PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Gauge fixing for sequence-function relationships

Anna Posfai[1], **Juannan Zhou**[1,2], **David M. McCandlish**[1]*, **Justin B. Kinney**[1]*

1 Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **2** Department of Biology, University of Florida, Gainesville, Florida, United States of America

* mccandlish@cshl.edu (DMM); jkinney@cshl.edu (JBK)

OPEN ACCESS

Citation: Posfai A, Zhou J, McCandlish DM, Kinney JB (2025) Gauge fixing for sequence-function relationships. PLoS Comput Biol 21(3): e1012818. https://doi.org/10.1371/journal.pcbi.1012818

Editor: Kiran R. Patil, University of Cambridge, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: September 16, 2024

Accepted: January 22, 2025

Published: March 20, 2025

Data availability statement: All data and Python scripts used to generate the figures are available on GitHub at https://github.com/jbkinney/24_posfai1 and on Zenodo at https://doi.org/10.5281/zenodo.14811498.

Funding: This work was supported by NIH grant R35 GM133613 (AP, JZ, DMM), NIH grant R35 GM133777 (AP, JBK), NIH grant R01 HG011787 (DMM, JBK), NIH grant R35 GM154908 (JZ), the Alfred P. Sloan foundation (DMM), as well as additional funding from the

Abstract

Quantitative models of sequence-function relationships are ubiquitous in computational biology, e.g., for modeling the DNA binding of transcription factors or the fitness landscapes of proteins. Interpreting these models, however, is complicated by the fact that the values of model parameters can often be changed without affecting model predictions. Before the values of model parameters can be meaningfully interpreted, one must remove these degrees of freedom (called "gauge freedoms" in physics) by imposing additional constraints (a process called "fixing the gauge"). However, strategies for fixing the gauge of sequence-function relationships have received little attention. Here we derive an analytically tractable family of gauges for a large class of sequence-function relationships. These gauges are derived in the context of models with all-order interactions, but an important subset of these gauges can be applied to diverse types of models, including additive models, pairwise-interaction models, and models with higher-order interactions. Many commonly used gauges are special cases of gauges within this family. We demonstrate the utility of this family of gauges by showing how different choices of gauge can be used both to explore complex activity landscapes and to reveal simplified models that are approximately correct within localized regions of sequence space. The results provide practical gauge-fixing strategies and demonstrate the utility of gauge-fixing for model exploration and interpretation.

Author summary

Biophysics and other areas of quantitative biology rely heavily on mathematical models that predict biological activities from DNA, RNA, or protein sequences. Interpreting the parameters of these models, however, is not trivial. Here we address a core challenge for model interpretation–the presence of "gauge freedoms", i.e., directions in parameter space that do not affect model predictions and therefore cannot be constrained by data. Our results provide an explicit mathematical method for removing these unconstrained degrees of freedom–a process called "fixing the gauge"–that can be applied to a wide

Simons Center for Quantitative Biology at CSHL (DMM, JBK) and the College of Liberal Arts and Sciences at the University of Florida (JZ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

range of commonly use models of sequence-function relationships, including models that describe interactions of arbitrarily high order. These results unify diverse gauge fixing methods that have been previously described in the literature for specific types of models. We further show how our gauge-fixing approach can be used to simplify complex models in user-specified regions of sequence space. This work thus overcomes a major obstacle in the interpretation of quantitative sequence-function relationships.

Introduction

One of the central challenges of biology is to understand how functionally relevant information is encoded within DNA, RNA, and protein sequences. Unlike the genetic code, most sequence-function relationships are quantitative in nature, and understanding them requires finding mathematical functions that, upon being fed unannotated sequences, return values that quantify sequence activity [1]. Multiplex assays of variant effects (MAVEs), functional genomics methods, and other high-throughput techniques are rapidly increasing the ease with which sequence-function relationships can be experimentally studied. And while quantitative modeling efforts based on these high-throughput data are becoming increasingly successful, in that they yield models with ever-increasing predictive ability, major open questions remain about how to interpret both the parameters [2–12] and the predictions [13–17] of the resulting models. One major open question is how to deal with the presence of gauge freedoms.

Gauge freedoms are directions in parameter space along which changes in model parameters have no effect on model predictions [18]. Not only can the values of model parameters along gauge freedoms not be determined from data, differences in parameters along gauge freedoms have no biological meaning even in principle. Many commonly used models of sequence-function relationships exhibit numerous gauge freedoms [19–35], and interpreting the parameters of these models requires imposing additional constraints on parameter values, a process called "fixing the gauge".

The gauge freedoms of sequence-function relationships are most completely understood in the context of additive models (commonly used to describe transcription factor binding to DNA [19,22,35]) and pairwise-interaction models (commonly used to describe proteins [20,21,23–34]). Recently, some gauge-fixing strategies have been described for all-order interaction models, again in the context of protein sequence-function relationships [30,31,34]. However, a unified gauge-fixing strategy applicable to diverse models of sequence-function relationships has yet to be developed.

Here we provide a general treatment of the gauge fixing problem for sequence-function relationships, focusing on the important case where the set of gauge-fixed parameters form a vector space. These "linear gauges" predominate in the literature (though there are exceptions [36,37]), and have the useful property that differences between vectors of gauge-fixed parameter values are directly interpretable. We first demonstrate the relationship between these linear gauges and L_2 regularization on parameter vectors, and then derive a mathematically tractable family of gauges for the all-order interaction model. Importantly, a subset of these gauges–the "hierarchical gauges"–can be applied to diverse models beyond just the all-order interaction model (including additive models, pairwise-interaction models, and higher-order interaction models) and include as special cases two types of gauges that are commonly used in practice

("zero-sum gauges" [23,28] and "wild-type gauges" [9,23,33]). We then illustrate the properties of this family of gauges by analyzing two example sequence-function relationships: a simulated all-order interaction landscape on short binary sequences, and an empirical pairwise-interaction landscape for the B1 domain of protein G (GB1). The GB1 analysis, in particular, shows how different hierarchical gauges can be used to explore, simplify, and interpret complex functional landscapes. A companion paper [38] further explores the mathematical origins of gauge freedoms in models of sequence-function relationships, and shows how gauge freedoms arise as a consequence of the symmetries of sequence space.

Results

Preliminaries and background

In this section we review how gauge freedoms arise in commonly used models of sequence-function relationships, as well as strategies commonly used to fix the gauge. In doing so, we establish notation and concepts that are used in subsequent sections.

Linear models. We define quantitative models of sequence-function relationships as follows. Let $\mathcal{A}$ denote an alphabet comprising α distinct characters (written $c_1, \ldots, c_\alpha$), let $\mathcal{S}$ denote the set of sequences of length L built from these characters, and let $N = \alpha^L$ denote the number of sequences in $\mathcal{S}$. A quantitative model of a sequence-function relationship (henceforth "model") is a function $f(s;\vec{\theta})$ that maps each sequence s in $\mathcal{S}$ to a real number. The vector $\vec{\theta}$ represents the parameters on which this function depends and is assumed to comprise M real numbers. s_l denotes the character at position l of sequence s. We use l, l', etc. to index positions (ranging from 1 to L) in a sequence and c, c', etc. to index characters in $\mathcal{A}$.

A linear model is a model that is a linear function of $\vec{\theta}$. Linear models have the form

$$f(s;\vec{\theta}) = \vec{\theta} \cdot \vec{x}(s) = \sum_{i=1}^{M} \theta_i x_i(s), \tag{1}$$

where $\vec{x}(\cdot)$ is a vector of M distinct sequence features and each sequence feature $x_i(\cdot)$ is a function that maps sequences to the real numbers. We refer to the space $\mathbb{R}^M$ in which $\vec{x}(\cdot)$ lives as feature space, and the specific vector $\vec{x}(s)$ as the embedding of sequence s in feature space. We use E to denote the vector space spanned by the set of embeddings $\vec{x}(s)$ for all sequences s in $\mathcal{S}$. We emphasize that E is often a proper subspace of $\mathbb{R}^M$ (i.e., has dimension less than M). Indeed, this is what causes f to have gauge freedoms.

One-hot models. One-hot models are linear models based on sequence features that indicate the presence or absence of specific characters at specific positions within a sequence [1]. Such models play a central role in scientific reasoning concerning sequence-function relationships because their parameters can be interpreted as quantitative contributions to the measured function due to the presence of specific biochemical entities (e.g. nucleotides or amino acids) at specific positions in the sequence. These one-hot models include additive models, pairwise-interaction models, all-order interaction models, and more. Additive models have the form

$$f_{\text{add}}(s) = \theta_0 x_0(s) + \sum_l \sum_c \theta_l^c x_l^c(s), \tag{2}$$

where $x_0(s)$ is the constant feature (equal to one for every sequence s) and $x_l^c(s)$ is an additive feature (equal to one if sequence s has character c at position l and equal to zero otherwise; note that c is used here as a superscript and not a power). Pairwise interaction models have the form

$$f_{\text{pair}}(s) = \theta_0 x_0(s) + \sum_l \sum_c \theta_l^c x_l^c(s) + \sum_{l<l'} \sum_{c,c'} \theta_{ll'}^{cc'} x_{ll'}^{cc'}(s), \tag{3}$$

where $x_{ll'}^{cc'}(s)$ is a pairwise feature (equal to one if s has character c at position l and character c' at position l', and equal to zero otherwise). All-order interaction models include interactions of all orders and have the form

$$f_{\text{all}}(s) = \sum_{K=0}^{L} \sum_{l_1<\ldots<l_K} \sum_{c_1,\ldots,c_K} \theta_{l_1\ldots l_K}^{c_1\ldots c_K} x_{l_1\ldots l_K}^{c_1\ldots c_K}(s), \tag{4}$$

where $x_{l_1 l_2 \ldots l_K}^{c_1 c_2 \ldots c_K}(s)$ is a K-order feature (equal to one if s has character c_k at position l_k for all k, and equal to zero otherwise; $K = 0$ corresponds to the constant feature).

Gauge freedoms. Gauge freedoms are transformations of model parameters that leave all model predictions (i.e., the values $f(s)$ at all sequences s) unchanged. The gauge freedoms of a general sequence-function relationship $f(\cdot,\cdot)$ are vectors $\vec{g}$ in $\mathbb{R}^M$ that satisfy

$$f(s;\vec{\theta}) = f(s;\vec{\theta}+\vec{g}) \quad \text{for all} \quad s \in \mathcal{S}. \tag{5}$$

For linear models, gauge freedoms $\vec{g}$ satisfy

$$X\vec{g} = \vec{0}, \tag{6}$$

where X is the $N \times M$ design matrix having rows $\vec{x}(s)$ for $s \in \mathcal{S}$. In linear models, gauge freedoms thus arise when sequence features (i.e., the columns of X) are not linearly independent. In such cases, the space E spanned by sequence embeddings is a proper subspace of $\mathbb{R}^M$, the space G of gauge freedoms is also a proper subspace, and G is orthogonal to E.

Each linear relation between multiple columns of X yields a gauge freedom. For example, additive models have L gauge freedoms arising from the L linear relations,

$$x_0(s) = \sum_c x_l^c(s), \tag{7}$$

for all positions l. Pairwise models have L gauge freedoms arising from the L additive model linear relations in Eq (7), and $\binom{L}{2}(2\alpha - 1)$ additional gauge freedoms arising from the linear relations

$$x_l^c(s) = \sum_{c'} x_{ll'}^{cc'}(s) \quad \text{and} \quad x_{l'}^{c'}(s) = \sum_c x_{ll'}^{cc'}(s) \tag{8}$$

for all characters c, c' and all positions l and l', with $l < l'$ (see 2 for details). More generally, the gauge freedoms of one-hot models arise from the fact that summing any K-order feature $x_{l_1\ldots l_K}^{c_1\ldots c_K}$ over all characters c_k at any chosen position l_k yields a feature of order K–1. A proof that all gauge freedoms arise from such constraints is given in our companion paper [38].

Parameter values depend on choice of gauge. Gauge freedoms pose problems for the interpretation of model parameters (e.g., when interpreting attribution maps from genomic AI models [40]) because, when gauge freedoms are present, different choices of model parameters can give the exact same model predictions. Thus, unless constraints are placed on the values of allowable parameters, individual parameters will have little biological meaning

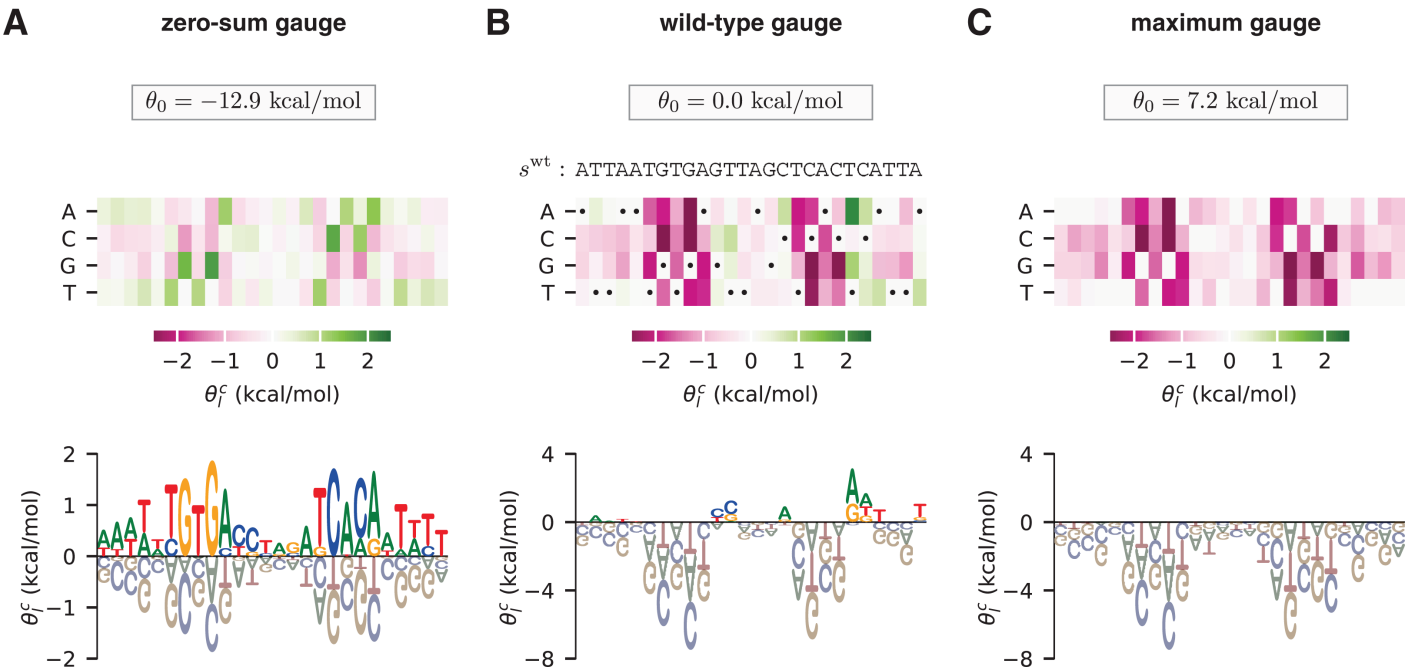

Fig 1. Choice of gauge impacts model parameters. (A–C) Parameters, expressed in three different gauges, for an additive model describing the (negative) binding energy of the *E. coli* transcription factor CRP to DNA. Model parameters are from [37]. In each panel, additive parameters θ_f^c are shown using both (top) a heat map and (bottom) a sequence logo [39]. The value of the constant parameter θ_0 is also shown. (A) The zero-sum gauge, in which the additive parameters at each position sum to zero. (B) The wild-type gauge, in which the additive parameters at each position quantify activity differences with respect to a wild-type sequence, s^{wt}. The wild-type sequence used here (indicated by dots on the heat map) is the CRP binding site present at the *E. coli* lac promoter. (C) The maximum gauge, in which the additive parameters at each position quantify differences with respect to the optimal character at that position. Note that, while the value of each additive parameter θ_f^c varies between panels A-C, differences of the form $\theta_f^c - \theta_f^{c'}$ are preserved.

https://doi.org/10.1371/journal.pcbi.1012818.g001

when viewed in isolation. To interpret model parameters, one therefore needs to adopt constraints that eliminate gauge freedoms and, as a result, make the values of model parameters unique. Geometrically, this means restricting model parameters to a subspace Θ, called "the gauge", on which these constraints are satisfied. This process of choosing constraints (i.e., choosing Θ) is called "fixing the gauge". There are many different gauge-fixing strategies. For example, Fig 1 shows an additive model of the DNA binding energy of CRP (an important transcription factor in *Escherichia coli* [41]) expressed in three different choices of gauge.

Fig 1A shows parameters expressed in the "zero-sum gauge" [23,28] (also called the "Ising gauge" [28], or the "hierarchical gauge" [9]). In the zero-sum gauge, the constant parameter is the mean sequence activity and the additive parameters quantify deviations from this mean activity. The name of the gauge comes from the fact that the additive parameters at each position sum to zero. The zero-sum gauge is commonly used in additive models of protein-DNA binding [35,42–47]. As we will see, zero-sum gauges are readily defined for models with pairwise and higher-order interactions as well.

Fig 1B shows parameters expressed in the "wild-type gauge" [9,23,33] (also called the "lattice-gas gauge" [28] or the "mismatch gauge" [35]). In the wild-type gauge, the constant parameter is equal to the activity of a chosen wild-type sequence (denoted s^{wt}), and additive parameters are the changes in activity that result from mutations away from the wild-type sequence. The wild-type gauge is commonly used to visualize the results of mutational scanning experiments on proteins [48–52] or on long DNA regulatory sequences [53–58]. As we

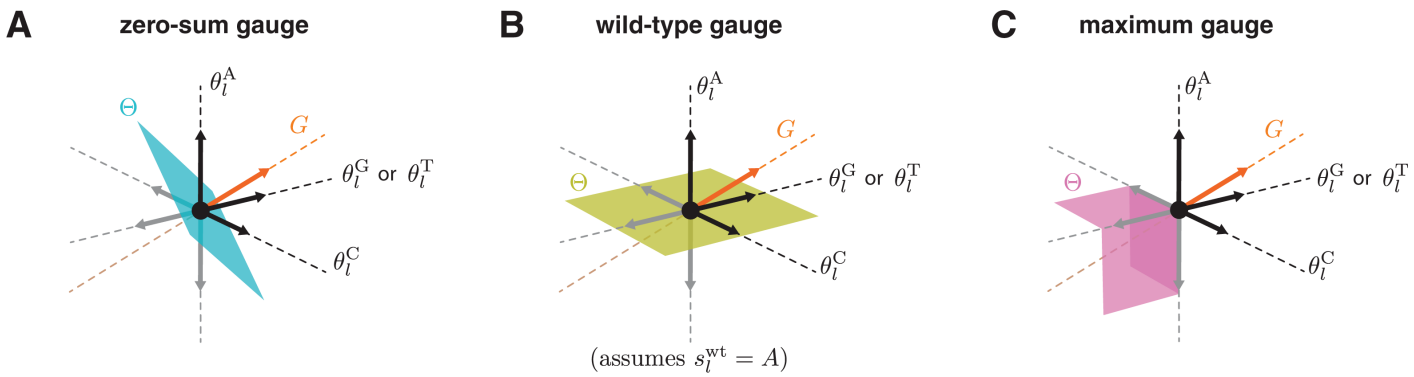

Fig 2. Geometry of gauge spaces for additive one-hot models. (A–C) Geometric representation of the gauge space Θ to which the additive parameters at each position l are restricted in the corresponding panel of Fig 1. Each of the four sequence features (θ_l^{A}, θ_l^{C}, θ_l^{G}, and θ_l^{T}) corresponds to a different axis. Note that the two axes for θ_l^{G} and θ_l^{T} are shown as one axis to enable 3D visualization. Black and gray arrows respectively denote unit vectors pointing in the positive and negative directions along each axis. G indicates the space of gauge transformations.

https://doi.org/10.1371/journal.pcbi.1012818.g002

will see, wild-type gauges are also readily defined for models with pairwise and higher-order interactions.

Fig 1C shows parameters expressed in what we call the "maximum gauge". In the maximum gauge, the constant parameter is equal to the activity of the highest-activity sequence, and additive parameters are the changes in activity that result from mutations away from this sequence. The maximum gauge is less common in the literature than the zero-sum gauge or wild-type gauge, but has been used in multiple publications [36,37].

Linear gauges. Here and throughout the rest of this paper we focus on linear gauges, i.e., choices of Θ that are linear subspaces of feature space. For example, the zero-sum gauge and wild-type gauge (Fig 2A and 2B) are two commonly used linear gauges, whereas the maximum gauge (Fig 2C) is not a linear gauge. Linear gauges are the most mathematically tractable family of gauges. Linear gauges also have the attractive property that the difference between any two parameter vectors in Θ is also in Θ. This property makes the comparison of models within the same gauge straight-forward.

Parameters can be fixed to any chosen linear gauge via a corresponding linear projection. Formally, for any linear gauge Θ there exists an $M \times M$ projection matrix P that projects any vector $\vec{\theta}_{\mathrm{init}}$ along the gauge space G to an equivalent vector $\vec{\theta}_{\mathrm{fixed}}$ that lies in Θ, i.e.

$$\vec{\theta}_{\mathrm{fixed}} = P\vec{\theta}_{\mathrm{init}}. \tag{9}$$

See S1 Text Sec 3 for a proof. We emphasize that P depends on the choice of Θ, and that P is an orthogonal projection only for the specific choice $\Theta = E$.

Parameters can also be gauge-fixed through a process of constrained optimization. Let Λ be any positive-definite $M \times M$ matrix, and let $\vec{y} = X\vec{\theta}_{\mathrm{init}}$ be the N-dimensional vector of model predictions on all sequences. Then Λ specifies a unique gauge-fixed set of parameters that preserves $\vec{y}$ via

$$\vec{\theta}_{\mathrm{fixed}} = \underset{\vec{\theta}\,:\,X\vec{\theta}=\vec{y}}{\mathrm{argmin}} \|\vec{\theta}\|_{\Lambda^2}, \quad \text{where} \quad \|\vec{\theta}\|_{\Lambda^2} = \vec{\theta}^{\mathsf{T}}\Lambda\vec{\theta}. \tag{10}$$

We call Λ the "penalization matrix" because it determines how much each direction in parameter space is penalized in Eq (10). The resulting gauge space comprises the set of vectors that minimize the Λ-norm $\|\vec{\theta}\|_\Lambda$ in each gauge orbit, where the gauge orbit of a parameter vector $\vec{\theta}$ is the set of equivalent vectors $\vec{\theta} + \vec{g}$ for all $g \in G$. The corresponding projection matrix is

$$P = \Lambda^{-1/2}(X\Lambda^{-1/2})^{+}X, \tag{11}$$

where '+' indicates the Moore-Penrose pseudoinverse. See S1 Text Sec 3 for a proof. In what follows, the connection between the penalization matrix Λ and the projection matrix P will be used to help interpret the constraints imposed by the gauge space Θ.

One consequence of Eq (10) is that parameter inference carried out using a positive-definite L_2 regularizer Λ on model parameters will result in gauge-fixed model parameters in the specific linear gauge determined by Λ (see S1 Text Sec 3). While it might then seem that L_2 regularization on parameter values during inference solves the gauge fixing problem, it is important to understand that such regularization will also change model predictions (i.e., the value of f), whereas gauge-fixing itself influences only the values of parameters while keeping the model predictions fixed. In addition, we show in S1 Text Sec 3 that, for any desired positive-definite regularizer on model predictions and choice of linear gauge Θ, we can construct a penalization matrix Λ that imposes the desired regularization on model predictions and yields inferred parameters in the desired gauge. Thus while L_2 regularization during parameter inference can simultaneously fix the gauge and regularize model predictions, the regularization imposed on model predictions does not constrain the choice of gauge.

Unified approach to gauge fixing

We now derive strategies for fixing the gauge of the all-order interaction model. We first introduce a geometric formulation of the all-order interaction model embedding. We then construct a parametric family of gauges for the all-order interaction model, and derive formulae for the corresponding projection and penalization matrices. Next, we highlight specific gauges of interest in this parametric family. We focus in particular on the "hierarchical gauges", which can be applied to a variety of commonly used models in addition to the all-order interaction model. The results provide explicit gauge-fixing formulae that can be applied to diverse quantitative models of sequence-function relationships.

All-order interaction models. To aid in our discussion of the all-order interaction model [Eq (4)], we define an augmented alphabet $\mathcal{A}' = \{*, c_1, \ldots, c_\alpha\}$, where $c_1, \ldots, c_\alpha$ are the characters in $\mathcal{A}$ and $*$ is a wild-card character that is interpreted as matching any character in $\mathcal{A}$. Let $\mathcal{S}'$ denote the set of sequences of length L comprising characters from $\mathcal{A}'$. For each augmented sequence $s' \in \mathcal{S}'$, we define the sequence feature $x_{s'}(s)$ to be 1 if a sequence s matches the pattern described by s' and to be 0 otherwise. In this way, each augmented sequence s' serves as a regular expression against which bona fide sequences are compared.

Assigning one parameter $\theta_{s'}$ to each of the $M = (\alpha + 1)^L$ augmented sequences s', the all-order interaction model can be expressed compactly as

$$f_{\text{all}}(s;\vec{\theta}) = \sum_{s' \in \mathcal{S}'} \theta_{s'} x_{s'}(s). \tag{12}$$

In this notation, the constant parameter θ_0 is written $\theta_{*\ldots*}$, each additive parameter θ_l^c is written $\theta_{*\ldots c\ldots*}$, each pairwise-interaction parameter $\theta_{ll'}^{cc'}$ is written $\theta_{*\ldots c\ldots c'\ldots*}$, and so on.

(Here c occurs at position l, c' occurs at position l', and $\cdots$ denotes a run of $*$ characters). We thus see that augmented sequences provide a convenient way to index the features and parameters of the all-order interaction model.

Next we observe that $x_{s'}$ can be expressed in a form that factorizes across positions. For each position l, we define $x_l^*(s) = 1$ for all sequences s and take $x_l^{c_1}, \ldots, x_l^{c_\alpha}$ to be the standard one-hot sequence features. $x_{s'}$ can then be written in the factorized form,

$$x_{s'}(s) = \prod_{l=1}^{L} x_l^{s'_l}(s). \tag{13}$$

From this it is seen that the embedding for the all-order interaction model, $\vec{x}_{\text{all}}(s)$, can be formulated geometrically as a tensor product:

$$\vec{x}_{\text{all}}(s) = \bigotimes_{l=1}^{L} \vec{x}'_l(s), \quad \text{where } \vec{x}'_l(s) = \begin{pmatrix} x_l^*(s) \\ x_l^{c_1}(s) \\ \vdots \\ x_l^{c_\alpha}(s) \end{pmatrix}. \tag{14}$$

See S1 Text Sec 4 for details.

Parametric family of gauges. We now define a useful parametric family of gauges for the all-order interaction model. As we will show, this family includes all of the most commonly used gauges in the literature (but not some less commonly used gauges, e.g., the maximum gauge [36,37]). Each gauge in this family is defined by two parameters, λ and p. λ is a non-negative real number that governs how much higher-order versus lower-order sequence features are penalized [in the sense of Eq (10)]. p is a probability distribution on sequence space that governs how strongly the specific characters at each position are penalized. This distribution is assumed to have the form

$$p(s) = p_1^{s_1} p_2^{s_2} \cdots p_L^{s_L}, \tag{15}$$

where p_l^c denotes the probability of character c at position l. This assumption excludes distributions that have correlations between positions. But as we show below, choosing appropriate values for λ and p nevertheless recovers the most commonly used linear gauges, including the zero-sum gauge, the wild-type gauge, and more.

Gauges in the parametric family have analytically tractable projection matrices because each gauge can be expressed as a tensor product of single-position gauge spaces. Let $\Theta_l^{\lambda,p}$ be the α-dimensional subspace of $\mathbb{R}^{\alpha+1}$ defined by

$$\Theta_l^{\lambda,p} = V_\lambda \oplus V_\perp^{p_l}, \tag{16}$$

where V_λ (a 1-dimensional subspace) and $V_\perp^{p_l}$ [an $(\alpha - 1)$-dimensional subspace] are defined by

$$V_\lambda = \text{span}\left\{ \begin{pmatrix} \lambda \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\}, \quad V_\perp^{p_l} = \left\{ \begin{pmatrix} 0 \\ v_{c_1} \\ \vdots \\ v_{c_\alpha} \end{pmatrix} : \sum_{i=1}^{\alpha} p_l^{c_i} v_{c_i} = 0 \right\}. \tag{17}$$

The full parametric gauge, denoted by $\Theta^{\lambda,p}$, is defined to be the tensor product of these single-position gauges:

$$\Theta^{\lambda,p} = \bigotimes_{l=1}^{L} \Theta_l^{\lambda,p}. \tag{18}$$

As detailed in S1 Text Sec 5, the corresponding projection matrix $P^{\lambda,p}$ is found to have elements given by

$$P^{\lambda,p}_{s't'} = \prod_{\substack{l \text{ s.t.} \\ s'_l \in \mathcal{A} \\ t'_l \in \mathcal{A}}} \left(\delta_{s'_l t'_l} - p_l^{t'_l}\eta\right) \times \prod_{\substack{l \text{ s.t.} \\ s'_l = * \\ t'_l \in \mathcal{A}}} \left(p_l^{t'_l}\eta\right) \times \prod_{\substack{l \text{ s.t.} \\ s'_l \in \mathcal{A} \\ t'_l = *}} (1-\eta) \times \prod_{\substack{l \text{ s.t.} \\ s'_l = * \\ t'_l = *}} \eta, \tag{19}$$

where $\eta = \lambda/(1+\lambda)$ and where the augmented sequences s' and t' respectively index rows and columns. We thus obtain an explicit formula for the projection matrix needed to project any parameter vector into any gauge in the parametric family.

Gauges in the parametric family also have penalization matrices of a simple diagonal form. Specifically, if $0 < \lambda < \infty$ and $p(s') > 0$ everywhere, Eq (10) is satisfied by the penalization matrix Λ having elements

$$\Lambda_{s't'} = p(s')\lambda^{o(s')}\delta_{s't'}, \tag{20}$$

where $o(s')$ denotes the order of interaction described by s' (i.e., the number of non-star characters in s') and $p(s')$ is defined as in Eq (15) but with $p_l^{s'_l} = 1$ when $s'_l = *$. See S1 Text Sec 5 for a proof. Note that, although Eq (20) does not hold when $\lambda = 0$, when $\lambda = \infty$, or when $p_l^c = 0$ for any choice of c and l, one can still interpret $\Theta^{\lambda,p}$ [which is well-defined in Eq (18) and Eq (19)] as arising from Eq (10) under a limiting series of penalization matrices Λ.

Trivial gauge. Choosing $\lambda = 0$ yields what we call the "trivial gauge". In the trivial gauge, $\theta_{s'} = 0$ if s' contains one or more star characters [by Eq (19)], and so the only nonzero parameters correspond to interactions of order L. As a result,

$$f_{\text{all}}(s, \vec{\theta}) = \theta_s \tag{21}$$

for every sequence $s \in \mathcal{S}$. Note in particular that the trivial gauge is unaffected by p. Thus, the trivial gauge essentially represents sequence-function relationships as catalogs of activity values, one value for every sequence. See S1 Text Sec 6 for details.

Euclidean gauge. Choosing $\lambda = \alpha$ and choosing p to be the uniform distribution recovers what we call the "euclidean gauge". In the euclidean gauge, the penalizing norm in Eq (10) is the standard euclidean norm, i.e.

$$||\vec{\theta}||_\Lambda^2 = \sum_{s'} \theta_{s'}^2. \tag{22}$$

In S1 Text Sec 6 we show that the euclidean gauge is equal to the embedding space E and that parameter inference using standard L_2 regularization (i.e. choosing Λ to be a positive multiple of the identity matrix) will yield parameters in E.

Equitable gauge. Choosing $\lambda = 1$ and letting p vary recovers what we call the "equitable gauge". In the equitable gauge, the penalizing norm is

$$||\vec{\theta}||_{\Lambda}^2 = \sum_{s'} p(s')\theta_{s'}^2 = \sum_{s'} \langle f_{s'}^2 \rangle_p = \sum_{s'} ||f_{s'}||_p^2, \tag{23}$$

where $f_{s'} = \theta_{s'} x_{s'}$ denotes the contribution to the activity landscape corresponding to the sequence feature s', $\langle \cdot \rangle_p$ denotes an average over sequences drawn from p, and $||f||_p^2 = \sum_{s \in \mathcal{S}} p(s) f(s)^2$ is the squared norm of a function f on sequence space with respect to p. The equitable gauge thus penalizes each parameter $\theta_{s'}$ in proportion to the fraction of sequences that parameter applies to. Equivalently, the equitable gauge can be thought of as minimizing the sum of the squared norms of the landscape contributions $||f_{s'}||_p^2$ rather than the squared norm of the parameter values themselves. Unlike the euclidean gauge, the equitable gauge accounts for the fact that different model parameters can affect vastly different numbers of sequences and can thereby have vastly different impacts on the activity landscape. See S1 Text Sec 6 for details.

Hierarchical gauge. Choosing arbitrary p and taking $\lambda \to \infty$ yields what we call the "hierarchical gauge". When expressed in the hierarchical gauge, model parameters obey the marginalization property,

$$\sum_{c_k} p_{l_k}^{c_k} \theta_{l_1 \ldots l_K}^{c_1 \ldots c_K} = 0, \tag{24}$$

for all interaction orders K, all choices of K positions $l_1, \ldots, l_K$, all choices of characters $c_1, \ldots, c_K$ at these positions, and all choices of index $k = 1, \ldots, K$. This marginalization property has important consequences that we now summarize. See S1 Text Sec 7 for proofs of these results.

A first consequence of Eq (24) is that, when parameters are expressed in the hierarchical gauge, the mean activity among sequences matched by an augmented sequence can be expressed as a simple sum of parameters. For example,

$$\langle f_{\text{all}} \rangle_p = \theta_0, \tag{25}$$

$$\langle f_{\text{all}} \,|\, c \text{ at } l \rangle_p = \theta_0 + \theta_l^c, \tag{26}$$

$$\langle f_{\text{all}} \,|\, c \text{ at } l, c' \text{ at } l' \rangle_p = \theta_0 + \theta_l^c + \theta_{l'}^{c'} + \theta_{ll'}^{cc'}, \tag{27}$$

and so on. Consequently, the parameters themselves can also be expressed in terms of differences of these average values. For instance, $\theta_l^c = \langle f_{\text{all}} \,|\, c \text{ at } l \rangle_p - \langle f_{\text{all}} \rangle_p$. Because p factorizes by position, conditioning on having particular characters in a subset of positions is equivalent to the probability distribution produced by drawing sequences from p and then fixing those positions in the drawn sequences to those specific characters. Thus, θ_l^c can also be interpreted as the average effect of mutating position l to character c when sequences are drawn from p. Similarly, $\theta_{ll'}^{cc'}$ is the average effect, when drawing sequences from p, of fixing the character at position l to c and the one at l' to c' beyond what would be expected based on the effects of changing l to c and l' to c' individually (i.e., epistasis). Higher-order coefficients have a similar interpretation. The hierarchical gauge thus provides an ANOVA-like decomposition of activity landscapes.

A second consequence of Eq (24) is that the activity landscape, when expressed in the hierarchical gauge, naturally decomposes into mutually orthogonal components. Let σ denote a set comprising all augmented sequences that have the same pattern of star and non-star

positions, and let $f_\sigma = \sum_{s' \in \sigma} \theta_{s'} x_{s'}$ be the corresponding component of f_{all}. These landscape components are p-orthogonal when expressed in the hierarchical gauge:

$$\langle f_\sigma f_\tau \rangle_p = \delta_{\sigma\tau} \sum_{s' \in \sigma} p(s') \theta_{s'}^2, \tag{28}$$

where σ and τ represent any two such sets of augmented sequences. One implication of this orthogonality relation is that the variance of the landscape (with respect to p) is the sum of contributions from interactions of different orders:

$$\text{var}_p[f] = \sum_{k=0}^{L} \text{var}_p[f_k], \tag{29}$$

where f_k denotes the sum of all k-order terms that contribute to f_{all}. Another implication is that the hierarchical gauge minimizes the variance attributable to different orders of interaction in a hierarchical manner: higher-order terms are prioritized for variance minimization over lower-order terms, and within a given order parameters are penalized in proportion to the fraction of sequences they apply to.

A third consequence of Eq (24) is that hierarchical gauges preserve the form of a large class of one-hot models that are equivalent to all-order interaction models with certain parameters fixed at zero (specifically, these models satisfy the condition that if a parameter for a sequence feature is fixed at zero, all higher-order sequence features contained within that sequence feature also have their parameters fixed at zero). These models, which we call the "hierarchical models," include all-order interaction models in which the parameters above a specified order are zero (e.g., additive models and pairwise-interaction models), but also include other models, such as nearest-neighbor interaction models. Projecting onto the hierarchical gauge (but not other parametric family gauges) is guaranteed to produce a parameter vector where the appropriate entries are still fixed to be zero.

Zero-sum gauge. The zero-sum gauge (illustrated in Figs 1A and 2A) is the hierarchical gauge for which p is the uniform distribution. The name of this gauge comes from the fact that, when p is uniform, Eq (24) becomes

$$\sum_{c_k} \theta_{l_1 \ldots l_K}^{c_1 \ldots c_K} = 0. \tag{30}$$

Prior studies [12,15] have characterized the zero-sum gauge for the all-order interaction model. Our formulation of the hierarchical gauge extends those findings and generalizes them to gauges defined by non-uniformly weighted sums of parameters.

Wild-type and generalized wild-type gauges. The wild-type gauge (illustrated in Figs 1B and 2B) is a hierarchical gauge that arises in the limit as p approaches an indicator function for some wild-type sequence, s^{wt}. In the wild-type gauge, only the parameters $\theta_{s'}$ for which s' matches s^{wt} receive any penalization, and all these penalized $\theta_{s'}$ (except for θ_0) are therefore driven to zero by minimization of the Λ-norm. Consequently, θ_0 quantifies the activity of the wild-type sequence, each θ_l^c quantifies the effect of a single mutation to the wild-type sequence, each $\theta_{ll'}^{cc'}$ quantifies the epistatic effect of two mutations to the wild-type sequence, and so on. However, seeing the wild-type gauge as a special case of the hierarchical gauge provides the possibility of generalizing the wild-type gauge by using a p that is not the indicator function on a single sequence but rather defines a distribution over one or more alleles per position that can be considered as being "wild-type" (equivalently, the frequencies of some

subset of position-specific characters are set to zero). Examples illustrating the utility of different choices for p are provided below. These gauges all inherit the property from the hierarchical gauge that their coefficients relate to the average effect of taking draws from the probability distribution defined by p and setting a subset of positions to the characters specified by that coefficient. More rigorously, these gauges are defined by considering the limit as $\epsilon \to 0^+$ of the hierarchical gauge with factorizable distribution

$$p_\epsilon(s) = \prod_l \left[(1-\epsilon) p_l^{s_l} + \frac{\epsilon}{\alpha} \right], \tag{31}$$

where the $p_l^{s_l} \geq 0$ are the position-specific factors of the desired nonnegative vector of probabilities p.

Applications

We now demonstrate the utility of our results on two example models of complex sequence-function relationships. First, we study how the parameters of the all-order interaction model behave under different parametric gauges in the context of a simulated landscape on short binary sequences. Although a number of studies have reported combinatorially complete landscapes in diverse biological systems (e.g., [46,47,59–65], focusing on this small simulated landscape allows us to better observe the nontrivial collective behavior that model parameters exhibit across different choices of gauge. Second, we examine the parameters of an empirical pairwise-interaction model for protein GB1 using the zero-sum and multiple generalized wild-type gauges. We observe how these different hierarchical gauges enable different interpretations of model parameters and facilitate the derivation of simplified models that are approximately correct in different localized regions of sequence space. The results provide intuition for the behavior of the various parametric gauges, and show in particular how hierarchical gauges can be used to explore and interpret real sequence-function relationships.

Gauge-fixing a simulated landscape on short binary sequences. To illustrate the consequences of choosing gauges in the parametric family, we consider a simulated random landscape on short binary sequences. Consider sequences of length $L = 3$ built from the alphabet $\mathcal{A} = \{0, 1\}$, and assume that the activities of these sequences are as shown in Fig 3A. The corresponding all-order interaction model has $(\alpha + 1)^L = 27$ parameters, which we index using augmented sequences: 1 constant parameter (θ_{***}), 6 additive parameters ($\theta_{0**}, \theta_{1**}, \theta_{*0*}, \theta_{*1*}, \theta_{**0}, \theta_{**1}$), 12 pairwise parameters ($\theta_{00*}, \theta_{01*}, \theta_{10*}, \theta_{11*}, \theta_{0*0}, \theta_{0*1}, \theta_{1*0}, \theta_{1*1}, \theta_{*00}, \theta_{*01}, \theta_{*10}, \theta_{*11}$), and 8 third-order parameters ($\theta_{000}, \theta_{001}, \theta_{010}, \theta_{011}, \theta_{100}, \theta_{101}, \theta_{110}, \theta_{111}$).

We now consider what happens to the values of these 27 parameters when they are expressed in different parametric gauges, $\Theta^{\lambda,p}$. Specifically, we assume that p is the uniform distribution (though analogous results hold for other choices of p) and vary the parameter λ from 0 to ∞ (equivalently η varies from 0 to 1). Note that each entry in the projection matrix $P^{\lambda,p}$ (Eq 19) is a cubic function of η due to $L = 3$. Consequently, each of the 27 gauge-fixed model parameters is a cubic function of η (Fig 3B). In the trivial gauge ($\lambda = 0,\ \eta = 0$), only the 8 third-order parameters are nonzero and the values of these parameters correspond to the values of the landscape at the 8 corresponding sequences. In the equitable gauge ($\lambda = 1,\ \eta = 1/2$), the spread of the 8 third-order parameters about zero is larger than that of the 12 pairwise parameters, which is larger than that of the 6 additive parameters, which is larger than that of the constant parameter. In the euclidean gauge ($\lambda = 2, \eta = 2/3$), the parameters of all orders exhibit a similar spread about zero. In the hierarchical gauge ($\lambda = \infty,\ \eta = 1$), the spread of the 8 third-order parameters about zero is smaller than that of the 12 pairwise parameters,

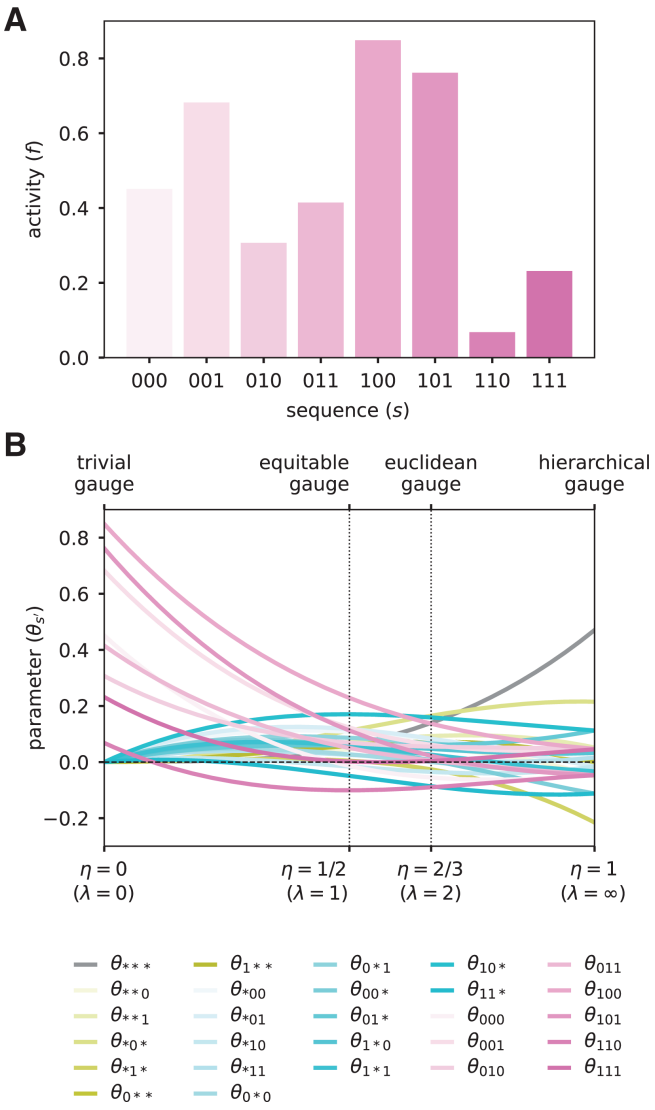

Fig 3. Binary landscape expressed in various parametric family gauges. (A) Simulated random activity landscape for binary sequences of length $L = 3$. (B) Parameters of the all-order interaction model for the binary landscape as functions of $\eta = \lambda/(1+\lambda)$. Values of η corresponding to different named gauges are indicated. Note: because the uniform distribution is assumed in all these gauges, the hierarchical gauge is also the zero-sum gauge.

https://doi.org/10.1371/journal.pcbi.1012818.g003

which is smaller than that of the 6 additive parameters, which is smaller than that of the constant parameter. Moreover, the marginalization and orthogonality properties of the hierarchical gauge fix certain parameters to be equal or opposite to each other, e.g., $\theta_{1**} = -\theta_{0**}$ and the third order parameters are all equal up to their sign, which depends only on whether the corresponding sequence feature has an even or odd number of "1"s.

This example illustrates generic features of the parametric gauges. For any all-order interaction model on sequences of length L, the entries of the projection matrix $P^{\lambda,p}$ will be L-order polynomials in η. Consequently, the values of model parameters, when expressed in the gauge $\Theta^{\lambda,p}$, will also be L-order polynomials in η. In the trivial gauge, only the highest-order parameters will be nonzero. In the equitable gauge, the spread about zero will tend to

be smaller for lower-order parameters relative to higher-order parameters. In the euclidean gauge, parameters of all orders will exhibit similar spread about zero. In the zero-sum gauge, the spread about zero will tend to be minimized for higher-order parameters relative to lower-order parameters. The nontrivial quantitative behavior of model parameters in different parametric gauges thus underscores the importance of choosing a specific gauge before quantitatively interpreting parameter values.

Hierarchical gauges of an empirical landscape for protein GB1. Projecting model parameters onto different hierarchical gauges can facilitate the exploration and interpretation of sequence-function relationships. To demonstrate this application of gauge fixing, we consider an empirical sequence-function relationship describing the binding of the GB1 protein to immunoglobulin G (IgG). Wu et al. [60] performed a deep mutational scanning experiment that measured how nearly all $20^4 = 160,000$ amino acid combinations at positions 39, 40, 41, and 54 of GB1 affect GB1 binding to IgG. These data report $\log_2$ enrichment values for each assayed sequence relative to the wild-type sequence at these positions, VDGV (Fig 4A and 4B). Using these data and least-squares regression, we inferred a pairwise-interaction model for $\log_2$ enrichment as a function of protein sequence at these $L = 4$ variable positions. The resulting model comprises 1 constant parameter, 80 additive parameters, and 2400 pairwise parameters (Fig 4C). While the model fits the data reasonably well (Fig 4D; $R^2 = 0.82$), the deviation from measurements is still greater than that expected by experimental uncertainty and can be further reduced by using a more complex model (e.g., one that includes a global epistasis nonlinearity [9,67]). Nevertheless, the pairwise-interaction model serves well to illuminate the utility of different gauge-fixing strategies. To understand the structure of the activity landscape described by the pairwise interaction model, we now examine the values of model parameters in multiple hierarchical gauges. Explicit formulae for implementing hierarchical gauges for pairwise-interaction models are given in S1 Text Sec 8.

Fig 4C shows the parameters of the pairwise interaction model expressed in the hierarchical gauge corresponding to a uniform probability distribution on sequence space (i.e., the zero-sum gauge). In the zero-sum gauge, the constant parameter θ_0 equals the average activity of all sequences. We observe $\theta_0 = -4.68$, indicating that a typical random sequence is depleted approximately 20-fold relative to the wild-type sequence, which the pairwise interaction model assigns a score of –.21. This finding confirms the expectation that a random sequence should be substantially less functional than the wild-type sequence.

The additive parameters in the zero-sum gauge are shown in the rectangular heat map in Fig 4C, and each additive parameter is equal to the difference between the mean activity of the set of sequences containing the corresponding amino acid at the relevant position relative to the mean activity of random sequences. We observe that the wild-type sequence receives positive or near-zero contributions at every position, including a contribution from the most positive additive parameter, corresponding to G at position 41. The additive parameters at positions 39, 40, and 54 that contribute to the wild-type sequence, however, are not the largest additive parameters at these positions. Moreover, the additive parameters that contribute to the wild-type sequence only sum to 2.32, meaning that even in the zero-sum gauge (which minimizes the variance due to pairwise parameters), of the total difference (4.47) between the wild-type score and the average sequence score, almost half (2.15) is due to contributions from pairwise parameters.

The pairwise parameters in the zero-sum gauge are shown in the triangular heat map in Fig 4C. Here, each pairwise parameter is equal to the difference between (i) the observed mean of the sequences containing the specified pair of characters at the specified pair of positions, and (ii) the expected mean activity based on the mean activity of sequences containing

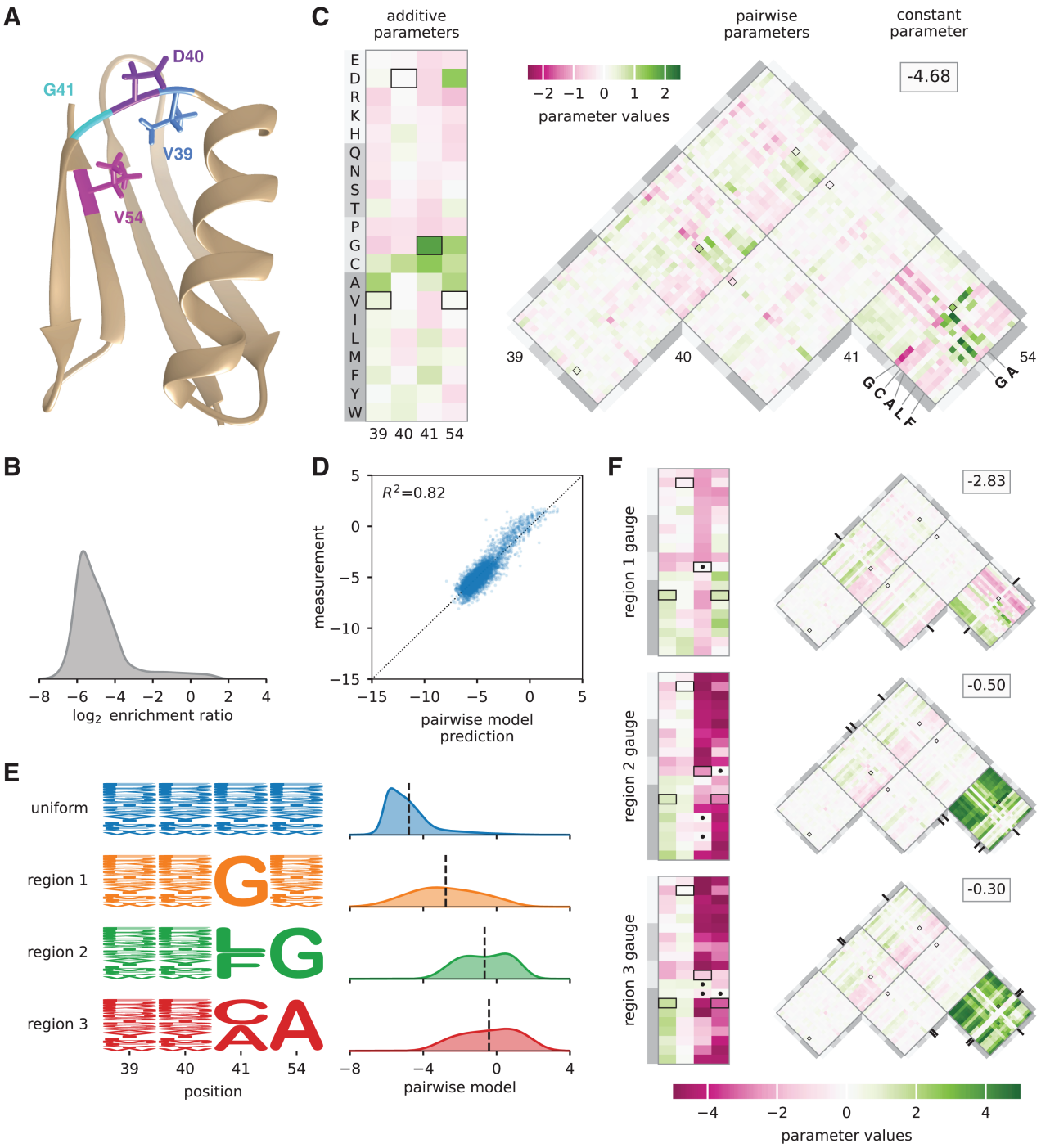

Fig 4. Landscape exploration using hierarchical gauges. (A) NMR structure of GB1, with residues V39, D40, G41, and V54 shown (PDB: 3GB1, from [66]). (B) Distribution of $\log_2$ enrichment relative to wild-type measured by [60] for nearly all 160,000 GB1 variants having mutations at positions 39, 40, 41, and 54. (C) Pairwise interaction model parameters inferred from the data of [60], expressed in the uniform hierarchical gauge (i.e., the zero-sum gauge). Boxes indicate parameters contributing to the wild-type sequence, VDGV. (D) Performance of pairwise-interaction model. Axes reflect $\log_2$ enrichment values relative to wild-type. Each dot represents a randomly chosen variant GB1 protein assayed by [60]. For clarity, only 5,000 of the ∼160,000 assayed GB1 variants are shown. (E) Probability logos [39] for uniform, region 1, region 2, and region 3 sequence distributions. Distributions of pairwise interaction model predictions for each region are also shown. (F) Model parameters expressed in the region 1, region 2, and region 3 hierarchical gauges. Dots and tick marks indicate region-specific constraints. Probability densities (panels B and D) were estimated using DEFT [45]. Pairwise interaction model parameters were inferred by least-squares regression using MAVE-NN [39]. Regions 1, 2, and 3 were defined based on [64]. NMR: nuclear magnetic resonance. GB1: domain B1 of protein G.

https://doi.org/10.1371/journal.pcbi.1012818.g004

the individual characters at those positions together with the grand mean activity. We observe that the three largest-magnitude pairwise contributions to the wildtype sequence are from the pair G41V54 (1.25), V39G41 (0.91), and D40G41 (-0.44), indicating that position 41 is a major hub of epistatic interactions contributing to the wild-type sequence. Moving to the landscape as a whole, we observe that the largest magnitude pairwise interactions link positions 41 and 54. Moreover, the strongest positive pairwise contributions are obtained when a small amino acid (G or A) is present at position 54, and a G, C, A, L, or P is present at position 41 (see also [49]). This finding provides insight into the chemical nature of the epistatic interactions that facilitate wild-type GB1 binding to IgG.

Previous work [64,68] identified three disjoint regions of sequence space (region 1, region 2, and region 3) that contain high-activity sequences as judged by the GB1 measurements of Wu et al. [60]. Region 1 comprises sequences with G at position 41; region 2 comprises sequences with L or F at position 41 and G at position 54; and region 3 comprises sequences with C or A at position 41 and A at position 54. To investigate the structure of the GB1 landscape within these three regions, we defined probability distributions that were uniform in each region of sequence space and zero outside (Fig 4E; see S1 Text Sec 8 for formal definitions of these regions). We then examined the values of the parameters of the pairwise-interaction model, with the parameters expressed in the hierarchical gauges corresponding to the probability distribution $p(s)$ for each of the three regions (the "region 1 hierarchical gauge", "region 2 hierarchical gauge", and "region 3 hierarchical gauge"). Since some characters at positions 41 and 54 have their frequencies set to zero, these hierarchical gauges are in fact generalized wild-type gauges, and the additive and pairwise parameters can be interpreted in terms of the mean effects of introducing mutations to these specific regions of sequences space.

In the region 1 hierarchical gauge (Fig 4F, top), the additive parameters for position 41 quantify the effect of mutations away from G, and the additive parameters for positions 39, 40, and 54 quantify the average effect of mutations conditional on G at position 41. From the additive parameters at position 54, we observe that cysteine (C) and hydrophobic residues (A, V, I, L, M, or F) increase binding, and that proline (P) and charged residues (E, D, R, K) decrease binding. From the additive parameters at position 40, we observe that amino acids with a 5-carbon or 6-carbon ring (H, F, Y, W) increase binding, suggesting the presence of structural constraints on side chain shape, rather than constraints on hydrophobicity or charge. The largest pairwise parameters all involve mutations from G at position 41 to another amino acid, and careful inspection of these pairwise parameters show that they are roughly equal and opposite to the additive effects of mutations at the other three positions. This indicates a classical form of masking epistasis, where the typical effect of a mutation at position 41 results in a more or less complete loss of function, after which mutations at the remaining three positions no longer have a substantial effect.

In the region 2 hierarchical gauge (Fig 4F, middle), the additive parameters at position 54 quantify the average effect of mutations away from G contingent on L or F at position 41, the additive parameters at position 41 quantify the average effects of mutations away from L or F contingent on G at position 54, and the additive parameters at positions 39 and 40 quantify the average effects of mutations contingent on L or F at position 41 and on G at position 54. From the values of the additive parameters, we observe that mutations away from L or F at position 41 in the presence of G at position 54 are typically strongly deleterious (mean effect –3.39), and that mutations away from G at position 54 in the presence of L or F at position 41 are also strongly deleterious (mean effect –3.75). However, the pairwise parameters linking positions 41 and 54 are strongly positive (mean effect 2.85), again indicating a masking effect where the first deleterious mutation at position 41 or 54 results in a more or less complete loss

of function, so that an additional mutation at the other position has little effect. Note also the similar but less extreme pattern of masking between the large effect mutations at positions 41 and 54 with the milder mutations at positions 40 and 41, whose interaction coefficients are of the opposite sign of the additive effects at positions 40 and 41. Similar results hold for the region 3 hierarchical gauge, where mutations at positions 41 and 54 have masking effects on each other as well as on mutations in the other two positions (Fig 4F, bottom). However, we can also contrast patterns of mutational effects between these regions. For example, mutating position 54 to G (a mutation leading towards region 2) on average has little effect in region 1 but would be deleterious in region 3. Similarly, if we consider mutations leading from region 2 to region 3, we can see that mutating 41 to C in region 2 typically has little effect whereas mutating 41 to A is more deleterious.

Besides using the interpretation of hierarchical gauge parameters as average effects of mutations to understand how mutational effects differ in different regions of sequence space, we hypothesized that by applying different hierarchical gauges to the pairwise interaction model, one might be able to obtain simple additive models that are accurate in different regions of sequence space. Our hypothesis was motivated by the fact that the parameters of all-order interaction models in the zero-sum gauge are chosen to maximize the fraction of variance in the sequence-function relationship that is explained by lower-order parameters. To test our hypothesis, we defined an additive model for each of the four hierarchical gauges described above (uniform, region 1, region 2, and region 3) by projecting pairwise interaction model parameters onto the hierarchical gauge for that region, then setting all the pairwise parameters to zero. We then evaluated the predictions of each additive model on sequences randomly drawn from each of the four corresponding probability distributions (uniform, region 1, region 2, and region 3). The results (Fig 5) show that the activities of sequences sampled uniformly from sequence space are best explained by the additive model derived from the zero-sum gauge, that the activities of region 1 sequences are best explained by the additive model derived from the region 1 hierarchical gauge, and so on for regions 2 and 3. In particular, additive models derived using region-specific gauges are far more accurate in their respective regions than is the additive model derived using the uniform (i.e., zero-sum) gauge. This shows that projecting a pairwise interaction model (or other hierarchical one-hot model) onto the hierarchical gauge corresponding to a specific region of sequence space can sometimes be used to obtain simplified models that approximate predictions by the original model in that region.

Discussion

Here we report a unified strategy for fixing the gauge of commonly used models of sequence-function relationships. First we defined a family of analytically tractable gauges for the all-order interaction model. We then derived explicit formulae for imposing any of these gauges on model parameters, and used these formulae to investigate the mathematical properties of these gauges. The results show that these linear gauges include all of the most commonly used gauges in the literature (even though most possible gauges, both linear and nonlinear, are not members of this family). We also find that a subset of these gauges (the hierarchical gauges) can be applied to diverse lower-order models including additive models, pairwise-interaction models, and higher-order interaction models.

Next, we demonstrated the family of gauges in two contexts: a simulated all-order interaction landscape on short binary sequences, and an empirical pairwise-interaction landscape for the protein GB1. The GB1 results, in particular, show how applying different hierarchical gauges can facilitate the biological interpretation of complex models of sequence-function

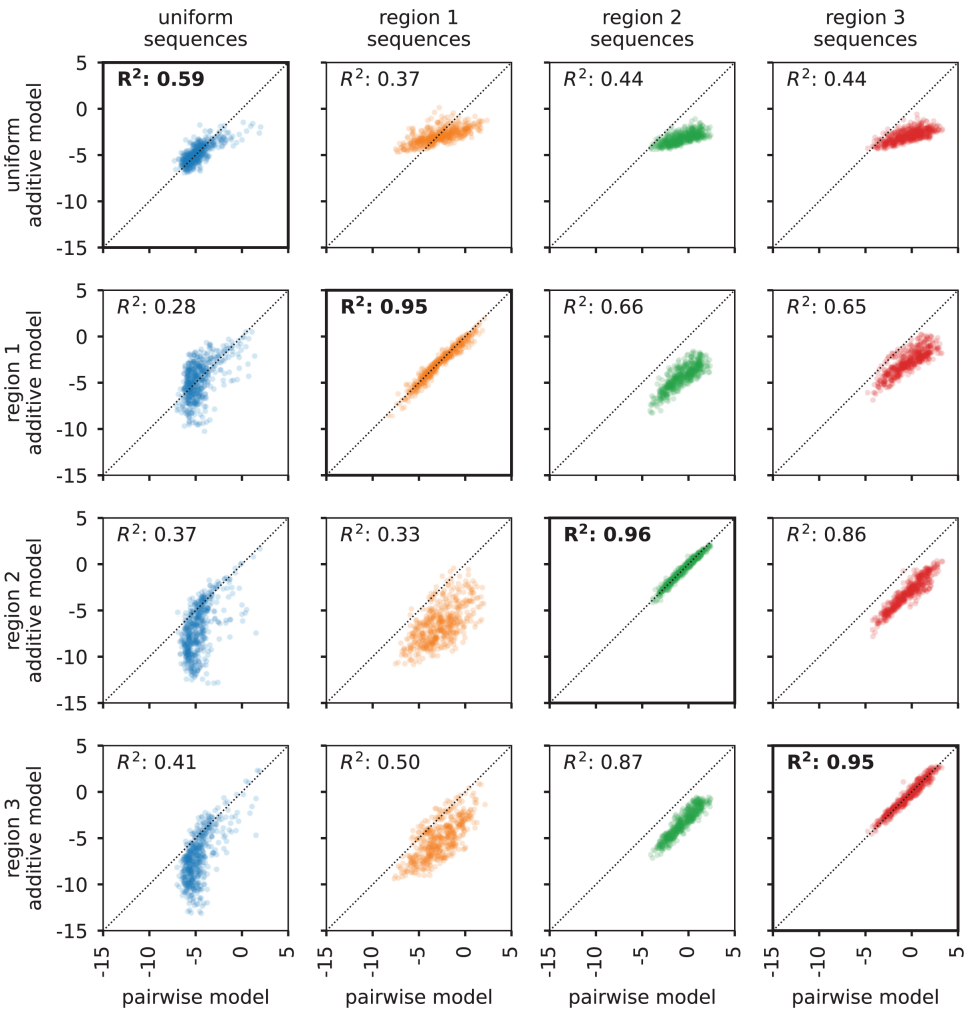

Fig 5. Model coarse-graining using hierarchical gauges. Shown are data for 500 random 4 aa sequences generated using each of the four distributions listed in Fig 4E (i.e., uniform, region 1, region 2, and region 3). Vertical axes show $\log_2$ enrichment (relative to wild-type) as predicted by additive models of GB1 derived by model truncation using region-specific zero-sum gauges (from Fig 4C and 4F). Horizontal axes show predictions of the full pairwise-interaction model. Diagonals indicate equality. GB1: domain B1 of protein G.

https://doi.org/10.1371/journal.pcbi.1012818.g005

relationships and the derivation of simplified models that are approximately correct in localized regions of sequence space.

Our study was limited to linear models of sequence-function relationships. Although linear models are used in many computational biology applications, more complex models are becoming increasingly common. For example, linear-nonlinear models (which include global epistasis models [9,67,69,70] and thermodynamic models [37,39,71–74]) are commonly used to describe fitness landscapes and/or sequence-dependent biochemical activities. The gauge-fixing strategies described here remain applicable to the linear part of linear-nonlinear models. We note, however, that such models often have additional gauge freedoms, such as diffeomorphic modes [75,76], that also need to be fixed before parameter values can be meaningfully interpreted.

Sloppy modes are another important issue to address when interpreting quantitative models of sequence-function relationships. Sloppy modes are directions in parameter space that (unlike gauge freedoms) do affect model predictions but are nevertheless poorly constrained by data [77,78]. Understanding the mathematical structure of sloppy modes, and developing systematic methods for fixing these modes, is likely to be more challenging than understanding gauge freedoms. This is because sloppy modes arise from a confluence of multiple factors: the mathematical structure of a model, the distribution of data in feature space, and measurement uncertainty. Nevertheless, understanding sloppy modes is likely to be as important in many applications as understanding gauge freedoms. We believe the study of sloppy modes in quantitative models of sequence-function relationships is an important direction for future research.

Deep neural network (DNN) models present perhaps the biggest challenge for parameter interpretation. DNN models have had remarkable success in quantitatively modeling biological sequence-function relationships, most notably in the context of protein structure prediction [79,80], but also in the context of other processes including transcriptional regulation [81–83], epigenetics [84–86], and mRNA splicing [87,88]. It remains unclear, however, how researchers might gain insights into the molecular mechanisms of biological processes from inferred DNN models. DNNs are by nature highly over-parameterized [89–91], making the direct interpretation of DNN parameters infeasible. Instead, a variety of attribution methods have been developed to facilitate DNN model interpretations [92–95]. Existing attribution methods can often be thought of as providing additive models that approximate DNN models in localized regions of sequence space [96], and the presence of gauge freedoms in these additive models needs to be addressed when interpreting attribution method output (as in [40,97]). We anticipate that, as DNN models become more widely adopted for mechanistic studies in biology, there will be a growing need for attribution methods that provide more complex quantitative models that approximate DNN models in localized regions of sequence space [16]. If so, a comprehensive mathematical understanding of gauge freedoms in parametric models of sequence-function relationships will be needed to aid in these DNN model interpretations.

Supporting information

S1 Text. Supporting information text.
(PDF)

Acknowledgments

We thank Peter Koo for helpful conversations and Samantha Petti for helpful comments on the manuscript.

Author contributions

Conceptualization: Anna Posfai, David M. McCandlish, Justin B. Kinney.

Formal analysis: Anna Posfai, Juannan Zhou, David M. McCandlish, Justin B. Kinney.

Funding acquisition: Juannan Zhou, David M. McCandlish, Justin B. Kinney.

Investigation: Anna Posfai, Juannan Zhou, David M. McCandlish, Justin B. Kinney.

Software: Anna Posfai, Justin B. Kinney.

Writing – original draft: Anna Posfai, David M. McCandlish, Justin B. Kinney.

Writing – review & editing: Anna Posfai, Juannan Zhou, David M. McCandlish, Justin B. Kinney.

References

1. Kinney JB, McCandlish DM. Massively parallel assays and quantitative sequence-function relationships. Annu Rev Genomics Hum Genet. 2019;20:99–127. https://doi.org/10.1146/annurev-genom-083118-014845 PMID: 31091417
2. Weinberger ED. Fourier and Taylor series on fitness landscapes. Biol Cybern. 1991;65:321–30.
3. Stadler PF. Landscapes and their correlation functions. J Math Chem. 1996;20:1–45.
4. Weinreich DM, Lan Y, Wylie CS, Heckendorn RB. Should evolutionary geneticists worry about higher-order epistasis?. Curr Opin Genet Dev. 2013;23(6):700–7. https://doi.org/10.1016/j.gde.2013.10.007 PMID: 24290990
5. Poelwijk FJ, Krishna V, Ranganathan R. The context-dependence of mutations: a linkage of formalisms. PLoS Comput Biol. 2016;12(6):e1004771. https://doi.org/10.1371/journal.pcbi.1004771 PMID: 27337695
6. Ferretti L, Schmiegelt B, Weinreich D, Yamauchi A, Kobayashi Y, Tajima F, et al. Measuring epistasis in fitness landscapes: the correlation of fitness effects of mutations. J Theor Biol. 2016;396:132–43.
7. Bank C, Matuszewski S, Hietpas RT, Jensen JD. On the (un)predictability of a large intragenic fitness landscape. Proc Natl Acad Sci U S A. 2016;113(49):14085–90. https://doi.org/10.1073/pnas.1612676113 PMID: 27864516
8. Poelwijk FJ, Socolich M, Ranganathan R. Learning the pattern of epistasis linking genotype and phenotype in a protein. Nat Commun. 2019;10(1):4213. https://doi.org/10.1038/s41467-019-12130-8 PMID: 31527666
9. Tareen A, Kooshkbaghi M, Posfai A, Ireland WT, McCandlish DM, Kinney JB. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. Genome Biol. 2022;23(1):98. https://doi.org/10.1186/s13059-022-02661-7 PMID: 35428271
10. Brookes D, Aghazadeh A, Listgarten J. On the sparsity of fitness functions and implications for learning. Proc Natl Acad Sci U S A. 2022;119:e2109649118. https://doi.org/10.1073/pnas.2109649118
11. Faure AJ, Lehner B, Miró Pina V, Serrano Colome C, Weghorn D. An extension of the Walsh-Hadamard transform to calculate and model epistasis in genetic landscapes of arbitrary shape and complexity. PLoS Comput Biol. 2024;20(5):e1012132. https://doi.org/10.1371/journal.pcbi.1012132 PMID: 38805561
12. Metzger BPH, Park Y, Starr TN, Thornton JW. Epistasis facilitates functional evolution in an ancient transcription factor. Elife. 2024;12:RP88737. https://doi.org/10.7554/eLife.88737 PMID: 38767330
13. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet. 2023;24(2):125–37. https://doi.org/10.1038/s41576-022-00532-2 PMID: 36192604
14. Koo PK, Majdandzic A, Ploenzke M, Anand P, Paul SB. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. PLoS Comput Biol. 2021;17(5):e1008925. https://doi.org/10.1371/journal.pcbi.1008925 PMID: 33983921
15. Park Y, Metzger B, Thornton J. The simplicity of protein sequence-function relationships. Nat Commun. 2024;15:7953 https://doi.org/10.1038/s41467-024-51895-5
16. Seitz EE, McCandlish DM, Kinney JB, Koo PK. Interpreting cis-regulatory mechanisms from genomic deep neural networks using surrogate models. Nat Mach Intell. 2024;6(6):701–13. https://doi.org/10.1038/s42256-024-00851-5 PMID: 39950082
17. Dupic T, Phillips AM, Desai MM. Protein sequence landscapes are not so simple: on reference-free versus reference-based inference. bioRxiv. 2024. https://doi.org/10.1101/2024.01.29.577800 PMID: 38352387
18. Jackson J, Okun L. Historical roots of gauge invariance. Rev Mod Phys. 2001;73(4):663–93.
19. Kinney JB, Tkacik G, Callan CG Jr. Precise physical models of protein-DNA interaction from high-throughput data. Proc Natl Acad Sci U S A. 2007;104(2):501–6. https://doi.org/10.1073/pnas.0609908104 PMID: 17197415
20. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A. 2009;106(1):67–72. https://doi.org/10.1073/pnas.0805923106 PMID: 19116270

21. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS ONE. 2011;6(12):e28766. https://doi.org/10.1371/journal.pone.0028766
22. Stormo GD. Maximally efficient modeling of DNA sequence motifs at all levels of complexity. Genetics. 2011;187(4):1219–24. https://doi.org/10.1534/genetics.110.126052 PMID: 21300846
23. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys. 2013;87(1):012707. https://doi.org/10.1103/PhysRevE.87.012707 PMID: 23410359
24. Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. J Comput Phys. 2014;276:341–56. https://doi.org/10.1016/j.jcp.2014.07.024
25. Stein RR, Marks DS, Sander C. Inferring pairwise interactions from biological data using maximum-entropy probability models. PLoS Comput Biol. 2015;11(7):e1004182. https://doi.org/10.1371/journal.pcbi.1004182 PMID: 26225866
26. Barton JP, De Leonardis E, Coucke A, Cocco S. ACE: adaptive cluster expansion for maximum entropy graphical model inference. Bioinformatics. 2016;32(20):3089–97. https://doi.org/10.1093/bioinformatics/btw328 PMID: 27329863
27. Haldane A, Flynn WF, He P, Levy RM. Coevolutionary landscape of kinase family proteins: Sequence probabilities and functional motifs. Biophys J. 2018;114:21–31. https://doi.org/10.1016/j.bpj.2017.10.028 PMID: 29320688
28. Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. Inverse statistical physics of protein sequences: a key issues review. Rep Prog Phys. 2018;81(3):032601. https://doi.org/10.1088/1361-6633/aa9965 PMID: 29120346
29. Haldane A, Levy RM. Influence of multiple-sequence-alignment depth on Potts statistical models of protein covariation. Phys Rev E. 2019;99(3–1):032405. https://doi.org/10.1103/PhysRevE.99.032405 PMID: 30999494
30. Zamuner S, Rios P. Interpretable neural networks based classifiers for categorical inputs. arXiv preprint 2021. http://arxiv.org/abs/2102.03202
31. Feinauer C, Meynard-Piganeau B, Lucibello C. Interpretable pairwise distillations for generative protein sequence models. PLoS Comput Biol. 2022;18(6):e1010219. https://doi.org/10.1371/journal.pcbi.1010219 PMID: 35737722
32. Gerardos A, Dietler N, Bitbol A-F. Correlations from structure and phylogeny combine constructively in the inference of protein partners from sequences. PLoS Comput Biol. 2022;18(5):e1010147. https://doi.org/10.1371/journal.pcbi.1010147 PMID: 35576238
33. Hsu C, Nisonoff H, Fannjiang C, Listgarten J. Learning protein fitness models from evolutionary and assay-labeled data. Nat Biotechnol. 2022;40(7):1114–22. https://doi.org/10.1038/s41587-021-01146-5 PMID: 35039677
34. Feinauer C, Borgonovo E. Mean dimension of generative models for protein sequences. bioRxiv preprint 2022. https://doi.org/10.1101/2022.12.12.520028
35. Rube HT, Rastogi C, Feng S, Kribelbauer JF, Li A, Becerra B, et al. Prediction of protein-ligand binding affinity from sequencing data with interpretable machine learning. Nat Biotechnol. 2022;40(10):1520–7. https://doi.org/10.1038/s41587-022-01307-0 PMID: 35606422
36. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol. 1987;193(4):723–50. https://doi.org/10.1016/0022-2836(87)90354-8 PMID: 3612791
37. Kinney JB, Murugan A, Callan CG Jr, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc Natl Acad Sci U S A. 2010;107(20):9158–63. https://doi.org/10.1073/pnas.1004290107 PMID: 20439748
38. Posfai A, McCandlish DM, Kinney JB. Symmetry, gauge freedoms, and the interpretability of sequence-function relationships. bioRxiv. 2024. https://doi.org/10.1101/2024.05.12.593774 PMID: 38798625
39. Tareen A, Kinney JB. Logomaker: beautiful sequence logos in Python. Bioinformatics. 2020;36(7):2272–4. https://doi.org/10.1093/bioinformatics/btz921 PMID: 31821414
40. Majdandzic A, Rajesh C, Koo PK. Correcting gradient-based interpretations of deep neural networks for genomics. Genome Biol. 2023;24(1):109. https://doi.org/10.1186/s13059-023-02956-3 PMID: 37161475
41. Busby S, Ebright RH. Transcription activation by catabolite activator protein (CAP). J Mol Biol. 1999;293(2):199–213. https://doi.org/10.1006/jmbi.1999.3161 PMID: 10550204
42. Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics. 2006;22(14):e141-9. https://doi.org/10.1093/bioinformatics/btl223 PMID: 16873464

43. Rube H, Rastogi C, Kribelbauer J, Bussemaker H. A unified approach for quantifying and interpreting DNA shape readout by transcription factors. Molecul Syst Biol. 2018;14:e7902. https://doi.org/10.15252/msb.20177902 PMID: 29472273
44. Hu Y, Tareen A, Sheu Y, Ireland W, Speck C, Li H, et al. Evolution of DNA replication origin specification and gene silencing mechanisms. Nat Commun. 2020;11:5175. https://doi.org/10.1038/s41467-020-18964-x PMID: 33056978
45. Chen W-C, Tareen A, Kinney JB. Density estimation on small data sets. Phys Rev Lett. 2018;121(16):160605. https://doi.org/10.1103/PhysRevLett.121.160605 PMID: 30387642
46. Skalenko KS, Li L, Zhang Y, Vvedenskaya IO, Winkelman JT, Cope AL, et al. Promoter-sequence determinants and structural basis of primer-dependent transcription initiation in Escherichia coli. Proc Natl Acad Sci U S A. 2021;118:e2106388118. https://doi.org/10.1073/pnas.2106388118 PMID: 34187896
47. Pukhrambam C, Molodtsov V, Kooshkbaghi M, Tareen A, Vu H, Skalenko KS, et al. Structural and mechanistic basis of σ-dependent transcriptional pausing. Proc Natl Acad Sci U S A. 2022;119(23):e2201301119. https://doi.org/10.1073/pnas.2201301119 PMID: 35653571
48. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, et al. High-resolution mapping of protein sequence-function relationships. Nat Methods. 2010;7(9):741–6. https://doi.org/10.1038/nmeth.1492 PMID: 20711194
49. Olson CA, Wu NC, Sun R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Curr Biol. 2014;24(22):2643–51. https://doi.org/10.1016/j.cub.2014.09.072 PMID: 25455030
50. Adams RM, Mora T, Walczak AM, Kinney JB. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. Elife. 2016;5:e23156. https://doi.org/10.7554/eLife.23156 PMID: 28035901
51. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol. 2019;20(1):223. https://doi.org/10.1186/s13059-019-1845-6 PMID: 31679514
52. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell. 2020;182(5):1295-1310.e20. https://doi.org/10.1016/j.cell.2020.08.012 PMID: 32841599
53. Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol. 2009;27(12):1173–5. https://doi.org/10.1038/nbt.1589 PMID: 19915551
54. Patwardhan R, Hiatt J, Witten D, Kim M, Smith R, May D, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012;30:265–70. https://doi.org/10.1038/nbt.2136 PMID: 22371081
55. Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci U S A. 2012;109(47):19498–503. https://doi.org/10.1073/pnas.1210678109 PMID: 23129659
56. Julien P, Minana B, Baeza-Centurion P, Valcarcel J, Lehner B. The complete local genotype-phenotype landscape for the alternative splicing of a human exon. Nat Commun. 2016;7:11558. https://doi.org/10.1038/ncomms11558 PMID: 27161764
57. Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell R, et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat Commun. 2019;10:3583. https://doi.org/10.1038/s41467-019-11356-5 PMID: 31395865
58. Urtecho G, Insigne K, Tripp A, Brinck M, Lubock N, Kim H, et al. Genome-wide functional characterization of Escherichia coli promoters and regulatory elements responsible for their function. eLife. 2023;12:RP92558. https://doi.org/10.7554/eLife.92558
59. Podgornaia AI, Laub MT. Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. Science. 2015;347(6222):673–7. https://doi.org/10.1126/science.1257360 PMID: 25657251
60. Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. Adaptation in protein fitness landscapes is facilitated by indirect paths. Elife. 2016;5:e16965. https://doi.org/10.7554/eLife.16965 PMID: 27391790
61. Winkelman JT, Vvedenskaya IO, Zhang Y, Zhang Y, Bird JG, Taylor DM, et al. Multiplexed protein-DNA cross-linking: scrunching in transcription start site selection. Science. 2016;351(6277):1090–3. https://doi.org/10.1126/science.aad6881 PMID: 26941320
62. Wong MS, Kinney JB, Krainer AR. Quantitative activity profile and context dependence of all human 5' splice sites. Mol Cell. 2018;71(6):1012-1026.e3. https://doi.org/10.1016/j.molcel.2018.07.033 PMID: 30174293

63. Baeza-Centurion P, Miñana B, Schmiedel JM, Valcárcel J, Lehner B. Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. Cell. 2019;176(3):549-563.e23. https://doi.org/10.1016/j.cell.2018.12.010 PMID: 30661752
64. Zhou J, McCandlish DM. Minimum epistasis interpolation for sequence-function relationships. Nat Commun. 2020;11(1):1782. https://doi.org/10.1038/s41467-020-15512-5 PMID: 32286265
65. Zhou J, Wong MS, Chen W-C, Krainer AR, Kinney JB, McCandlish DM. Higher-order epistasis and phenotypic prediction. Proc Natl Acad Sci U S A. 2022;119(39):e2204233119. https://doi.org/10.1073/pnas.2204233119 PMID: 36129941
66. Kuszewski J, Gronenborn A, Clore G. Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. J Am Chem Soc. 1999;121:2337–8.
67. Otwinowski J, McCandlish DM, Plotkin JB. Inferring the shape of global epistasis. Proc Natl Acad Sci U S A. 2018;115(32):E7550–8. https://doi.org/10.1073/pnas.1804015115 PMID: 30037990
68. Rozhonová H, Martí-Gómez C, McCandlish D, Payne J. Protein evolvability under rewired genetic codes. PLoS Biology. 2024;22(5):e3002594. https://doi.org/10.1371/journal.pbio.3002594
69. Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, et al. Local fitness landscape of the green fluorescent protein. Nature. 2016;533(7603):397–401. https://doi.org/10.1038/nature17995 PMID: 27193686
70. Sailer ZR, Harms MJ. Detecting high-order epistasis in nonlinear genotype-phenotype maps. Genetics. 2017;205(3):1079–88. https://doi.org/10.1534/genetics.116.195214 PMID: 28100592
71. Mogno I, Kwasnieski JC, Cohen BA. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. Genome Res. 2013;23(11):1908–15. https://doi.org/10.1101/gr.157891.113 PMID: 23921661
72. Otwinowski J. Biophysical inference of epistasis and the effects of mutations on protein stability and function. Mol Biol Evol. 2018;35(10):2345–54. https://doi.org/10.1093/molbev/msy141 PMID: 30085303
73. Belliveau NM, Barnes SL, Ireland WT, Jones DL, Sweredoski MJ, Moradian A, et al. Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. Proc Natl Acad Sci U S A. 2018;115(21):E4796–805. https://doi.org/10.1073/pnas.1722055115 PMID: 29728462
74. Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. Nature. 2022;604(7904):175–83. https://doi.org/10.1038/s41586-022-04586-4 PMID: 35388192
75. Kinney JB, Atwal GS. Parametric inference in the large data limit using maximally informative models. Neural Comput. 2014;26(4):637–53. https://doi.org/10.1162/NECO_a_00568 PMID: 24479782
76. Atwal G, Kinney J. Learning quantitative sequence-function relationships from massively parallel experiments. J Statist Phys. 2016;162:1203–43. https://doi.org/10.1007/s10955-015-1398-3
77. Machta BB, Chachra R, Transtrum MK, Sethna JP. Parameter space compression underlies emergent theories and predictive models. Science. 2013;342(6158):604–7. https://doi.org/10.1126/science.1238723 PMID: 24179222
78. Transtrum MK, Machta BB, Brown KS, Daniels BC, Myers CR, Sethna JP. Perspective: sloppiness and emergent theories in physics, biology, and beyond. J Chem Phys. 2015;143(1):010901. https://doi.org/10.1063/1.4923066 PMID: 26156455
79. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844
80. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123–30. https://doi.org/10.1126/science.ade2574 PMID: 36927031
81. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18(10):1196–203. https://doi.org/10.1038/s41592-021-01252-x PMID: 34608324
82. Karbalayghareh A, Sahin M, Leslie CS. Chromatin interaction-aware gene regulatory modeling with graph attention networks. Genome Res. 2022;32(5):930–44. https://doi.org/10.1101/gr.275870.121 PMID: 35396274
83. de Almeida BP, Reiter F, Pagani M, Stark A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat Genet. 2022;54(5):613–24. https://doi.org/10.1038/s41588-022-01048-5 PMID: 35551305
84. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat Genet. 2021;53(3):354–66. https://doi.org/10.1038/s41588-021-00782-6 PMID: 33603233

85. Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet. 2022;54(7):940–9. https://doi.org/10.1038/s41588-022-01102-2 PMID: 35817977

86. Toneyan S, Tang Z, Koo PK. Evaluating deep learning for predicting epigenomic profiles. Nat Mach Intell. 2022;4(12):1088–100. https://doi.org/10.1038/s42256-022-00570-9 PMID: 37324054

87. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015 PMID: 30661751

88. Cheng J, Çelik MH, Kundaje A, Gagneur J. MTSplice predicts effects of genetic variants on tissue-specific splicing. Genome Biol. 2021;22(1):94. https://doi.org/10.1186/s13059-021-02273-7 PMID: 33789710

89. Raghu M, Poole B, Kleinberg J, Ganguli S, Dickstein JS. On the expressive power of deep neural networks. Proc Mach Learn Res. 2017;70:2847–54. https://doi.org/10.48550/arXiv.1905.12207

90. Kaplan J, McCandlish S, Henighan T, Brown T, Chess B, Child R. Scaling laws for neural language models. arXiv preprint 2020. http://arxiv.org/abs/2001.08361

91. Nakkiran P, Kaplun G, Bansal Y, Yang T, Barak B, Sutskever I. Deep double descent: where bigger models and more data hurt. J Statist Mech. 2021;2021:124003.

92. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint 2013. http://arxiv.org/abs/1312.6034

93. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. Proc Mach Learn Res. 2017;70:3145–53. https://doi.org/10.48550/arXiv.1704.02685

94. Lundberg S, Lee S. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4768–77. https://doi.org/10.48550/arXiv.1705.07874

95. Jha A, K Aicher J, R Gazzara M, Singh D, Barash Y. Enhanced Integrated Gradients: improving interpretability of deep learning models using splicing codes as a case study. Genome Biol. 2020;21(1):149. https://doi.org/10.1186/s13059-020-02055-7 PMID: 32560708

96. Han T, Srinivas S, Lakkaraju H. Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations. arXiv preprint 2022. http://arxiv.org/abs/2206.01254

97. Sasse A, Chikina M, Mostafavi S. Quick and effective approximation of in silico saturation mutagenesis experiments with first-order Taylor expansion. iScience 2024;20:110807. https://doi.org/10.1016/j.isci.2024.110807 PMID: 39286491

Supplemental Information for: Gauge fixing for sequence-function relationships

Anna Posfai, Juannan Zhou, David M. McCandlish, Justin B. Kinney

February 5, 2025

Contents

1 Sequences and embeddings

Definition 1. *The alphabet $\mathcal{A}$ is an ordered set of α characters $(c_1, \ldots, c_\alpha)$.*

Definition 2. *Sequence space $\mathcal{S}$ is the set of all sequences of length L built from characters in the alphabet $\mathcal{A}$. We use N to denote the number of sequences in $\mathcal{S}$, and s_l to denote the character at position l in sequence s.*

Definition 3. *An embedding $\vec{x}$ is a mapping from $\mathcal{S}$ to a real vector space $V = \mathbb{R}^M$. We use M throughout to denote the dimension of V.*

Definition 4. *The embedding space E of an embedding $\vec{x} : \mathcal{S} \to V$ is a vector space defined by $E \equiv \text{span}(\{\vec{x}(s) : s \in \mathcal{S}\})$. E is also called the span of $\vec{x}$.*

Definition 5. *The design matrix X of an embedding $\vec{x} : \mathcal{S} \to V$ is an $N \times M$ matrix having elements $X_{ij} = [\vec{x}(s_i)]_j$, where $i = 1, \ldots, N$ indexes all sequences in $\mathcal{S}$ (the specific order does not matter) and $j = 1, \ldots, M$ indexes the dimensions of V.*

2 Gauge freedoms

Definition 6. *The space of gauge freedoms (a.k.a. freedom space) G of an embedding $\vec{x} : \mathcal{S} \to V$ is a vector space defined by*

$$G \equiv \left\{\vec{g} \in V : \vec{g}^\top \vec{x}(s) = 0 \;\forall\; s \in \mathcal{S}\right\}. \tag{1}$$

We use γ to denote the dimension of G.

Claim 1. *Let $\vec{x}$ be an embedding, E be the embedding space of $\vec{x}$, and G be the freedom space of $\vec{x}$. Then G is the orthogonal complement of E, i.e., $G = E^\perp$.*

Proof. First we show that $G \subseteq E^\perp$. Consider any $\vec{g} \in G$. By Definition 6, $\vec{g}^\top \vec{x}(s) = 0$ for every $s \in \mathcal{S}$, and so $\vec{g}$ is orthogonal to E, and thus $\vec{g} \in E^\perp$. This establishes that $G \subseteq E^\perp$. Next we show that $E^\perp \subseteq G$. If $\vec{v} \in E^\perp$, then $\vec{v}^\top \vec{w} = 0$ for every $\vec{w} \in S$. In particular, $\vec{v}^\top \vec{x}(s) = 0$ for every $s \in \mathcal{S}$, implying that $\vec{v} \in G$. This establishes that $E^\perp \subseteq G$, proving the claim. □

Gauge freedoms of pairwise-interaction models. Eq. 7 in the main text appears to gives 2α linear relations for every pair of positions $l < l'$, namely

$$x_l^c(s) = \sum_{c' \in \mathcal{A}} x_{ll'}^{cc'}(s) \quad (\alpha \text{ linear relations}), \tag{2}$$

$$x_{l'}^{c'}(s) = \sum_{c \in \mathcal{A}} x_{ll'}^{cc'}(s) \quad (\alpha \text{ linear relations}). \tag{3}$$

However, these linear relations are not independent, as summing Eq. 2 over all $c \in \mathcal{A}$ gives the same linear relation as summing Eq. 3 over all $c' \in \mathcal{A}$:

$$1 = \sum_{c,c' \in \mathcal{A}} x_{ll'}^{cc'}(s). \tag{4}$$

Eq. 4 is in fact the only dependency between the 2α linear relations in Eq. 2 and Eq. 3. Therefore, there are actually $2\alpha - 1$ independent gauge freedoms per pair of positions $l < l'$, and there are $\binom{L}{2}$ such pairs of positions. We thus find a total of

$$\gamma_{\text{pairwise}} = L + \binom{L}{2}(2\alpha - 1) \tag{5}$$

gauge freedoms for the pairwise-interaction model.

3 Linear gauges

Definition 7. *A linear gauge space Θ for an embedding $\vec{x} : \mathcal{S} \to V$ having freedom space G is a vector space such that, for all $\vec{v} \in V$, there is a unique decomposition $\vec{v} = \vec{\theta} + \vec{g}$ where $\vec{\theta} \in \Theta$ and $\vec{g} \in G$. All gauge spaces discussed in what follows are assumed to be linear. Note that* $\dim \Theta = M - \gamma$.

Definition 8. *A projection matrix P that projects from a vector space V into a vector space V_1 along a vector space V_2 is a matrix such that, for all $\vec{v} \in V$, $P\vec{v} \in V_1$ and $\vec{v} - P\vec{v} \in V_2$.*

Claim 2. *Let $\vec{x} : \mathcal{S} \to V$ be an embedding, let G be the freedom space of $\vec{x}$, and let Θ be a gauge space of $\vec{x}$. Then there is a unique a projection matrix P that projects V into Θ along G. Moreover, the matrix $Q \equiv I_M - P$ is the unique matrix that projects V into G along Θ.*

Proof. Let $\vec{e}_1, \ldots, \vec{e}_\gamma$ be a basis for G, and $\vec{f}_1, \ldots, \vec{f}_{M-\gamma}$ be a basis for Θ. Using these basis vectors as columns, define the $M \times \gamma$ matrix $E \equiv (\vec{e}_1, \ldots, \vec{e}_\gamma)$ and the $M \times (M - \gamma)$ matrix $F \equiv (\vec{f}_1, \ldots, \vec{f}_{M-\gamma})$. Choose any $\vec{v} \in V$. By Definition 7, $\vec{v}$ can be uniquely decomposed as

$$\vec{v} = \vec{\theta} + \vec{g} \tag{6}$$

where $\vec{\theta} \in \Theta$ and $\vec{g} \in G$. Because the columns of E and F provide bases for G and Θ, respectively, there is a γ-dimensional vector $\vec{a}$ and an $(M - \gamma)$-dimensional vector $\vec{b}$ such that $\vec{g} = E\vec{a}$ and $\vec{\theta} = F\vec{b}$. Therefore,

$$\vec{v} = E\vec{a} + F\vec{b} \tag{7}$$

$$= (E \;\; F) \begin{pmatrix} \vec{a} \\ \vec{b} \end{pmatrix} \tag{8}$$

$$\Rightarrow \begin{pmatrix} \vec{a} \\ \vec{b} \end{pmatrix} = (E \;\; F)^{-1} \vec{v} \tag{9}$$

$$\Rightarrow \vec{\theta} = (0_{M\times\gamma} \;\; F) \begin{pmatrix} \vec{a} \\ \vec{b} \end{pmatrix} \tag{10}$$

$$= (0_{M\times\gamma} \;\; F) (E \;\; F)^{-1} \vec{v}, \tag{11}$$

where $(E \;\; F)$ is the $M \times M$ matrix given by horizontally concatenating E and F, and $0_{M\times\gamma}$ is an $M \times \gamma$ matrix of zeros. A projection matrix P from V into Θ along G therefore exists and is given by

$$P = (0_{M\times\gamma} \;\; F) (E \;\; F)^{-1}. \tag{12}$$

To see that Q projects V into G along Θ, simply note that, for any $\vec{v} \in V$, $Q\vec{v} = \vec{v} - P\vec{v} \in G$ and $\vec{v} - Q\vec{v} = P\vec{v} \in \Theta$. To prove that P is unique, assume that there is another matrix $P' \neq P$ that projects into Θ along G. There must therefore be a $\vec{v} \in V$ such that $P'\vec{v} \neq P\vec{v}$. By Definition 8, $P'\vec{v} \in \Theta$ and $\vec{v} - P'\vec{v} \in G$. But $P\vec{v} \in \Theta$ and $\vec{v} - P\vec{v} \in G$ as well, and by Definition 7, the decomposition of $\vec{v}$ into a component in Θ and a component in G is unique, implying that $P\vec{v} = P'\vec{v}$, which is a contradiction. An analogous proof shows that Q is unique. □

Definition 9. *A metric Λ on an M-dimensional vector space V is a symmetric positive-definite $M \times M$ matrix. The Λ-inner product of two vectors $\vec{v}, \vec{w} \in V$ is defined to be*

$$\langle \vec{v}, \vec{w} \rangle_\Lambda \equiv \vec{v}^\top \Lambda \vec{w}. \tag{13}$$

The Λ-norm of a vector $\vec{v} \in V$ is defined to be

$$||\vec{v}||_\Lambda \equiv \sqrt{\langle \vec{v}, \vec{v} \rangle_\Lambda} = \sqrt{\vec{v}^\top \Lambda \vec{w}}. \tag{14}$$

$\vec{v}$ and $\vec{w}$ are Λ-orthogonal if and only if $\langle \vec{v}, \vec{w} \rangle_\Lambda = 0$.

Definition 10. *An orthogonalizing metric, Λ, for a gauge space Θ and freedom space G is a metric for which all $\vec{g} \in G$ are Λ-orthogonal to all $\vec{\theta} \in \Theta$.*

Claim 3. *Let $\vec{x} : \mathcal{S} \to V$ be an embedding, G be the freedom space of $\vec{x}$, and Θ be a gauge space of $\vec{x}$. Then there exists an orthogonalizing metric Λ for Θ and G.*

Proof. Let P be the projection matrix from V into Θ along G, let $Q = I_M - P$, and define the matrix $\Lambda \equiv P^\top P + Q^\top Q$. It is a simple matter to show that Λ is symmetric and positive-definite, and is thus a metric. Using $P^2 = P$, $Q^2 = Q$ and $Q = I_M - P$, it is readily shown that PQ are both equal to the zero matrix and hence Λ satisfies $\Lambda = P^\top \Lambda P + Q^\top \Lambda Q$. Therefore, by Claim 25 in the Appendix, Λ orthogonalizes Θ and G. □

Claim 4. *Let $\vec{x} : \mathcal{S} \to V$ be an embedding, X be the design matrix of $\vec{x}$, G be the freedom space of $\vec{x}$, Θ be a gauge space of $\vec{x}$, and Λ be a metric that orthogonalizes Θ and G. Then for any vector $\vec{\theta}_{\text{init}} \in V$, the vector $\vec{\theta}^*$ in the gauge orbit of $\vec{\theta}_{\text{init}}$ that has minimal Λ-norm lies in Θ.*

Proof. Let θ^* be the unique element of Θ that lies in the gauge orbit of $\vec{\theta}_{\text{init}}$. Because Λ orthogonalizes Θ and G, $\vec{g}^\top \Lambda \vec{\theta}^* = 0$ for all $\vec{g} \in G$. By Claim 25,

$$||\vec{\theta}^* + \vec{g}||_\Lambda^2 \geq ||\vec{\theta}^*||_\Lambda^2, \tag{15}$$

with equality obtaining only when $\vec{g} = \vec{0}$. This proves that $\vec{\theta}^*$ is the unique vector with the smallest Λ-norm of any vector in the gauge orbit of $\vec{\theta}_{\text{init}}$. □

Claim 5. *Let $\vec{x} : \mathcal{S} \to V$ be an embedding, X be the design matrix of $\vec{x}$, G be the freedom space of $\vec{x}$, Θ be a gauge space of $\vec{x}$, and Λ be a metric that orthogonalizes Θ and G. Then the matrix P that projects along G into Θ is given by $P = \Lambda^{-1/2}(X\Lambda^{-1/2})^+ X$.*

Proof. Consider the transformed embedding $\vec{x}' = \Lambda^{-1/2}\vec{x}$, which corresponds to the transformed design matrix $X' = X\Lambda^{-1/2}$ as well as the transformed embedding space $E' = \Lambda^{-1/2}E$. If a parameter vector $\vec{\theta} \in V$ yields model predictions $X\vec{\theta}$), then $\vec{\theta}' = \Lambda^{1/2}\vec{\theta}$ yields the same model predictions $X'\vec{\theta}'$. Consequently, $G' = \Lambda^{1/2}G$ is the transformed freedom space and $\Theta' = \Lambda^{-1/2}\Theta$ is the transformed gauge space. Because Θ and G are Λ-orthogonal, Θ' and G' are orthogonal in the Euclidean sense. The transformed embedding space E' is also orthogonal to G', and so $\Theta' = E'$ The matrix P' that projects along G' and onto Θ' is therefore the orthogonal projection matrix onto the space spanned by the rows of X', and is given by $P' = (X')^+ X$. Now consider an initial parameter $\vec{\theta}_{\text{init}} \in V$, its gauge-fixed counterpart $\vec{\theta}_{\text{fixed}} = P\vec{\theta}_{\text{init}} \in \Theta$ as well as the transformed versions of these vectors, $\vec{\theta}'_{\text{init}} = \Lambda^{-1/2}\vec{\theta}_{\text{init}}$ and $\vec{\theta}'_{\text{fixed}} = \Lambda^{-1/2}\vec{\theta}_{\text{fixed}} \in \Theta'$. One can readily verify that

$$X\vec{\theta}_{\text{init}} = X\vec{\theta}_{\text{fixed}} = X'\vec{\theta}'_{\text{init}} = X'\vec{\theta}'_{\text{fixed}}. \tag{16}$$

Therefore,

$$\begin{aligned}
\vec{\theta}_{\text{fixed}} &= \Lambda^{-1/2}\vec{\theta}'_{\text{fixed}}, && (17)\\
&= \Lambda^{-1/2}P'\vec{\theta}'_{\text{init}}, && (18)\\
&= \Lambda^{-1/2}(X')^+X'\vec{\theta}'_{\text{init}}, && (19)\\
&= \Lambda^{-1/2}(X\Lambda^{-1/2})^+X\Lambda^{-1/2}\Lambda^{1/2}\vec{\theta}_{\text{init}}, && (20)\\
&= P\vec{\theta}_{\text{init}} && (21)
\end{aligned}$$

where $P = \Lambda^{-1/2}(X\Lambda^{-1/2})^+X$. This proves the claim. □

Definition 11. *A loss function $\mathcal{L}$ is said to be L_2-regularized by a metric Λ iff it has the form*

$$\mathcal{L}(\vec{\theta}) = \mathcal{L}_{\text{data}}(X\vec{\theta}) + \frac{\beta}{2}\vec{\theta}^\top \Lambda \vec{\theta}, \tag{22}$$

where X is the design matrix of an embedding $\vec{x}$, $\mathcal{L}_{\text{data}}$ is a data-dependent loss function that depends only on model predictions $X\vec{\theta}$, and β is a positive scalar.

Claim 6. *Let $\vec{x} : \mathcal{S} \to V$ be an embedding, G be the freedom space of $\vec{x}$, Θ be a gauge space of $\vec{x}$, Λ be a metric that orthogonalizes Θ and G, $\mathcal{L}$ be a loss function that is L_2-regularized by Λ, and $\vec{\theta}^*$ be a minimum of $\mathcal{L}$. Then $\vec{\theta}^* \in \Theta$.*

Proof.

$$\begin{aligned}
\vec{\theta}^* &= \text{argmin}_{\vec{v}\in V}\ \mathcal{L}(\vec{v}) && (23)\\
&= \text{argmin}_{\vec{v}\in V}\ \left[\mathcal{L}_{\text{data}}(X\vec{v}) + \frac{\beta}{2}\vec{v}^\top\Lambda\vec{v}\right] && (24)\\
&= \text{argmin}_{\{\vec{v}=\vec{\theta}+\vec{g}:\vec{\theta}\in\Theta,\,\vec{g}\in G\}}\ \left[\mathcal{L}_{\text{data}}(X(\vec{\theta}+\vec{g})) + \frac{\beta}{2}(\vec{\theta}+\vec{g})^\top\Lambda(\vec{\theta}+\vec{g})\right] && (25)\\
&= \text{argmin}_{\vec{\theta}\in\Theta}\ \left[\mathcal{L}_{\text{data}}(X\vec{\theta}) + \frac{\beta}{2}\vec{\theta}^\top\Lambda\vec{\theta}\right] + \text{argmin}_{\vec{g}\in G}\ \left[\frac{\beta}{2}\vec{g}^\top\Lambda\vec{g}\right] && (26)\\
&= \text{argmin}_{\vec{\theta}\in\Theta}\ \left[\mathcal{L}_{\text{data}}(X\vec{\theta}) + \frac{\beta}{2}\vec{\theta}^\top\Lambda\vec{\theta}\right]. && (27)
\end{aligned}$$

In Eq. 25 we used the fact that $\vec{v} \in V$ can be expressed uniquely as $\vec{v} = \vec{\theta} + \vec{g}$ for $\vec{\theta} \in \Theta$ and $\vec{g} \in G$ (by Definition 7), and in Eq. 26 we used the fact that $X\vec{g} = \vec{0}$ for any $\vec{g} \in G$ (by Definition 6), together with the assumption that Λ orthogonalizes Θ and G. We thus find that $\vec{\theta}^* \in \Theta$. □

Claim 7. *Let $\vec{x} : \mathcal{S} \to V$ be an embedding, G be the freedom space of $\vec{x}$, Θ be a gauge space of $\vec{x}$, $\mathcal{L}_{\text{data}}$ be a data-dependent loss function that depends only on model predictions $X\vec{\theta}$, and Δ be an $N \times N$ positive definite matrix. Then there exists a metric Λ defining a loss $\mathcal{L}$ that is L_2-regularized by Λ that has the following property: every minimum $\vec{\theta}^*$ of $\mathcal{L}$ lies within Θ and satisfies*

$$\mathcal{L}(\theta^*) = \mathcal{L}_{\text{data}}(X\vec{\theta}^*) + \frac{\beta}{2}\left(X\vec{\theta}^*\right)^\top\Delta\left(X\vec{\theta}^*\right). \quad (28)$$

Proof. By Claim 6 it suffices to construct a Λ that is an orthogonalizing metric for Θ and G and which satisfies Equation 28. Let P be the projection matrix from V into Θ along G, let $Q = I_M - P$, and define the matrix $\Lambda \equiv P^\top X^\top \Delta X P + Q^\top Q$. It is a simple matter to show that Λ is symmetric and positive-definite, and is thus a metric. Using $P^2 = P$ and $Q^2 = Q$, and $Q = I_M - P$, it is readily shown that PQ are both equal to the zero matrix and hence Λ satisfies $\Lambda = P^\top\Lambda P + Q^\top\Lambda Q$. Therefore, by Claim 25 in the Appendix, Λ orthogonalizes Θ and G. Then by Claim 6, we have $\vec{\theta}^* \in \Theta$ and hence $P\vec{\theta}^* = \vec{\theta}^*$ and $\vec{\theta}^*$ is in the null space of Q. Consequently,

$$\begin{aligned}
\mathcal{L}(\theta^*) &= \mathcal{L}_{\text{data}}(X\vec{\theta}^*) + \frac{\beta}{2}\vec{\theta}^{*\top}\left(P^\top X^\top\Delta X P + Q^\top Q\right)\vec{\theta}^* && (29)\\
&= \mathcal{L}_{\text{data}}(X\vec{\theta}^*) + \frac{\beta}{2}\left(\vec{\theta}^{*\top}P^\top X^\top\Delta X P\vec{\theta}^* + \vec{\theta}^{*\top}Q^\top Q\vec{\theta}^*\right) && (30)\\
&= \mathcal{L}_{\text{data}}(X\vec{\theta}^*) + \frac{\beta}{2}\left(X\vec{\theta}^*\right)^\top\Delta\left(X\vec{\theta}^*\right) && (31)
\end{aligned}$$

as required. □

Practical recipe for fixing a linear gauge using L_2 regularization. Claim 7 shows that for any choice of linear gauge Θ and any desired positive definite L_2 regularizer Δ on model predictions, we can construct a positive definite L_2 regularizer Λ on model parameters that penalizes the vector of predictions according to Δ and results in an inferred parameter vector $\vec{\theta}^*$ that is guaranteed to be a member of our desired gauge Θ. Practically, the steps to calculate Λ are:

1. Find a basis $\vec{e}_1, \ldots, \vec{e}_\gamma$ for the null space of the design matrix X. This can be done, for instance, by using Gaussian elimination to column reduce the block matrix $\begin{pmatrix} X \\ I_M \end{pmatrix}$, where the resulting matrix will have γ non-zero columns whose top N entries are 0 and whose bottom N entries will each serve as one vector in the desired basis. See see ref. [**?**] for analytical methods to determine this basis for the all-order interaction model and related models.
2. Find a basis $\vec{f}_1, \ldots, \vec{f}_{M-\gamma}$ for the desired linear gauge space Θ. Any linearly independent set of $M - \gamma$ vectors that are members of Θ will suffice.
3. Using these basis vectors as columns, define the $M \times \gamma$ matrix $E \equiv (\vec{e}_1, \ldots, \vec{e}_\gamma)$ and the $M \times (M - \gamma)$ matrix $F \equiv (\vec{f}_1, \ldots, \vec{f}_{M-\gamma})$. Then calculate the projection matricies $P = (0_{M\times\gamma}\ F)\,(E\ \ F)^{-1}$ and $Q = I_M - P$.
4. Set $\Lambda \equiv P^\top X^\top\Delta X P + Q^\top Q$.
5. Minimize $\mathcal{L}(\vec{\theta}) \equiv \mathcal{L}_{\text{data}}(X\vec{\theta}) + \frac{\beta}{2}\vec{\theta}^\top\Lambda\vec{\theta}$ for some choice of regularization parameter $\beta > 0$.

Note similarly that given a specified L_2 regularizer Λ on model parameters, the induced L_2 regularizer on model predictions is given by $\Delta \equiv (PX^+)^\top\,\Lambda\,(PX^+)$ which satisfies $x^\top\Delta x > 0$ for all nonzero x in the column space of X.

4 All-order interaction models

Definition 12. *The position-specific augmented embedding $\vec{x}'_l : \mathcal{S} \to V_l$, where $V_l = \mathbb{R}^{\alpha+1}$, is given by*

$$\vec{x}'_l(s) = \begin{pmatrix} x_l^*(s) \\ x_l^{c_1}(s) \\ \dots \\ x_l^{c_\alpha}(s) \end{pmatrix} \quad \textit{for all} \quad s \in \mathcal{S}, \tag{32}$$

where, as in the main text, $x_l^*(s) = 1$ for all $s \in \mathcal{S}$ and, for all $c \in \mathcal{A}$, $x_l^c(s)$ is one if $s_l = c$ and is zero otherwise. In what follows, we use G_l to denote the freedom space of $\vec{x}'_l$ and E_l to denote the embedding space of $\vec{x}'_l$.

Claim 8. *E_l has dimension α and is given by $\left\{(\beta^*, \beta^{c_1}, \dots, \beta^{c_\alpha})^\top : \beta^* = \sum_{c\in\mathcal{A}} \beta^c\right\}$; G_l has dimension 1 and is spanned by the vector $(-1, 1, \dots, 1)^\top$.*

Proof. Let E_l denote the span of $\vec{x}'_l$. By definition, any vector $\vec{v} \in E_l$ can be written as a linear combination of vectors $\vec{x}'_l(s)$ over $s \in S$, i.e.,

$$\vec{v} = \sum_{s\in\mathcal{S}} \beta(s) \begin{pmatrix} x_l^*(s) \\ x_l^{c_1}(s) \\ \vdots \\ x_l^{c_\alpha}(s) \end{pmatrix} \tag{33}$$

for some mapping $\beta : \mathcal{S} \to \mathbb{R}$. Defining $\beta^c \equiv \sum_{s\in\mathcal{S}} \beta(s) x_l^c(s)$ for all $c \in \mathcal{A}'$, we get

$$\vec{v} = \begin{pmatrix} \beta^* \\ \beta^{c_1} \\ \vdots \\ \beta^{c_\alpha} \end{pmatrix}. \tag{34}$$

The $\alpha + 1$ values $\{\beta^c\}_{c\in\mathcal{A}}$ are arbitrary except for one constraint arising from the fact that $x_l^*(s) = \sum_{c\in\mathcal{A}} x_l^c(s)$ for all $s \in \mathcal{S}$:

$$\begin{aligned} \sum_{c\in\mathcal{A}} \beta^c &= \sum_{c\in\mathcal{A}} \sum_{s\in\mathcal{S}} \beta(s) x_l^c(s) && (35)\\ &= \sum_{s\in\mathcal{S}} \beta(s) \sum_{c\in\mathcal{A}} x_l^c(s) && (36)\\ &= \sum_{s\in\mathcal{S}} \beta(s) x_l^*(s) && (37)\\ &= \beta^*. && (38) \end{aligned}$$

We therefore see that E_l has dimension α and comprises the vectors stated in Claim 8.

Now let $\vec{g} = (-1, 1, \dots, 1)^\top$. Taking the dot product of $\vec{g}$ with $\vec{x}'_l$ gives

$$\vec{g}^\top \vec{x'}(s) = -x_l^*(s) + \sum_{c\in\mathcal{A}} x_l^c(s) = 0 \quad \text{for all} \quad s \in \mathcal{S}, \tag{39}$$

This shows that $\vec{g} \in G_l$ and, in light of the fact that $\dim G_l = \dim V - \dim E_l = 1$, proves that G_l is spanned by $\vec{g}$. □

Definition 13. *The all-order embedding $\vec{x}_{\text{all}} : \mathcal{S} \to V_{\text{all}}$ ($V_{\text{all}} = \mathbb{R}^M$ where $M = (\alpha+1)^L$) is defined by the tensor product*

$$\vec{x}_{\text{all}}(s) = \bigotimes_{l=1}^{L} \vec{x}'_l(s) \quad \textit{for all} \quad s \in \mathcal{S}. \tag{40}$$

We use E_{all} to denote the embedding space of $\vec{x}_{\text{all}}$, and G_{all} to denote the freedom space of $\vec{x}_{\text{all}}$.

Claim 9. *The embedding space of $\vec{x}_{\text{all}}$ is given by*

$$E_{\text{all}} = \bigotimes_{l=1}^{L} E_l. \tag{41}$$

Proof. For each $c \in \mathcal{A}$, define α distinct $(\alpha+1)$-dimensional vectors $\vec{e}_c$, one for each $c \in \mathcal{A}$ and with elements indexed by $c' \in \mathcal{A}'$ given by

$$[\vec{e}_c]_{c'} = \begin{cases} 1 & \text{if } c' = * \text{ or } c' = c \\ 0 & \text{otherwise} \end{cases} \tag{42}$$

Note that, for any $s \in \mathcal{S}$, $\vec{e}_{s_l} = \vec{x}'_l(s)$. Next, for each position l, define the set of vectors $\mathcal{B}_l = \{\vec{e}_c\}_{c\in\mathcal{A}}$. From the proof of Claim 8, $\mathcal{B}_l$ forms a linearly independent basis for E_l. By definition of the tensor product of vector spaces, a basis for the space $E \equiv \bigotimes_{l=1}^{L} E_l$ is given by

$$\begin{aligned}
\mathcal{B} &\equiv \left\{ \bigotimes_{l=1}^{L} \vec{v}_l : \vec{v}_1 \in \mathcal{B}_1, \ldots, \vec{v}_L \in \mathcal{B}_L \right\} && (43)\\
&= \left\{ \bigotimes_{l=1}^{L} \vec{e}_{c_l} : c_1, \ldots, c_L \in \mathcal{A} \right\} && (44)\\
&= \left\{ \bigotimes_{l=1}^{L} \vec{e}_{s_l} : s \in \mathcal{S} \right\} && (45)\\
&= \left\{ \bigotimes_{l=1}^{L} \vec{x}'_l(s) : s \in \mathcal{S} \right\} && (46)\\
&= \{\vec{x}_{\text{all}}(s) : s \in \mathcal{S}\}. && (47)
\end{aligned}$$

$\mathcal{B}$ is therefore also a basis for E_{all}. Since E and E_{all} share the same basis, they are the same vector space. Note that we have learned, in the process, that all vectors $\vec{x}_{\text{all}}(s)$, $s \in \mathcal{S}$, are linearly independent. □

Claim 10. *The freedom space of $\vec{x}_{\text{all}}$ is given by*

$$G_{\text{all}} = \bigoplus_{(R_1,\ldots,R_L)\in\mathcal{R}} \left[\bigotimes_{l=1}^{L} R_l \right], \tag{48}$$

where

$$\mathcal{R} = \{(R_1, \ldots, R_L) : R_l \in \{E_l, G_l\} \textit{ for all } l \textit{ and } R_l = G_l \textit{ for at least one } l\}. \tag{49}$$

Proof.

$$\begin{aligned}
V_{\text{all}} &= \bigotimes_{l=1}^{L} \mathrm{V}_l && (50)\\
&= \bigotimes_{l=1}^{L} [E_l \oplus G_l] && (51)\\
&= \left[\bigotimes_{l=1}^{L} E_l \right] \oplus \left[\bigoplus_{\{R_1,\ldots,R_L\}\in\mathcal{R}} \bigotimes_{l=1}^{L} R_l \right] && (52)\\
&= E_{\text{all}} \oplus W && (53)
\end{aligned}$$

where

$$W \equiv \bigoplus_{\{R_1,\ldots,R_L\}\in\mathcal{R}} \bigotimes_{l=1}^{L} R_l, \tag{54}$$

This shows that $\dim W = \dim V_{\text{all}} - \dim E_{\text{all}} = \dim G_{\text{all}}$. However, this does not yet show that $W = G_{\text{all}}$.

To see that $W = G_{\text{all}}$, let $\{\vec{u}_l^c\}_{c\in\mathcal{A}'}$ be a basis for V_l. Each $\vec{u}_l^c$ has a unique decomposition $\vec{u}_l^c = \vec{\theta}_l^c + \vec{g}_l^c$ where $\vec{\theta}_l^c \in E_l$ and $\vec{g}_l^c \in G_l$. For every augmented sequence $s' \in \mathcal{S}'$, there is a corresponding vector in V_{all} given by $\vec{u}_{s'} = \bigotimes_{l=1}^{L} \vec{u}_l^{s_l}$. Moreover, the set $\{\vec{u}_{s'}\}_{s'\in\mathcal{S}'}$ is a basis for V_{all}. The vector $\vec{u}_{s'}$ can be decomposed into a component in E_{all} and a component in W via

$$\vec{u}_{s'} = \bigotimes_{l=1}^{L} \vec{u}_l^{s'_l} \tag{55}$$

$$= \bigotimes_{l=1}^{L} \left[\vec{\theta}_l^{s'_l} + \vec{g}_l^{s'_l}\right] \tag{56}$$

$$= \bigotimes_{l=1}^{L} \vec{\theta}_l^{s'_l} + \sum_{(\vec{r}_1,\ldots,\vec{r}_L)\in\rho_{s'}} \bigotimes_{l=1}^{L} \vec{r}_l \tag{57}$$

$$= \vec{\theta}_{s'} + \vec{r}_{s'} \tag{58}$$

where

$$\vec{\theta}_{s'} \equiv \bigotimes_{l=1}^{L} \vec{\theta}_l^{s'_l}, \tag{59}$$

$$\vec{r}_{s'} \equiv \sum_{(\vec{r}_1,\ldots,\vec{r}_L)\in\rho_{s'}} \bigotimes_{l=1}^{L} \vec{r}_l, \tag{60}$$

$$\rho_{s'} \equiv \left\{(\vec{r}_1,\ldots,\vec{r}_L) : \vec{r}_l \in \left\{\vec{\theta}_l^{s'_l}, \vec{g}_l^{s'_l}\right\} \text{ for all } l, \text{ and } \vec{r}_l = \vec{g}_l^{s'_l} \text{ for at least one } l\right\}. \tag{61}$$

Observe that $r_{s'} \in W$ for every $s' \in \mathcal{S}'$. Moreover, since $\{\vec{u}_{s'}\}_{s'\in\mathcal{S}'}$ is a basis for $\vec{V}_{\text{all}}$, $\{\vec{r}_{s'}\}_{s'\in\mathcal{S}'}$ is a basis for W. Finally, observe that for every $s \in \mathcal{S}$, the dot product of $\vec{x}_{\text{all}}(s)$ with $\vec{r}_{s'}$ is zero, i.e.,

$$[\vec{x}_{\text{all}}(s)]^\top \vec{r}_{s'} = \sum_{(\vec{r}_1,\ldots,\vec{r}_L)\in\rho_{s'}} \prod_{l=1}^{L} [\vec{x}'_l(s)]^\top \vec{r}_l = 0, \tag{62}$$

because $\vec{r}_l \in G_l$ for at least one l in each product. Therefore, $\vec{r}_{s'} \in G_{\text{all}}$ for every $s' \in \mathcal{S}'$. Since $\{\vec{r}_{s'}\}_{s'\in\mathcal{S}'}$ is a basis for W, we conclude that $W = G_{\text{all}}$.

□

Claim 11. *For each position l, let Θ_l be a gauge space for $\vec{x}'_l$, let P_l be projection matrix into Θ_l along G_l, and let Λ_l be a metric that orthogonalizes Θ_l and G_l. Then $\Theta \equiv \bigotimes_l \Theta_l$ is a gauge space of $\vec{x}_{\text{all}}$, $P \equiv \bigotimes_l P_l$ projects into Θ along G_{all}, and $\Lambda \equiv \bigotimes_l \Lambda_l$ is a metric that orthogonalizes Θ and G_{all}.*

Proof. For each position $l = 1,\ldots,L$, let $\vec{v}_l = \vec{\theta}_l + \vec{g}_l$ where $\vec{v}_l \in V_l$, $\vec{\theta}_l \in \Theta_l$, and $\vec{g}_l \in G_l$. Defining $\vec{v} = \bigotimes_{l=1}^{L} \vec{v}_l$, we get

$$\vec{v} = \bigotimes_{l=1}^{L} \left[\vec{\theta}_l + \vec{g}_l\right] = \vec{\theta} + \vec{g} \tag{63}$$

where $\vec{\theta} = \bigotimes_{l=1}^{L} \vec{\theta}_l$ lives in the space $\Theta \equiv \bigotimes_{l=1}^{L} \Theta_l$, and

$$\vec{g} = \sum_{(\vec{r}_1,\ldots,\vec{r}_L)\in\rho} \bigotimes_{l=1}^{L} \vec{r}_l, \tag{64}$$

$$\rho \equiv \left\{(\vec{r}_1,\ldots,\vec{r}_L) : \vec{r}_l \in \left\{\vec{\theta}_l, \vec{g}_l\right\} \text{ for all } l, \text{ and } \vec{r}_l = \vec{g}_l \text{ for at least one } l\right\}, \tag{65}$$

lives in the space

$$G \equiv \bigoplus_{(R_1,\ldots,R_L)\in\mathcal{R}} \left[\bigotimes_{l=1}^{L} R_l\right], \tag{66}$$

where

$$\mathcal{R} = \{(R_1,\ldots,R_L) : R_l \in \{\Theta_l, G_l\} \text{ for all } l \text{ and } R_l = G_l \text{ for at least one } l\}. \tag{67}$$

By arguments similar to those in Claim 10, it is readily seen that $\vec{x}_{\text{all}}(s)^\top \vec{g} = \vec{0}$ for every $s \in \mathcal{S}$. Moreover, by arguments similar to those in Claim 10, it is readily seen that this construction is able to yield a basis of vectors $\vec{g}$ for G, proving that that $G = G_{\text{all}}$ is the freedom space of $\vec{x}_{\text{all}}$. Therefore, for every vector of the form $\vec{v} = \bigotimes_{l=1}^{L} \vec{v}_l$, there is a unique decomposition $\vec{v} = \vec{\theta} + \vec{g}$, where $\vec{\theta} \in \Theta$ and $\vec{g} \in G_{\text{all}}$. Because a basis of such vectors $\vec{v}$ can be found for V, we see that any vector $\vec{v} \in V$ can be uniquely decomposed as $\vec{v} = \vec{\theta} + \vec{g}$, where $\vec{\theta} \in \Theta$ and $\vec{g} = G_{\text{all}}$. This proves that Θ is a gauge space of $\vec{x}_{\text{all}}$.

Defining $P \equiv \bigotimes_{l=1}^{L} P_l$ and applying this to $\vec{v}$, we find that

$$P\vec{v} = \bigotimes_{l=1}^{L} P_l \vec{v}_l = \bigotimes_{l=1}^{L} \vec{\theta}_l = \vec{\theta} \tag{68}$$

is in Θ, which implies that $\vec{v} - P\vec{v} = \vec{g}$ is in G_{all}. Since a basis of vectors $\vec{v}$ for V can be found that decompose in this way, all vectors in V decompose in this manner. This proves that P projects into Θ along G_{all}

Now define $\Lambda \equiv \bigotimes_{l=1}^{L} \Lambda_l$. It is readily seen that Λ is symmetric and positive definite from the fact that every Λ_l is symmetric and positive definite. Moreover,

$$\begin{aligned}
\Lambda &= \bigotimes_{l=1}^{L} (P_l^\top \Lambda_l P_l + Q_l^\top \Lambda_l Q_l) && (69)\\
&= \bigotimes_{l=1}^{L} P_l^\top \Lambda_l P_l + \sum_{(R_1,\ldots,R_L)\in\mathcal{R}} \bigotimes_{l=1}^{L} R_l^\top \Lambda_l R_l && (70)\\
&= \left[\bigotimes_{l=1}^{L} P_l\right]^\top \left[\bigotimes_{l=1}^{L} \Lambda_l\right] \left[\bigotimes_{l=1}^{L} P_l\right] + \sum_{(R_1,\ldots,R_L)\in\mathcal{R}} \left[\bigotimes_{l=1}^{L} R_l\right]^\top \left[\bigotimes_{l=1}^{L} \Lambda_l\right] \left[\bigotimes_{l=1}^{L} R_l\right] && (71)\\
&= \left[\bigotimes_{l=1}^{L} P_l\right]^\top \left[\bigotimes_{l=1}^{L} \Lambda_l\right] \left[\bigotimes_{l=1}^{L} P_l\right] + \left[\sum_{(R_1,\ldots,R_L)\in\mathcal{R}} \bigotimes_{l=1}^{L} R_l\right]^\top \left[\bigotimes_{l=1}^{L} \Lambda_l\right] \left[\sum_{(R_1,\ldots,R_L)\in\mathcal{R}} \bigotimes_{l=1}^{L} R_l\right] && (72)\\
&= P^\top \Lambda P + Q^\top \Lambda Q. && (73)
\end{aligned}$$

where $Q = I_M - P$ and the set $\mathcal{R}$ of ordered operator sets is defined to be

$$\mathcal{R} \equiv \{(R_1,\ldots,R_L) : R_i = P_i \text{ or } R_i = Q_i \text{ for all } i = 1,\ldots,L, \text{ with } R_i = Q_i \text{ for at least one } i\}. \tag{74}$$

In Eq. 69, we used the fact that Λ_l satisfies

$$\Lambda_l = P_l^\top \Lambda_l P_l^\top + Q_l^\top \Lambda_l Q_l, \tag{75}$$

where $Q_l \equiv I_{(\alpha+1)} - P_l$ (see Claim 25). In going from Eq. 71 to Eq. 72, we used the fact that, if two ordered sets of operators $(R_1,\ldots,R_L)$ and $(R'_1,\ldots,R'_L)$ in $\mathcal{R}$ are different, then

$$\left[\bigotimes_{l=1}^{L} R_l\right]^\top \left[\bigotimes_{l=1}^{L} \Lambda_l\right] \left[\bigotimes_{l=1}^{L} R'_l\right] = \bigotimes_{l=1}^{L} R_l^\top \Lambda R'_l = 0 \tag{76}$$

since there will be an l such that $R_l^\top \Lambda_l R'_l$ is either $P_l^\top \Lambda_l Q_l = 0$ or $Q_l \Lambda_l P_l = 0$. In going from Eq. 72 to Eq. 73, we used the fact that

$$\sum_{(R_1,\ldots,R_L)\in\mathcal{R}} \bigotimes_{l=1}^{L} R_l = \bigotimes_{i=1}^{L} (P_l + Q_l) - \bigotimes_{i=1}^{L} P_l = I_M - P = Q. \tag{77}$$

The fact that P projects into Θ along G_{all}, together with the result $\Lambda = P^\top \Lambda P + Q^\top \Lambda Q$ in Eq. 73 proves (by Claim 25) that Λ orthogonalizes Θ and G_{all}.

□

5 Parametric family of gauges

Definition 14. *A probability distribution p on $\mathcal{S}$ is positive iff $p(s) > 0$ for all $s \in \mathcal{S}$.*

Definition 15. *A factorizable probability distribution p on $\mathcal{S}$ is one that can be written $p(s) = \prod_{l=1}^{L} p_l^{s_l}$ for all $s \in \mathcal{S}$, where p_l is a probability distribution over the α possible characters at position l, i.e., $p(s) = p_l^{s_l}$.*

Definition 16. *An augmented sequence s' is a sequence built from the augmented alphabet $\mathcal{A}' = \{*, c_1, \ldots, c_\alpha\}$, where $c_1, \ldots, c_\alpha$ are the characters in $\mathcal{A}$ and $*$ is a wild-card character that is interpreted as matching any character in $\mathcal{A}$. The set of all augmented sequences is denoted $\mathcal{S}'$. Note that every augmented sequence s' is can be interpreted as a subset of $\mathcal{S}$ that comprises sequences matching the pattern defined by s'. For this reason we will sometimes use expressions like $s \in s'$ and $s' \subseteq t'$ (for $s \in \mathcal{S}$ and $s', t' \in \mathcal{S}$).*

To aid in our discussion of the all-order interaction model, we define an augmented alphabet $\mathcal{A}' = \{*, c_1, \ldots, c_\alpha\}$, where $c_1, \ldots, c_\alpha$ are the characters in $\mathcal{A}$ and $*$ is a wild-card character that is interpreted as matching any character in $\mathcal{A}$. Let $\mathcal{S}'$ denote the set of sequences of length L comprising characters from $\mathcal{A}'$. For each augmented sequence $s' \in \mathcal{S}'$, we define the sequence feature $x_{s'}(s)$ to be 1 if a sequence s matches the pattern described by s' and to be 0 otherwise. In this way, each augmented sequence s' serves as a regular expression against which bona fide sequences are compared.

Definition 17. *Given a non-negative real number λ, a factorizable probability distribution p on $\mathcal{S}$, and a sequence position l, the position-specific parametric gauge $\Theta_l^{\lambda,p}$ is defined as*

$$\Theta_l^{\lambda,p} \equiv V_\lambda \oplus V_\perp^{p_l}, \tag{78}$$

where $V_\lambda \equiv \text{span}\,\{(\lambda, 1, \cdots, 1)^\top\}$ and $V_\perp^{p_l} \equiv \left\{(0, v_{c_1}, \ldots, v_{c_\alpha})^\top : \sum_{c\in\mathcal{A}} p_l^c v_c = 0\right\}$.

Claim 12. *The matrix $P_l^{\lambda,p}$ that projects V_l along G_l and onto $\Theta_l^{\lambda,p}$ is an $(\alpha+1)\times(\alpha+1)$ matrix given by*

$$P_l^{\lambda,p} = \begin{pmatrix} \eta & p_l^{c_1}\eta & \cdots & p_l^{c_\alpha}\eta \\ 1-\eta & 1-p_l^{c_1}\eta & \cdots & -p_l^{c_\alpha}\eta \\ \vdots & \vdots & \ddots & \vdots \\ 1-\eta & -p_l^{c_1}\eta & \cdots & 1-p_l^{c_\alpha}\eta \end{pmatrix}, \tag{79}$$

where $\eta \equiv \lambda/(1+\lambda)$.

Proof. Any $(\alpha+1)$-dimensional vector $\vec{v} \in V_l$ can be decomposed as

$$\vec{v} = \vec{\theta} + \vec{g}, \tag{80}$$

where $\vec{\theta} \in \Theta_l^{\lambda,p}$ and $\vec{g} \in G_l$. From the definitions of $\Theta_l^{\lambda,p}$ and G, $\vec{\theta}$ and $\vec{g}$ must have the forms

$$\vec{\theta} = c_\lambda \vec{v}_\lambda + \vec{\theta}_\perp \quad \text{and} \quad \vec{g} = c_{-1}\vec{e}_{-1}, \tag{81}$$

where

$$\vec{e}_\lambda = \begin{pmatrix} \lambda \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \vec{e}_{-1} = \begin{pmatrix} -1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \vec{\theta}_\perp = \begin{pmatrix} 0 \\ \theta_l^{c_1} \\ \vdots \\ \theta_l^{c_\alpha} \end{pmatrix}, \tag{82}$$

and where c_λ and c_{-1} are real numbers that depend on $\vec{v}$. The projection matrix $Q_l^{\lambda,p} \equiv I - P_l^{\lambda,p}$ projects V_l into G_l along $\Theta_l^{\lambda,p}$, and so is related to the scalar c_{-1} via

$$Q_l^{\lambda,p}\vec{v} = c_{-1}\vec{e}_{-1}. \tag{83}$$

We now compute c_{-1} as a function of $\vec{v}$. First we define the metric

$$\Lambda \equiv \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \alpha p_l^{c_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha p_l^{c_\alpha} \end{pmatrix}, \tag{84}$$

and compute

$$\vec{e}_\lambda^\top \Lambda \vec{e}_\lambda = \lambda^2 + \sum_{i=1}^{\alpha} \alpha p_l^{c_i} = \alpha + \lambda^2, \tag{85}$$

$$\vec{e}_{-1}^\top \Lambda \vec{e}_\lambda = -\lambda + \sum_{i=1}^{\alpha} \alpha p_l^{c_i} = \alpha - \lambda, \tag{86}$$

$$\vec{e}_{-1}^\top \Lambda \vec{e}_{-1} = 1 + \sum_{i=1}^{\alpha} \alpha p_l^{c_i} = \alpha + 1, \tag{87}$$

$$\vec{e}_\lambda^\top \Lambda \vec{\theta}_\perp = 0 + \sum_{i=1}^{\alpha} \alpha p_l^{c_i} \theta_l^{c_i} = 0, \tag{88}$$

$$\vec{e}_{-1}^{\top}\Lambda\vec{\theta}_{\perp} = 0 + \sum_{c}\sum_{i=1}^{\alpha}\alpha p_l^{c_i}\theta_l^{c_i} = 0. \quad (89)$$

Next we solve for c_{-1} via

$$\begin{aligned}
\vec{e}_{\lambda}^{\top}\Lambda\vec{v} &= c_\lambda(\vec{e}_{\lambda}^{\top}\Lambda\vec{e}_{\lambda}) + c_{-1}(\vec{e}_{\lambda}^{\top}\Lambda\vec{e}_{-1}) + (\vec{e}_{\lambda}^{\top}\Lambda\vec{\theta}_{\perp}) && (90)\\
&= c_\lambda(\alpha+\lambda^2) + c_{-1}(\alpha-\lambda). && (91)\\
\vec{e}_{-1}^{\top}\Lambda\vec{v} &= c_\lambda(\vec{e}_{-1}^{\top}\Lambda\vec{e}_{\lambda}) + c_{-1}(\vec{e}_{-1}^{\top}\Lambda\vec{e}_{-1}) + (\vec{e}_{-1}^{\top}\Lambda\vec{\theta}_{\perp}) && (92)\\
&= c_\lambda(\alpha-\lambda) + c_{-1}(\alpha+1). && (93)\\
\Rightarrow (\alpha-\lambda)\vec{e}_{\lambda}^{\top}\Lambda\vec{v} - (\alpha+\lambda^2)\vec{e}_{-1}^{\top}\Lambda\vec{v} &= c_{-1}\left[(\alpha-\lambda)^2 - (\alpha+1)(\alpha+\lambda^2)\right] && (94)\\
&= c_{-1}\left[-2\alpha\lambda - \alpha(1+\lambda^2)\right] && (95)\\
&= -\alpha(1+2\lambda+\lambda^2)c_{-1} && (96)\\
&= -\alpha(1+\lambda)^1 c_{-1} && (97)\\
\Rightarrow c_{-1} &= -\frac{(\alpha-\lambda)}{\alpha(1+\lambda)^2}\vec{e}_{\lambda}^{\top}\Lambda\vec{v} + \frac{(\alpha+\lambda^2)}{\alpha(1+\lambda)^2}\vec{e}_{-1}^{\top}\Lambda\vec{v}. && (98)
\end{aligned}$$

This shows that $Q_l^{\lambda,p}$ is given by

$$Q_l^{\lambda,p} = -\frac{(\alpha-\lambda)}{\alpha(1+\lambda)^2}\vec{e}_{-1}\vec{e}_{\lambda}^{\top}\Lambda + \frac{(\alpha+\lambda^2)}{\alpha(1+\lambda)^2}\vec{e}_{-1}\vec{e}_{-1}^{\top}\Lambda. \quad (99)$$

Next we compute the matrix elements $[Q_l^{\lambda,p}]_{ij}$ for all $i,j \in \{0,1,\ldots,\alpha\}$. Setting $i=0$, $j=0$ gives

$$\begin{aligned}
[Q_l^{\lambda,p}]_{00} &= -\frac{(\alpha-\lambda)}{\alpha(1+\lambda)^2}(-\lambda) + \frac{(\alpha+\lambda^2)}{\alpha(1+\lambda)^2}(1) && (100)\\
&= \frac{(\alpha-\lambda)\lambda + (\alpha+\lambda^2)}{\alpha(1+\lambda)^2} && (101)\\
&= \frac{\alpha(1+\lambda)}{\alpha(1+\lambda)^2} && (102)\\
&= \frac{1}{1+\lambda} && (103)\\
&= 1-\eta. && (104)
\end{aligned}$$

Setting $i=0$, $j>0$ gives

$$\begin{aligned}
[Q_l^{\lambda,p}]_{0j} &= -\frac{(\alpha-\lambda)}{\alpha(1+\lambda)^2}(-\alpha p_l^{c_j}) + \frac{(\alpha+\lambda^2)}{\alpha(1+\lambda)^2}(-\alpha p_l^{c_j}) && (105)\\
&= \frac{\alpha(\alpha-\lambda) - \alpha(\alpha+\lambda^2)}{\alpha(1+\lambda)^2}p_l^{c_j} && (106)\\
&= -\frac{\lambda(1+\lambda)}{(1+\lambda)^2}p_l^{c_j} && (107)\\
&= -\frac{\lambda}{1+\lambda}p_l^{c_j} && (108)\\
&= -\eta p_l^{c_j}. && (109)
\end{aligned}$$

Setting $i>0$, $j=0$ gives

$$\begin{aligned}
[Q_l^{\lambda,p}]_{i0} &= -\frac{(\alpha-\lambda)}{\alpha(1+\lambda)^2}(\lambda) + \frac{(\alpha+\lambda^2)}{\alpha(1+\lambda)^2}(-1) && (110)\\
&= \frac{-\lambda(\alpha-\lambda) - (\alpha+\lambda^2)}{\alpha(1+\lambda)^2} && (111)\\
&= -\frac{\alpha(1+\lambda)}{\alpha(1+\lambda)^2} && (112)\\
&= -\frac{1}{1+\lambda} && (113)
\end{aligned}$$

$$= \quad \eta - 1. \tag{114}$$

Setting $i > 0$, $j > 0$ gives

$$\begin{aligned}
[Q_l^{\lambda,p}]_{ij} &= -\frac{(\alpha-\lambda)}{\alpha(1+\lambda)^2}(\alpha p_l^{c_j}) + \frac{(\alpha+\lambda^2)}{\alpha(1+\lambda)^2}(\alpha p_l^{c_j}) && (115)\\
&= \frac{-(\alpha-\lambda)+(\alpha+\lambda^2)}{(1+\lambda)^2} p_l^{c_j} && (116)\\
&= \frac{\lambda(1+\lambda)}{(1+\lambda)^2} p_l^{c_j} && (117)\\
&= \frac{\lambda}{1+\lambda} p_l^{c_j} && (118)\\
&= \eta p_l^{c_j}. && (119)
\end{aligned}$$

We therefore obtain

$$Q_l^{\lambda,p} = \begin{pmatrix} 1-\eta & -p_l^{c_1}\eta & \cdots & -p_l^{c_\alpha}\eta \\ \eta-1 & p_l^{c_1}\eta & \cdots & p_l^{c_\alpha}\eta \\ \vdots & \vdots & \ddots & \vdots \\ \eta-1 & p_l^{c_1}\eta & \cdots & p_l^{c_\alpha}\eta \end{pmatrix}. \tag{120}$$

Finally, plugging this result into $P_l^{\lambda,p} = I - Q_l^{\lambda,p}$ gives

$$P_l^{\lambda,p} = \begin{pmatrix} \eta & p_l^{c_1}\eta & \cdots & p_l^{c_\alpha}\eta \\ 1-\eta & 1-p_l^{c_1}\eta & \cdots & -p_l^{c_\alpha}\eta \\ \vdots & \vdots & \ddots & \vdots \\ 1-\eta & -p_l^{c_1}\eta & \cdots & 1-p_l^{c_\alpha}\eta \end{pmatrix}, \tag{121}$$

which proves the claim. □

Claim 13. *$\Theta^{\lambda,p} \equiv \bigotimes_l \Theta_l^{\lambda,p}$ is a valid gauge space for the embedding $\vec{x}_{\text{all}}$.*

Proof. This follows from Claim 11 and the fact that $\Theta_l^{\lambda,p}$ is a valid gauge space for $\vec{x}_l$. □

Claim 14. *The matrix $P^{\lambda,p}$ that projects V along G_{all} and into $\Theta^{\lambda,p}$ is an $M \times M$ matrix with elements*

$$P^{\lambda,p}_{s't'} = \prod_{\substack{l \ s.t. \\ s'_l \in \mathcal{A} \\ t'_l \in \mathcal{A}}} \left(\delta_{s'_l t'_l} - p_l^{t'_l}\eta\right) \times \prod_{\substack{l \ s.t. \\ s'_l = * \\ t'_l \in \mathcal{A}}} \left(p_l^{t'_l}\eta\right) \times \prod_{\substack{l \ s.t. \\ s'_l \in \mathcal{A} \\ t'_l = *}} (1-\eta) \times \prod_{\substack{l \ s.t. \\ s'_l = * \\ t'_l = *}} \eta. \tag{122}$$

Proof. By Claim 11, $\Theta^{\lambda,p} \equiv \bigotimes_l \Theta_l^{\lambda,p}$ imples that $P^{\lambda,p} \equiv \bigotimes_l P_l^{\lambda,p}$. Eq. 122 follows from expressing the elements of $P_l^{\lambda,p}$ in Eq. 79 as

$$[P^{\lambda,p}]_{cc'} = \begin{cases} \delta_{cc'} - p_l^{c'}\eta & \text{if } c \in \mathcal{A}, c' \in \mathcal{A}, \\ p_l^{c'}\eta & \text{if } c = *, c' \in \mathcal{A}, \\ 1-\eta & \text{if } c \in \mathcal{A}, c' = *, \\ \eta & \text{if } c = *, c' = *, \end{cases} \tag{123}$$

then taking the product across positions l using $c = s'_l$ and $c' = t'_l$. □

Claim 15. *If λ and p are positive, the $\alpha \times \alpha$ metric*

$$\Lambda_l^{\lambda,p} \equiv \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda p_l^{c_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda p_l^{c_\alpha} \end{pmatrix} \tag{124}$$

orthogonalizes $\Theta_l^{\lambda,p}$ and G_l.

Proof. Assume that $\Lambda_l^{\lambda,p}$ is diagonal, and thus has the form

$$\Lambda_l^{\lambda,p} = \begin{pmatrix} d_0 & 0 & \cdots & 0 \\ 0 & d_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_\alpha \end{pmatrix}, \tag{125}$$

for some set of scalars $d_0, \ldots, d_\alpha$. The requirement that $\Lambda_l^{\lambda,p}$ be a metric (and thus positive definite) implies that all $d_0, \ldots, d_\alpha$ must be positive. We now solve for $d_0, \ldots, d_\alpha$ using the matrix equation (Claim 25):

$$\Lambda_l^{\lambda,p} = \left(P_l^{\lambda,p}\right)^\top \Lambda_l^{\lambda,p} \left(P_l^{\lambda,p}\right) + \left(Q_l^{\lambda,p}\right)^\top \Lambda_l^{\lambda,p} \left(Q_l^{\lambda,p}\right), \tag{126}$$

$$\Rightarrow d_i \delta_{ij} = \sum_{k=0}^{\alpha} d_k [P_l^{\lambda,p}]_{ki} [P_l^{\lambda,p}]_{kj} + \sum_{k=0}^{\alpha} d_k [Q_l^{\lambda,p}]_{ki} [Q_l^{\lambda,p}]_{kj}, \tag{127}$$

where $Q_l^{\lambda,p} \equiv I - P_l^{\lambda,p}$. We now solve Eq. 127 for different choices of i and j using the matrix elements of $P_l^{\lambda,p}$ and $Q_l^{\lambda,p}$ computed in the proof of Claim 12. For $i = j = 0$, we get

$$\begin{aligned}
d_0 &= \left[d_0 [P_l^{\lambda,p}]_{00}^2 + \sum_{k=1}^{\alpha} d_k [P_l^{\lambda,p}]_{k0}^2\right] + \left[d_0 [Q_l^{\lambda,p}]_{00}^2 + \sum_{k=1}^{\alpha} d_k [Q_l^{\lambda,p}]_{k0}^2\right] && (128)\\
&= \left[d_0 \eta^2 + \sum_{k=1}^{\alpha} d_k (1-\eta)^2\right] + \left[d_0 (1-\eta)^2 + \sum_{k=1}^{\alpha} d_k (\eta - 1)^2\right] && (129)\\
&= [d_0 \eta^2 + a(1-\eta)^2] + [d_0 (1-\eta)^2 + a(\eta-1)^2] && (130)\\
&= d_0 [2\eta^2 - 2\eta + 1] + 2a(1-\eta)^2, && (131)\\
\Rightarrow 2a(2-\eta)^2 &= 2d_0 \eta (1-\eta), && (132)\\
\Rightarrow a &= d_0 \frac{\eta}{1-\eta} && (133)\\
&= d_0 \lambda, && (134)
\end{aligned}$$

where in Eq. 134 we have used $\lambda = \eta/(1-\eta)$. For $i = 0$, $j > 0$, we get

$$\begin{aligned}
0 &= \left[d_0 [P_l^{\lambda,p}]_{00} [P_l^{\lambda,p}]_{0j} + \sum_{k=1}^{\alpha} d_k [P_l^{\lambda,p}]_{k0} [P_l^{\lambda,p}]_{kj}\right] + \left[d_0 [Q_l^{\lambda,p}]_{00} [Q_l^{\lambda,p}]_{0j} + \sum_{k=1}^{\alpha} d_k [Q_l^{\lambda,p}]_{k0} [Q_l^{\lambda,p}]_{kj}\right] && (135)\\
&= \left[d_0 p_j \eta^2 + \sum_{k=1}^{\alpha} d_k (1-\eta)(\delta_{jk} - p_j \eta)\right] + \left[-d_0 p_j \eta (1-\eta) + \sum_{k=1}^{\alpha} d_k (\eta - 1)(p_j \eta)\right] && (136)\\
&= 2d_0 p_j \eta^2 - d_0 p_j \eta - 2a p_j \eta (1-\eta) + d_j (1-\eta) && (137)\\
&= 2(d_0 + a) p_j \eta^2 - (2a + d_0) p_j \eta + d_j (1-\eta) && (138)\\
&= 2d_0 (1+\lambda) \frac{\lambda}{1+\lambda} p_j (\eta - 1) + d_0 p_j \eta + d_j (1-\eta), && (139)\\
\Rightarrow d_j &= 2d_0 p_j \lambda + d_0 p_j \frac{\eta}{\eta - 1} && (140)\\
&= 2d_0 p_j \lambda - d_0 p_j \lambda && (141)\\
&= d_0 p_j \lambda. && (142)
\end{aligned}$$

The resulting equation,

$$d_i = d_0 p_i \lambda, \quad \text{for all} \quad i = 1, \ldots, \alpha, \tag{143}$$

determines all elements of $\Lambda_l^{\lambda,p}$ up to a multiplicative factor d_0. However, we must still verify that Eq. 143 is consistent with the constraints placed on the other elements of $\Lambda_l^{\lambda,p}$ by Eq. 127. For $i > 0$, $j = 0$, we get the same result as for $i = 0$, $j > 0$, because Eq. 127 is symmetric. For $i > 0$, $j > 0$, $i = j$, we get

$$\begin{aligned}
d_i &= \left[d_0 [P_l^{\lambda,p}]_{0i}^2 + \sum_{k=1}^{\alpha} d_k [P_l^{\lambda,p}]_{ki}^2\right] + \left[d_0 [Q_l^{\lambda,p}]_{0i}^2 + \sum_{k=1}^{\alpha} d_k [Q_l^{\lambda,p}]_{ki}^2\right] && (144)\\
&= \left[d_0 p_i^2 \eta^2 + \sum_{k=1}^{\alpha} d_k (\delta_{ki} - p_i \eta)^2\right] + \left[d_0 p_i^2 \eta^2 + \sum_{k=1}^{\alpha} d_k p_i^2 \eta^2\right] && (145)
\end{aligned}$$

$$
\begin{aligned}
&= \left[d_0 p_i^2 \eta^2 + d_i - 2 d_i p_i \eta + a p_i^2 \eta^2\right] + \left[d_0 p_i^2 \eta^2 + a p_i^2 \eta^2\right] && (146)\\
&= d_i + 2[(d_0 + a) p_i^2 \eta^2 - d_i p_i \eta], && (147)\\
\Rightarrow d_i &= (d_0 + a) p_i \eta && (148)\\
&= d_0 (1+\lambda) \frac{\lambda}{1+\lambda} p_i && (149)\\
&= d_0 p_i \lambda, && (150)
\end{aligned}
$$

which is consistent with Eq. 143. For $i > 0$, $j > 0$, $i \neq j$, we get

$$
\begin{aligned}
0 &= \left[d_0 [P_l^{\lambda,p}]_{0i} [P_l^{\lambda,p}]_{0j} + \sum_{k=1}^{\alpha} d_k [P_l^{\lambda,p}]_{ki} [P_l^{\lambda,p}]_{kj}\right]\\
&\quad + \left[d_0 [Q_l^{\lambda,p}]_{0i} [Q_l^{\lambda,p}]_{0j} + \sum_{k=1}^{\alpha} d_k [Q_l^{\lambda,p}]_{ki} [Q_l^{\lambda,p}]_{kj}\right] && (151)\\
&= \left[d_0 p_i p_j \eta^2 + \sum_{k=1}^{\alpha} d_k (\delta_{ki} - p_i \eta)(\delta_{kj} - p_j \eta)\right] + \left[d_0 p_i p_j + \sum_{k=1}^{\alpha} d_k p_i p_j \eta^2\right] && (152)\\
&= 2 d_0 p_i p_j \eta^2 + 2 a p_i p_j \eta^2 - d_i p_j \eta - p_i d_j \eta && (153)\\
&= 2(d_0 + a) p_i p_j \eta^2 - d_i p_j \eta - p_i d_j \eta, && (154)\\
\Rightarrow d_i p_j + p_i d_j &= 2 d_0 (1+\lambda) \frac{\lambda}{1+\lambda} && (155)\\
&= 2 d_0 \lambda, && (156)
\end{aligned}
$$

which is also consistent with Eq. 143. We therefore see that Eq. 143 is indeed consistent with Eq. 127. Thus we find that

$$
\Lambda_l^{\lambda,p} = d_0 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda p_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda p_\alpha \end{pmatrix}, \tag{157}
$$

orthogonalizes $\Theta_l^{\lambda,p}$ and G_l, and is determined up to an unknown positive scalar d_0. This proves the claim. □

Claim 16. *If λ and p are positive, the $M \times M$ metric*

$$
\Lambda_{s't'} \equiv p(s') \lambda^{o(s')} \delta_{s't'} \tag{158}
$$

orthogonalizes $\Theta^{\lambda,p}$ and G_{all}.

Proof. By Claim 11, $\Theta^{\lambda,p} \equiv \bigotimes_l \Theta_l^{\lambda,p}$ implies that $\Lambda^{\lambda,p} \equiv \bigotimes_l \Lambda_l^{\lambda,p}$. Eq. 158 follows from expressing the elements of $\Lambda_l^{\lambda,p}$ in Eq. 157 as

$$
[\Lambda_l^{\lambda,p}]_{cc'} = \delta_{cc'} p_l^c \times \begin{cases} 1 & \text{if } c = *, \\ \lambda & \text{if } c \in \mathcal{A}, \end{cases} \tag{159}
$$

then taking the product across positions l using $c = s'_l$ and $c' = t'_l$. □

6 Trivial gauge, Euclidean gauge, and equitable gauge

Claim 17. *The trivial gauge $\Theta^{0,p}$ is unaffected by the probability distribution p.*

Proof. Setting $\lambda = 0$ gives $\eta = 0$, and thus

$$
\begin{aligned}
P^{0,p}_{s't'} &= \prod_{\{l : s'_l \in \mathcal{A} \text{ and } t'_l \in \mathcal{A}\}} \delta_{s'_l t'_l} \times \prod_{\{l : s'_l = *\}} 0 && (160)\\
&= \prod_{\{l : s'_l \in \mathcal{A} \text{ and } t'_l \in \mathcal{A}\}} \delta_{s'_l t'_l} \times \begin{cases} 1 & \text{if } s' \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases} && (161)\\
&= \prod_{\{l : t'_l \in \mathcal{A}\}} \delta_{s'_l t'_l} \times \begin{cases} 1 & \text{if } s' \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases} && (162)
\end{aligned}
$$

$$\Rightarrow P^{0,p}_{s't'} = \begin{cases} 1 & \text{if } s' \in \mathcal{S} \text{ and } x_{t'}(s') = 1 \\ 0 & \text{otherwise} \end{cases}, \tag{163}$$

which does not depend on p. □

Claim 18. *The Euclidean gauge is equal to the embedding space, i.e.,* $\Theta_{\text{eucl}} = E_{\text{all}}$. *Moreover standard* L_2 *regularization yields parameters in* Θ_{eucl}.

Proof. In the Euclidean gauge $p(s) = \alpha^{-L}$ for all $s \in \mathcal{S}$. Consequently, $p(s') = \alpha^{-o(s')}$ for all $s' \in \mathcal{S}'$. Because $\lambda = \alpha$ in this gauge as well, the metric Λ is given by

$$\Lambda_{s't'} = \delta_{s't'} p(s') \lambda^{o(s')} = \delta_{s't'} \alpha^{-o(s')} \alpha^{o(s')} = \delta_{s't'}. \tag{164}$$

Λ is therefore the Euclidean metric. Because Θ_{eucl} is Λ-orthogonal to G_{all},

$$\Theta_{\text{eucl}} = G^{\perp}_{\text{all}} = E_{\text{all}}. \tag{165}$$

And because standard L_2 parameter regularization uses a penalty of $\sum_{s'} \theta^2_{s'} = ||\vec{\theta}||^2$, which is the Euclidean norm of $\vec{\theta}$, inference using standard L_2 regularization yields parameters in the Euclidean gauge. This completes the proof. □

Claim 19. *In the equitable gauge,* $||\vec{\theta}||^2_{\Lambda} = \sum_{s'} p(s') \theta^2_{s'} = \sum_{s'} \left\langle f^2_{s'} \right\rangle_p$ *where* $f_{s'} \equiv \theta_{s'} x_{s'}$.

Proof. The equitable gauge is defined by $\lambda = 1$. Setting $\lambda = 1$ gives $\Lambda_{s't'} = \delta_{s't'} p(s')$, and so

$$\begin{aligned} ||\theta||^2_{\Lambda} &= \sum_{s',t'} \Lambda_{s't'} \theta_{s'} \theta_{t'} && (166) \\ &= \sum_{s'} p(s') \theta^2_{s'} && (167) \\ &= \sum_{s'} \left\langle x_{s'} \right\rangle_p \theta^2_{s'} && (168) \\ &= \sum_{s'} \left\langle x^2_{s'} \right\rangle_p \theta^2_{s'} && (169) \\ &= \sum_{s'} \left\langle (\theta_{s'} x_{s'})^2 \right\rangle_p && (170) \\ &= \sum_{s'} \left\langle f^2_{s'} \right\rangle_p && (171) \end{aligned}$$

This completes the proof. □

7 Hierarchical gauges

Definition 18. *The order* $o(s')$ *of an augmented sequence* $s' \in \mathcal{S}'$ *is defined to be the number of non-star characters in* s', *i.e., the number of positions* l *for which* $s_l \in \mathcal{A}$.

Definition 19. *For any two augmented sequences* $s', t' \in \mathcal{S}'$, *we write* $t' \subseteq s'$ *if the sequences matched by* t' *form a subset of those matched by* s'. *More formally,* $t' \subseteq s'$ *iff, for all positions* l, $s'_l \in \mathcal{A} \Rightarrow t'_l = s'_l$. *Note that* $t' \subseteq s'$ *implies that* $o(t') \geq o(s')$.

Definition 20. *For any two augmented sequences* $s', t' \in \mathcal{S}'$, *we write* $t' \succeq s'$ *iff, for all positions* l, $s'_l \in \mathcal{A} \Rightarrow t'_l \in \mathcal{A}$, *or equivalently,* $t'_l = * \Rightarrow s'_l = *$. *Note that* $t' \succeq s'$ *implies that* $o(t') \geq o(s')$. *Also note that* $t' \subseteq s' \Rightarrow t' \succeq s'$, *but that the reverse is not true, since* $t' \succeq s'$ *does not require that* $t'_l = s'_l$ *when* $s'_l \in \mathcal{A}$.

For example, if $L = 8$ and $\mathcal{A} = \{\text{A}, \text{C}, \text{G}, \text{T}\}$ is the set of DNA bases, then

$$\begin{aligned} **\text{AT}**\text{C}* &\succeq ***\text{T}**\text{C}*, && (172) \\ **\text{AT}**\text{C}* &\subseteq ***\text{T}**\text{C}*, && (173) \end{aligned}$$

whereas,

$$\begin{aligned} **\text{AT}**\text{G}* &\succeq ***\text{T}**\text{C}*, && (174) \\ **\text{AT}**\text{G}* &\not\subseteq ***\text{T}**\text{C}*. && (175) \end{aligned}$$

Definition 21. *A hierarchical model is a an all-order interaction model in which $\theta_{s'} = 0$ for all augmented sequences s' in a set $\mathcal{Z}' \subseteq \mathcal{S}'$ (the "zero set" of the model) that has the following property: for every $s' \in \mathcal{Z}'$ and $t' \in \mathcal{S}'$, $t' \succeq s'$ implies that $t' \in \mathcal{Z}'$.*

Claim 20. *Hierarchical gauges preserve the form of hierarchical models. Specifically, if $\vec{\theta}$ are the parameters of a hierarchical model having zero set $\mathcal{Z}'$, then $\vec{\theta}_{\text{fixed}} = P^{\infty,p}\vec{\theta}$ are also parameters of a hierarchical model with zero set $\mathcal{Z}'$.*

Proof. Setting $\lambda = \infty$ in Eq. 22 of the main text, we see that the elements of $P^{\infty,p}_{s't'}$ can written

$$P^{\infty,p}_{s't'} = \prod_{\substack{l \text{ s.t.} \\ s'_l \in \mathcal{A} \\ t'_l \in \mathcal{A}}} \left(\delta_{s'_l t'_l} - p_l^{t'_l}\right) \times \prod_{\substack{l \text{ s.t.} \\ s'_l = * \\ t'_l \in \mathcal{A}}} p_l^{t'_l} \times \prod_{\substack{l \text{ s.t.} \\ s'_l \in \mathcal{A} \\ t'_l = *}} 0, \tag{176}$$

$$= \prod_{\substack{l \text{ s.t.} \\ s'_l \in \mathcal{A} \\ t'_l \in \mathcal{A}}} \left(\delta_{s'_l t'_l} - p_l^{t'_l}\right) \times \prod_{\substack{l \text{ s.t.} \\ s'_l = * \\ t'_l \in \mathcal{A}}} p_l^{t'_l} \times \begin{cases} 1 & \text{if } t' \succeq s' \\ 0 & \text{otherwise} \end{cases} \tag{177}$$

$$= \prod_{\{l : s'_l \in \mathcal{A}\}} \left(\delta_{s'_l t'_l} - p_l^{t'_l}\right) \times \prod_{\{l : s'_l = *\}} p_l^{t'_l} \times \begin{cases} 1 & \text{if } t' \succeq s' \\ 0 & \text{otherwise} \end{cases}. \tag{178}$$

Here we used Def. 20 in going from Eq. 176 to Eq. 177, and we used the fact that $t' \succeq s'$ implies that $s'_l \in \mathcal{A} \Rightarrow t'_l \in \mathcal{A}$, as well as the fact that $t'_l = * \Rightarrow p_l^{t'_l} = 1$, in going from Eq. 177 to Eq. 178. Now assume $\vec{\theta}$ are the parameters of a hierarchical model with zero set $\mathcal{Z}'$, and choose any $s' \in \mathcal{Z}'$. Then in the hierarchical gauge $\Theta^{\infty,p}_{s't'}$,

$$\theta^{\text{fixed}}_{s'} = \sum_{t' \in \mathcal{S}'} P^{\infty,p}_{s't'} \theta_{t'}. \tag{179}$$

$$= \sum_{t' \succeq s'} \theta_{t'} \prod_{\{l : s'_l \in \mathcal{A}\}} \left(\delta_{s'_l t'_l} - p_l^{t'_l}\right) \prod_{\{l : s'_l = *\}} p_l^{t'_l} \tag{180}$$

$$= 0 \tag{181}$$

because $t' \succeq s'$ implies that $t' \in \mathcal{Z}'$ and thus that $\theta_{t'} = 0$ for all t' in the sum. We conclude that $\theta^{\text{fixed}}_{s'} = 0$ for every $s' \in \mathcal{Z}'$, i.e., $\vec{\theta}_{\text{fixed}}$ are the parameters of a hierarchical model defined by zero set $\mathcal{Z}'$. □

Claim 21. *Parameters $\vec{\theta}_{\text{fixed}}$ in the hierarchical gauge $\Theta^{\infty,p}$ satisfy the marginalization constraint*

$$\sum_{c_k} p_{l_k}^{c_k} \theta^{c_1 \ldots c_K}_{l_1 \ldots l_K, \text{fixed}} = 0 \tag{182}$$

for every $K = 1, \ldots, L$, every subset of positions $\{l_1, \ldots, l_K\}$, and every choice of $k = 1, \ldots, K$.

Proof. From Eq. 180 in the proof of Claim 20, parameters $\vec{\theta}$ in the hierarchical gauge $\Theta^{\infty,p}$ are given in terms of unfixed parameters $\vec{\theta}$ via

$$\theta^{\text{fixed}}_{s'} = \sum_{t' \in \mathcal{S}'} P^{\infty,p}_{s't'} \theta_{t'} = \sum_{t' \succeq s'} \theta_{t'} \prod_{\{l : s'_l \in \mathcal{A}\}} \left(\delta_{s'_l t'_l} - p_l^{t'_l}\right) \prod_{\{l : s'_l = *\}} p_l^{t'_l}. \tag{183}$$

Now choose any $K \in \{1, \ldots, L\}$, any set of positions $\sigma = \{l_1, \ldots, l_K\}$, and any index $k \in \{1, \ldots, K\}$. Define $u' \in \mathcal{S}'$ to be the augmented sequence for which $u'_{l_i} = c_i$ for all $i = 1, \ldots, K$, and that has $u'_l = *$ for all $l \notin \sigma$. Further define $\mathcal{S}_{u',k} \subseteq \mathcal{S}'$ to bet the set of augmented sequences obtained by replacing the character at position k in u' with the α different characters in $\mathcal{A}$. To reduce the notational burden, we use i as a synonym for l_i when $i = 1, \ldots, K$, and use $i = K+1, \ldots, L$ to denote positions not in σ. We find that

$$\sum_{c_k} p_k^{c_k} \theta^{c_1 \ldots c_K}_{1 \ldots K, \text{fixed}} = \sum_{s' \in \mathcal{S}_{u',k}} p_k^{s'_k} \theta_{s'} \tag{184}$$

$$= \sum_{s' \in \mathcal{S}_{u',k}} p_k^{s'_k} \sum_{t' \subseteq s'} \theta_{t'} \prod_{i=1}^{K} \left(\delta_{s'_i t'_i} - p_i^{t'_i}\right) \prod_{i=K+1}^{L} p_i^{t'_i} \tag{185}$$

$$= \sum_{t' \subseteq u'} \sum_{s' \in \mathcal{S}_{u',k}} \theta_{t'} p_k^{s'_k} \prod_{i=1}^{K} \left(\delta_{s'_i t'_i} - p_i^{t'_i}\right) \prod_{i=K+1}^{L} p_i^{t'_i} \tag{186}$$

$$= \sum_{t' \subseteq u'} \sum_{c \in \mathcal{A}} \theta_{t'} p_k^c (\delta_{ct'_k} - p_k^{t'_k}) \prod_{\substack{i=1 \\ i \neq k}}^{K} \left(\delta_{u'_i t'_i} - p_i^{t'_i} \right) \prod_{i=K+1}^{L} p_i^{t'_i} \tag{187}$$

$$= \sum_{t' \subseteq u'} \theta_{t'} \left[\sum_{c \in \mathcal{A}} p_k^c (\delta_{ct'_k} - p_k^{t'_k}) \right] \prod_{\substack{i=1 \\ i \neq k}}^{K} \left(\delta_{u'_i t'_i} - p_i^{t'_i} \right) \prod_{i=K+1}^{L} p_i^{t'_i} \tag{188}$$

$$= 0. \tag{189}$$

In going from Eq. 184 to Eq. 185 we used Eq. 183. In going from Eq. 185 to Eq. 186 we used the fact that $t' \subseteq s'$ and $t' \subseteq u'$ are the same condition on t' when $s' \in \mathcal{S}_{u',k}$. In going from Eq. 186 to Eq. 187, we eliminated s' by separating the case $i = k$ out of the product over $i = 1, \ldots, K$, by replacing $s'_k \to c$, and by replacing $s'_i \to u'_i$ for all $i \neq k$. In going from Eq. 187 to Eq. 188, we collected in brackets all quantities that depend on c. And in going from Eq. 188 to Eq. 189, we use the fact that the term in brackets vanishes:

$$\sum_{c \in \mathcal{A}} p_k^c \left(\delta_{ct'_k} - p_k^{t'_k} \right) = \left(\sum_{c \in \mathcal{A}} p_k^c \delta_{ct'_k} \right) - \left(\sum_{c \in \mathcal{A}} p_k^c \right) p_k^{t'_k} = p_k^{t'_k} - p_k^{t'_k} = 0. \tag{190}$$

This proves the claim. □

Definition 22. *The $\mathcal{A}$-positions of an augmented sequence s' are the positions l such that $s'_l \in \mathcal{A}$. Similarly, the $*$-positions of s' are the positions l such that $s'_l = *$.*

Definition 23. *An augmented sequence orbit σ is a set comprising all augmented sequences that have a specified set of $\mathcal{A}$-positions (or equivalently, a specified set of $*$-positions). The order of the orbit, $o(\sigma)$, is defined to be the order of all $s' \in \sigma$. The term "orbit" comes from the fact that such sets are formed from the orbit of s' under the group of position-specific character permutations; see ref. [?].*

Claim 22. *Let $f(s) = \sum_{t'} \theta_{t'} x_{t'}(s)$ be an activity landscape and p be a positive probability distribution. Define the expectation value of f with respect to p conditioned on $s' \in \mathcal{S}'$ to be $\langle f|s' \rangle_p = \frac{1}{p(s')} \sum_{s \in s'} p(s) f(s)$. Then when $\vec{\theta}$ is in the hierarchical gauge,*

$$\langle f|s' \rangle_p = \sum_{t' \supseteq s'} \theta_{t'}. \tag{191}$$

This claim is readily extended to non-positive probability distributions p by defining $\langle f|s' \rangle_p = \lim_{\epsilon \to 0^+} \langle f|s' \rangle_{p_\epsilon}$, where p_ϵ is a regularized version of p given by

$$p_\epsilon(s) = \prod_l \left[(1 - \epsilon) p_l^{s_l} + \frac{\epsilon}{\alpha} \right], \tag{192}$$

with p_l being the position-specific factors of p.

Proof. Assume that p is positive, and that $\vec{\theta}$ is in the hierarchical gauge. Then,

$$\langle f|s' \rangle_p = \frac{1}{p(s')} \sum_{s \in s'} p(s) \sum_{t'} \theta_{t'} x_{t'}(s) \tag{193}$$

$$= \frac{1}{p(s')} \sum_{t'} \theta_{t'} \sum_{s \in s'} p(s) x_{t'}(s) \tag{194}$$

$$= \frac{1}{p(s')} \sum_{t'} \theta_{t'} \sum_{s \in s' \cap t'} p(s) \tag{195}$$

$$= \frac{1}{p(s')} \sum_{t'} p(s' \cap t') \theta_{t'} \tag{196}$$

$$= \frac{1}{p(s')} \sum_{\tau} \sum_{t' \in \tau} p(s' \cap t') \theta_{t'}. \tag{197}$$

where $\sum_\tau$ denotes a sum over all augmented sequence orbits τ. Now let $l_1, \ldots, l_K, m_1, \ldots, m_J$ denote the $\mathcal{A}$-positions of s', let $l_1, \ldots, l_K, n_1, \ldots, n_I$ denote the $\mathcal{A}$-positions of τ, and assume $l_1, \ldots, l_K, m_1, \ldots, m_J, n_1, \ldots, n_I$ are distinct. Then for each orbit τ,

$$\sum_{t' \in \tau} p(s' \cap t') \theta_{t'} = \sum_{t'_{l_1} \cdots t'_{l_K}} \sum_{t'_{n_1} \cdots t'_{n_I}} p_{l_1}^{s'_{l_1}} \cdots p_{l_K}^{s'_{l_K}} p_{m_1}^{s'_{m_1}} \cdots p_{m_J}^{s'_{m_J}} p_{n_1}^{t'_{n_1}} \cdots p_{n_I}^{t'_{n_I}} \delta_{s'_{l_1} t'_{l_1}} \cdots \delta_{s'_{l_K} t'_{l_K}} \theta_{l_1 \cdots l_K n_1 \cdots n_I}^{t'_{l_1} \cdots t'_{l_K} t'_{n_1} \cdots t'_{n_I}} \tag{198}$$

$$= p_{l_1}^{s'_{l_1}} \cdots p_{l_K}^{s'_{l_K}} p_{m_1}^{s'_{m_1}} \cdots p_{m_J}^{s'_{m_J}} \sum_{t'_{l_1} \cdots t'_{l_K}} \delta_{s'_{l_1} t'_{l_1}} \cdots \delta_{s'_{l_K} t'_{l_K}} \sum_{t'_{n_1} \cdots t'_{n_I}} p_{n_1}^{t'_{n_1}} \cdots p_{n_I}^{t'_{n_I}} \theta_{l_1 \cdots l_K n_1 \cdots n_I}^{t'_{l_1} \cdots t'_{l_K} t'_{n_1} \cdots t'_{n_I}}. \quad (199)$$

Noting that

$$\delta_{s'_{l_1} t'_{l_1}} \cdots \delta_{s'_{l_K} t'_{l_K}} = \begin{cases} 1 & \text{if } s'_l = t'_l \text{ for all } l = l_1, \ldots, l_K, \\ 0 & \text{otherwise,} \end{cases} \quad (200)$$

and that by Claim 21,

$$\sum_{t'_{n_1} \cdots t'_{n_I}} p_{n_1}^{t'_{n_1}} \cdots p_{n_I}^{t'_{n_I}} \theta_{l_1 \cdots l_K n_1 \cdots n_I}^{t'_{l_1} \cdots t'_{l_K} t'_{n_1} \cdots t'_{n_I}} = \begin{cases} \theta_{l_1 \cdots l_K}^{t'_{l_1} \cdots t'_{l_K}} & \text{if } I = 0, \\ 0 & \text{otherwise,} \end{cases} \quad (201)$$

we see that,

$$\sum_{t' \in \tau} p(s' \cap t') \theta_{t'} = p_{l_1}^{s'_{l_1}} \cdots p_{l_K}^{s'_{l_K}} p_{m_1}^{s'_{m_1}} \cdots p_{m_J}^{s'_{m_J}} \sum_{t'_{l_1} \cdots t'_{l_K}} \theta_{l_1 \cdots l_K}^{t'_{l_1} \cdots t'_{l_K}} \times \begin{cases} 1 \text{ if } t'_l \in \mathcal{A} \Rightarrow s'_l = t'_l, \text{ for all } l = l_1, \ldots, l_K, \\ 0 \text{ otherwise.} \end{cases} \quad (202)$$

$$= p(s') \sum_{t' \in \tau} \theta_{t'} \begin{cases} 1 \text{ if } s' \subseteq t', \\ 0 \text{ otherwise.} \end{cases} \quad (203)$$

Consequently,

$$\langle f | s' \rangle_p = \frac{1}{p(s')} \sum_{\tau} p(s') \sum_{t' \in \tau} \theta_{t'} \begin{cases} 1 \text{ if } s' \subseteq t', \\ 0 \text{ otherwise,} \end{cases} \quad (204)$$

$$= \sum_{t'} \theta_{t'} \begin{cases} 1 \text{ if } s' \subseteq t', \\ 0 \text{ otherwise,} \end{cases} \quad (205)$$

$$= \sum_{t' \supseteq s'} \theta_{t'}. \quad (206)$$

This completes the proof for the case where p is positive. The proof for non-positive p follows from the definition and the continuity of projection matrix elements $P_{s't'}^{\infty,p}$ (Eq. 176) with respect to p and the definition $\langle f | s' \rangle_p = \lim_{\epsilon \to 0^+} \langle f | s' \rangle_{p_\epsilon}$. □

Definition 24. *The orbital component of an activity landscape $f = \sum_{s'} \theta_{s'} x_{s'}$ corresponding to an augmented sequence orbit σ is defined to be $f_\sigma(s) = \sum_{s' \in \sigma} \theta_{s'} x_{s'}(s)$.*

Claim 23. *Given an activity landscape f, and augmented sequence orbits σ and τ,*

$$\langle f_\sigma \rangle_p = \begin{cases} \theta_0 & \text{if } o(\sigma) = 0, \\ 0 & \text{otherwise,} \end{cases} \quad (207)$$

and $\langle f_\sigma f_\tau \rangle_p = \delta_{\sigma\tau} \langle f_\sigma^2 \rangle_p$ where

$$\langle f_\sigma^2 \rangle_p = \sum_{s' \in \sigma} p(s') \theta_{s'}^2. \quad (208)$$

Proof. Eq. 207 follows directly from Claim 21:

$$\langle f_\sigma \rangle_p = \sum_{s' \in \sigma} p(s') \theta_{s'} \begin{cases} \theta_0 & \text{if } o(\sigma) = 0, \\ 0 & \text{otherwise,} \end{cases} \quad (209)$$

Let $l_1, \ldots, l_K, m_1, \ldots, m_J$ denote the $\mathcal{A}$-positions of σ, let $l_1, \ldots, l_K, n_1, \ldots, n_I$ denote the $\mathcal{A}$-positions of τ, and assume $l_1, \ldots, l_K, m_1, \ldots, m_J, n_1, \ldots, n_I$ are distinct. Then,

$$\langle f_\sigma f_\tau \rangle_p = \sum_u p(u) \left[\sum_{s' \in \sigma} \theta_{s'} x_{s'}(u) \right] \left[\sum_{t' \in \tau} \theta_{t'} x_{t'}(u) \right] \quad (210)$$

$$= \sum_{s' \in \sigma} \sum_{t' \in \tau} \theta_{s'} \theta_{t'} \sum_u p(u) x_{s' \cap t'}(u) \quad (211)$$

$$= \sum_{s' \in \sigma} \sum_{t' \in \tau} p(s' \cap t') \theta_{s'} \theta_{t'} \quad (212)$$

$$= \sum_{s'_{l_1} \cdots s'_{l_K}} \sum_{s'_{m_1} \cdots s'_{m_J}} \sum_{t'_{l_1} \cdots t'_{l_K}} \sum_{t'_{n_1} \cdots t'_{n_I}} p_{l_1}^{s'_{l_1}} \cdots p_{l_K}^{s'_{l_K}} p_{m_1}^{s'_{m_1}} \cdots p_{m_J}^{s'_{m_J}} p_{n_1}^{t'_{n_1}} \cdots p_{n_I}^{t'_{n_I}} \times \quad (213)$$

$$\delta_{s'_{l_1}t'_{l_1}}\cdots\delta_{s'_{l_K}t'_{l_K}}\theta^{s'_{l_1}\cdots s'_{l_K}s'_{m_1}\cdots s'_{m_J}}_{l_1\cdots l_K m_1\cdots m_J}\theta^{t'_{l_1}\cdots t'_{l_K}t'_{n_1}\cdots t'_{n_I}}_{l_1\cdots l_K n_1\cdots n_I} \tag{214}$$

$$= \sum_{s'_{l_1}\cdots s'_{l_K}}\sum_{t'_{l_1}\cdots t'_{l_K}} p^{s'_{l_1}}_{l_1}\cdots p^{s'_{l_K}}_{l_K}\delta_{s'_{l_1}t'_{l_1}}\cdots\delta_{s'_{l_K}t'_{l_K}}\times \tag{215}$$

$$\left[\sum_{s'_{m_1}\cdots s'_{m_J}} p^{s'_{m_1}}_{m_1}\cdots p^{s'_{m_J}}_{m_J}\theta^{s'_{l_1}\cdots s'_{l_K}s'_{m_1}\cdots s'_{m_J}}_{l_1\cdots l_K m_1\cdots m_J}\right]\times\left[\sum_{t'_{n_1}\cdots t'_{n_I}} p^{t'_{n_1}}_{n_1}\cdots p^{t'_{n_I}}_{n_I}\theta^{t'_{l_1}\cdots t'_{l_K}t'_{n_1}\cdots t'_{n_I}}_{l_1\cdots l_K n_1\cdots n_I}\right]. \tag{216}$$

Because

$$\sum_{s'_{m_1}\cdots s'_{m_J}} p^{s'_{m_1}}_{m_1}\cdots p^{s'_{m_J}}_{m_J}\theta^{s'_{l_1}\cdots s'_{l_K}s'_{m_1}\cdots s'_{m_J}}_{l_1\cdots l_K m_1\cdots m_J} = \begin{cases}\theta^{s'_{l_1}\cdots s'_{l_K}}_{s_1\cdots s_K} & \text{if } J=0,\\ 0 & \text{otherwise,}\end{cases} \tag{217}$$

and

$$\sum_{t'_{n_1}\cdots t'_{n_I}} p^{t'_{n_1}}_{n_1}\cdots p^{t'_{n_I}}_{n_I}\theta^{t'_{l_1}\cdots t'_{l_K}t'_{n_1}\cdots t'_{n_I}}_{l_1\cdots l_K n_1\cdots n_I} = \begin{cases}\theta^{t'_{l_1}\cdots t'_{l_K}}_{l_1\cdots l_K} & \text{if } I=0,\\ 0 & \text{otherwise,}\end{cases} \tag{218}$$

we see that the summand in Eq. 215 vanishes unless $J=0$ and $I=0$. This is equivalent to the requirement that $\sigma=\tau$. Therefore,

$$\langle f_\sigma f_\tau\rangle_p = \delta_{\sigma\tau}\sum_{s'_{l_1}\cdots s'_{l_K}}\sum_{t'_{l_1}\cdots t'_{l_K}} p^{s'_{l_1}}_{l_1}\cdots p^{s'_{l_K}}_{l_K}\delta_{s'_{l_1}t'_{l_1}}\cdots\delta_{s'_{l_K}t'_{l_K}}\theta^{s'_{l_1}\cdots s'_{l_K}}_{s_1\cdots s_K}\theta^{t'_{l_1}\cdots t'_{l_K}}_{l_1\cdots l_K} \tag{219}$$

$$= \delta_{\sigma\tau}\sum_{s'\in\sigma}\sum_{t'\in\tau} p(s')\delta_{s't'}\theta_{s'}\theta_{t'} \tag{220}$$

$$= \delta_{\sigma\tau}\sum_{s'\in\sigma} p(s')\theta^2_{s'} \tag{221}$$

$$= \delta_{\sigma\tau}\left\langle f^2_\sigma\right\rangle_p \tag{222}$$

where

$$\left\langle f^2_\sigma\right\rangle_p = \sum_{s'\in\sigma} p(s')\theta^2_{s'}. \tag{223}$$

□

Definition 25. *Given $k\in\{0,1,\ldots,L\}$, the k'the order component of an activity landscape $f=\sum_{s'}\theta_{s'}x_{s'}$ is defined to be*

$$f_k(s) = \sum_{s':o(s')=k}\theta_{s'}x_{s'}(s). \tag{224}$$

Claim 24. *Given an activity landscape $f=\sum_{s'}\theta_{s'}x_{s'}$, and parameters $\vec{\theta}$ expressed in the hierarchical gauge,*

$$\mathrm{var}_p[f] = \sum_{k=0}^{L}\mathrm{var}_p[f_k], \quad \textit{where} \quad \mathrm{var}_p[f_k] = \begin{cases}\sum_{s':o(s')=k} p(s')\theta^2_{s'} & \textit{if } k\geq 1,\\ 0 & \textit{if } k=0.\end{cases} \tag{225}$$

Proof. First we decompose each f_k into a sum of f_σ over augmented sequence orbits σ of order k:

$$f_k = \sum_{\sigma:o(\sigma)=k} f_\sigma. \tag{226}$$

next, by Claim 23,

$$\langle f_k\rangle_p = \sum_{\sigma:o(\sigma)=k}\langle f_\sigma\rangle_p, \tag{227}$$

and

$$\left\langle f^2_k\right\rangle_p = \left\langle\sum_{\sigma:o(\sigma)=k} f_\sigma\sum_{\tau:o(\tau)=k} f_\tau\right\rangle_p \tag{228}$$

$$= \sum_{\sigma:o(\sigma)=k}\sum_{\tau:o(\tau)=k}\langle f_\sigma f_\tau\rangle_p \tag{229}$$

$$= \sum_{\sigma:o(\sigma)=k} \sum_{\tau:o(\tau)=k} \delta_{\sigma\tau} \left\langle f_\sigma^2 \right\rangle \tag{230}$$

$$= \sum_{\sigma:o(\sigma)=k} \left\langle f_\sigma^2 \right\rangle \tag{231}$$

$$= \sum_{\sigma:o(\sigma)=k} \sum_{s'\in\sigma} p(s')\theta_{s'}^2 \tag{232}$$

$$= \sum_{s':o(s')=k} p(s')\theta_{s'}^2. \tag{233}$$

Consequently

$$\mathrm{var}_p[f_k] = \left\langle f_k^2 \right\rangle_p - \langle f_k \rangle_p^2 \tag{234}$$

$$= \sum_{s':o(s')=k} p(s')\theta_{s'}^2 - \delta_{k0}\theta_0^2 \tag{235}$$

$$= \begin{cases} \sum_{s':o(s')=k} p(s')\theta_{s'}^2 & \text{if } k \geq 1, \\ 0 & \text{if } k = 0. \end{cases} \tag{236}$$

Finally,

$$\mathrm{var}_p[f] = \left\langle f^2 \right\rangle_p - \langle f \rangle_p^2 \tag{237}$$

$$= \left\langle \sum_\sigma f_\sigma \sum_\tau f_\tau \right\rangle_p - \left\langle \sum_\sigma f_\sigma \right\rangle_p^2 \tag{238}$$

$$= \sum_\sigma \sum_\tau \langle f_\sigma f_\tau \rangle - \sum_\sigma \langle f_\sigma \rangle_p^2 \tag{239}$$

$$= \sum_\sigma \left[\left\langle f_\sigma^2 \right\rangle_p - \langle f_\sigma \rangle^2 \right] \tag{240}$$

$$= \sum_{k=0}^{L} \sum_{\sigma:o(\sigma)=k} \left[\left\langle f_\sigma^2 \right\rangle_p - \langle f_\sigma \rangle^2 \right] \tag{241}$$

$$= \sum_{k=0}^{L} \mathrm{var}_p[f_k]. \tag{242}$$

This completes the proof. □

Gauge-fixing formula for the all-order interaction model Using the formula for $P^{\infty,p}$ in Eq. 176 to compute

$$\vec{\theta}_{\text{fixed}} = P^{\infty,p}\vec{\theta} \tag{243}$$

for an all-order interaction model in which only the zero-order, first-order, and second-order parameters are nonzero, one finds that

$$\theta_{0,\text{fixed}} = \theta_0 + \sum_l \sum_c p_l^c \theta_l^c + \sum_l \sum_{l'>l} \sum_{c,c'} p_l^c p_{l'}^{c'} \theta_{ll'}^{cc'}, \tag{244}$$

$$\theta_{l,\text{fixed}}^c = \sum_{c'} (\delta_{cc'} - p_l^{c'})\theta_l^c + \sum_{l'<l} \sum_{c',c''} (\delta_{cc'} - p_l^{c'}) p_{l'}^{c''} \theta_{ll'}^{cc'} + \sum_{l'>l} \sum_{c',c''} (\delta_{cc'} - p_l^{c'}) p_{l'}^{c''} \theta_{l'l}^{c'c}, \tag{245}$$

$$\theta_{ll',\text{fixed}}^{cc'} = \sum_{c'',c'''} (\delta_{cc''} - p_l^{c''})(\delta_{c'c'''} - p_{l'}^{c'''}) \theta_{ll'}^{c''c'''}, \tag{246}$$

$$\theta_{l_1\ldots l_K,\text{fixed}}^{c_1\ldots c_K} = 0 \quad \text{for all} \quad K = 3, \ldots, L. \tag{247}$$

Ignoring the formula for parameters of order three or greater, one thus obtains the gauge-fixing formulae for the parameters of the pairwise-interaction model. These are the formulas used for the computations in Fig. 4 and Fig. 5 of the main text. The specific choices for p used in these figures are given below.

Region-specific distributions. In Fig. 4D, the probability distributions $p(s) = \prod_{l=1}^{4} p_l^{s_l}$ for the four different regions (global, region 1, region 2, region 3) were defined as follows.

- Uniform: For $l \in \{1,2,3,4\}$, $p_l^c = \frac{1}{20}$.
- Region 1: For $l \in \{1,2,4\}$, $p_l^c = \frac{1}{20}$; for $l = 3$, $p_l^c = \delta_{c\mathrm{G}}$.
- Region 2: For $l \in \{1,2\}$, $p_l^c = \frac{1}{20}$; for $l = 3$, $p_l^c = \frac{1}{2}\delta_{c\mathrm{L}} + \frac{1}{2}\delta_{c\mathrm{F}}$; for $l = 4$, $p_l^c = \delta_{c\mathrm{G}}$.
- Region 3: For $l \in \{1,2\}$, $p_l^c = \frac{1}{20}$; for $l = 3$, $p_l^c = \frac{1}{2}\delta_{c\mathrm{C}} + \frac{1}{2}\delta_{c\mathrm{A}}$; for $l = 4$, $p_l^c = \delta_{c\mathrm{A}}$.

Here $l = 1,2,3,4$ are used to denote protein positions $39, 40, 41, 54$, respectively, and c indexes all $\alpha = 20$ possible amino acids in $\mathcal{A}$.

8 Appendix

Claim 25. *Let V_1 and V_2 be two subspaces of a vector space V such that any vector in V can be uniquely decomposed into the sum of a vector in V_1 and a vector in V_2. Let P_1 be the projection into V_1 along V_2, and P_2 be the projection into V_2 along V_1. Let Λ be a symmetric positive definite matrix acting on V. Then the following three statements are equivalent.*

1. *V_1 and V_2 are Λ-orthogonal, i.e., $\vec{v}_1^{\top}\Lambda\vec{v}_2 = 0$ for all $\vec{v}_1 \in V_1$ and $\vec{v}_2 \in V_2$.*
2. *For any fixed $\vec{v_1} \in V_1$, $\operatorname{argmin}_{\vec{v}_2 \in V_2}(\vec{v}_1 + \vec{v}_2)^{\top}\Lambda(\vec{v}_1 + \vec{v}_2) = \vec{0}$.*
3. *$\Lambda = P_1^{\top}\Lambda P_1 + P_2^{\top}\Lambda P_2$.*

Proof. We prove equivalence of the three statements (denoted 1, 2, and 3) as follows.

- $1 \Rightarrow 2$: Assume that 1 is true. Then for all $\vec{v}_1 \in V_1$ and $\vec{v}_2 \in V_2$, $(\vec{v}_1 + \vec{v}_2)^{\top}\Lambda(\vec{v}_1 + \vec{v}_2) = \vec{v}_1^{\top}\Lambda\vec{v}_1 + \vec{v}_2^{\top}\Lambda\vec{v}_2 \geq \vec{v}_1^{\top}\Lambda\vec{v}_1$. Because equality obtains only when $\vec{v}_2 = \vec{0}$, $\vec{0}$ is the unique $\vec{v}_2 \in V$ that minimizes $(\vec{v}_1 + \vec{v}_2)^{\top}\Lambda(\vec{v}_1 + \vec{v}_2)$. This proves 2, thereby establishing that $1 \Rightarrow 2$.
- $2 \Rightarrow 1$: Assume that 1 is not true, i.e., there exists $\vec{v}_1 \in V_1$ and $\vec{v}_2 \in V_2$ such that $\vec{v}_1^{\top}\Lambda\vec{v}_2 \neq 0$. Then $\frac{d}{d\epsilon}(\vec{v}_1 + \epsilon\vec{v}_2)^{\top}\Lambda(\vec{v}_1 + \epsilon\vec{v}_2) = 2\vec{v}_1^{\top}\Lambda\vec{v}_2 + 2\epsilon\vec{v}_2^{\top}\Lambda\vec{v}_2$ is nonzero at $\epsilon = 0$. This contradicts 2, thereby establishing that $\neg 1 \Rightarrow \neg 2$ and hence $2 \Rightarrow 1$.
- $1 \Rightarrow 3$: Assume that 1 is true, and choose any $\vec{v}, \vec{w} \in V$. Then $\vec{v}^{\top}\Lambda\vec{w} = (P_1\vec{v}+P_2\vec{v})^{\top}\Lambda(P_1\vec{w}+P_2\vec{w}) = (P_1\vec{v})^{\top}\Lambda(P_1\vec{w})+(P_2\vec{v})^{\top}\Lambda(P_2\vec{w}) = \vec{v}^{\top}\left[P_1^{\top}\Lambda P_1 + P_2^{\top}\Lambda P_2\right]\vec{w}$. This proves 3, thereby establishing that $1 \Rightarrow 3$.
- $3 \Rightarrow 1$: Assume that 3 is true. Then given any $\vec{v}_1 \in V_1$ and $\vec{v}_2 \in V_2$, $\vec{v_1}^{\top}\Lambda\vec{v}_2 = (P_1\vec{v_1})^{\top}\Lambda(P_1\vec{v}_2)+(P_2\vec{v_1})^{\top}\Lambda(P_2\vec{v}_2) = 0$ since $P_1\vec{v}_2 = P_2\vec{v_1} = \vec{0}$. This proves 1, thereby establishing that $3 \Rightarrow 1$.

□