Salgado, H., et al., *RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12*. Nucleic Acids Res, 2024. **52**(D1): p. D255-D264.
20. Plumbridge, J.A., *Repression and induction of the nag regulon of Escherichia coli K-12: the roles of nagC and nagA in maintenance of the uninduced state*. Mol Microbiol, 1991. **5**(8): p. 2053-62.
21. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics, 2010. **26**(6): p. 841-2.
22. Bintu, L., et al., *Transcriptional regulation by the numbers: models*. Curr Opin Genet Dev, 2005. **15**(2): p. 116-24.
23. Brewster, R.C., D.L. Jones, and R. Phillips, *Tuning promoter strength through RNA polymerase binding site design in Escherichia coli*. PLoS Comput Biol, 2012. **8**(12): p. e1002811.

24. Milo, R., et al., *BioNumbers—the database of key numbers in molecular and cell biology (BNID 100050)*. Nucleic Acids Research, 2010. **38**(suppl_1): p. D750-D753.
25. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching*. Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
26. Tareen, A. and J.B. Kinney, *Logomaker: beautiful sequence logos in Python*. Bioinformatics, 2020. **36**(7): p. 2272-2274.
27. Rube, H.T., et al., *Prediction of protein-ligand binding affinity from sequencing data with interpretable machine learning*. Nat Biotechnol, 2022. **40**(10): p. 1520-1527.
28. Avsec, Z., et al., *Base-resolution models of transcription-factor binding reveal soft motif syntax*. Nat Genet, 2021. **53**(3): p. 354-366.
29. Shrikumar, A., et al., *Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5*. arXiv preprint arXiv:1811.00416, 2018.
30. Lundberg, S., *A unified approach to interpreting model predictions*. arXiv preprint arXiv:1705.07874, 2017.
31. Keilwagen, J., S. Posch, and J. Grau, *Accurate prediction of cell type-specific transcription factor binding*. Genome Biology, 2019. **20**(1): p. 9.
32. Keilwagen, J., S. Posch, and J. Grau, *Learning from mistakes: Accurate prediction of cell type-specific transcription factor binding*. bioRxiv, 2017: p. 230011.
33. Wade, J.T., *High-throughput determination of in vivo DNA sequence preferences for Cas protein binding using Library-ChIP*. Methods Enzymol, 2019. **616**: p. 117-132.
34. Gazon, C., et al., *A progesterone biosensor derived from microbial screening*. Nat Commun, 2020. **11**(1): p. 1276.
35. Sankar, K., et al., *An Allosteric Transcription Factor DNA-Binding Electrochemical Biosensor for Progesterone*. ACS Sensors, 2022. **7**(4): p. 1132-1137.
36. Quail, M.A. and J.R. Guest, *Purification, characterization and mode of action of PdhR, the transcriptional repressor of the pdhR-aceEF-lpd operon of Escherichia coli*. Mol Microbiol, 1995. **15**(3): p. 519-29.
37. Weeramange, C.J., et al., *The strengths and limitations of using biolayer interferometry to monitor equilibrium titrations of biomolecules*. Protein Science, 2020. **29**(4): p. 1004-1020.
38. Alipanahi, B., et al., *Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning*. Nat Biotechnol, 2015. **33**(8): p. 831-8.
39. Shrikumar, A., P. Greenside, and A. Kundaje. *Learning important features through propagating activation differences*. PMIR.
40. Quang, D. and X. Xie, *FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data*. Methods, 2019. **166**: p. 40-47.