bioRxiv preprint doi: https://doi.org/10.1101/835942; this version posted February 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 of approaching this problem. Our strategy is based on the observation that two large classes of biophysical models of 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 inferred, 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 a promising 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 [2–5]. These applications remain somewhat controversial, however, due to the difficulty of mechanistically interpreting neural network models trained on functional genomics data. Multiple attribution strategies, which seek to extract meaning post-hoc from neural networks that have rather generic architectures, have been proposed for addressing this interpretability problem [6–8]. However, there remains a substantial gap between the outputs of such 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 cis-regulatory sequences function. This strategy is based on two key observations. First, two large classes of biophysical models can be formulated as three-layer neural networks that have a stereotyped form and in which nodes and weights have explicit physiochemical interpretations. This is true of thermodynamic models, which rely on a quasi-equilibrium assumption [9–14], as well as kinetic models, which are more complex but do not make such assumptions [15–17]. Second, existing deep learning frameworks such as TensorFlow [18] are able to rapidly infer such models from the data produced by certain classes of MPRAs.

The idea of using neural networks to model the biophysics of gene regulation goes back to at least [19]. To our knowledge, however, the fact that all thermodynamic models of gene regulation can be represented by a simple three-layer architecture has not been previously reported. We are not aware of any prior work that uses neural networks to represent kinetic models or models involving King-Altman diagrams. There is a growing literature on modeling MPRA data using neural networks [20–24]. However, such modeling has yet to be advocated for MPRAs that dissect the mechanisms of single cis-regulatory sequences, such as in [25–29], which is our focus here. Finally, some recent work has used deep learning frameworks to infer parametric models of cis-regulation [30, 31]. These studies did not, however, infer the type of explicit biophysical quantities, such as ΔG values, that our approach recovers.

All thermodynamic models of cis-regulation can be represented as three-layer neural networks as follows. First one defines a set of molecular complexes, or "states", which we index using s. Each state has both a Gibbs free energy ΔG_s

Presented at the 14th conference on Machine Learning in Computational Biology (MLCB 2019), Vancouver, Canada.

bioRxiv preprint doi: https://doi.org/10.1101/835942; this version posted February 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

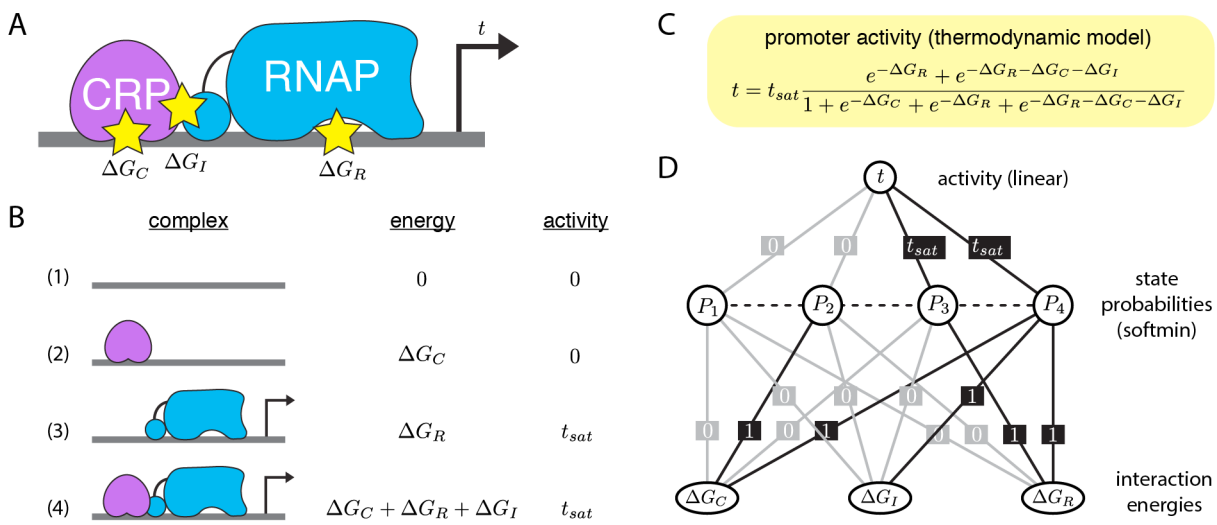

Figure 1: A thermodynamic model of transcriptional regulation. (A) Transcriptional activation at the *E. coli lac* promoter is regulated by two proteins, CRP and σ^{70} RNA polymerase (RNAP). CRP is a transcriptional activator that up-regulates transcription by stabilizing RNAP-DNA binding. ΔG_C and 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 model formulated as a three-layer neural network. First layer nodes represent interaction energies, second layer nodes represent state probabilities, and third layer nodes represent transcriptional 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. All thermodynamic models of cis-regulation can be represented using this general three-layer form.

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]

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

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 an MPRA performed on an appropriate sequence library [26]. Indeed, a number of MPRAs have been performed with this explicit purpose in mind [26, 27, 33–35]. 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

[1]To reduce notational burden, all ΔG values are assumed 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.

bioRxiv preprint doi: https://doi.org/10.1101/835942; this version posted February 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

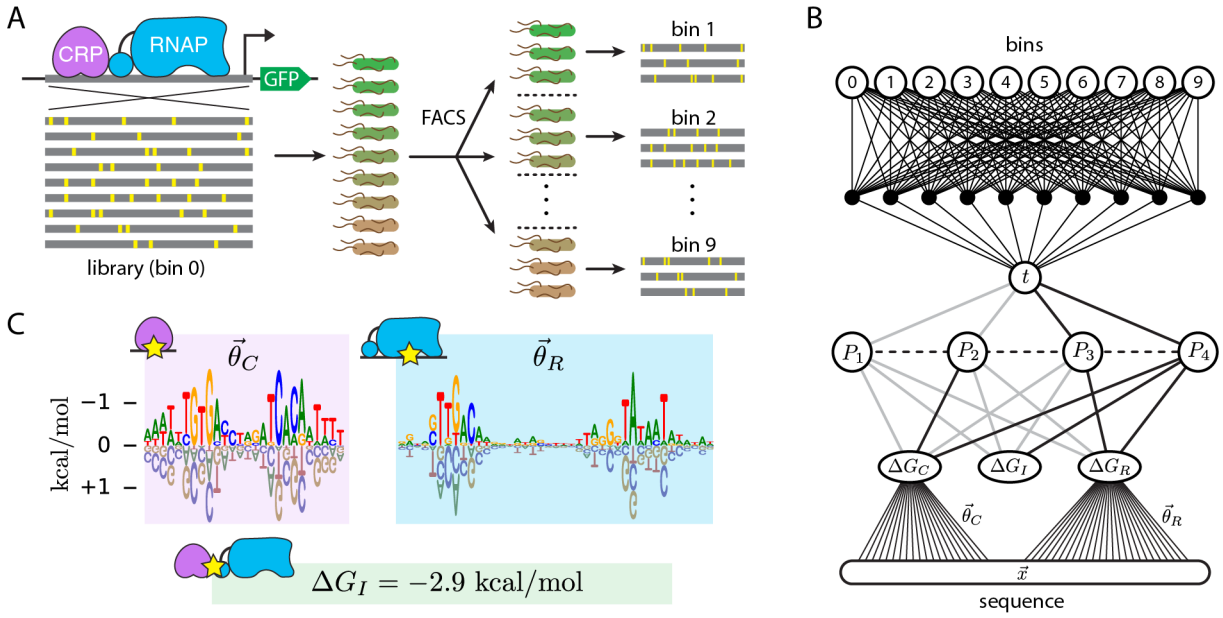

Figure 2: Inference of a thermodynamic model from MPRA data. (A) Schematic of the sort-seq MPRA of [26]. A 75 bp region of the *E. coli lac* promoter was mutagenized at 12% per nucleotide. Variant promoters were then used to drive the expression of GFP. Cells carrying these expression constructs were then sorted using FACS, and the variant sequences in each bin were sequenced. This yielded data on about 5×10^4 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 linking nodes P_3 and P_4 to node t were constrained to have the same value t_{sat}. (C) The parameter values inferred from the MPRA data of [26]. Shown are 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 [32].

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 (e.g. [20, 21, 23, 24, 31]) are unlikely to satisfy this modeling assumption.

Fig. 2A summarizes the sort-seq MPRA described in [26]. *Lac* promoter variants were used to drive GFP expression in *E. coli*, cells were sorted into 10 bins using fluorescence-activated cell sorting, and the variant promoters within each bin were sequenced. This yielded data comprising about 5×10^4 variant *lac* promoter sequences, each associated with one of 10 bins. 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.

Here we used TensorFlow to infer the same model formulated as a deep neural network. Specifically, we augmented the network in Fig. 1D by making ΔG_C and ΔG_R sequence-dependent as in [26]. To link t to the MPRA measurements, we introduced a feed-forward network with one hidden layer and a softmax output layer corresponding to the 10 bins into which cells were sorted. Model parameters were then fit to the MPRA dataset using stochastic gradient descent and early stopping. The results agreed well with those reported in [26]. In particular, the parameters in the energy matrices for CRP ($\vec{\theta}_C$) and RNAP ($\vec{\theta}_R$) exhibited respective Pearson correlation coefficients of 0.986 and 0.994 with those reported in [26]. 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 [26] 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 [26], 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 [34, 35].

bioRxiv preprint doi: https://doi.org/10.1101/835942; this version posted February 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

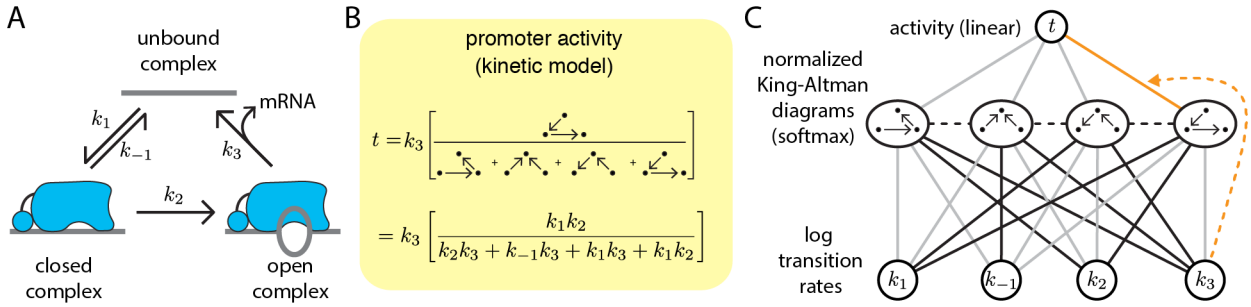

Figure 3: A kinetic model for transcriptional initiation by *E. coli* RNAP. (A) In this model, promoter DNA can participate in three complexes: unbound, closed, and open [39, 40]. Transitions between these complexes are governed by four rate constants: k_1, k_{-1}, k_2, and k_3. (B) A formula for the steady-state rate of mRNA production can be obtained using King-Altman diagrams [41, 42]. (C) This formula can be represented using the three-layer neural network shown, where layer 1 represents log transition rates, layer 2 (softmax activation) represents normalized King-Altman diagrams, and layer 3 (linear activation) represents promoter activity. Black lines indicate weight 1; gray lines indicate weight 0. Note that the single nonzero weight connecting layer 2 to layer 3 (orange) is actually the transition rate k_3 from layer 1. All kinetic models of cis-regulation share this general three-layer form.

Also of note is the fact that in [26] the authors inferred models using information maximization. Specifically, the authors fit the parameters of $t(\vec{x})$ by maximizing the mutual information $I[t; \mathrm{bin}]$ between model predictions and observed bins. One difficulty with this strategy is the need to estimate mutual information. Instead, we used maximum likelihood to infer the parameters of $t(\vec{x})$ as well as the experimental transfer function (i.e., noise model) $p(\mathrm{bin}|t)$, 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 $I[t; \mathrm{bin}]$ are the same as the parameters one obtains when maximizing likelihood over the parameters of *both* t and $p(\mathrm{bin}|t)$; see [36–38].

A shortcoming of thermodynamic models is that they ignore non-equilibrium processes. Kinetic models address this problem by 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. Fig. 3A shows a three-state kinetic model of transcriptional initiation consisting of unbound promoter DNA, an RNAP-DNA complex in the closed conformation, and the RNAP-DNA complex in the open conformation [39, 40]. The rate k_3 going from the open state to the unbound state represents transcript initiation. The overall transcription rate in steady state is therefore k_3 times the occupancy of the open complex.

King-Altman diagrams [41, 42], a technique from mathematical enzymology, provide a straight-forward way to compute steady-state occupancy in kinetic models. Specifically, each state's occupancy is proportional to the sum of directed spanning trees (a.k.a. King-Altman diagrams) 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. Every such kinetic model can be represented by a three-layer neural network (e.g., Fig. 3C) in which first layer nodes represent log transition rates, second layer nodes (after a softmax activation) represent normalized King-Altman diagrams, and third layer nodes represent the activities of interest.

Here we have shown how both thermodynamic and kinetic models of gene regulation can be formulated as three-layer deep neural networks in which nodes and weights have explicit biophysical meaning. This represents a new strategy for interpretable deep learning in the study of gene regulation, one complementary to existing post-hoc attribution methods. We have further demonstrated that a neural-network-based thermodynamic model can be rapidly inferred from MPRA data using TensorFlow. This was done in the context of a well-characterized bacterial promoter because previous studies of this system 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 amenable to MPRAs, including transcriptional regulation and alternative mRNA splicing in higher eukaryotes.

Code Availability and Acknowledgements: The neural network model shown in Fig. 2C, as well as the scripts used to infer it from the data of [26], are available at `https://github.com/jbkinney/19_mlcb`. We thank Anand Murugan, Yifei Huang, Peter Koo, Alan Moses, and Mahdi Kooshkbaghi for helpful discussions, as well as and three anonymous referees for providing constructive criticism. This project was supported by NIH grant 1R35GM133777 and a grant from the CSHL/Northwell Health partnership.

bioRxiv preprint doi: https://doi.org/10.1101/835942; this version posted February 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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," *Nat Methods*, vol. 12, pp. 931–934, Aug. 2015.

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

[4] 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, pp. 535–548.e24, Jan. 2019.

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

[6] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[7] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3145–3153, JMLR.org, 2017.

[8] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian, "Neural network attributions: A causal perspective," *arXiv preprint arXiv:1902.02302*, 2019.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

[18] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283, 2016.

[19] E. Mjolsness, "On cooperative quasi-equilibrium models of transcriptional regulation," *J Bioinform Comput Biol*, vol. 5, no. 2b, pp. 467–490, 2007.

[20] 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.

bioRxiv preprint doi: https://doi.org/10.1101/835942; this version posted February 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[21] 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, no. 12, pp. 2015 – 2024, 2017.

[22] R. Movva, P. Greenside, G. K. Marinov, S. Nair, A. Shrikumar, and A. Kundaje, "Deciphering regulatory dna sequences and noncoding genetic variants using neural network models of massively parallel reporter assays," *PLoS ONE*, vol. 14, no. 6, p. e0218073, 2019.

[23] 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, no. 7, pp. 803–809, 2019.

[24] N. Bogard, J. Linder, A. B. Rosenberg, and G. Seelig, "A deep neural network for predicting and engineering alternative polyadenylation," *Cell*, vol. 178, no. 1, pp. 91–106.e23, 2019.

[25] R. P. Patwardhan, C. Lee, O. Litvin, D. L. Young, D. Pe'er, and J. Shendure, "High-resolution analysis of dna regulatory elements by synthetic saturation mutagenesis," *Nat Biotechnol*, vol. 27, no. 12, pp. 1173 – 1175, 2009.

[26] 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.

[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] 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.

[29] R. P. Patwardhan, J. B. Hiatt, D. M. Witten, M. J. Kim, R. P. Smith, D. May, C. Lee, J. M. Andrie, S.-I. Lee, G. M. Cooper, N. Ahituv, L. A. Pennacchio, and J. Shendure, "Massively parallel functional dissection of mammalian enhancers in vivo," *Nat Biotechnol*, vol. 30, no. 3, pp. 265 – 270, 2012.

[30] Y. Liu, K. Barr, and J. Reinitz, "Fully interpretable deep learning model of transcriptional control," *bioRxiv preprint doi:10.1101/655639*, May 2019.

[31] C. G. d. Boer, E. D. Vaishnav, R. Sadeh, E. L. Abeyta, N. Friedman, and A. Regev, "Deciphering eukaryotic gene-regulatory logic with 100 million random promoters," *Nat Biotechnol*, pp. 1–10, 2019.

[32] A. Tareen and J. B. Kinney, "Logomaker: beautiful sequence logos in python," *Bioinformatics*, Dec. 2019. btz921.

[33] 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.

[34] 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.

[35] 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.

[36] J. B. Kinney, G. Tkačik, and C. G. Callan, "Precise physical models of protein–DNA interaction from high-throughput data," *Proc Natl Acad Sci USA*, vol. 104, pp. 501–506, Jan. 2007.

[37] 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.

[38] 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.

[39] W. R. McClure, "Rate-limiting steps in rna chain initiation," *Proc Natl Acad Sci USA*, vol. 77, no. 10, pp. 5634 – 5638, 1980.

bioRxiv preprint doi: https://doi.org/10.1101/835942; this version posted February 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[40] W. R. McClure, "Mechanism and control of transcription initiation in prokaryotes," *Annu Rev Biochem*, vol. 54, no. 1, pp. 171 – 204, 1985.

[41] E. King and C. Altman, "A schematic method of deriving the rate laws for enzyme-catalyzed reactions," *J Phys Chem*, vol. 60, no. 10, pp. 1375–1378, 1956.

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