bioRχiv, 2026, 1–6
doi:

GaugeFixer: overcoming parameter non-identifiability in models of sequence-function relationships

Carlos Martí-Gómez,[1] David M. McCandlish[1] and Justin B. Kinney[1,*]

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, 1 Bungtown Rd., Cold Spring Harbor, 11724, New York, United States

*Corresponding author: jkinney@cshl.edu

Abstract

Mathematical models that describe sequence-function relationships are widely used in computational biology. A key challenge when interpreting these models is that their parameters are not uniquely determined: many different parameter choices can encode the same sequence-function landscape. These ambiguities, known as "gauge freedoms," must be resolved before parameter values can be meaningfully interpreted. Resolving gauge freedoms requires imposing mathematical constraints on parameters that remove these degrees of freedom, a procedure called "fixing the gauge." We recently developed mathematical methods for fixing the gauge of a large class of commonly used models, but the direct computational implementation of these methods is often impractical due to the need for projection matrices whose memory requirements scale quadratically with the number of parameters. Here we introduce GaugeFixer, a Python package that exploits the specific mathematical structure of gauge-fixing projections to achieve linear scaling, thus enabling application to models with millions of parameters. To demonstrate GaugeFixer, we analyze the local structure of peaks in an empirical fitness landscape for translation initiation. GaugeFixer reveals striking similarities, but also fine-scaled variation, in ribosome binding preferences at different positions relative to the start codon, thereby facilitating the interpretation of an otherwise unwieldy fitness landscape. GaugeFixer thus fills an unmet need in the computational tools available for biologically interpreting sequence-function relationships.

Availability and implementation: GaugeFixer is compatible with Python $\geq$ 3.10 and can be installed using the pip package manager. Documentation is provided at `https://gaugefixer.readthedocs.io`. Source code is available at `https://github.com/jbkinney/gaugefixer`, as are the scripts used to carry out the analyses presented here.

Introduction

Computational biology routinely involves the use of models that describe the quantitative relationship between biological sequences (DNA, RNA, or protein) and their biological activities [Kinney and McCandlish, 2019]. For example, quantitative models of sequence-function relationships have been used to predict the locations of transcription factor binding sites in promoters and enhancers [Stormo, 2013], the locations of splice sites in pre-mRNA transcripts [Yeo and Burge, 2004], structural contacts between residues in folded proteins [Marks et al., 2012], and the effects of human genetic variation on protein function [Hopf et al., 2017]. One application of growing interest is the modeling of data from high-throughput mutagenesis experiments [Kinney et al., 2010, Mogno et al., 2013, Belliveau et al., 2018, Otwinowski et al., 2018, Faure et al., 2022, Zhou et al., 2022, Tareen et al., 2022, Faure and Lehner, 2024, Martí-Gómez et al., 2026, Zhou et al., 2025].

A common way to mathematically represent sequence-function relationships is through the use of generalized one-hot models. In these models, each sequence is represented as a set of binary features indicating the presence or absence of specific subsequences at specific sets of positions, and each feature has a corresponding parameter that quantifies its effect. An important challenge when interpreting these parameters is that they are not uniquely determined by the fitness landscape they describe, as different choices of model parameters can yield the same landscape. These extra degrees of freedom in parameter space are called "gauge freedoms" (see Supplementary Information (SI), Secs. 1–2, for formal definitions). Before the numerical values of model parameters can be meaningfully interpreted, gauge freedoms must first be removed by a process called "fixing the gauge" [Posfai et al., 2025b,a].

Gauge freedoms have long played a central role in theoretical physics [Jackson and Okun, 2001], but they have received

© The Author 2026. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

remarkably little attention in the context of sequence-function relationships. A scattered body of work has described technical methods for handling gauge freedoms in specific sequence-to-function modeling applications [Kinney et al., 2007, Stormo, 2011, Rube et al., 2022, Weigt et al., 2009, Marks et al., 2011, Ekeberg et al., 2013, Barton et al., 2016, Cocco et al., 2018, Haldane and Levy, 2019, Gerardos et al., 2022, Hsu et al., 2022, Zamuner and Rios, 2021, Feinauer et al., 2022, Feinauer and Borgonovo, 2022]. Other work has explored ways of eliminating these freedoms through alternative parameterizations of genetic interactions [Poelwijk et al., 2019, Metzger et al., 2024, Park et al., 2024b, Dupic et al., 2024, Park et al., 2024a, Faure et al., 2024]. However, a unified approach for handling gauge freedoms had been missing. To address this need, we recently developed a mathematical theory of gauge freedoms for generalized one-hot models [Posfai et al., 2025a]; see also [Posfai et al., 2025b]. This work showed that a family of gauges that encompasses nearly all of the gauges used in the prior literature can be imposed by multiplying parameter vectors by an appropriate projection matrix. Despite its simplicity, directly implementing this gauge-fixing procedure is often impractical for models with more than a few thousand parameters due to the required projection matrices being very large (and not sparse).

Here we introduce GaugeFixer, an open-source Python package for fixing the gauge of generalized one-hot models of sequence-function relationships using the mathematical methods developed by Posfai et al. [2025a]. By exploiting the mathematical structure of generalized one-hot models and their corresponding gauge-fixing projection matrices, GaugeFixer achieves linear scaling in both memory and computation time, enabling the application to models with millions of parameters. To demonstrate its utility, we apply GaugeFixer here to a fitness landscape for Shine-Dalgarno sequences described by a model with nearly 2 million parameters [Kuo et al., 2020, Martí-Gómez et al., 2026]. The results illustrate how GaugeFixer can be used to reveal phenomena that are otherwise difficult to discern within complex sequence-function landscapes.

Background

Generalized one-hot models

We consider generalized one-hot models that predict a scalar function $f(s)$ for sequences s of fixed length L. Of particular interest among these models are all-order models, which describe interactions of every order from 0 through L. Such models can represent any possible sequence-function relationship, but the number of parameters they require grows exponentially with sequence length. It is therefore common to use models that include only interactions up to a specified order (K-order models) or only between nearby positions (nearest-neighbor or K-adjacent models). These are examples of hierarchical models, a class of generalized one-hot models that remain tractable for longer sequences while still capturing important biological dependencies (SI Sec. 2).

Minimal gauge fixing example

To motivate the need for gauge fixing, consider a simple landscape over sequences of length $L = 1$ built using a two-character alphabet $\{\mathrm{A}, \mathrm{B}\}$ and defined by the fitness values $f(A) = 0$, $f(B) = 2$. To model this landscape we adopt a generalized one-hot model having three parameters—a constant-effect parameter (θ_0) and two additive parameters (θ_1^{A} and θ_1^{B}) that describe the effect of either character at position 1 in the sequence. One choice of parameters that describes the landscape is $\vec{\theta} = [\theta_0, \theta_1^{\mathrm{A}}, \theta_1^{B}] = [0, 0, 2]$ (Figure 1A, orange dot). There are, however, an infinite number of equivalent parameter vectors we can choose, namely vectors of the form $[z, -z, 2-z]$ where z is any real number (red line). Fixing the gauge involves specifying an additional constraint that selects a single point on this line and thus specifies a unique choice of parameter values. The blue plane in Figure 1A represents one such constraint: the "zero-sum gauge," which requires that the additive parameters sum to zero at each position. In the present example this restricts parameters to the plane defined by $\theta_1^{\mathrm{A}} + \theta_1^{\mathrm{B}} = 0$. The resulting gauge-fixed parameters lie at the intersection of the line of equivalent parameters and the zero-sum plane ($\vec{\theta}_{\mathrm{fixed}} = [1, -1, 1]$; green dot). Note that no choice of parameters in the zero-sum gauge other than $\vec{\theta}_{\mathrm{fixed}}$ will yield the same landscape.

Families of gauges

In previous work [Posfai et al., 2025b], we introduced a family of gauges for the all-order interaction model. These gauges are parameterized by two quantities: a non-negative number λ and a probability distribution π over sequences. λ controls how much explanatory power is allocated across interaction orders, while π specifies the distribution of sequences against which parameter effects are expressed. This λ, π family of gauges has multiple attractive properties. In particular, it encompasses the most commonly used gauges in the literature: these include the trivial gauge, the Euclidean gauge, the equitable gauge, the zero-sum gauge, and the wild-type gauge. See Posfai et al. [2025b] and SI Sec. 3 for details.

A particularly useful subset of gauges in this family are the hierarchical gauges, which are obtained in the limit where λ approaches infinity. In these gauges, lower-order terms explain as much of the landscape's variance as possible, while higher-order interaction terms capture only residual variation that cannot be attributed to lower-order effects. Unlike other gauges in the λ, π family, hierarchical gauges can be applied to a much larger class of models called the "hierarchical models." The values of model parameters in hierarchical gauges also have a natural interpretation: each parameter represents the average effect of introducing specific characters at specific positions, compared to the effect expected from lower-order terms, when sequences are drawn from the distribution π. By choosing different distributions π, one can observe how sequence-function relationships vary across different regions of sequence space.

Results

The GaugeFixer algorithm

Fixing the gauge requires projecting parameter vectors onto lower-dimensional subspaces, a computation performed by multiplying the parameter vector by a projection matrix. But when models have more than about 10^4 parameters, such direct matrix multiplication becomes impractical owing to the need to generate and apply large projection matrices (which are not sparse). To overcome this limitation for all-order models, GaugeFixer exploits the fact that the projection matrices for all-order models can be written as Kronecker products of L much smaller matrices, one matrix for each sequence position [Posfai et al., 2025b]. This allows projections to be computed without ever constructing the full projection matrix [Martí-Gómez et al., 2026], thus reducing both

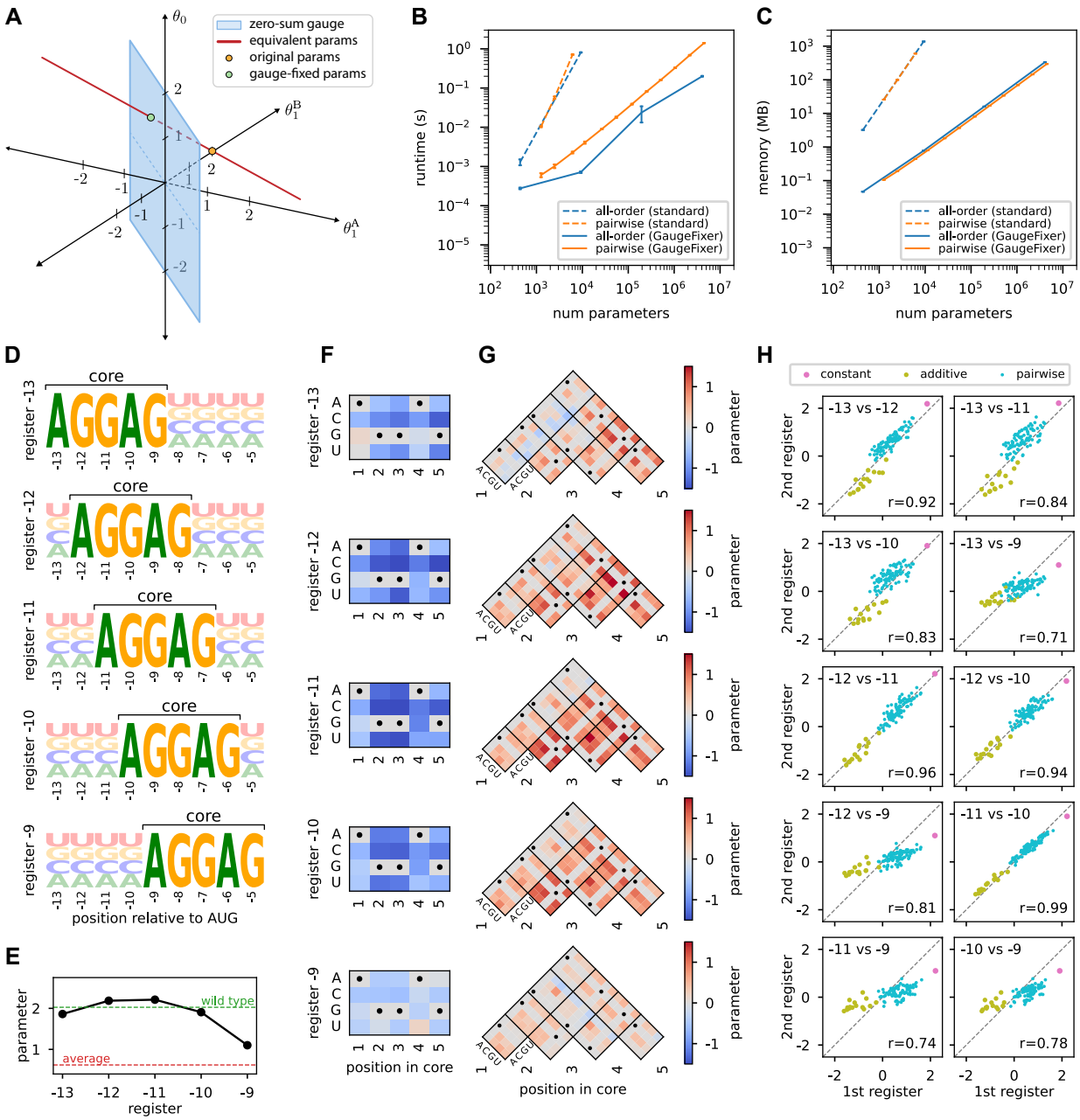

Figure 1: Illustration of gauge-fixing and demonstration of GaugeFixer. (A) Minimal example of gauge fixing. Shown is the space of all possible three-dimensional parameter vectors, $\vec{\theta} = [\theta_0, \theta_1^{\mathrm{A}}, \theta_1^{\mathrm{B}}]$, for the all-order interaction model on sequences of length $L = 1$ built from the alphabet $\{A, B\}$. Orange dot: initial choice of $\vec{\theta} = [0, 0, 2]$, which yields $f(A) = 0$, $f(B) = 2$. Red line: parameter vectors equivalent to $\vec{\theta}$. Blue plane: parameter vectors in the zero-sum gauge, i.e., that satisfy $\theta_1^{\mathrm{A}} + \theta_1^{\mathrm{B}} = 0$. Green dot: gauge-fixed parameters $\vec{\theta}_{\mathrm{fixed}} = [1, -1, 1]$. (B,C) Algorithm performance when fixing the gauge of all-order and pairwise protein models with increasing number of parameters. Shown are the runtime (B) and peak memory (C) requirements for GaugeFixer and for standard matrix multiplication on a laptop computer. Error bars represent the standard deviation in runtime and memory usage over 10 projections of parameter vectors drawn i.i.d. from a normal distribution. (D) Sequence logos [Tareen and Kinney, 2019] showing the probability distributions π corresponding to different peaks in the Shine-Dalgarno fitness landscape measured by Kuo et al. [2020] and modeled by Martí-Gómez et al. [2026]. Each distribution has a fixed AGGAG motif at a specific register relative to the start codon. (E,F,G) Gauge-fixed constant (E), additive (F), and pairwise (G) parameters in the hierarchical gauge associated with each register. Horizontal dashed lines in (E) represent the average phenotype across all possible sequences (red) and the phenotype of the wild-type sequence AAGGAGGUG (green), which occurs in the 5'UTR of the *dmsC* gene of *Escherichia coli*. (H) Scatterplots comparing the values of gauge-fixed parameters observed in the core region at different registers.

the memory requirements and computation time from $O(M^2)$ to $O(M)$ where M is the number of parameters. For the more general class of hierarchical models, GaugeFixer decomposes the model into a sum of all-order models restricted to subsets of positions, applies the efficient projection algorithm to each, and sums the results together (SI Secs. 4–5). Using these strategies, GaugeFixer achieves orders of magnitude improvement in both runtime and memory usage compared to gauge-fixing by direct matrix multiplication (Figures 1B,C). This dramatic improvement enables gauge-fixing computations to be performed on models with millions of parameters in just a few seconds on a standard laptop computer.

Analysis of the Shine-Dalgarno fitness landscape

To illustrate the utility of GaugeFixer, we analyzed a fitness landscape for the Shine-Dalgarno (SD) sequence, a motif in bacterial messenger RNA that facilitates translation initiation by base-pairing with the 3' tail of the 16S ribosomal RNA [Shine and Dalgarno, 1975]. Based on measurements of the translational activity of nearly every possible 9-nucleotide RNA sequence [Kuo et al., 2020], we previously inferred an all-order model having 1,953,125 parameters. The resulting fitness landscape contains multiple prominent fitness peaks corresponding to the canonical AGGAG motif positioned in different registers relative to the start codon [Zhou et al., 2022, Martí-Gómez et al., 2026].

Here, we used GaugeFixer to quantitatively characterize the local structure around the fitness peaks of this model. For each peak we defined a distribution π that fixes the AGGAG core motif in the corresponding register while randomizing the characters at the remaining positions (Figure 1D). We then imposed the hierarchical gauge corresponding to each choice of π and examined the resulting parameter values. In this gauge, the constant, additive, and pairwise parameters can be respectively interpreted as the average phenotype, average single-nucleotide effect, and average epistatic effect observed when mutations are introduced into sequences randomly drawn from π. We emphasize that these different parameter values, corresponding to different gauges, are simply different representations of the same model.

The constant term, θ_0, represents the mean fitness of sequences with AGGAG in the specified register. We find that this value is highest for registers -12 and -11, consistent with known optimal spacing requirements for translation initiation (Figure 1E). In register -9, by contrast, the mean fitness is much lower, indicating markedly lower translation on average. The additive parameters in the core region, shown in Figure 1F, reveal how individual nucleotide mutations away from the AGGAG motif affect translational efficiency in each register. As expected, these mutations are overwhelmingly deleterious. The effects are also remarkably consistent across registers, though some differences emerge near the boundaries. The pairwise interaction parameters in the core region, shown in Figure 1G, capture the effects of mutating pairs of nucleotides within the AGGAG motif beyond what additive effects alone predict. These interactions are also remarkably consistent across registers. Predominantly positive values are observed, indicating that combinations of mutations tend to be less deleterious than expected from their individual effects, one hallmark of global epistasis [Otwinowski et al., 2018, Kinney and McCandlish, 2019].

Finally, we compared the constant, additive, and pairwise parameters in the core region across registers (Figure 1H). The results show that neighboring registers tend to have more similar parameters, while more distant registers show greater divergence. This smooth variation suggests that ribosomal binding preferences change gradually as a function of distance from the start codon.

Discussion

Here we introduced GaugeFixer, a Python package that efficiently removes gauge freedoms from generalized one-hot models that describe sequence-function relationships. By exploiting the Kronecker factorization of projection operators and the specific mathematical structure of the models, GaugeFixer dramatically reduces runtime and memory requirements, thus enabling rapid gauge fixing for models with millions of parameters.

We wish to emphasize that gauge fixing is fundamentally different from parameter inference. Both procedures are an essential part of modeling sequence-function relationships, but they serve orthogonal purposes. Inference is the process of identifying parameters that make a model's predictions best fit one's data, but without regard to parameter interpretation. Gauge fixing, by contrast, alters model parameters in ways that have no effect on model predictions, but which are important for the interpretation of those parameters. GaugeFixer makes this distinction explicit by providing utilities to convert precomputed parameter vectors into any chosen gauge. A related but more subtle point is that some regularization schemes and/or prior distributions used during inference yield parameters in specific gauges [Posfai et al., 2025b, Petti et al., 2025]. This does not, however, eliminate the need for post-inference gauge fixing if model parameters are to be interpreted.

Different gauges have different advantages when interpreting model parameters. For instance, the influence of global epistasis is more evident when parameters are represented in the wild-type gauge, whereas using additive models to illustrate the specificities of transcription factors, or computing contact maps in pairwise interaction models, typically requires the models to be in the zero-sum gauge [Kinney et al., 2007, Stormo, 2011, Rube et al., 2022, Marks et al., 2011, Ekeberg et al., 2013]. Hierarchical gauges, on the other hand, facilitate the investigation of how mutational effects and their interactions change across different regions of sequence space, providing an intuitive way to interpret higher-order interactions.

Although GaugeFixer is designed to work with specific classes of linear models (all-order models and hierarchical models), it can also be applied to nonlinear and nonparametric models, such as neural networks or Gaussian processes, by representing the predicted landscape using an all-order model and applying GaugeFixer to the parameters of that model. While GaugeFixer requires exhaustive enumeration of all model parameters, recent theoretical results suggest the possibility of directly computing the posterior distribution over specific gauge-fixed parameters using Gaussian process models, thus bypassing the need to construct an equivalent all-order model with an astronomically large number of parameters [Petti et al., 2025]. Gauge fixing is also conceptually adjacent to neural network interpretability methods such as "global importance analysis" [Koo et al., 2021] or "in silico marginalization" [Schreiber, 2025, Liu et al., 2025].

GaugeFixer thus provides a high-performance software library for fixing the gauge of commonly used sequence-to-function models, thereby filling an important gap in the computational tools available for interpreting sequence-function relationships.

Competing interests

The authors declare no competing interests.

Author contributions statement

C.M-G. and J.B.K. conceived the project, wrote the package, and performed the research. C.M-G., D.M.M., and J.B.K. wrote the manuscript. D.M.M. and J.B.K. supervised the research.

Acknowledgments

This work was supported by NIH grants R01HG011787 (J.B.K., D.M.M.), R35GM133777 (J.B.K.), and R35GM133613 (D.M.M., C. M-G.). Computational equipment was supported by NIH grant S10OD028632.

References

J P Barton, E De Leonardis, A Coucke, and S Cocco. ACE: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics*, 32(20):3089–3097, 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw328.

Nathan M. Belliveau, Stephanie L. Barnes, William T. Ireland, Daniel L. Jones, Michael J. Sweredoski, Annie Moradian, Sonja Hess, Justin B. Kinney, and Rob Phillips. Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proceedings of the National Academy of Sciences*, 115(21):201722055, 05 2018. ISSN 0027-8424. doi: 10.1073/pnas.1722055115.

Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, R Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81(3):032601, 2018. ISSN 0034-4885. doi: 10.1088/1361-6633/aa9965.

Thomas Dupic, Angela M. Phillips, and Michael M. Desai. Protein sequence landscapes are not so simple: on reference-free versus reference-based inference. *bioRxiv*, page 2024.01.29.577800, 2024. doi: 10.1101/2024.01.29.577800.

Magnus Ekeberg, Cecilia L ovkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1): 012707, 2013. ISSN 1539-3755. doi: 10.1103/physreve.87.012707.

Andre J. Faure and Ben Lehner. MoCHI: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis, and allostery from deep mutational scanning data. *Genome Biology*, 25(1):303, 2024. ISSN 1474-7596. doi: 10.1186/s13059-024-03444-y.

Andre J. Faure, Júlia Domingo, Jörn M. Schmiedel, Cristina Hidalgo-Carcedo, Guillaume Diss, and Ben Lehner. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature*, 604(7904):175–183, 2022. ISSN 0028-0836. doi: 10.1038/s41586-022-04586-4.

Andre J. Faure, Ben Lehner, Verónica Miró Pina, Claudia Serrano Colome, and Donate Weghorn. An extension of the walsh-hadamard transform to calculate and model epistasis in genetic landscapes of arbitrary shape and complexity. *PLOS Computational Biology*, 20(5):e1012132, 2024. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1012132.

Christoph Feinauer and Emanuele Borgonovo. Mean Dimension of Generative Models for Protein Sequences. *bioRxiv*, page 2022.12.12.520028, 2022. doi: 10.1101/2022.12.12.520028.

Christoph Feinauer, Barthelemy Meynard-Piganeau, and Carlo Lucibello. Interpretable pairwise distillations for generative protein sequence models. *PLoS Computational Biology*, 18(6): e1010219, 2022. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1010219.

Andonis Gerardos, Nicola Dietler, and Anne-Florence Bitbol. Correlations from structure and phylogeny combine constructively in the inference of protein partners from sequences. *PLoS Computational Biology*, 18(5):e1010147, 2022. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1010147.

Allan Haldane and Ronald M. Levy. Influence of multiple-sequence-alignment depth on Potts statistical models of protein covariation. *Physical Review E*, 99(3):032405, 2019. ISSN 2470-0045. doi: 10.1103/physreve.99.032405.

Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta P I Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, February 2017. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.3769.

Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Learning protein fitness models from evolutionary and assay-labeled data. *Nature Biotechnology*, 40(7):1114–1122, 2022. ISSN 1087-0156. doi: 10.1038/s41587-021-01146-5.

J. D. Jackson and L. B. Okun. Historical roots of gauge invariance. *Reviews of Modern Physics*, 73(3):663–680, 2001. ISSN 0034-6861. doi: 10.1103/revmodphys.73.663.

Justin B Kinney and David M McCandlish. Massively parallel assays and quantitative sequence-function relationships. *Annual Review of Genomics and Human Genetics*, 20(1):99–127, 08 2019. ISSN 1527-8204. doi: 10.1146/annurev-genom-083118-014845.

Justin B. Kinney, Gasper Tkacik, and Curtis G. Callan. Precise physical models of protein-DNA interaction from high-throughput data. *Proceedings of the National Academy of Sciences*, 104(2):501–506, 01 2007. ISSN 0027-8424. doi: 10.1073/pnas.0609908104.

Justin B. Kinney, Anand Murugan, Curtis G. Callan, and Edward C. Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*, 107(20): 9158–9163, 05 2010. ISSN 0027-8424. doi: 10.1073/pnas.1004290107.

Peter K. Koo, Antonio Majdandzic, Matthew Ploenzke, Praveen Anand, and Steffan B. Paul. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLOS Computational Biology*, 17(5):e1008925, May 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008925.

Syue Ting Kuo, Ruey Lin Jahn, Yuan Ju Cheng, Yi Lan Chen, Yun Ju Lee, Florian Hollfelder, Jin Der Wen, and Hsin Hung David Chou. Global fitness landscapes of the Shine-Dalgarno sequence. *Genome Research*, 30(5):711–723, 2020. ISSN 15495469. doi: 10.1101/gr.260182.119.

Betty B. Liu, Selin Jessa, Samuel H. Kim, Yan Ting Ng, Soon Il Higashino, Georgi K. Marinov, Derek C. Chen, Benjamin E. Parks, Li Li, Tri C. Nguyen, Austin T. Wang, Sean K. Wang, Serena Y. Tan, Michael Kosicki, Len A. Pennacchio, Eyal Ben-David, Anca M. Pasca, Anshul Kundaje, Kyle K. H. Farh, and William J. Greenleaf. Dissecting regulatory syntax in human development with scalable multiomics and deep learning, 2025. URL `https://www.biorxiv.org/content/10.1101/2025.04.30.651381v1`.

Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS ONE*, 6(12):e28766, 2011. doi: 10.1371/journal.pone.0028766.

Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11):1072–1080, November 2012. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.2419.

Carlos Martí-Gómez, Juannan Zhou, Wei-Chia Chen, Arlin Stoltzfus, Justin B Kinney, and David M McCandlish. Inference and Visualization of Complex Genotype–Phenotype Maps. *Molecular Biology and Evolution*, 43(2):msag023, 2026. ISSN 1537-1719. doi: 10.1093/molbev/msag023.

Brian PH Metzger, Yeonwoo Park, Tyler N Starr, and Joseph W Thornton. Epistasis facilitates functional evolution in an ancient transcription factor. *Elife*, 12:RP88737, 2024.

Ilaria Mogno, Jamie C Kwasnieski, and Barak A Cohen. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res*, 23(11):1908 – 1915, 2013. ISSN 1088-9051. doi: 10.1101/gr.157891.113.

Jakub Otwinowski. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol Biol Evol*, 35(10):2345 – 2354, 10 2018. ISSN 0737-4038. doi: 10.1093/molbev/msy141.

Jakub Otwinowski, David M McCandlish, and Joshua B Plotkin. Inferring the shape of global epistasis. *Proc Natl Acad Sci USA*, 115(32):E7550 – E7558, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1804015115.

Yeonwoo Park, Brian P.H. Metzger, and Joseph W. Thornton. On the analysis of protein genetic architecture: Response to "protein sequence landscapes are not so simple". *bioRxiv*, 2024a. doi: 10.1101/2024.09.17.613512. URL `https://www.biorxiv.org/content/early/2024/12/06/2024.09.17.613512`.

Yeonwoo Park, Brian PH Metzger, and Joseph W Thornton. The simplicity of protein sequence-function relationships. *Nature Communications*, 15(1):7953, 2024b.

Samantha Petti, Carlos Martí-Gómez, Justin B Kinney, Juannan Zhou, and David M McCandlish. On learning functions over biological sequence space: relating gaussian process priors, regularization, and gauge fixing. *arXiv*, 2025. doi: 10.48550/arxiv.2504.19034.

Frank J Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature Communications*, 10(1):4213, 2019. doi: 10.1038/s41467-019-12130-8.

Anna Posfai, David M. McCandlish, and Justin B. Kinney. Symmetry, gauge freedoms, and the interpretability of sequence-function relationships. *Physical Review Research*, 7(2):023005, 2025a. doi: 10.1103/physrevresearch.7.023005.

Anna Posfai, Juannan Zhou, David M. McCandlish, and Justin B. Kinney. Gauge fixing for sequence-function relationships. *PLOS Computational Biology*, 21(3):e1012818, 2025b. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1012818.

H. Tomas Rube, Chaitanya Rastogi, Siqian Feng, Judith F. Kribelbauer, Allyson Li, Basheer Becerra, Lucas A. N. Melo, Bach Viet Do, Xiaoting Li, Hammaad H. Adam, Neel H. Shah, Richard S. Mann, and Harmen J. Bussemaker. Prediction of protein-ligand binding affinity from sequencing data with interpretable machine learning. *Nature Biotechnology*, 40 (10):1520–1527, 2022. ISSN 1087-0156. doi: 10.1038/s41587-022-01307-0.

Jacob Schreiber. tangermeme: A toolkit for understanding cis-regulatory logic using deep learning models. *bioRxiv*, 2025. doi: 10.1101/2025.08.08.669296. URL `https://www.biorxiv.org/content/early/2025/08/12/2025.08.08.669296`.

J Shine and L Dalgarno. Determinant of cistron specificity in bacterial ribosomes. *Nature*, 254:34–38, 1975.

Gary D Stormo. Maximally efficient modeling of DNA sequence motifs at all levels of complexity. *Genetics*, 187(4):1219 – 1224, 2011. ISSN 0016-6731. doi: 10.1534/genetics.110.126052.

Gary D. Stormo. Modeling the specificity of protein-DNA interactions. *Quantitative Biology*, 1(2):115–130, June 2013. ISSN 2095-4689, 2095-4697. doi: 10.1007/s40484-013-0012-4.

Ammar Tareen and Justin B Kinney. Logomaker: beautiful sequence logos in python. *Bioinformatics*, 36(7):2272–2274, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz921.

Ammar Tareen, Mahdi Kooshkbaghi, Anna Posfai, William T. Ireland, David M. McCandlish, and Justin B. Kinney. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome Biology*, 23(1):98, 2022. ISSN 1474-7596. doi: 10.1186/s13059-022-02661-7.

Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0805923106.

Gene Yeo and Christopher B Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. In *Journal of Computational Biology*, volume 11, pages 377–394, March 2004. doi: 10.1089/1066527041410418.

Stefano Zamuner and Paolo De Los Rios. Interpretable Neural Networks based classifiers for categorical inputs. *arXiv*, 2021. doi: 10.48550/arxiv.2102.03202.

Juannan Zhou, Mandy S Wong, Wei-Chia Chen, Adrian R Krainer, Justin B Kinney, and David M McCandlish. Higher-order epistasis and phenotypic prediction. *Proc. Natl. Acad. Sci. USA*, 119(39), 2022. doi: https://doi.org/10.1073/pnas.2204233119.

Juannan Zhou, Carlos Martí-Gómez, Samantha Petti, and David M. McCandlish. Learning sequence-function relationships with scalable, interpretable Gaussian processes. *eLife*, 14, November 2025. doi: 10.7554/eLife.108964.1.