Biophysical models of cis-regulation as interpretable neural networks

Ammar Tareen
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724
tareen@cshl.edu

Justin B. Kinney
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724
jkinney@cshl.edu

Abstract

The adoption of deep learning techniques in genomics has been hindered by the difficulty of mechanistically interpreting the models that these techniques produce. In recent years, a variety of post-hoc attribution methods have been proposed for addressing this neural network interpretability problem in the context of gene regulation. Here we describe a complementary way to address this problem. Our approach is based on the observation that two large classes of biophysical models for cis-regulatory mechanisms can be expressed as deep neural networks in which nodes and weights have explicit physiochemical interpretations. We also demonstrate how such biophysical networks can be rapidly learned, using modern deep learning frameworks, from the data produced by certain types of massively parallel reporter assays (MPRAs). These results suggest a scalable strategy for using MPRAs to systematically characterize the biophysical basis of gene regulation in a wide range of biological contexts. They also highlight gene regulation as an ideal venue for the development of scientifically interpretable approaches to deep learning.

Deep learning – the use of large multi-layer neural networks in machine learning applications – is revolutionizing information technology [1]. There is currently a great deal of interest in applying deep learning techniques to problems in genomics, especially for understanding gene regulation [2–6]. These applications of deep learning remain somewhat controversial, however, due to the difficulty of mechanistically interpreting trained neural networks.

Multiple strategies have been proposed for addressing this neural network interpretability problem [7–13]. A common feature of these attribution strategies is that they seek to extract meaning post-hoc from neural networks that have arbitrary architectures. However, there remains a substantial gap between the outputs of these attribution methods and fully mechanistic models of gene regulation.

Here we advocate for a complementary approach: the inference of neural network models whose architecture reflects explicit biophysical hypotheses for how a cis-regulatory sequence of interest might work. This strategy is based on two key observations. First, explicit biophysical models can be naturally formulated as deep neural networks in which nodes and weights have explicit physiochemical interpretations. This is true of standard thermodynamic models (which rely on a quasi-equilibrium assumption) [14–19] as well as fully kinetic models [20–22], and requires no mathematical approximations (c.f. [23, 24]). Second, existing deep learning frameworks are able to rapidly infer such models from the data produced by certain classes of massively parallel reporter assays (MPRAs).

Thermodynamic models are specified by a set of molecular complexes, or "states", which we index using s. Each state has both a Gibbs free energy ΔG_s and an associated activity α_s. These energies determine the probability P_s of each state occurring in thermodynamic equilibrium via the Boltzmann distribution,[1]

$$P_s = \frac{e^{-\Delta G_s}}{\sum_{s'} e^{-\Delta G_{s'}}}. \tag{1}$$

[1]To reduce notational burden, all ΔG values are assume to be in thermal units. At 37°C, one thermal unit is $1\,k_B T = 0.62$ kcal/mol, where k_B is Boltzmann's constant and T is temperature.

1st Conference on Machine Learning in Computational Biology, (MLCB 2019), Vancouver, Canada.

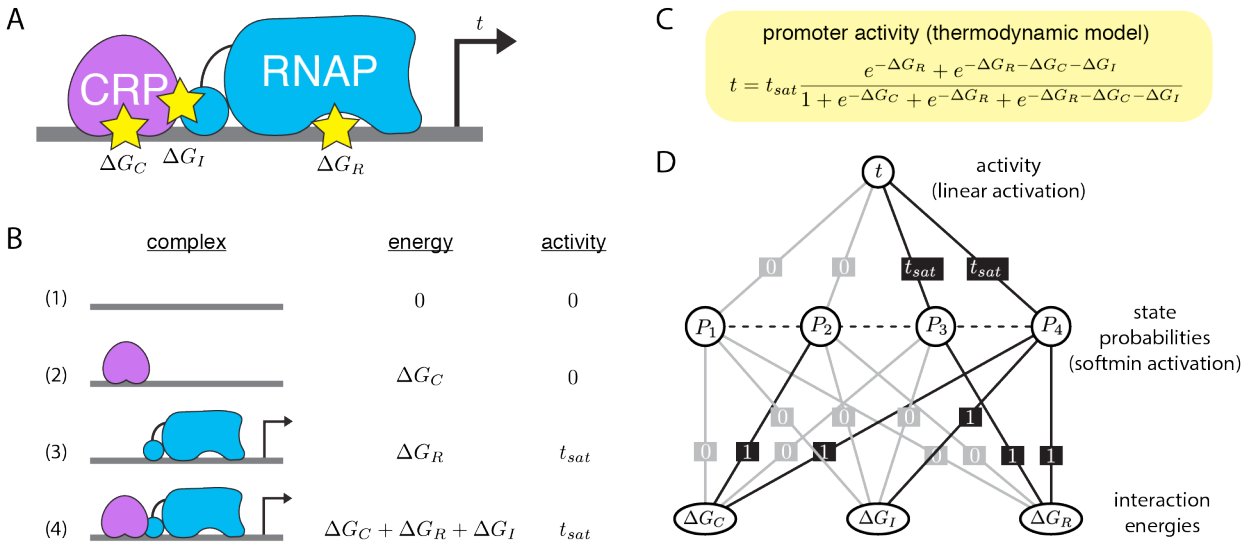

Figure 1: A thermodynamic model of transcriptional regulation (A) Transcriptional activation at the *E. coli lac* promoter is regulated by two proteins, CRP and σ^{70} RNA polymerase (RNAP). CRP is a transcriptional activator that up-regulates transcription by stabilizing RNAP-DNA binding. ΔG_C and ΔG_R respectively denote the Gibbs free energies of the CRP-DNA and RNAP-DNA interactions, while ΔG_I denotes the Gibbs free energy of interaction between CRP and RNAP. (B) Like all thermodynamic models of gene regulation, this model consists of a set of states, each state having an associated Gibbs free energy and activity. The probability of each state is assumed to follow the Boltzmann distribution. (C) The corresponding activity predicted by such thermodynamic models is the state-specific activity averaged together using these Boltzmann probabilities. (D) This thermodynamic model formulated as a neural network. First layer nodes present interaction energies, second layer nodes represent state probabilities, and third layer nodes represent activity. The values of weights are indicated; gray lines correspond to zero weights. The second layer has a softmin activation, while the third has a linear activation.

The energy of each state is, in turn, computed using integral combinations of the individual interaction energies ΔG_j that occur in that state. We can therefore write $\Delta G_s = \sum_j \omega_{sj} \Delta G_j$, where ω_{sj} is the number of times that interaction j occurs in state s. The resulting activity predicted by the model is given by the activities α_s of the individual states averaged over this distribution, i.e., $t = \sum_s \alpha_s P_s$.

Fig. 1 illustrates a thermodynamic model for transcriptional activation at the *E. coli lac* promoter. This model involves two proteins, CRP and RNAP, as well as three interaction energies: ΔG_C, ΔG_R, and ΔG_I. The rate of transcription t is further assumed to be proportional to the fraction of time that RNAP is bound to DNA (Fig. 1A). This model is summarized by four different states, two of which lead to transcription and two of which do not (Fig. 1B). Fig. 1C shows the resulting formula for t in terms of model parameters. This model is readily formulated as a feed-forward neural network (Fig. 1D). Indeed, all thermodynamic models of cis-regulation can be formulated as three-layer neural networks: layer 1 represents molecular interaction energies, layer 2 (which uses a softmin activation) represents state probabilities, and layer 3 (using linear activation) represents the biological activity of interest, which in this case is transcription rate.

We can infer thermodynamic models like these for a cis-regulatory sequence of interest (the wild-type sequence) from the data produced by a massively parallel reporter assay (MPRA) performed on an appropriate sequence library [25]. Indeed, a number of MPRAs have been performed with this explicit purpose in mind [25, 27–30]. To this end, such MPRAs are generally performed using libraries that consist of sequence variants that differ from the wild-type sequence by a small number of single nucleotide polymorphisms (SNPs). The key modeling assumption that motivates using libraries of this form is that the assayed sequence variants will form the same molecular complexes as the wild-type sequence, but with Gibbs free energies and state activities whose values vary from sequence to sequence. By contrast, variant libraries that contain insertions, deletions, or large regions of random DNA are unlikely to satisfy this modeling assumption.

Fig. 2A summarizes the sort-seq MPRA described in [25]. *Lac* promoter variants were used to drive GFP expression in *E. coli*. GFP-expressing cells were then sorted into 10 bins using fluorescence-activated cell sorting, after which variant

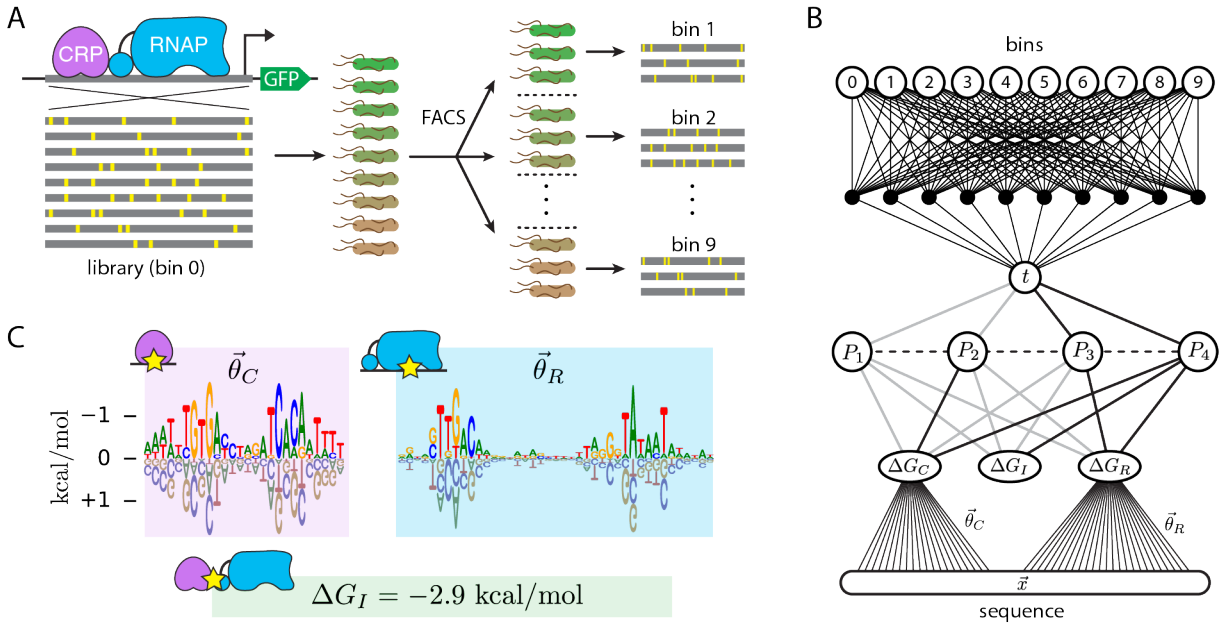

Figure 2: Inference of a thermodynamic model from MPRA data. (A) Schematic of the sort-seq MPRA of [25]. A 75bp region of the *E. coli lac* promoter was mutagenized at 12% per nucleotide. Variant promoters were then used to drive the expression of GFP. Cells carrying these expression constructs were then sorted using FACS, and the variant sequences in each bin were sequenced. This yielded data on about 50K variant promoters across 10 bins. (B) The neural network from Fig. 1D, but with ΔG_C and ΔG_R expressed as linear functions of the DNA sequence $\vec{x}$, as well as a dense feed-forward network mapping activity t to bins via a probability distribution $p(\text{bin}|t)$. Gray lines indicate weights fixed at 0. The weights in the second and third hidden layers have additional, hardcoded, constraints. (C) The parameter values inferred for the CRP energy matrix $\vec{\theta}_C$, the RNAP energy matrix $\vec{\theta}_R$, and the CRP-RNAP interaction energy ΔG_I. Since increasingly negative energy corresponds to stronger binding, they y-axis in the logo plots is inverted. Logos were generated using Logomaker [26].

promoters within each bin were sequenced, yielding about 50K sequences in total. The authors then fit the biophysical model shown in Fig. 1C, but under the assumption that $\Delta G_C = \vec{\theta}_C \cdot \vec{x} + \mu_C$ and $\Delta G_R = \vec{\theta}_R \cdot \vec{x} + \mu_R$, where $\vec{x}$ is a one-hot encoding of promoter DNA sequence.

For this study, we used TensorFlow to infer the same model formulated as a deep neural network. Specifically, we embellished the network in Fig. 1D by using same sequence-dependence for ΔG_C and ΔG_R as in [25]. To link t to the MPRA measurements, we introduced a feed-forward network with one hidden layer (having softmin activation) and a softmin output layer corresponding to the 10 bins into which cells were sorted. All model parameters were fit to the MPRA dataset using stochastic gradient descent and early stopping. The results agreed well with those reported in [25]. In particular, the parameters in the energy matrices for CRP ($\vec{\theta}_C$) and RNAP ($\vec{\theta}_R$) exhibited Pearson correlation coefficients of 0.986 and 0.994, respectively, with those reported in [25]. The protein-protein interaction energy that we found, $\Delta G_I = -2.9$ kcal/mol, was also compatible with the previously reported value $\Delta G_I = -3.3 \pm 0.4$ kcal/mol.

A major difference between our results and those of [25] is the ease with which they were obtained. Training of the network in Fig. 2B consistently took about 15 minutes on a standard laptop computer. The model fitting procedure in [25], by contrast, relied on a custom Parallel Tempering Monte Carlo algorithm that took about a week to run on a multi-node computer cluster (personal communication), and more recent efforts to train biophysical models on MPRA data have encountered similar computational bottlenecks [29, 30].

Also of note is the fact that in [25] the authors inferred models using information maximization. Specifically, the authors fit the parameters of $t(\vec{x})$ by maximizing the mutual information between model predictions and observed bins $I[t, \text{bin}]$. One practical difficulty with this strategy is the need to estimate mutual information from finite data. Instead, we used maximum likelihood to infer the parameters of $t(\vec{x})$ as well as the experimental transfer function (i.e., noise model) $p(\text{bin}|t)$, which was modeled by a dense feed-forward network with one hidden layer. These two inference methods, however, are essentially equivalent: in the large data regime, the parameters of t that maximize $I[t, \text{bin}]$ are the same as the parameters one obtains when maximizing likelihood over the parameters of *both* t and $p(\text{bin}|t)$; see [31, 32].

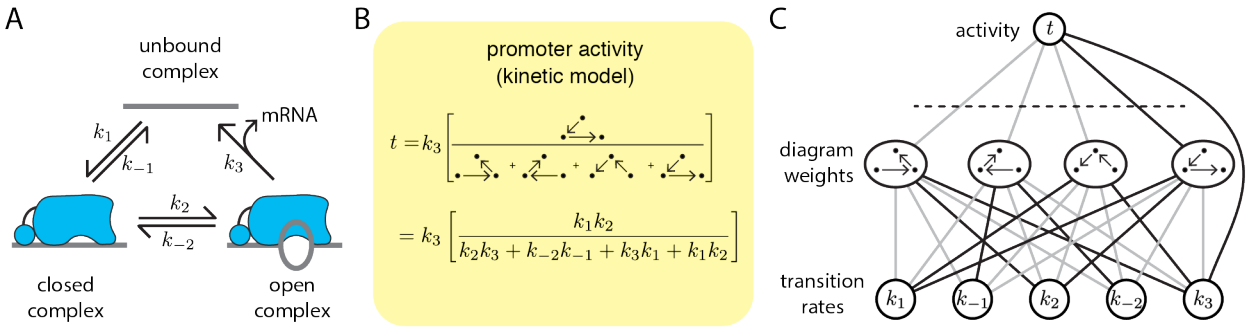

Figure 3: A kinetic model for transcript initiation by *E. coli* RNAP. (A) In this model, promoter DNA can participate in three complexes: unbound, closed, and open. Transitions between these complexes are governed by rate constants k. (B) A formula for the steady-state rate of mRNA production can be obtained using King-Altman diagrams [33, 34]. (C) This formula is naturally represented using the three-layer neural network, where layer 1 represents transition rates, layer 2 represents the weights of distinct King-Altman diagrams, and layer 3 represents promoter activity. Here, black lines indicate weight 1, gray lines indicate weight 0, all activations are log-linear, and the dashed line represents L_1 normalization of layer 2 output.

A shortcoming of thermodynamic models is that they ignore non-equilibrium processes. Kinetic models address this problem, providing a fully non-equilibrium characterization of steady state activity. Kinetic models are specified by listing explicit state-to-state transition rates rather than Gibbs free energies. An example is shown in Fig. 3. Fig. 3A shows a three-state kinetic model of transcript initiation consisting of unbound promoter DNA, an RNAP-DNA complex in the closed conformation, and the RNAP-DNA complex in the open conformation. The rate k_3 going from the open to the unbound states is the rate corresponding to transcript initiation. The resulting transcription rate in steady state is therefore k_3 times the occupancy of the open complex.

King-Altman diagrams [33, 34], a technique from mathematical enzymology, provide a straight-forward way to compute the steady-state occupancy of any individual state. Specifically, each state's occupancy is proportional to the sum of spanning trees that flow to that state, where each spanning tree's value is given by the product of rates comprising that tree. Fig. 3B illustrates this procedure for the kinetic model in Fig. 3A.

Here we have shown how thermodynamic and kinetic models of transcriptional regulation can be formulated as deep neural networks in which nodes and weights have explicit physiochemical meaning. We have further demonstrated how a thermodynamic model can be rapidly inferred from MPRA data using existing deep learning frameworks. These results suggest a new strategy for interpretable deep learning in the study of gene regulation, one complementary to existing post-hoc attribution methods.

Our approach can be applied to a wide variety of gene regulatory systems in both prokaryotes and eukaryotes. We demonstrated this approach in the context of a well-characterized bacterial promoter in order to avoid any uncertainty about the underlying biology, and because previous quantitative studies have established concrete results against which we could compare our inferred model. The same modeling approach, however, should be readily applicable to a wide variety of biological systems, including transcriptional regulation and alternative mRNA splicing in higher eukaryotes.

Some MPRA datasets are more amenable to this modeling strategy than others. Our approach is best-suited to mutagenesis studies of individual cis-regulatory sequences, rather than genome-wide MPRA datasets. By assaying SNP-containing variants of a specific cis-regulatory sequence, one preserves both the overall location of regulatory protein binding sites as well as the resulting protein-protein interactions. This greatly simplifies the biophysical models that one needs to infer. And indeed, such MPRA data have already been generated in a variety of eukaryotic systems [35, 36, 27, 37]. The trade-off is that the resulting neural network model can only be expected to work in a narrow region of sequence space.

It may eventually be possible to relax this restriction while retaining biophysical interpretability via the use of convolutional or recurrent neural networks. Indeed, [24] has recently explored the possibility of using recurrent neural networks for biophysically modeling gene expression in flies. In principle, it should be possible to apply similar strategies to MPRAs performed on genome-wide or random sequence libraries [38, 4, 39, 40, 6]. But across all of biology, very few individual cis-regulatory sequences have been characterized to the level that our modeling strategy aims to elucidate. We therefore suggest that, at least in the short term, this approach should be used for modeling individual cis-regulatory sequences rather than genome-wide cis-regulatory codes.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[2] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature Methods*, vol. 12, pp. 931–934, Aug. 2015.

[3] A. Rosenberg, R. Patwardhan, J. Shendure, and G. Seelig, "Learning the sequence determinants of alternative splicing from millions of random sequences," *Cell*, vol. 163, no. 3, pp. 698 – 711, 2015.

[4] J. T. Cuperus, B. Groves, A. Kuchina, A. B. Rosenberg, N. Jojic, S. Fields, and G. Seelig, "Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences.," *Genome Res*, vol. 27, pp. 2015–2024, Dec. 2017.

[5] K. Jaganathan, S. K. Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, and K. K.-H. Farh, "Predicting splicing from primary sequence with deep learning," *Cell*, vol. 176, no. 3, pp. 535 – 548.e24, 2019.

[6] P. J. Sample, B. Wang, D. W. Reid, V. Presnyak, I. J. McFadyen, D. R. Morris, and G. Seelig, "Human 5' UTR design and variant effect prediction from a massively parallel translation assay.," *Nat Biotechnol*, vol. 37, pp. 803–809, July 2019.

[7] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of dna- and rna-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, pp. 831–838, Jul 2015.

[8] B. Sundararaman, L. Zhan, S. Blue, R. Stanton, K. Elkins, S. Olson, X. Wei, E. VanNostrand, G. Pratt, S. Huelga, B. Smalec, X. Wang, E. Hong, J. Davidson, E. Lécuyer, B. Graveley, and G. Yeo, "Resources for the comprehensive discovery of functional rna elements," *Molecular Cell*, vol. 61, no. 6, pp. 903 – 913, 2016.

[9] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," *arXiv e-prints*, Apr 2017.

[10] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian, "Neural Network Attributions: A Causal Perspective," *arXiv e-prints*, Feb 2019.

[11] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv e-prints*, Dec 2013.

[12] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: new computational modelling techniques for genomics," *Nature Reviews Genetics*, vol. 20, no. 7, pp. 389–403, 2019.

[13] P. K. Koo, P. Anand, S. B. Paul, and S. R. Eddy, "Inferring sequence-structure preferences of rna-binding proteins with convolutional residual networks," *bioRxiv*, 2018.

[14] G. Ackers, A. Johnson, and M. Shea, "Quantitative model for gene regulation by lambda phage repressor," *Proc Natl Acad Sci USA*, vol. 79, pp. 1129–1133, Feb. 1982.

[15] M. A. Shea and G. K. Ackers, "The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation.," *J Mol Biol*, vol. 181, pp. 211–230, Jan. 1985.

[16] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips, "Transcriptional regulation by the numbers: applications.," *Curr Opin Genet Dev*, vol. 15, pp. 125–135, Apr. 2005.

[17] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, "Transcriptional regulation by the numbers: models.," *Curr Opin Genet Dev*, vol. 15, pp. 116–124, Apr. 2005.

[18] E. Segal and J. Widom, "From DNA sequence to transcriptional behaviour: a quantitative approach.," *Nat Rev Genet*, vol. 10, pp. 443–456, July 2009.

[19] M. S. Sherman and B. A. Cohen, "Thermodynamic state ensemble models of cis-regulation.," *PLoS Comput Biol*, vol. 8, no. 3, p. e1002407, 2012.

[20] J. Estrada, F. Wong, A. DePace, and J. Gunawardena, "Information integration and energy expenditure in gene regulation.," *Cell*, vol. 166, pp. 234–244, June 2016.

[21] C. Scholes, A. H. DePace, and A. Sanchez, "Combinatorial Gene Regulation through Kinetic Control of the Transcription Cycle.," *Cell Syst*, vol. 4, pp. 97–108.e9, Jan. 2017.

[22] J. Park, J. Estrada, G. Johnson, B. J. Vincent, C. Ricci-Tam, M. D. Bragdon, Y. Shulgina, A. Cha, Z. Wunderlich, J. Gunawardena, and A. H. DePace, "Dissecting the sharp response of a canonical developmental enhancer reveals multiple sources of cooperativity.," *eLife*, vol. 8, p. 2787, June 2019.

[23] E. Mjolsness, "On cooperative quasi-equilibrium models of transcriptional regulation," *Journal of Bioinformatics and Computational Biology*, vol. 05, no. 02b, pp. 467–490, 2007.

[24] Y. Liu, K. Barr, and J. Reinitz, "Fully Interpretable Deep Learning Model of Transcriptional Control," *bioRxiv*, vol. 538, pp. 20–11, May 2019.

[25] J. B. Kinney, A. Murugan, C. G. Callan, and E. C. Cox, "Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence.," *Proc Natl Acad Sci USA*, vol. 107, pp. 9158–9163, May 2010.

[26] A. Tareen and J. B. Kinney, "Logomaker: Beautiful sequence logos in python," *bioRxiv*, vol. doi: http://dx.doi.org/10.1101/635029., May 2019.

[27] A. Melnikov, A. Murugan, X. Zhang, T. Tesileanu, L. Wang, P. Rogov, S. Feizi, A. Gnirke, C. G. Callan, J. B. Kinney, M. Kellis, E. S. Lander, and T. S. Mikkelsen, "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.," *Nat Biotechnol*, vol. 30, pp. 271–277, Feb. 2012.

[28] M. Razo-Mejia, J. Q. Boedicker, D. Jones, A. DeLuna, J. B. Kinney, and R. Phillips, "Comparison of the theoretical and real-world evolutionary potential of a genetic circuit.," *Phys Biol*, vol. 11, p. 026005, Apr. 2014.

[29] N. M. Belliveau, S. L. Barnes, W. T. Ireland, D. L. Jones, M. J. Sweredoski, A. Moradian, S. Hess, J. B. Kinney, and R. Phillips, "Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria.," *Proc Natl Acad Sci USA*, vol. 115, pp. E4796–E4805, May 2018.

[30] S. L. Barnes, N. M. Belliveau, W. T. Ireland, J. B. Kinney, and R. Phillips, "Mapping DNA sequence to transcription factor binding energy in vivo.," *PLoS Comput Biol*, vol. 15, p. e1006226, Feb. 2019.

[31] J. B. Kinney and G. S. Atwal, "Parametric inference in the large data limit using maximally informative models.," *Neural Comput*, vol. 26, pp. 637–653, Apr. 2014.

[32] G. S. Atwal and J. B. Kinney, "Learning quantitative sequence–function relationships from massively parallel experiments," *J Stat Phys*, vol. 162, no. 5, pp. 1203–1243, 2016.

[33] E. King and C. Altman, "A schematic method of deriving the rate laws for enzyme-catalyzed reactions," *The Journal of physical chemistry*, vol. 60, no. 10, pp. 1375–1378, 1956.

[34] T. L. Hill, *Free Energy Transduction and Biochemical Cycle Kinetics*. New York: Springer-Verlag, 1989.

[35] I. Liachko, R. A. Youngblood, U. Keich, and M. J. Dunham, "High-resolution mapping, characterization, and optimization of autonomously replicating sequences in yeast.," *Genome Res*, vol. 23, pp. 698–704, Apr. 2013.

[36] J. C. Kwasnieski, I. Mogno, C. A. Myers, J. C. Corbo, and B. A. Cohen, "Complex effects of nucleotide variants in a mammalian cis-regulatory element.," *Proc Natl Acad Sci USA*, vol. 109, pp. 19498–19503, Nov. 2012.

[37] S. Ke, V. Anquetil, J. R. Zamalloa, A. Maity, A. Yang, M. A. Arias, S. Kalachikov, J. J. Russo, J. Ju, and L. A. Chasin, "Saturation mutagenesis reveals manifold determinants of exon definition.," *Genome Res*, vol. 28, pp. 11–24, Jan. 2018.

[38] A. B. Rosenberg, R. P. Patwardhan, J. Shendure, and G. Seelig, "Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences," *Cell*, vol. 163, pp. 698–711, Oct. 2015.

[39] C. de Boer, R. Sadeh, N. Friedman, and A. Regev, "Deciphering cis-regulatory logic with 100 million synthetic promoters," *bioRxiv*, vol. doi: http://dx.doi.org/10.1101/224907, Nov. 2017.

[40] N. Bogard, J. Linder, A. B. Rosenberg, and G. Seelig, "A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation," *Cell*, pp. 1–43, June 2019.