nature portfolio

Peer Review File

Predictive Biophysical Neural Network Modeling of a Compendium of *in vivo* Transcription Factor DNA Binding Profiles for *Escherichia coli*

Corresponding Author: Professor James Galagan

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The manuscript, “Predictive Biophysical Neural Network Modeling of a Compendium of in vivo Transcription Factor DNA Binding Profiles for Escherichia coli”, describes an extensive study of defined E. coli transcriptional regulators using native and induced expression and a ChIP-seq approach. With this data, including weak affinity sites, they could train their novel neural network, Boltznet, to predict transcription factor binding energies to most any sequence, thereby allowing a fuller description of DNA-binding affinities and occupancy prediction throughout an organism’s genome.

This is a fairly well-written manuscript describing a new method of analyzing ChIP-seq data, developed from a considerable amount of data. As with other artificial intelligence approaches, it provides a useful alternative to current methods (e.g. SELEX; MEME), especially when identifying and designing sequences with defined binding affinities. The data provided are sound and provide sufficient support for most all of their conclusions. Thus, this manuscript is a worthy contribution, one that could be of benefit to many investigators. However, we did have a few concerns that, if addressed, would make for an even better paper suitable for publication in the journal Nature Communications.

[Introduction, para 3] The application of deep-learning/neural networks to transcription factor binding is not a new concept. In fact, it was part of the ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge in 2016. Several approaches have been described to address this question: BPNet, Catchitt, DeepBind, FactorNet, and ProBound, among many others. The authors mention in passing that these approaches are overly complex, beyond what is necessary for prokaryotic transcription factors. While it is true that they may not be as tied to physical parameters (e.g., binding energies) as BoltzNet, for the general investigator interested in ascertaining gene regulatory networks, are they that much inferior? Minimally, a greater discussion of their strengths and weaknesses would be warranted.

[Introduction, para 5] The authors describe the global mapping of in vivo DNA binding for 139 E. coli transcription factors using ChIP-Seq. Bravo – a truly impressive body of work. However, there are an estimated 304 putative transcription factors in E. coli. Why did the authors only investigate this subset? Were there transcription factors that were investigated but did not work? Likewise, they were able to generate high-confidence DNA-binding models for 125 of the 139 transcription factors successfully investigated by ChIP-seq. A further elaboration on the choice of transcription factors investigated and what characteristics best predict success would be most informative for those who wish to use BoltzNet.

[Results, para 1] The ordering of the figures vis-à-vis their first appearance in the text is hardly conventional. For example, the first figure cited is S2. Figure S1 only appears in the Acknowledgements. Likewise, their analysis pipeline is supposedly shown in Figure S4 and global mapping data S10. These are S5 and S11, respectively. Please carefully check these misattributions and correct them throughout.

[Results, para 2] Figures 1A and S11 are the same data, just different presentations. Necessary? Also, while 1A explains different gene name coloration, S11 does not. This is a common problem throughout the manuscript, where information that should be in the figure or at least its legend is either absent or only available in the text (e.g., outlined circles in Figure 1E). These must be improved.

[Results, para 2] The results of their extensive ChIP-seq studies are supposedly given in Table S1. A simplified distillation of their findings, without any explanations, perhaps. Rather, we strongly support disseminating their original data (sequences, abundance, etc.) in a format that permits facile data mining for others. Such would be quite useful.

[Results, para 3] It is stated that the binding regions were highly enriched within 150 bp upstream and 50 bp downstream of gene starts (Figure 1G). That is qualitatively true if one is comparing only upstream sites or downstream sites amongst themselves, not upstream to downstream sites. In absolute terms, sites from -500 upstream to +1 are greater than those from +1 to +50 bp.

[Figure 2] Visual impairments aside, I am hardly a fan of the color schemes used in these figures. Perhaps there is a better way (e.g., Figure 2B - using shades of the same primary color to indicate replicates)? Also, one needs to be consistent throughout (e.g., native 1 & 2 in 1B and 1C). Such would make interpretation far easier for the reader. Also, using the same color scheme for different data in a single figure is completely unacceptable (Figure 2J, both Native 1 and in vitro 10X 2 are purple).

[Figure 2] Once again, the order of the figure subparts is not standard. E-J-F? Aesthetics should be secondary to consistency.

[Figure 2F, G, H, I] (a) BoltzNet results for which transcription factors are presented in these figures? Best, provide in the figure itself; minimally, in legend. (b) Different scales are used for the different transcription factors. This leads one to assume they are of equal significance, but actually they are not. (c) Affinity scores of binding sites are usually ~2x greater than background. This is far less than the observed or predicted data would indicate. Why is that? Tyranny of the weak?

[Figure 2J] Nine traces are visible but eleven are indicated by symbols. Complete overlaps? Needs to be clarified.

[Figure 3] (A) Rather dense figure. Comparing BoltzNet models with RegulonDB motifs? Why do the BoltzNet models differ from the PFMs in Table S2 (e.g., argP)? (B, C, D, E) Color schemes for the Coverage Accuracy are not defined. Cyan highlighting in sequences not defined.

[Results, para 5] Why is the weight matrix only 25 bp? While many transcription factor binding sites may be adequately defined by such, definitely not all.

[Figure 4] (B, C, D). R2 all < 0.9. Why are these considered “highly correlated” in the text? (D) Why are the triangles bolded? (Legend) the text “predicted BoltzNet predictions” is a redundancy. “predicted” can be removed.

[Results, para 8] “PFMs are not capable of predicting ChIP-Seq coverage” is a very strong statement, especially given that you only provide data for one transcription factor (pdhR). How true is this for all the transcription factors you have studied? Have you compared other methods for generating PFMs – XSTREME or MEME-ChIP?

[Figure 5] Why are the measured and predicted scales so different? Any idea why the correlations are better for some transcription factors (PdhR, GlnG) but not for others (Nac, UlaR)?

[BLI data] Figure 5 provides a comparison of BLI experimental and BoltzNet predicted binding energies. But where is the original BLI data? There should be Supplemental data that shows the association and dissociation curves for each of these.

[Figure 6] (a) What do the different color traces correspond to? Different in vivo ChIP experiments? More details are needed in the legend. (b) Why this particular alphabetical order (A, C, E, B, D, F)? Can be confusing. (c) (F) Quite an interesting observation, but this Figure is not referenced in the text.

[Discussion, para 5] The importance of weak sites is a major theme throughout this manuscript. Transcription factor overexpression, extensive ChIP-Seq, and BoltzNet permit their identification. But are they physiologically relevant, as alluded? This paper is woefully lacking in biology, and it would have been much more impactful if there had been more ties to biological implications. Future studies in gene expression (para 6) allude to this. Perhaps, if their original ChIP-Seq data are made freely accessible, others may find them, too.

[Methods, ChIP-Seq Pipeline] Not Figure S4A. Figure S5A & B?

[References] Some are not complete (e.g., #15).

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The manuscript describes a computational method for predicting ChIP-seq coverage values for transcription factor (TF) binding sites in E. coli using a machine learning approach informed by a mathematical model of binding affinity. Using ChIP-seq data from 139 TFs, they trained a two-part model which first predicts binding affinity, then predicts ChIP-seq coverage under each of the measured experimental conditions. The key innovation is in incorporating a biophysical model of TF-DNA binding into the model architecture, which enables intermediate predictions of binding affinity. The inferred binding affinity was validated with native and engineered DNA sequences for a handful of TFs. Overall, the model does well in predicting ChIP-seq coverage and may be useful in characterizing TF binding in continuous, rather than discrete, terms, as well as designing new or modified binding sites with quantitative predictive power. The manuscript omits some key points which would improve the impact and rigor of the results, including a comparison with existing methods, and would be improved by greater clarity in some areas. Overall, the work is novel and interesting, and may be of interest to those working on bioinformatics, biophysics, and systems biology.

1. The manuscript does not present any comparison with existing methods that also aim to predict ChIP-seq signatures, such as ProBound and BPNet. The authors mention some differences in target systems which may complicate this analysis, but some comparison on high-level metrics would be helpful to contextualize whether this model is an advance in state-of-the-art performance as well as in interpretability.
2. I found the description of the convolutional layer somewhat unclear, in particular:
1) What approach is taken at the ends of the 101-bp sequence when 25bps are not available – e.g., is the data padded with a 'null' nucleotide?
2) When the forward and reverse strands are concatenated, it seems that at the middle of the vector the filter is convolving paired bases, but at the ends of the vector it's convolving bases on a single strand - is this correct, and does it pose any obstacle to learning an informative weight matrix? It's possible I misunderstand the approach but would appreciate more clarity in the text on this.
3. The manuscript mentions a "circular shift permutation" to increase the model's generalization abilities, but it would be helpful to have supporting data to show that this adjustment did in fact improve the model's generalization.
4. A few unsuccessful cases are mentioned where the model failed to achieve satisfactory performance for some TFs – do the authors have any reasoning or data to explain why the model broke down in these cases? Some further discussion of this data would be informative for understanding the model limitations.

Minor points:
1. Figures 3 and 6 have several panels with data grouped by color, but no legends for them.
2. Equation 5 may have a missing negative sign in the exponent.
3. On page 12, the phrase "the denominator in equation 4" appears to be a typo and I suspect should reference equation 5 instead.
4. More details are needed in the description of "grey symbols" in Figure 1G. Are they means and standard deviations or not? It will also be helpful to highlight the overrepresented bins.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
This study provides a comprehensive profiling of TF binding sites in E.coli. The authors uniformly performed large-scale mapping of TF binding sites in E.coli covering 134 TFs, and demonstrated ability in replicating known sites and discovering novel binding regions. A shallow and interpretable non-linear model was further developed to associated TF-DNA binding physics to ChIP-seq bioinformatics signals. Compared to other recent physics-inspired TF-binding models like ProBound, which built model TF-DNA binding from SELEX, the design in BoltzNet fits scenarios of the ChIP-seq data type and investigated prokaryote. The study provides a comprehensive resource of E.coli TF binding sites and the manuscript and supplements include well-written details in TF-DNA binding dynamics. However, following are a few major concerns:
Major:
1. The authors should include description about how the 134 TFs are selected from the ~300 TFs for E. coli.
2. It would be better if the authors show a quantitative replication correlation between Native and inducible replicates for all TFs with replicates, besides PdhR (Fig1E), so that it is clear which TFs are with high replication correlation and which do not (on top of Fig1F and FigS13). Also, the number of overlapping TF binding sites between replicates shall be shown. Furthermore, does Fig1H implicate most novel regions are weaker binding sites? Could it be due to differences in thresholds in peak calling/statistical region enrichment of the ChIP-seq data processing step between re-analyses of data from other studies and data from this study? Did the authors use the same ChIP-seq parameter in FigS5 to generate the known sites by reanalyzing some TF ChIP-seq data from other studies?
3. Model design: first, the BoltzNet model implements a two-component model that associates TF-DNA binding affinity score with ChIP-seq experiment signals. The TF-DNA binding dynamics layers summarizes DNA sequences into pooled affinity scores. The affinity scores are layer fitted to ChIP-seq coverage from experiments of the corresponding TF. The major

concern is model overfitting, since it is mapping a low dimension affinity score to scale up to genomic coverage. This imbalance of number of parameters and number of data points may cause loss of model stability, since the neural network weights can have many possibilities to fit the data (although L1 regularization was mentioned to put constraint on middle layer). Thus, ideally there shall be investigation in model stability. It seems the cross validation was only done for TFs with many binding sites detected (Fig2 PdhR). Second, based on Figure 1A and supplementary data, the number of binding sites is at maximum hundreds, and many TFs have binding sites below hundred. For those TFs, can the model reach high accuracy? It will be a imbalanced training case with less than hundred of positive binding scores and 5000 control scores. Thus, it is recommended to consider the canonical approaches to benchmark machine learning model accuracy per TF, using metrics like ROC curves. A comprehensive accuracy benchmarking will be critical to support the motifs interpreted in Fig3. Third, author shall also describe the reason to select 140 nodes in the middle layer.
4. The fitted models are able to predict coverage from DNA sequences and interpret motif PWM/PFM. Have the authors considered comparison with deep learning models of TF binding like DeepLift etc., where motifs can also be learned from ChIP-seq sequences? One interesting aspect of approaches used this study is the possible ability to reveal TF-DNA binding dynamics and weakly binding sites, which may be lacking when using top peaks called from ChIP-seq data. Should there be investigation about differences of the inferred motifs between strong/known binding sites and weakly binding sites?
Minor:
- Page 2, line 2 contains extra comma.
- Fig1, Fig4 panels display different boxes visually.
- Uniformly processed data and TF parameters can be deposited/stored in publicly available resources like Zenodo.

(Remarks on code availability)
It is recommended to add dependency/installation information in README.
Since the test data demonstrated in code base follow pre-fixed formats which is pulled from Galagan lab database, there shall be formatting functions or guidelines provided for other ChIP-seq data from standard ChIP-seq formats like bam/wiggle/peaks.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
A novel neural network, Boltznet, is described. It allows the prediction of transcription factor binding to almost any sequence. The authors clearly demonstrate that Boltznet is, in many ways, a superior approach to those previously described (e.g., BPNet, Catchitt, DeepBind, FactorNet, and ProBound). Thus, this manuscript is a worthy contribution to the field of transcriptional regulation, one that should benefit many investigators. The revised manuscript well addresses all of my prior concerns. It should be a valuable contribution to the scientific literature and worthy of publication in Nature Communications.

[Signed] Michael Van Dyke, Professor of Biochemistry, Kennesaw State University

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have properly addressed my questions.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
The authors have addressed the concern in model robustness and hyperparameter selections of the neural network model, providing comprehensive evaluations in the revision. However, from the perspective of machine learning, the model in effect functions by fitting the training data using shallow neural networks, which is with less novelty and involvement of Boltzman distributions. It is suggested that the author explain the method with simplicity especially "Boltzman factor layers" in a straightforward way since the model code uses a straightforward average pooling layer. For example, can K_D and [TF]_eq be inferred and interpreted by the trained models for each TF, generating reusable biophysics parameters? Moreover, the open-access data accessibility of rich TF binding dataset and training results of affinity spectrum will form a key contribution to the field.

(Remarks on code availability)
The author provided source code of the model and associated transcription factor data. It is suggested how the code can be easily adapted to other ChIP-seq data and associated parameters in standard ChIP-seq formats.

Author's Response to the Reviewer's Comments

We thank the reviewers for their thorough and careful consideration of our paper and for their helpful comments. We have responded to each of their comments and revised our manuscript accordingly. These changes have further strengthened our paper, for which we are very grateful. Below are point-by-point responses.

Reviewer 1:

1. ***Provide a comparison of your approach with other methods such as BPNet, Catchitt, DeepBind, FactorNet, and ProBound.***

We agree that this is valuable and important, and we have now performed this comprehensive analysis. We attempted to compare BoltzNet to all the tools suggested by the reviewers: ProBound, BPNet, Catchitt, DeepBind, DeepLift and FactorNet. Not all tools were available, and different tools provided different capabilities (Table S10). 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 neural network tools: 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, AllR, 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 tools. ProBound is the only tool compared that predicts an explicit binding energy and binding energy matrix. 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. For ProBound, we compared binding energy weight matrices, the relationship between sequence and binding affinity across the genome, and the accuracy of predicting binding sequence energy for the sequences measured by BLI in Figure 5 (Figure S59). ProBound matrixes for PdhR, Nac, and UlaR captured positive binding energy contributions for the expected core binding sequences and relative positions, although bases with negative weights were often more uniform. Matrices for GlnG and AllR failed to capture the expected core sequence. Predicted binding energy vs coverage displayed the expected sigmoidal relationship, with the exception of GlnG. Most importantly, ProBound performed poorly in predicting the binding energy of sequences measured by BLI (Figure 5). In nearly all cases, ProBound performed poorly on this test ($0 < R^2$ 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, AllR displayed the lowest variance relationship between coverage an ΔΔG on the genome, but the lowest accuracy on the independent test (Figure S59). The results reveal that ProBound could differentiate between expected peaks and non-specific sequences, but often failed to predict difference in affinity between specific sequences and between non-specific sequences.

BPNet is a deep convolutional neural network designed to predict genome coverage from sequence [1]. 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[2] is used to predict a PFM binding model, and SHAP [3] is used to assign to each nucleotide an importance value. In the absence of a predicted affinity score for sequences, we calculated the sum of SHAP scores (see Supplement). Moreover, BPNet could only be trained on one experiment at a time, so we generated one model per experiment. We first compared the PFM derived from the BoltzNet binding affinity matrix with the TF-MoDISco PFM (Figure S60A). 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 on the test set (Figure S60B). In the vast majority of cases, coverage across the genome could be modeled as a sigmoidal function of SHAP Sum. Finally, we tested the ability of BPNet to predict binding affinity on sequences never seen during training (Figure S61). To do so, we ran predictions using all BPNet models on the novel designed sequences from Figure 5. BPNet performed worse than BoltzNet in all comparison for all TFs. As with ProBound, BPNet could differentiate between expected peaks and non-specific sequences, but often failed to predict difference in affinity between specific sequences and between non-specific sequences.

Catchitt, unlike BoltzNet, ProBound and BPNet, is not a tool for binding motif/weight matrix discovery [4]. For training, Catchitt 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. To generate Catchitt predictions, we provided a PFM model derived from MEME run on called peaks for each TF, and applied Catchitt to called peaks for each experiment. The results reveal that the Catchitt probability of binding was not predictive of peak binding coverage (Figure 57). 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 have included these new results in both the Main Text and in Supplementary Material.

2. ***Provide a further elaboration on the choice of transcription factors investigated and what characteristics best predicted success.***

We regret our lack of clarity on this point. We have now provided a more complete description of ChIP-Seq experiments performed, and the criteria used to select the 139 TFs with high quality data that we report. In short, we targeted all 318 proteins in *E. coli* that we curated to be potential transcription factors (Table S1). For a TF to be included in our high-quality set, we then required at least 2 replicate ChIP-Seq experiments that passed all the quality control filters described in Methods and Supplementary Material.

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 one ChIP-Seq experiment that passed quality control filters (22 of these had only one native promoter experiment and 42 had one successful inducible experiment). In all 64 cases, additional attempts were made to generate replicates. 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 a stability analysis in which we built models from sub-selections 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*.

We have added this discussion to both the Main Text Discussion section and to Supplementary Material.

3. ***Check and correct the ordering of figures vis-à-vis their first appearance in the text.***

We regret that the citations to our figures were not in the correct ordering with respect to the appearance in the text. In some cases, there is a conflict in the order in which Supplementary Figures are cited in the Main Text relative to the order in which they are introduced in the Supplementary Material. We have elected to keep the ordering of Supplemental Figures consistent with the introductions in Supplementary Material. And we have worked to ensure that Main Figures are introduced in the correct order in the Main Text.

4. ***Is it necessary to provide the same information in different ways in Figures 1A and S11?***

We regret that this caused confusion. Figure 1A, Figure S11 do indeed present the same information but are ordered differently (by number of peaks or alphabetically, respectively). During our analysis of the data, we found that both figures were useful. For example, it is far easier to identify the results for a TF by name in Figure S11. However, we are happy to remove this figure if it is more confusing than useful.

5. ***Update figures to ensure that all necessary information is present.***

We regret that our figures were not as clear as they should have been. We have now worked to ensure that all our figures are fully labeled.

6. ***All original ChIP-Seq data should be disseminated in a format that permits facile data mining for others.***

We completely agree that all our data must be publicly released. And we regret that we were not clearer that this is indeed our goal. We have submitted all ChIP-Seq data for the 139 TFs with replicate data to GEO. We have now included a table of GEO accessions for all submitted data (Table S11). These data are currently embargoed but will be released prior to publication. After publication, we will also make summary data available through RegulonDB. We have chosen not to release data for TFs for which we were able to generate only one ChIP-Seq data set, but not a high-quality replicate. Our experience suggests that in the absence of a replicate, single datasets are potentially erroneous and misleading. We are concerned that releasing these data may cause more harm than good.

7. ***Figure 2: improve the color scheme used, particularly to ensure that the same color is not used for different data. Be consistent with the color scheme throughout the paper.***

We have now updated the color scheme for different data and different data types (e.g., the type of experiments). And we have worked to ensure that this color scheme is used in all figures in both the Main Text and the Supplementary Material.

8. ***Figure 2: ordering of figure subparts is not standard.***

We have revised the ordering of figure subparts to be more standard.

9. ***Figure 2F,G,H,I: Clarify which transcription factors are being presented.***

We regret this confusion. All data in Figure 2 (B-J) is for PdhR. We have now modified the caption to be more explicit about this. And panels F-I (which have now been reordered per the previous comment) display different binding regions for PdhR. We have also clarified this in the caption.

10. ***Figure 2F,G,H,I: Why are different scales used in each panel?***

These panels display multiple binding sites for PdhR that differ by several orders of magnitude in coverage between sites and between different experiments for the same site. The peaks consequently also differ significantly in affinity score and predicted coverage. Our goal with these panels is to demonstrate the spatial accuracy of BoltzNet predictions. If the same scale was used throughout, the weaker sites and experiments, and the accuracy of the predictions, would be difficult if not impossible to discern. We have modified the panels to make the y axes labels larger. We have also modified the caption to indicate that peaks and experiments are plotted on different y-scales to facilitate visualizing prediction concordance within a site.

11. ***Figure 2F,G,H,I: Affinity scores of binding sites are usually ~2x greater than background. This is far less than the observed or predicted data would indicate. Why is that?***

This is indeed one of the key findings of the paper. As shown in Figure 6, very small binding energy differences can lead to very large changes in coverage. Also, by design, and as confirmed in Figure 5, the BoltzNet affinity score of a sequence is linearly related to the binding energy. The scaling factors of this linear relationship is such that differences in binding energy are roughly the same order of magnitude as the differences in affinity score. Thus, small changes in this score are also associated with large changes in coverage.

These finding addresses one of the key questions of ChIP-Seq: to what extent is weak binding "physiologically relevant". As discussed in the Main Text and in more detail in response to a comment below, these results indicate the weak binding differs from strong binding by ~ 4 k_bT which is an energy difference that can be realized by many cellular processes, including physiological binding to different sites by LacI.

We do not have a full explanation for why such small energy differences can lead to large changes in coverage. But we expect that a large part of the explanation lies with the Boltzmann distribution: at thermal equilibrium the probability of different states depends on the *exponentiation of energy differences* (Equation 5 of Main Text and see also Supplement). Similarly, for unimolecular model of TF binding to a DNA sequence (Equation 8), the binding affinity of this reaction is also related to the exponentiation of the energy difference (Equations 13 and 14). Finally, we have previously shown that ChIP-Seq coverage can be accurately modeled as a convolution process in which the magnitude of the input signal for a single binding site (and thus the magnitude of the impulse response) is proportional to the probability of binding to this site [5, 6]. Taken together, these observations suggest that it is not biophysically unexpected that small changes in energy (or BoltzNet affinity score) can give rise to large

changes in affinity, large changes in the probability of being bound, and large changes in ChIP-Seq coverage.

12. ***Figure 2J: Nine traces are visible but eleven are indicated by symbols. Complete overlaps? Needs to be clarified.***

We have clarified this by noting the near complete overlap in the figure citation.

13. ***Why is the weight matrix only 25 bp? While many transcription factor binding sites may be adequately defined by such, definitely not all.***

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 known RegulonDB binding sites are less than 25 bp in length. Moreover, for the three that have longer motifs, an inspection suggests that it contains more than one copy of a core 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. We have updated the Methods to described this.

In addition, as described in detail below, to address this concern we performed an analysis of the impact of different matrix lengths on model accuracy. One conclusion from this analysis is that, past a threshold, longer weight matrices provide no additional predictive value and can in fact decrease accuracy. As a uniform weight matrix length applied to all TFs, 25 bp thus balances the preference for a shorter matrix while also encompassing the information from known motifs. We also note that for TFs for which we could not build models (see the discussion below), increasing the matrix length did not enable models to be found, indicating the motif length was not the primary issue in these cases.

Finally, we note that the weight matrix length is a hyper-parameter than can be selected when a model is built. Thus, users of BoltzNet are free to select a different length or build models for a range of lengths.

14. ***Figure 4B,C,D: R2 all < 0.9. Why are these considered "highly correlated" in the text? (D) Why are the triangles bolded? (Legend) the text "predicted BoltzNet predictions" is a redundancy. "predicted" can be removed.***

We have removed the adjective "highly" and now instead describe the R^2 values. We have also removed "predicted". In keeping with the requirements of Nature Communications, we have now plotted individual data points rather than means and error bars. The legend for the figure now describes all the symbols used.

15. ***PFMs are not capable of predicting ChIP-Seq coverage" is a very strong statement, especially given that you only provide data for one transcription factor (pdhR). How true is this for all the transcription factors you have studied? Have you compared other methods for generating PFMs – XSTREME or MEME-ChIP?***

We agree that this is a strong statement. To support this statement, we have now included an analysis of MEME, XSTREME, and MEME-ChIP applied to ChIP-Seq data for all TFs with ChIP-Seq. All three tools generate models that perform poorly in predicting coverage for all TFs (Figure S26 – S31).

This is expected given the nature of PFMs. 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 match 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.

Different MEME based approaches differ only slightly on how motif sequences are identified [7]. 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 (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. MEME tools do not take coverage as part of training.

When applied with 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 binding affinity and/or coverage, by contrast, is a regression problem. BoltzNet is explicitly designed to solve this regression.

We have now added a section in Supplementary Material the includes this discussion.

16. ***Figure 5 provides a comparison of BLI experimental and BoltzNet predicted binding energies. But where is the original BLI data?***

We have now included the original BLI data as Figures S54-S58. And we have cited this figure in Main Text.

17. ***Figure 6: What do the different color traces correspond to? Figure is not referenced in the text.***

We regret the lack of clarity in this figure and the missing reference. We have now added legends for the color traces. And we have ensured that Figure 6 is appropriately referenced in the main text.

18. ***The importance of weak sites is a major theme throughout this manuscript. Transcription factor overexpression, extensive ChIP-Seq, and BoltzNet permit their identification. But are they physiologically relevant, as alluded? If [the] original ChIP-Seq data are made freely accessible, others can investigate this question.***

We agree on the importance of the question of the physiological relevance of commonly discovered weak sites. As described above, all our data will be made publicly and freely available to enable others to investigate this question.

We note that physiological relevance is a broad term that can have multiple meanings. We consider the physiological relevance of weak binding in two stages. First, are weak binding sites "physiologically realizable" in the sense that they might be occupied under physiological conditions. Second, if occupied, do weak binding sites exert a transcriptional, or other functional, effect. Our analysis cannot speak to the second aspect - this is a topic of future research. But we believe our results do speak to the first aspect.

One of the key findings of our paper is that for the 5 TFs we calibrated, weak binding sites differ in binding energy from the strongest sites by typically ~4 k_bT (Figure 6). By comparison, the operators of

LacI differ by an estimated span of 5.8 k_bT [8-10]. The difference in binding energy between strong and weak called sites is thus well within a physiologically relevant range. More generally, strong and weak sites exist on a continuum with a significant background affinity for the genome as a whole (Figure 6). Our analysis is consistent with prior results suggesting that this background affinity is strong enough that most TFs are bound to the genome under physiological conditions. Consequently, even weak called binding sites - with higher affinity than the genome average - have sufficient affinity to be occupied in principle under physiological conditions.

19. ***Methods, ChIP-Seq Pipeline: Not Figure S4A. Figure S5A & B?***

We thank the reviewer for catching this typo. It has now been corrected.

20. ***Some references (e.g. #15) are not complete.***

We have corrected the formatting of this citation.

Reviewer 3:

1. ***The manuscript does not present any comparison with existing methods that also aim to predict ChIP-Seq signatures, such as ProBound and BPNet.***

As described in our response to Reviewer 1, we have now included the results of comparing BoltzNet to existing methods including ProBound and BPNet.

2. ***In the convolutional layer, what approach is taken at the ends of 101bp sequence when 25bps are not available?***

No padding is used at the end of the input sequences. The output of the convolution is thus M positions shorter than the length of the (concatenated) input sequences, where M is the length of the motif. We have clarified this in the Methods and also in Figure 2A.

3. ***When the forward and reverse strands are concatenated, it seems that at the middle of the vector the filter is convolving paired bases, but at the ends of the vector it's convolving bases on a single strand - is this correct, and does it pose any obstacle to learning an informative weight matrix?***

This is correct. The weight matrix is applied across the junction of the concatenated sequences. This does not seem to have posed an obstacle for learning informative weight matrixes, as confirmed by our results.

We expect that several factors explain this. First, the input sequences from ChIP-Seq are centered on the point of maximum coverage, and thus the high affinity sequences are expected to be near the centers of the sequences. Second, given a 25 bp matrix, the number of positions that significantly overlap the junction are small compared to non-overlapping positions. Third, the sequences at the junction are expected to be random across different input sequence, while the true binding sites will be a consistent signal. The junction is thus expected to add a small amount of variance, but no bias, to the learning signal.
Nonetheless, we acknowledge that this aspect of the architecture of BoltzNet warrants further examination in future work.

4. ***The manuscript mentions a "circular shift permutation" to increase the model's generalization abilities, but it would be helpful to have supporting data to show that this adjustment did in fact improve the model's generalization.***

We have tested the impact of the circular shift permutation by building models of all 5 TFs calibrated with BLI without this data augmentation. In the cases of PdhR, AllR, GlnG, and Nac, data augmentation did not dramatically alter the model learned and the accuracy on the training set, genome, and predicted sequence energies. In contrast, for UlaR a model satisfying our accuracy requirements could not be found during training without augmentation (despite multiple build attempts). Thus, data augmentation was necessary for accurate training in at least one example. We show these results in Figure S64 and have modified the Methods section to describe these.

5. ***A few unsuccessful cases are mentioned where the model failed to achieve satisfactory performance for some TFs – do the authors have any reasoning or data to explain why the model broke down in these cases.***

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 no other intermediate peaks. The meta-analysis of models built with subsamples of called peaks 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 the typical behavior of 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 demonstrates that BoltzNet can build models for TFs that bind with broader peaks.

We have now included a supplementary table listing the TFs that failed to generate models (Table S3), and we have included a section in Supplementary Material with the discussion above.

6. ***Figures 3 and 6 have several panels with data grouped by color, but no legends for them.***

We regret this lack of clarity and have now added legends.

7. ***Equation 5 may have a missing negative sign in the exponent.***

We are grateful to the Reviewer for catching this typo. It has now been corrected.

8. ***On page 12, the phrase "the denominator in equation 4" appears to be a typo and I suspect should reference equation 5 instead.***

We are grateful to the Reviewer for catching this typo. We have corrected this.

9. ***More details are needed in the description of "grey symbols" in Figure 1G. Are they means and standard deviations or not? It will also be helpful to highlight the overrepresented bins.***

Grey circles with black error bars indicate the expected distributions of binding sites assuming random locations. Grey bars are means. We have now made these symbols larger which we believe makes it more apparent which bins differ from random expectation (both over and under).

Reviewer 4:

1. ***The authors should include description about how the 134 TFs are selected from the ~300 TFs for E. coli.***

We regret our lack of clarity on this point. As described in the response to Reviewer 1, we have now included this description.

2. ***It would be better if the authors show a quantitative replication correlation between Native and inducible replicates for all TFs with replicates, besides PdhR (Fig1E), so that it is clear which TFs are with high replication correlation and which do not (on top of Fig1F and FigS13).***

We have now provided a Supplementary Figure in with this analysis. Figure S14 replicates Figure 1F for every TF, and highlights the specific points derived from that TF. The points are colored to indicate the type of comparison.

3. ***Does Fig1H implicate most novel regions are weaker binding sites? Could it be due to differences in thresholds in peak calling/statistical region enrichment of the ChIP-seq data processing step between re-analyses of data from other studies and data from this study?***

The Reviewer is correct, the majority of novel regions are weaker sites. The set of known sites was derived from RegulonDB v12.0 [11] based on sites for 38 TFs with strong classical binding evidence, and not dependent on other high-throughput datasets (Figure S19). No ChIP-Seq data from any other study was used in our analysis. This selection was made to unsure that only high-confidence binding sites were considered. We regret the lack of clarity on this point and have noted this in our methods.

4. ***Major concern is model overfitting, since it is mapping a low dimension affinity score to scale up to genomic coverage. This imbalance of number of parameters and number of data points may cause loss of model stability, since the neural network weights can have many possibilities to fit the data (although L1 regularization was mentioned to put constraint on middle layer). Thus, ideally there shall be investigation in model stability. It seems the cross validation was only done for TFs with many binding sites detected.***

We agree with the Reviewer that the question of model stability is crucial. We have now included the results of a comprehensive analysis demonstrating the broad stability of BoltzNet models.

First, we have added the results of cross-validation for AllR, GlnG, Nac, and UlaR (Figures S21-24). The results confirm model generalization and the potential for extrapolation, as with PdhR. We note that along with PdhR, models for these TFs were also previously validated for independent predictive accuracy by comparing predicted binding energies for novel sequences with independent experimental measurement with BLI using purified protein and DNA (Figure 5).

Second, for the five TFs calibrated with BLI (AllR, GlnG, Nac, PdhR, UlaR) we analyzed the impact of varying model hyperparameters and input data. To do so, we built models in which we varied (a) weight matrix length, (b) neural network width, (c) sub-selection of called peaks, and (d) sub-selection experiments. By performing this analysis for the five calibrated TFs, accuracy could be independently benchmarked 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.

BoltzNet models were robust to the selection of hyperparameters and input data. We summarize the results of these analyses here. We have also included new Results subsections (“BoltzNet models robust to hyperparameters” and “BoltzNet models robust to input data”) with these results and have included a detailed description of these results in Supplementary Material and Supplementary figures with the results.

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 (Figures S43-S38).

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 (FIGURES). 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 (Figures S39-S43). 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.

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 (Figures S44-S48). For each subsample, only the top N peaks were provided as positive samples. 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. For the majority of cases, accurate models could be built with as few as the five strongest called peaks (PdhR, GlnG, AllR, 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.

Subsampling Experiments. To assess the impact of different experiments, we built models on subsets of available experiments (Figures S49-S52). The subsets were selected to span the different types of experiments available for each TF and included multiple subsets of only two *in vivo* experiments as would be the case for the majority of the 124 models we report.

Models for PdhR, AllR, GlnG, and Nac were insensitive to the selection of experiment subsets, with equivalent accuracy on the prediction of binding energy as compared to BLI measurements. 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 (we discuss the behavior of UlaR in Supplementary Material).

5. ***Based on Figure 1A and supplementary data, the number of binding sites is at maximum hundreds, and many TFs have binding sites below hundred. For those TFs, can the model reach high accuracy? It will be an imbalanced training case with less than hundreds of positive binding scores and 5000 control scores.***

As demonstrated by the analysis of subsampling called peaks described above, at least for TFs with highly specific binding preferences, accurate models could be built on as few as 3 called peaks and 5000 negative samples (with moderate accuracy possible in some cases with 2 called peaks). For TFs like Nac that bind a large number of sites, a larger number of positive samples appears necessary. But, of course, in these cases a larger number of samples will be generated by ChIP.

6. ***Describe the reason for selecting 140 nodes in the middle layer.***

We selected 140 nodes to ensure that the neural network layer would have sufficient expressive power to model the relationship between binding energy and coverage in multiple experiments. As demonstrate by the analysis above, BoltzNet models are highly robust to this parameter across a range of 5 to 2000 nodes.

As described in the new Supplementary Material section, the results indicate that in all cases an expected sigmoid function was learned without any evidence of overfitting (e.g., memorizing a map from affinity score to observed coverage). Most importantly, the number nodes in the neural network layers did not impact the energy weight matrix learned. This is 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.

7. ***Have the authors considered comparison with deep learning models of TF binding like DeepLift etc.,***

As described above in our responses to Reviewer 1, we have now included a detailed comparison of BoltzNet to other deep learning models of TF binding.

8. ***One interesting aspect of approaches used this study is the possible ability to reveal TF-DNA binding dynamics and weakly binding sites, which may be lacking when using top peaks called from ChIP-seq data. Should there be investigation about differences of the inferred motifs between strong/known binding sites and weakly binding sites?***

We agree with the reviewer that only cataloging the strongest sites from ChIP-Seq data provides an incomplete description of binding preference. We believe this is an important conclusion from our results. In addition, our results argue against classifying sites as strong or weak. Our result indicate that strong and weak sites exist on a spectrum of binding to all sequences in the genome. A complete model must thus predict binding affinity to any sequence. The binding energy matrix learned by BoltzNet is sufficient to explain the differences between strong and weak sites as simply reflecting the difference in binding energy associated with the underlying sequence.

Two patterns emerge, however, that are frequently associated with stronger or weaker binding, as describe in the main text. First, many sequences contain clusters of binding sites, and these clusters can sum to an overall high effective affinity to the sequence. Second, core sequences are commonly conserved across a range of binding strengths, with differences in binding strength associated with accessory bases outside this core. This is also apparent from weight matrices, where core bases have the strongest relative contributions, but bases outside the core can be equally important, presumably by stabilizing or destabilizing contacts with the core motif.

9. ***Page 2, line 2 contains extra comma.***

We believe we have corrected this.

10. ***Uniformly processed data and TF parameters can be deposited/stored in publicly available resources.***

We completely agree. And as described above, all raw data will be made publicly available in GEO and processed data will be made available through RegulonDB.

11. ***It is recommended to add dependency/installation information in README. Since the test data demonstrated in code base follow pre-fixed formats which is pulled from Galagan lab database, there shall be formatting functions or guidelines provided for other ChIP-seq data.***

We have now included two different yml files in our Code Ocean repository. One is an environment description for a machine without CUDA, and the other for a machine with CUDA.

We have now also updated the README file in the Code Ocean repository with a detailed description of the input data dictionary.

Additional Text Updates

To further clarify and improve our manuscript, we have made the following additional changes.

- To be explicit about units, we have updated the derivation of the relationship between the thermodynamic model of TF binding, the standard Hill function, and the associated equilibrium constants K_D and K_A in the Methods and Supplementary Material (Relationship to Hill Function and Equilibrium Constant Relation to Free Energy). In particular, we now explicitly show the relationship between unitless energy $\Delta\varepsilon$ (e.g., energy normalized to k_bT), Gibbs free energy (ΔG, units of J/mol), K_D (units of mol^{-1}), and K_A (units of mol).

- The models used in Figure 3, 5, and 6 for AllR, GlnG, Nac, and UlaR now include both *in vivo* and *in vitro* data. We made this change to enable subsampling of all available data for the stability analysis requested by Reviewer 4. It also improves consistency with the PdhR model (which includes both *in vitro* and *in vivo* data)

- When reviewing the models that we report, we noticed that FhlA had been incorrectly counted. We have thus updated the number of models to 124 throughout instead of 125 (Figure 1 and Figure S11 did not include this error and are unchanged).

References

1. Avsec, Z., et al., *Base-resolution models of transcription-factor binding reveal soft motif syntax.* Nat Genet, 2021. **53**(3): p. 354-366.
2. 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.
3. Lundberg, S., *A unified approach to interpreting model predictions.* arXiv preprint arXiv:1705.07874, 2017.
4. Keilwagen, J., S. Posch, and J. Grau, *Accurate prediction of cell type-specific transcription factor binding.* Genome Biology, 2019. **20**(1): p. 9.
5. Gomes, A.L., et al., *Decoding ChIP-seq with a double-binding signal refines binding peaks to single-nucleotides and predicts cooperative interaction.* Genome Res, 2014.
6. Lun, D.S., et al., *A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data.* Genome Biol, 2009. **10**(12): p. R142.
7. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
8. Garcia, H.G. and R. Phillips, *Quantitative dissection of the simple repression input-output function.* Proc Natl Acad Sci U S A, 2011. **108**(29): p. 12173-8.
9. 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.
10. Kuhlman, T., et al., *Combinatorial transcriptional control of the lactose operon of Escherichia coli.* Proc Natl Acad Sci U S A, 2007. **104**(14): p. 6043-8.
11. 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.

Author's Response to the Reviewer's Comments

We remain grateful to all the reviewers for their thorough and careful consideration of our paper and for their helpful comments. We are gratified that our responses addressed their concerns and questions. The process has significantly strengthened our manuscript. We respond below to a question raised by Reviewer 4.

Reviewer 4: "Can K_D and [TF]_eq be inferred and interpreted by the trained models for each TF, generating reusable biophysics parameters"

We are grateful for this question. BoltzNet does indeed infer a biophysical model of TF-DNA binding energy which we demonstrate can be independently measured and confirmed using BLI measurements on purified protein and DNA for multiple TFs. In this respect, and by design, BoltzNet is learning a "reusable" biophysical parameter that has intrinsic physical meaning.