Symmetry, gauge freedoms, and the interpretability of sequence-function relationships

Anna Posfai, David M. McCandlish, and Justin B. Kinney*
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA

(Received 14 August 2024; accepted 25 March 2025; published 2 April 2025)

Quantitative models that describe how biological sequences encode functional activities are ubiquitous in modern biology. One important aspect of these models is that they commonly exhibit gauge freedoms, i.e., directions in parameter space that do not affect model predictions. In physics, gauge freedoms arise when physical theories are formulated in ways that respect fundamental symmetries. However, the connections that gauge freedoms in models of sequence-function relationships have to the symmetries of sequence space have yet to be systematically studied. In this work we study the gauge freedoms of models that respect a specific symmetry of sequence space: the group of position-specific character permutations. We find that gauge freedoms arise when model parameters transform under redundant irreducible matrix representations of this group. Based on this finding, we describe an "embedding distillation" procedure that enables both analytic calculation of the number of independent gauge freedoms and efficient computation of a sparse basis for the space of gauge freedoms. We also study how parameter transformation behavior affects parameter interpretability. We find that in many (and possibly all) nontrivial models, the ability to interpret individual model parameters as quantifying intrinsic allelic effects requires that gauge freedoms be present. This finding establishes an incompatibility between two distinct notions of parameter interpretability. Our work thus advances the understanding of symmetries, gauge freedoms, and parameter interpretability in models of sequence-function relationships.

DOI: 10.1103/PhysRevResearch.7.023005

I. INTRODUCTION

Understanding the quantitative nature of sequence-function relationships is a major goal of modern biology [1]. To study a sequence-function relationship of interest, researchers often propose a mathematical model, fit the parameters of the model to data, then biologically interpret the resulting parameter values. This interpretation step is complicated, however, by gauge freedoms—directions in parameter space along which model parameters can be changed without altering model predictions. When gauge freedoms are present in a model, the values of individual model parameters cannot be meaningfully interpreted without additional constraints. In standard Potts models of proteins, for example, the values of the parameters representing interactions between amino acids cannot be directly interpreted as quantifying interaction strength. This is because gauge freedoms make it possible to change any specific coupling parameter of interest without affecting model predictions by also making appropriate compensatory changes to other model parameters [2–6].

Researchers who study sequence-function relationships using quantitative models routinely encounter gauge freedoms. In practice, one of two methods is used to overcome the difficulties that gauge freedoms present. One method, called "gauge fixing," removes gauge freedoms by introducing additional constraints on model parameters [2–17]. Another method limits the mathematical models that one uses to models that do not have any gauge freedoms in the first place [18–23]. Despite being frequently encountered in the course of research, the gauge freedoms present in models of sequence-function relationships have received only limited attention (e.g., Refs. [2,4–6,12,24]). In particular, the mathematical properties of these gauge freedoms have yet to be systematically studied.

In physics, by contrast, gauge freedoms are well recognized as a topic of fundamental importance [25]. Gauge freedoms arise when a physical theory is expressed in a form that manifestly respects fundamental symmetries. For example, the classical theory of electricity and magnetism (E&M) is invariant to Lorentz transformations, i.e., relativistic changes in an observer's velocity [26]. Lorentz invariance is obscured, however, when the equations of E&M are expressed directly in terms of electric and magnetic fields. To express these equations in a form that is manifestly Lorentz invariant, one must instead formulate them in terms of an electromagnetic four-potential. Doing this introduces gauge freedoms because the four-potential, unlike electric and magnetic fields, is neither directly measurable nor uniquely determined by the configuration of a physical system [27]. Nevertheless, working with the four-potential simplifies the equations of E&M and can aid in both their solution and their physical interpretation.

Motivated by the connection between gauge freedoms and symmetries in physics, we asked whether gauge freedoms in

*Contact author: jkinney@cshl.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

models of sequence-function relationships have a connection to the symmetries of sequence space, i.e., the possible ways of transforming the space of sequences without altering the Hamming distances between sequences. In this work we study the gauge freedoms of linear models that are equivariant under a specific symmetry group of sequence space—the group of position-specific character permutations (PSCP). Here "linear models" are models that can be expressed as a sum of sequence features, each multiplied by a corresponding parameter; "PSCP" encompasses transformations that permute the identities of the individual characters (e.g., DNA bases or protein amino acids) at one sequence position, as well as transformations built from combinations of such permutations; and "equivariant" describes models for which linear transformations of the model parameters are able to compensate for the effects of PSCP transformations of sequences. Equivariant linear models include many of the most commonly used models in the literature, such as models with pairwise and/or higher-order interactions.

Using techniques from the theory of matrix representations of the symmetric group [28], we find that the gauge freedoms of these linear equivariant models arise when model parameters transform under redundant irreducible representations of PSCP. Based on this finding, we introduce an "embedding distillation" procedure that, for any linear equivariant model, facilitates both analytical and computational analyses of the vector space of gauge freedoms. We also study the connection between parameter interpretability and model transformation behavior. We find that in many (and possibly all) nontrivial models, the ability to interpret model parameters as quantifying the intrinsic effects of alleles requires that the model have gauge freedoms. This finding shows that models having gauge freedoms can have important advantages over models that have no gauge freedoms.

A companion paper [29] reports specific gauge-fixing strategies that can be applied to an important subset of the linear equivariant models, one that includes the most commonly used models of sequence-function relationships. It also describes specific ways of using these gauge-fixing strategies to assist in the development and biological interpretation of such models.

II. BACKGROUND

We now establish definitions and notation used under Results. We also review basic results regarding gauge freedoms in mathematical models of sequence-function relationships. Our companion paper [29] provides an expanded discussion of these results together with corresponding proofs.

A. Sequence-function relationships

Let $\mathcal{A}$ denote an alphabet comprising α distinct characters, let $\mathcal{S}$ denote the set of α^L sequences of length L built from these characters, and let $s_l \in \mathcal{A}$ denote the character at position l in any sequence $s \in \mathcal{S}$. A real-valued model of a sequence-function relationship, $f(s;\vec{\theta})$, is defined to be a function that maps each sequence s to a real number. The vector $\vec{\theta}$ denotes the parameters of the model and is assumed to comprise M real numbers. For technical reasons it is sometimes useful to consider complex-valued models of sequence-function relationships, which are defined analogously.

B. Linear models

Linear models of sequence-function relationships are linear in $\vec{\theta}$ and thus have the form

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

for all $s \in \mathcal{S}$. Here $\vec{x}(\cdot)$ is an M-dimensional vector of sequence features, and each feature $x_i(\cdot)$ is a function that maps $\mathcal{S}$ to $\mathbb{R}$. We refer to the space $\mathbb{R}^M$ in which $\vec{x}$ and $\vec{\theta}$ live as feature space [30].

An example of a linear model is the pairwise one-hot model, which has the form

$$f^{\text{ohe}}_{\text{pair}} = \theta_0 x_0 + \sum_{l}\sum_{c} \theta^c_l x^c_l + \sum_{l<l'}\sum_{c,c'} \theta^{cc'}_{ll'} x^{cc'}_{ll'}, \tag{2}$$

where the arguments of both the model and features have been kept implicit. In Eq. (2), $l, l' \in \{1, \ldots, L\}$ index the positions within each sequence, and $c, c' \in \mathcal{A}$ index the possible characters at these positions. We use the superscript "ohe" here and in what follows to indicate mathematical objects (such as embeddings, models, and representations) that are based on one-hot embeddings. Pairwise one-hot models, in particular, make use of the pairwise one-hot embedding $\vec{x}^{\,\text{ohe}}_{\text{pair}}(s)$, the elements of which represent three types of features: the constant feature, $x_0(s)$, which equals one for every sequence s; additive one-hot features, $x^c_l(s)$, which equal one if $s_l = c$ and equal zero otherwise; and pairwise one-hot features, $x^{cc'}_{ll'}(s)$, which equal one if both $s_l = c$ and $s_{l'} = c'$, and equal zero otherwise.

C. Gauge freedoms

Gauge freedoms are transformations of model parameters that do not affect model predictions. Formally, a gauge freedom is any vector $\vec{g} \in \mathbb{R}^M$ that satisfies

$$f(s;\vec{\theta}) = f(s;\vec{\theta} + \vec{g}) \quad \text{for all } s \in \mathcal{S}. \tag{3}$$

For linear sequence-function relationships the set of gauge freedoms, denoted by G, forms a vector space in $\mathbb{R}^M$. It is readily shown that G is the orthogonal complement of the space spanned by sequence embeddings [29]. In what follows, we use γ to represent the dimension of G, i.e., the number of (independent) gauge freedoms.

Gauge freedoms arise from linear dependencies among sequence features. By inspection we see that $f^{\text{ohe}}_{\text{pair}}$ has

$$M^{\text{ohe}}_{\text{pair}} = 1 + \alpha L + \binom{L}{2}\alpha^2 \tag{4}$$

parameters. However, it turns out that the space spanned by the corresponding embedding $\vec{x}^{\,\text{ohe}}_{\text{pair}}$ has only $1 + (\alpha - 1)L + \binom{L}{2}(\alpha - 1)^2$ dimensions. This difference reflects the presence of $L + \binom{L}{2}(2\alpha - 1)$ constraints on the features, namely $x_0 = \sum_{c'} x^{c'}_l$ for all positions l (yielding 1 constraint per position) and $x^c_l = \sum_{c'} x^{cc'}_{ll'}, x^{c'}_{l'} = \sum_c x^{cc'}_{ll'}$ for all pairs of positions

$l < l'$ and all choices of character c or c' (yielding $2\alpha - 1$ independent constraints per pair of positions). The model $f^{\text{ohe}}_{\text{pair}}$ therefore has

$$\gamma^{\text{ohe}}_{\text{pair}} = L + \binom{L}{2}(2\alpha - 1) \tag{5}$$

gauge freedoms. See our companion paper [29] for more details, as well as Refs. [2,4,6,10] for earlier treatments of gauge freedoms in the pairwise one-hot model.

D. Fixing the gauge

Fixing the gauge is the process of removing gauge freedoms by restricting $\vec{\theta}$ to a subset Θ of parameter space called "the gauge." For example, the commonly used "zero-sum gauge" [4,6] for the pairwise one-hot model is the subspace of parameter space in which the additive parameters at every position sum to zero when marginalized over characters ($\sum_c \theta^c_l = 0$ for every l) and the pairwise parameters at all pairs of positions sum to zero when marginalized over the characters at either position ($\sum_c \theta^{cc'}_{ll'} = 0$ and $\sum_{c'} \theta^{cc'}_{ll'} = 0$ for every l, l', c, c').

Linear gauges are choices of Θ that are vector spaces. The zero-sum gauge is one such linear gauge. A useful property of linear gauges is that gauge-fixing can be accomplished through linear projection. Specifically, for any linear gauge Θ, there exists a projection matrix P that projects each parameter vector $\vec{\theta} \in \mathbb{R}^M$ onto an equivalent parameter vector $\vec{\theta}_{\text{fixed}} \in \Theta$ via $\vec{\theta}_{\text{fixed}} = P\,\vec{\theta}$. Our companion paper describes a parametric family of linear gauges (including an explicit formula for the corresponding projection matrices) that includes as special cases many of the most commonly used gauges in the literature [29].

III. RESULTS

We begin this section by formally defining the group of PSCP transformations, as well as the notion of model equivariance under this group. We then illustrate, for two example pairwise-interaction models, how transformation behavior under PSCP impacts both gauge freedoms and a specific type of parameter interpretability, namely the ability to assign intrinsic effects to individual alleles. Next we formally investigate this relationship more generally using methods from the theory of group representations [28]. In doing so, we establish an "embedding distillation" procedure that, for any equivariant model, enables analytic calculation of the number of gauge freedoms. We also establish an algorithm that enables the efficient computation of a sparse basis for the space of gauge freedoms. We conclude by revisiting the issue of parameter interpretability in light of these results.

A. Position-specific character permutations (PSCP)

Different transformations of sequence space impact models of sequence-function relationships in different ways. Here we focus on PSCP transformations. These transformations of sequence space form a mathematical group, which we denote by H_{PSCP}. The action of a transformation $h \in H_{\text{PSCP}}$ on a sequence $s \in \mathcal{S}$ is written hs. H_{PSCP} is a symmetry group of sequence space in that its transformations preserve the Hamming distances between sequences. There are other symmetry groups of sequence space as well, but we ultimately find that these symmetry groups do not have the same connections to gauge freedoms that H_{PSCP} does [discussed below and in Supplemental Material (SM) Sec. 7].

(a)

	character at position l			feat.
	A	B	C	
one-hot embedding ($\vec{x}^{\text{ohe}}_l$)	$\begin{bmatrix}1\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\1\end{bmatrix}$	x^{A}_l x^{B}_l x^{C}_l
simplex embedding ($\vec{x}^{\text{sim}}_l$)	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}0\\1\end{bmatrix}$	$\begin{bmatrix}-1\\-1\end{bmatrix}$	x^1_l x^2_l

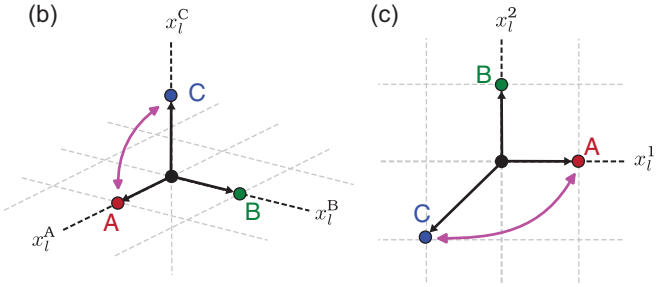

FIG. 1. Transformation behavior of two single-position embeddings. (a) Two possible embeddings of the characters at position l of a sequence built from the three-character alphabet $\mathcal{A} = \{\text{A}, \text{B}, \text{C}\}$: the three-dimensional one-hot embedding $\vec{x}^{\text{ohe}}_l$ and the two-dimensional simplex embedding $\vec{x}^{\text{sim}}_l$. The elements of $\vec{x}^{\text{ohe}}_l$ are the three one-hot sequence features x^A_l, x^B_l, and x^C_l. The two elements of $\vec{x}^{\text{sim}}_l$ are denoted x^1_l and x^2_l. (b) The three-dimensional embedding $\vec{x}^{\text{ohe}}$ for each possible character at position l. (c) The two-dimensional embedding $\vec{x}^{\text{sim}}_l$ for each possible character at position l. Pink arrows in panels (b) and (c) indicate the transformation of each embedding vector induced by permuting the characters A and C.

B. Equivariance

We also focus on equivariant linear models of sequence-function relationships. These are models for which both embeddings and parameters transform linearly under H_{PSCP}. The specific sets of matrices that encode these linear transformations are called "representations" [28]. In general, a representation R of a group H is a function that maps each $h \in H$ to a matrix $R(h)$ in a way that preserves the multiplicative structure of H, i.e., $R(h_1 h_2) = R(h_1)R(h_2)$ for any two group elements $h_1, h_2 \in H$. The degree of the representation R (denoted $\deg R$) is the dimension of the vector space on which R acts. Two different examples of representations for the same group are described below [see Eqs. (8) and (11)] and illustrated in Fig. 1.

Formally, we say that an embedding $\vec{x}$ is equivariant in H if and only if there is a representation R of H such that

$$\vec{x}(hs) = R(h)\,\vec{x}(s) \tag{6}$$

for all $h \in H$ and all $s \in \mathcal{S}$. We also say that a model is equivariant if and only if it has an equivariant embedding. For an

equivariant model whose embedding transforms as in Eq. (6), the transformation of $\mathcal{S}$ by any $h \in H$ can be compensated for by the transformation of $\vec{\theta}$ by $R(h)^{-1\top}$, in the sense that

$$f(s;\vec{\theta}\,) = f(hs; R(h)^{-1\top}\vec{\theta}\,) \tag{7}$$

for every $s \in \mathcal{S}$ and every $\vec{\theta} \in \mathbb{R}^M$ (see SM Sec. 3.2). Although linear models of sequence-function relationships can be equivariant in a variety of symmetry groups H, we use the term "equivariant" to specifically refer to equivariance under H_{PSCP} unless otherwise noted.

C. One-hot models

The most commonly used equivariant models are based on single-position one-hot embeddings. Such models are arguably the most intuitive, as their features are built from the indicator functions for single-position alleles (e.g., the nucleotides in a DNA sequence or the amino acids in a protein sequence). We denote the single-position one-hot embedding for position l as $\vec{x}_l^{\text{ohe}}$ and define it to be a binary vector of dimension α with features $x_l^{c_1}, \ldots, x_l^{c_\alpha}$, where $c_1, \ldots, c_\alpha$ is an ordering of the characters in $\mathcal{A}$. For example, Fig. 1(a) shows $\vec{x}_l^{\text{ohe}}$ for the three-character alphabet $\mathcal{A} = \{\text{A}, \text{B}, \text{C}\}$.

The embedding $\vec{x}_l^{\text{ohe}}$ transforms under what is known as a "permutation representation" [28]. We denote this representation as R_l^{ohe}. For example, consider the transformation $h_{\text{A}\leftrightarrow\text{C}}$ that exchanges characters A and C at every position in a sequence. The effect of this transformation on $\vec{x}_l^{\text{ohe}}$ [Fig. 1(b)] is equivalent to multiplying $\vec{x}_l^{\text{ohe}}$ by the matrix

$$R_l^{\text{ohe}}(h_{\text{A}\leftrightarrow\text{C}}) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}. \tag{8}$$

This and all other matrices in the representation R_l^{ohe} are permutation matrices, in that all matrix elements are 0 or 1, and each row and column contains a single 1. Consequently, multiplying a vector by one of these matrices changes the order of the elements in that vector but does not change the overall set values that those elements take. We refer to $\vec{x}_l^{\text{ohe}}$ and other embeddings that transform under permutation representations as permutation embeddings; their corresponding models are called permutation models.

The embeddings of many different models can be built by taking direct sums of Kronecker products of $\vec{x}_l^{\text{ohe}}$. For example, the pairwise one-hot model of Eq. (2) is based on the embedding

$$\vec{x}_{\text{pair}}^{\text{ohe}} = \vec{x}^{\text{triv}} \oplus \left\{ \bigoplus_l \vec{x}_l^{\text{ohe}} \right\} \oplus \left\{ \bigoplus_{l<l'} \vec{x}_{ll'}^{\text{ohe}} \right\}, \tag{9}$$

where $\vec{x}^{\text{triv}}$ denotes the trivial embedding (defined to be the one-dimensional vector [1] for all sequences) and the Kronecker product

$$\vec{x}_{ll'}^{\text{ohe}} = \vec{x}_l^{\text{ohe}} \otimes \vec{x}_{l'}^{\text{ohe}} \tag{10}$$

yields an α^2-dimensional embedding having elements $x_{ll'}^{cc'}(s) = x_l^c(s)\, x_{l'}^{c'}(s)$ for all $s \in \mathcal{S}$ and all characters $c, c' \in \mathcal{A}$. The direct sums in Eq. (9) yield $\vec{x}_{\text{pair}}^{\text{ohe}}$ by stacking the component embeddings on top of one another in the resulting column vector. Note that because $\vec{x}_l^{\text{ohe}}$ is a permutation embedding, so is $\vec{x}_{\text{pair}}^{\text{ohe}}$. In fact, any embedding constructed from a direct sum of Kronecker products of $\vec{x}_l^{\text{ohe}}$ is a permutation embedding. We call this class of models the "generalized one-hot models."

How a single-position embedding transforms has important consequences for how the parameters of models constructed from that embedding are interpreted. For the pairwise one-hot model, the fact that $\vec{x}_l^{\text{ohe}}$ transforms under a permutation representation implies that both $\vec{x}_{\text{pair}}^{\text{ohe}}$ and $\vec{\theta}_{\text{pair}}^{\text{ohe}}$ transform under permutation representations as well. A consequence of this is that the individual parameters in $\vec{\theta}_{\text{pair}}^{\text{ohe}}$ can be interpreted as quantifying the intrinsic effects of individual alleles. For example, the transformation $h_{\text{A}\leftrightarrow\text{C}}$ induces a permutation of parameters that exchanges $\theta_l^A \leftrightarrow \theta_l^C$ at all positions l, exchanges $\theta_{ll'}^{\text{AA}} \leftrightarrow \theta_{ll'}^{CC}$ at all pairs of positions $l < l'$, and so on. Model parameters therefore track their corresponding alleles: θ_l^A tracks sequences that have A at position l, $\theta_{ll'}^{\text{AA}}$ tracks sequences that have AA at positions l and l', etc.

The fact that $\vec{x}_l^{\text{ohe}}$ transforms under a permutation representation also means that the features therein are not linearly independent. For example, the three embedding vectors in Fig. 1(b) lie within a two-dimensional affine subspace defined by the constraint $x_l^A + x_l^B + x_l^C = 1$. As we will see, a consequence of such constraints is that embeddings (like $\vec{x}_{\text{pair}}^{\text{ohe}}$) that are built from direct sums of Kronecker products of single-position one-hot embeddings will yield models that have gauge freedoms. So although the parameters of generalized one-hot models can be interpreted as quantifying intrinsic allelic effects, the numerical values of individual parameters cannot generally be interpreted in the absence of gauge-fixing constraints.

D. Simplex models

Simplex embeddings mathematically represent alleles in a more compact but less intuitive way than the one-hot embeddings discussed above. Single-position simplex embeddings encode the α characters of $\mathcal{A}$ using zero-centered vectors of dimension $\alpha - 1$ and thus have fewer dimensions than corresponding alleles. Simplex embeddings can be defined in multiple ways that differ from one another by similarity transformations, i.e., change-of-basis transformations. Here we adopt a particularly convenient definition: we define $\vec{x}_l^{\text{sim}}(s)$ to be an $\alpha - 1$-dimensional vector, the ith element of which is $x_l^{c_i}(s)$ if $s_l \neq c_\alpha$ and -1 if $s_l = c_\alpha$. We use the superscript "sim" here and in what follows to indicate mathematical objects that are based on simplex embeddings. Figures 1(a) and 1(c) illustrate $\vec{x}_l^{\text{sim}}$ for the three-character alphabet. Unlike $\vec{x}_l^{\text{ohe}}$, $\vec{x}_l^{\text{sim}}$ transforms under a non-permutation representation, that we denote as R_l^{sim}. For example, the effect of $h_{\text{A}\leftrightarrow\text{C}}$ on $\vec{x}_l^{\text{sim}}$ is equivalent to multiplication by the matrix

$$R_l^{\text{sim}}(h_{\text{A}\leftrightarrow\text{C}}) = \begin{bmatrix} -1 & 0 \\ -1 & 1 \end{bmatrix}. \tag{11}$$

As with one-hot embeddings, the embeddings of many different models can be built from direct sums of direct products of $\vec{x}_l^{\text{sim}}$. For example, a simplex embedding analogous to $\vec{x}_{\text{pair}}^{\text{ohe}}$

can be constructed as

$$\vec{x}^{\,\mathrm{sim}}_{\mathrm{pair}} = \vec{x}^{\,\mathrm{triv}} \oplus \left\{ \bigoplus_{l} \vec{x}^{\,\mathrm{sim}}_{l} \right\} \oplus \left\{ \bigoplus_{l<l'} \vec{x}^{\,\mathrm{sim}}_{l} \otimes \vec{x}^{\,\mathrm{sim}}_{l'} \right\}. \tag{12}$$

The corresponding pairwise simplex model has the form

$$f^{\,\mathrm{sim}}_{\mathrm{pair}} = \theta_0 x_0 + \sum_{l} \sum_{i=1}^{\alpha-1} \theta^{i}_{l} x^{i}_{l} + \sum_{l<l'} \sum_{i,j=1}^{\alpha-1} \theta^{ij}_{ll'} x^{ij}_{ll'}, \tag{13}$$

where x^{i}_{l} denotes the ith element of $\vec{x}^{\,\mathrm{ohe}}_{l}$ and $x^{ij}_{ll'}(s) = x^{i}_{l}(s) x^{j}_{l'}(s)$ for all $s \in \mathcal{S}$. We use $\vec{\theta}^{\,\mathrm{sim}}_{\mathrm{pair}}$ to denote the parameters of this model. Note that individual parameters are indexed using numerical superscripts ranging from 1 to $\alpha - 1$, rather than by characters in $\mathcal{A}$.

Pairwise simplex models describe the same sequence-function relationships that pairwise one-hot models do, i.e., given a set of parameters for one of these models, there exists a corresponding set of parameters for the other model that yields the same predictions over all sequences. However, because $\vec{x}^{\,\mathrm{sim}}_{l}$ has lower dimension than $\vec{x}^{\,\mathrm{ohe}}_{l}$, $\vec{\theta}^{\,\mathrm{sim}}_{\mathrm{pair}}$ contains fewer parameters than $\vec{\theta}^{\,\mathrm{ohe}}_{\mathrm{pair}}$. Inspection of Eq. (12) shows that the number of parameters in $\vec{\theta}^{\,\mathrm{ohe}}_{\mathrm{pair}}$ is in fact

$$\dim \vec{x}^{\,\mathrm{sim}}_{\mathrm{pair}} = 1 + (\alpha - 1)L + \binom{L}{2}(\alpha - 1)^2. \tag{14}$$

This reduction in the number of parameters entirely eliminates gauge freedoms, as can be seen from

$$\gamma^{\,\mathrm{ohe}}_{\mathrm{pair}} = \dim \vec{x}^{\,\mathrm{ohe}}_{\mathrm{pair}} - \dim \vec{x}^{\,\mathrm{sim}}_{\mathrm{pair}}. \tag{15}$$

The lack of gauge freedoms in $f^{\,\mathrm{sim}}_{\mathrm{pair}}$ is one example of the fact that, as we will see, models defined using (nonredundant) simplex embeddings do not have gauge freedoms. In fact, multiple groups [19,21,22] have argued for the use of simplex models, rather than one-hot models, based on simplex models not having gauge freedoms.

We argue, however, that the parameters of simplex models are fundamentally more difficult to interpret as allelic effects than are the parameters of one-hot models. Because $\vec{x}^{\,\mathrm{sim}}_{l}$ does not transform under a permutation representation, neither does $\vec{x}^{\,\mathrm{sim}}_{\mathrm{pair}}$ and neither does $\vec{\theta}^{\,\mathrm{sim}}_{\mathrm{pair}}$. In the case of the three-character alphabet, one sees from Eq. (12) that $h_{\mathrm{A}\leftrightarrow\mathrm{C}}$ induces a transformation of model parameters that maps $\theta^{1}_{l} \to -\theta^{1}_{l}$, $\theta^{2}_{l} \to -\theta^{1}_{l} + \theta^{2}_{l}$, $\theta^{22}_{ll'} \to \theta^{11}_{ll'} - \theta^{12}_{ll'} - \theta^{21}_{ll'} + \theta^{22}_{ll'}$, and so on. The fact that these parameters change in ways described by nontrivial linear combinations means that individual parameters cannot be interpreted as quantifying individual allelic effects.

E. Maschke decomposition

We now use methods from the theory of group representations to formally investigate the general connection between model transformation behavior and gauge freedoms. Maschke's theorem, a foundational result in representation theory, states that every matrix representation of a finite group is equivalent to a direct sum of irreducible matrix representations. Here the term "equivalent" means that there is a similarity transformation (i.e., a change of basis) that maps one representation to another; we use the symbol $\simeq$ to denote equivalence in what follows. The term "irreducible" means that the representation has no proper invariant subspace. Consider for example R^{ohe}_{l}, the representation that describes how $\vec{x}^{\,\mathrm{ohe}}_{l}$ transforms under H_{PSCP}. The group H_{PSCP} is isomorphic to the symmetric group (i.e., the group of permutations), the representations of which are well understood [28]. In this context, R^{ohe}_{l} is called the "defining representation" and is well known to be reducicble. Specifically, R^{ohe}_{l} has two proper invariant subspaces. One subspace has dimension 1 and is spanned by the vector $[1\ 1\ \cdots\ 1]^{\top}$. The other subspace has dimension $\alpha - 1$ and consists of the set of α-dimensional vectors whose elements sum to zero. The first of these subspaces transforms under the "trivial representation", which is simply the 1×1 matrix [1] and which we denote by R^{triv}. The other subspace transforms (after an appropriate change of coordinates) under the representation R^{sim}_{l}. R^{sim}_{l} is called the "standard representation" and is well known to be irreducible. The Maschke decomposition of R^{ohe}_{l} is therefore given by

$$R^{\mathrm{ohe}}_{l} \simeq R^{\mathrm{triv}} \oplus R^{\mathrm{sim}}_{l}, \tag{16}$$

where the direct sum on the right-hand side yields a block diagonal matrix created from R^{triv} and R^{sim}_{l}.

Equivalently, we can think of Maschke decomposition in terms of embeddings. Thinking in terms of embeddings can be helpful for deriving the specific invertible matrix that performs the similarity transformation needed to express a Maschke decomposition as an equality instead of an equivalence. When multiplied by an appropriate similarity transformation matrix T, $\vec{x}^{\,\mathrm{ohe}}_{l}$ can be expressed as a direct sum of the trivial embedding $\vec{x}^{\,\mathrm{triv}}$ (which is simply the one-dimensional vector [1]) and the simplex embedding $\vec{x}^{\,\mathrm{sim}}_{l}$, i.e.,

$$T\vec{x}^{\,\mathrm{ohe}}_{l}(s) = \vec{x}^{\,\mathrm{triv}}(s) \oplus \vec{x}^{\,\mathrm{sim}}_{l}(s), \tag{17}$$

for all sequences s. This allows us to express the equivalence relation in Eq. (16) as an equality, as it implies that

$$T R^{\mathrm{ohe}}_{l}(h) T^{-1} = R^{\mathrm{triv}}(h) \oplus R^{\mathrm{sim}}_{l}(h) \tag{18}$$

for all group elements h. Based on the definition of the embeddings $\vec{x}^{\,\mathrm{ohe}}_{l}$ and $\vec{x}^{\,\mathrm{sim}}_{l}$ above, one can readily show that the similarity transformation matrix T is given by

$$T = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix}. \tag{19}$$

This matrix T will be used later when defining an algorithm for distilling general equivariant embeddings.

F. Decomposition of equivariant embeddings

Maschke's theorem implies that any representation R of H_{PSCP} can be expressed as

$$R \simeq \bigoplus_{k=1}^{K} Q_k R_k, \tag{20}$$

where the R_k are distinct irreducible representations of H_{PSCP} and each Q_k is a natural number that denotes the multiplicity of R_k in the direct sum. R is thus equivalent to a block-diagonal representation formed by placing Q_k copies of each

R_k along the diagonal and setting all other matrix elements to zero [see Fig. 2(b)]. One consequence of Eq. (20) is that any embedding $\vec{x}$ that transforms under R can be decomposed as

$$\vec{x} \simeq \bigoplus_{k=1}^{K} \bigoplus_{q=1}^{Q_k} \vec{x}_{kq}, \tag{21}$$

where each $\vec{x}_{kq}$ is an embedding that transforms under R_k. In what follows, we say that embeddings like $\vec{x}_{kq}$ are irreducible because they transform under irreducible representations. We also assume that all $\vec{x}_{kq}$ are nonzero, but this assumption can be removed without fundamentally changing our results; see SM Sec. 5.2 for details. The Maschke decompositions of R and $\vec{x}$ are illustrated in Figs. 2(a) and 2(b).

G. Distillation of equivariant embeddings

We now describe an "embedding distillation" procedure that connects the Maschke decomposition of $\vec{x}$ to the gauge freedoms of the corresponding model. In SM Sec. 5.1 we prove the following:

Theorem 1. Any two nonzero sequence embeddings that transform under the same irreducible representation of H_{PSCP} are equal up to a constant of proportionality.

Using Theorem 1 we find that there is a similarity transformation matrix T_{decom} such that

$$T_{\mathrm{decom}}\,\vec{x} = \bigoplus_{k=1}^{K} Q_k \vec{x}_k, \tag{22}$$

where, for each k, $\vec{x}_k$ denotes any one of the irreducible embeddings $\vec{x}_{kq}$ in Eq. (21) and Q_k denotes the multiplicity of each term in the direct sum. Next we perform a similarity transformation (described by a matrix T_{thin}) that "thins out" the embedding by setting all except one copy of each $\vec{x}_k$ to zero. Finally, we perform a similarity transformation (described a matrix T_{sort}) that "sorts" the remaining nonzero embeddings, arranging them in series at the top of the resulting embedding vector. We thus find that applying the cumulative similarity transformation given by

$$T_{\mathrm{dist}} = T_{\mathrm{sort}}\,T_{\mathrm{thin}}\,T_{\mathrm{decom}} \tag{23}$$

to the embedding $\vec{x}$ yields

$$T_{\mathrm{dist}}\,\vec{x} = \vec{x}^{\,\mathrm{dist}} \oplus \vec{0}_\gamma, \tag{24}$$

where $\vec{0}_\gamma$ is a γ-dimensional vector of zeros and

$$\vec{x}^{\,\mathrm{dist}} = \bigoplus_{k=1}^{K} \vec{x}_k \tag{25}$$

is a "distilled embedding." When applied to the representation R, this distillation procedure yields

$$T_{\mathrm{dist}}\,R\,T_{\mathrm{dist}}^{-1} = \vec{R}^{\,\mathrm{dist}} \oplus \vec{R}^{\,\mathrm{redun}}, \tag{26}$$

where the "distilled representation," $R^{\mathrm{dist}} = \bigoplus_{k=1}^{K} R_k$, comprises one copy of each R_k present in Eq. (20), and where the redundant representation, $R^{\mathrm{redun}} = \bigoplus_{k=1}^{K} (Q_k - 1) R_k$, sweeps up the remaining copies of each R_k. The final distilled versions of R and $\vec{x}$ are illustrated in Fig. 2(c). Explicit formulas for constructing T_{decom}, T_{thin}, and T_{dist} are given in SM Sec. 8.

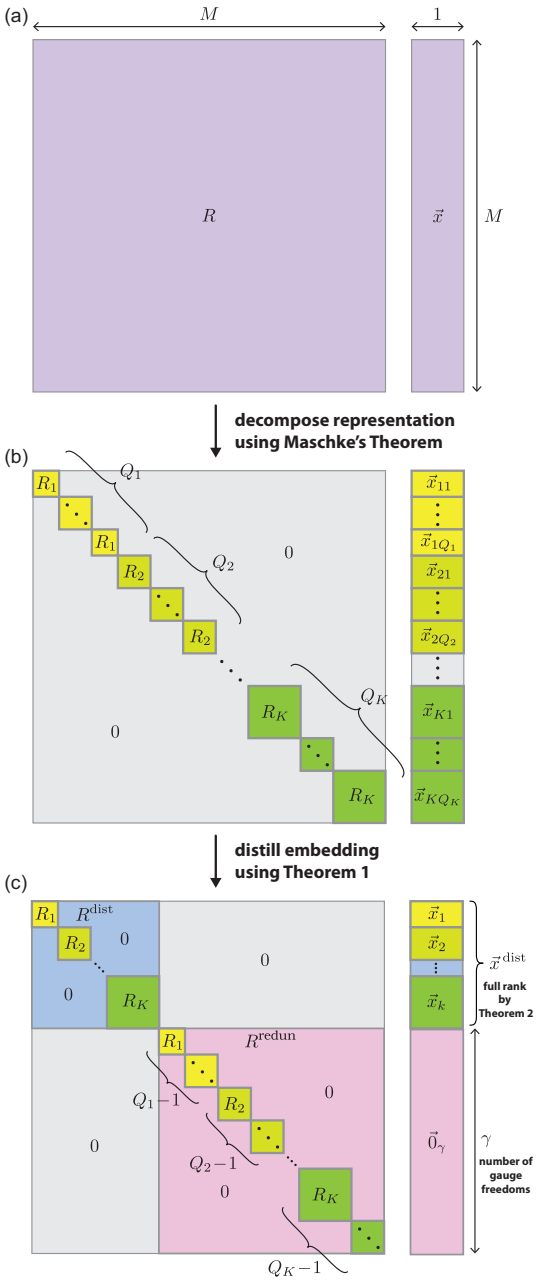

FIG. 2. Embedding distillation. (a) Given an M-dimensional embedding $\vec{x}$ that is equivariant under H_{PSCP}, let R be the representation that describes how $\vec{x}$ transforms. (b) By Maschke's theorem, R can be decomposed into a direct sum of irreducible representations, R_k ($k \in \{1, \ldots, K\}$), each of which occurs with multiplicity Q_k [Eq. (20)]. Similarly, $\vec{x}$ can be decomposed into a direct sum of irreducible embeddings $\vec{x}_{kq}$ ($q \in \{1, \ldots, Q_k\}$), where each $\vec{x}_{kq}$ transforms under R_k [Eq. (21)]. (c) By Theorem 1, an additional similarity transformation can be performed that, for each value of k, zeros out all but one $\vec{x}_{kq}$ and sorts the remaining embeddings; each remaining $\vec{x}_{kq}$ is denoted by $\vec{x}_k$. Consequently, $\vec{x}$ decomposes into a direct sum of a distilled embedding $\vec{x}^{\,\mathrm{dist}}$ and a zero vector $\vec{0}_\gamma$ having some dimension γ [Eq. (24)]. $\vec{x}^{\,\mathrm{dist}}$ is given by the direct sum of all $\vec{x}_k$ [Eq. (25)] and is full rank by Theorem 2. The distilled representation R^{dist} describes how $\vec{x}^{\,\mathrm{dist}}$ transforms and is given by a direct sum of one copy of each R_k. The redundant representation R^{redun} operates on $\vec{0}_\gamma$ and comprises the $Q_k - 1$ remaining copies of each R_k. The resulting number of gauge freedoms is γ [see Eq. (28)].

H. Identification of gauge freedoms in equivariant models

To identify the gauge freedoms of an equivariant model, we use the fact that $\vec{x}^{\,\text{dist}}$ [defined in Eq. (25)] is full rank. This is a consequence of the following theorem, which is proven in SM Sec. 3.4:

Theorem 2. For each $k \in \{1, \ldots, K\}$, let $\vec{x}_k$ be a nonzero embedding that transforms under an irreducible representation R_k of the group H_{PSCP}. Then the direct sum of all $\vec{x}_k$ is full rank if and only if all R_k are pairwise inequivalent.

Because $\vec{x}^{\,\text{dist}}$ is full rank, $\vec{g}^{\,\top}\vec{x}(s) = 0$ for all $s \in \mathcal{S}$ if and only if

$$\vec{g} = T_{\text{dist}}^{\top}[\vec{0}_{M-\gamma} \oplus \vec{g}_{\gamma}], \tag{27}$$

where T_{dist} is the distillation matrix in Eq. (23) and $\vec{g}_{\gamma}$ is any vector in $\mathbb{R}^{\gamma}$. The space of gauge transformations G is therefore given by the set of vectors $\vec{g}$ that have the form in Eq. (27). In particular, the number of gauge freedoms is seen to be

$$\gamma = \dim \vec{x} - \dim \vec{x}^{\,\text{dist}} = \deg R^{\text{redun}}. \tag{28}$$

We thus see that the number of gauge freedoms of an equivariant linear model is equal to the sum of the degrees of all the redundant irreducible representations under which that model's embedding (or, equivalently, that model's parameter vector) transforms.

I. Identification of all equivariant models

The mathematical structure of a group defines the models that transform equivariantly under that group. In the case of H_{PSCP}, the relatively simple group structure allows the straightforward identification of all inequivalent distilled embeddings and thus the identification of all equivariant linear models.

H_{PSCP} can be written as a direct product of simpler groups:

$$H_{\text{PSCP}} = H_1 \times \cdots \times H_L, \tag{29}$$

where each H_l denotes the group of character permutations at sequence position l. Each irreducible representation R_k of H_{PSCP} can therefore be expressed as the Kronecker product

$$R_k \simeq \bigotimes_{l=1}^{L} R_l^k, \tag{30}$$

where each R_l^k is an irreducible representation of H_l (see Theorem 1.11.3 of Ref. [28]). An embedding $\vec{x}_k$ that transforms under R_k will therefore have the form

$$\vec{x}_k \simeq \bigotimes_{l=1}^{L} \vec{x}_l^{\,k}, \tag{31}$$

where $\vec{x}_l^{\,k}$ is an irreducible embedding that transforms under H_l. In SM Sec. 4.3 we show that H_l supports only two inequivalent irreducible embeddings (regardless of alphabet size): $\vec{x}^{\,\text{triv}}$ and $\vec{x}_l^{\,\text{sim}}$. Each $\vec{x}_l^{\,k}$ must therefore be equivalent to one of these two embeddings. Ignoring the factors of $\vec{x}^{\,\text{triv}}$ because they do not impact Kronecker products, Eq. (31) becomes

$$\vec{x}_k \simeq \bigotimes_{l \in B_k} \vec{x}_l^{\,\text{sim}}, \tag{32}$$

where B_k is a subset of the positions $\{1, \ldots, L\}$. There are 2^L possible choices for each subset B_k, and thus 2^L inequivalent irreducible embeddings $\vec{x}_k$. Since each $\vec{x}_k$ can appear at most once on the left-hand side of Eq. (25), there are 2^{2^L} inequivalent distilled embeddings $\vec{x}^{\,\text{dist}}$.

For each choice of $\vec{x}^{\,\text{dist}}$ there are an infinite number of possible choices for T_{dist} and γ that can be used to define $\vec{x}$ [via Eq. (24)]. The number of possible equivariant embeddings $\vec{x}$, and thus the number of equivariant models f, is therefore infinite. However, all models corresponding to a specific $\vec{x}^{\,\text{dist}}$ have the same expressivity, i.e., the set of sequence-function relationships that each model describes (considered over all possible values of model parameters) is the same. We therefore consider these models to be equivalent and conclude that there are a total of 2^{2^L} inequivalent equivariant linear models on sequences of length L.

J. Analytical analysis of generalized one-hot models

We now use the embedding distillation procedure to compute the number of gauge freedoms of all generalized one-hot models. This derivation is based on the Maschke decomposition $\vec{x}_l^{\,\text{ohe}} \simeq \vec{x}^{\,\text{triv}} \oplus \vec{x}_l^{\,\text{sim}}$ from Eq. (17).

We first demonstrate this calculation on the pairwise one-hot model. Plugging the decomposition of $\vec{x}_l^{\,\text{ohe}}$ into the definition for $\vec{x}_{\text{pair}}^{\,\text{ohe}}$ in Eq. (9), then expanding the Kronecker products and grouping like terms, we find that

$$\vec{x}_{\text{pair}}^{\,\text{ohe}} \simeq \left[1 + L + \binom{L}{2}\right]\vec{x}^{\,\text{triv}} \oplus \times \left\{\bigoplus_{l} L\,\vec{x}_l^{\,\text{sim}}\right\} \oplus \left\{\bigoplus_{l<l'} \vec{x}_l^{\,\text{sim}} \otimes \vec{x}_{l'}^{\,\text{sim}}\right\}, \tag{33}$$

where the scalar coefficients correspond to the Q_k in Eq. (22). We derive the corresponding distilled embedding by replacing each of these coefficients with 1. Doing so reveals the distillation of $\vec{x}_{\text{pair}}^{\,\text{ohe}}$ to be $\vec{x}_{\text{pair}}^{\,\text{sim}}$. The result for $\gamma_{\text{pair}}^{\,\text{ohe}}$ in Eq. (5) is therefore just a manifestation of Eq. (28).

We now extend this approach to all generalized one-hot models. The embedding $\vec{x}_{\text{goh}}$ of any generalized one-hot model can be written as

$$\vec{x}_{\text{goh}} = \bigoplus_{j=1}^{J} \bigotimes_{l \in A_j} \vec{x}_l^{\,\text{ohe}}. \tag{34}$$

where $A_1, \ldots, A_J$ denote J (not necessarily distinct) sets of positions. Because the dimension of $\vec{x}_l^{\,\text{ohe}}$ is α, the number of corresponding model parameters is

$$M_{\text{goh}} = \dim \vec{x}_{\text{goh}} = \sum_{j=1}^{J} \alpha^{|A_j|}. \tag{35}$$

Decomposing $\vec{x}_l^{\,\text{ohe}}$ in terms of $\vec{x}^{\,\text{triv}}$ and $\vec{x}_l^{\,\text{sim}}$, expanding each Kronecker product, then grouping the resulting terms, we find that

$$\vec{x}_{\text{goh}}^{\,\text{dist}} = \bigoplus_{k=1}^{K} \bigotimes_{l \in B_k} \vec{x}_l^{\,\text{sim}}, \tag{36}$$

TABLE I. Analytical results for various generalized one-hot models, computed using Eqs. (35) and (38). See SM Sec. 6 for derivations of these results. K-adjacent models assume $K \geqslant 1$. ‡Only includes interactions among adjacent positions.

Model type	Interaction orders	Number of parameters (M_{goh})	Number of gauge freedoms (γ_{goh})
Constant	0	1	0
Additive	0, 1	$1 + L\alpha$	L
Pairwise	0, 1, 2	$1 + L\alpha + \binom{L}{2}\alpha^2$	$L + \binom{L}{2}(2\alpha - 1)$
Nearest-neighbor	0, 1, 2‡	$1 + L\alpha + (L-1)\alpha^2$	$L + (L-1)(2\alpha - 1)$
All-order	$0, 1, \ldots, L$	$(\alpha + 1)^L$	$(\alpha + 1)^L - \alpha^L$
All-adjacent	$0, 1, 2^\ddagger, \ldots, L^\ddagger$	$1 + \frac{\alpha}{(\alpha-1)^2}[\alpha^{L+1} - (L+1)\alpha + L]$	$1 + \frac{\alpha}{(\alpha-1)^2}[2\alpha^L - \alpha^{L-1} - (L+1)\alpha + L]$
K-order	K	$\binom{L}{K}\alpha^K$	$\binom{L}{K}\alpha^K - \sum_{k=0}^{K}\binom{L}{k}(\alpha-1)^k$
Hierarchical K-order	$0, 1, \ldots, K$	$\sum_{k=0}^{K}\binom{L}{k}\alpha^k$	$\sum_{k=0}^{K}\binom{L}{k}[\alpha^k - (\alpha-1)^k]$
K-adjacent‡	$K^\ddagger$	$(L - K + 1)\alpha^K$	$(L-K)\alpha^{K-1}$
Hierarchical K-adjacent‡	$0, 1, 2^\ddagger, \ldots, K^\ddagger$	$1 + \sum_{k=1}^{K}(L-k+1)\alpha^k$	$(L-K)\alpha^{K-1} + 1 + \sum_{k=1}^{K-1}(L-k+1)\alpha^k$

where $B_1, \ldots, B_K$ denote the distinct subsets of positions that occur among all the A_j. Because the dimension of $\vec{x}_l^{\,\text{sim}}$ is $\alpha - 1$,

$$\dim \vec{x}_{\text{goh}}^{\,\text{dist}} = \sum_{k=1}^{K}(\alpha-1)^{|B_k|}. \tag{37}$$

The number of gauge freedoms of the generalized one-hot model having embedding $\vec{x}_{\text{goh}}$ is therefore given by

$$\gamma_{\text{goh}} = \sum_{j=1}^{J}\alpha^{|A_j|} - \sum_{k=1}^{K}(\alpha-1)^{|B_k|}. \tag{38}$$

Table I reports the number of gauge freedoms calculated in this manner for a variety of generalized one-hot models (illustrated in Fig. 3). SM Sec. 6 provides expanded descriptions for each generalized one-hot model, as well as detailed derivations of the results in Table I.

A result of this analysis is that all generalized one-hot models have gauge freedoms, save models for which the direct sum in Eq. (34) includes only one term. To see this, observe that Eq. (22) gives

$$\dim \vec{x}_{\text{goh}} = \sum_{k=1}^{K} Q_k(\alpha-1)^{|B_k|}, \tag{39}$$

where each multiplicity value Q_k is equal to the number of sets A_j that contain B_k. Using this together with Eq. (36) and Eq. (28) gives

$$\gamma_{\text{goh}} = \sum_{k=1}^{K}(Q_k - 1)(\alpha-1)^{|B_k|}. \tag{40}$$

We thus find that $\gamma_{\text{goh}} = 0$ if and only if none of the Q_k are greater than 1. But since the empty set is a subset of every A_j, it will always be among the B_k, and the corresponding multiplicity value will be $Q_k = J$. Gauge freedoms are therefore present in all generalized one-hot models for which $J \geqslant 2$. Conversely, $\gamma_{\text{goh}} = 0$ when $J = 1$ because all B_k occur with multiplicity $Q_k = 1$. Gauge freedoms are therefore absent in all generalized one-hot models for which $J = 1$.

K. Computational analysis of generalized one-hot models

Embedding distillation also allows one to efficiently compute a sparse basis for the space of gauge freedoms G_{goh} of any generalized one-hot model. Equation (27) reveals that G_{goh} is spanned by the last γ_{goh} row vectors of T_{dist}. One can therefore compute a basis for G_{goh} by computing T_{dist}. This is notable because computing T_{dist} only requires keeping track of the similarity transformations needed to express $\vec{x}_{\text{goh}}$ in the distilled form shown in Eq. (24). This computation is far less demanding than computing a basis for G_{goh} using Gaussian elimination or singular value decomposition when (as is often the case) the number of possible sequences is far greater than the number of model parameters.

In Eq. (23) we described how to construct T_{dist} from a product of three matrices: T_{decom}, T_{thin}, and T_{sort}. Explicit formulas for computing these matrices, as well as their inverses, are provided in SM Sec. 8. For these formulas we observe that each matrix, as well as its inverse, is sparse in the large L limit when the maximal order of interaction described by the model is fixed. The resulting distillation matrix T_{dist} is therefore also sparse, as is its inverse. It also turns out that every nonzero element of T_{dist} is $+1$ or -1. Taking the last γ_{goh} rows of T_{dist} thus provides a basis for G_{goh} consisting of sparse vectors whose only nonzero elements are $+1$ and -1. Having sparse matrices for T_{dist} and T_{dist}^{-1} also allows us to compute a sparse gauge-fixing projection matrix P; see SM Sec. 8 for details. Figure 4 illustrates the actions of T_{decom}, T_{thin}, and T_{sort} on an example embedding vector for the all-order interaction model corresponding to $L = 3$ and $\alpha = 3$. Figure 4 also illustrates the corresponding distillation matrix T_{dist}.

L. Other symmetry groups

The proof of Theorem 1 in SM Sec. 5.1, and thus our embedding distillation procedure, applies only to the symmetry group H_{PSCP}. There are other symmetry groups of sequence space besides H_{PSCP}, however, and it is worth asking whether Theorem 1, and thus Eqs. (24)–(28), hold for those groups as well.

One symmetry group is the group of global character permutations, H_{GCP}. This group comprises transformations that

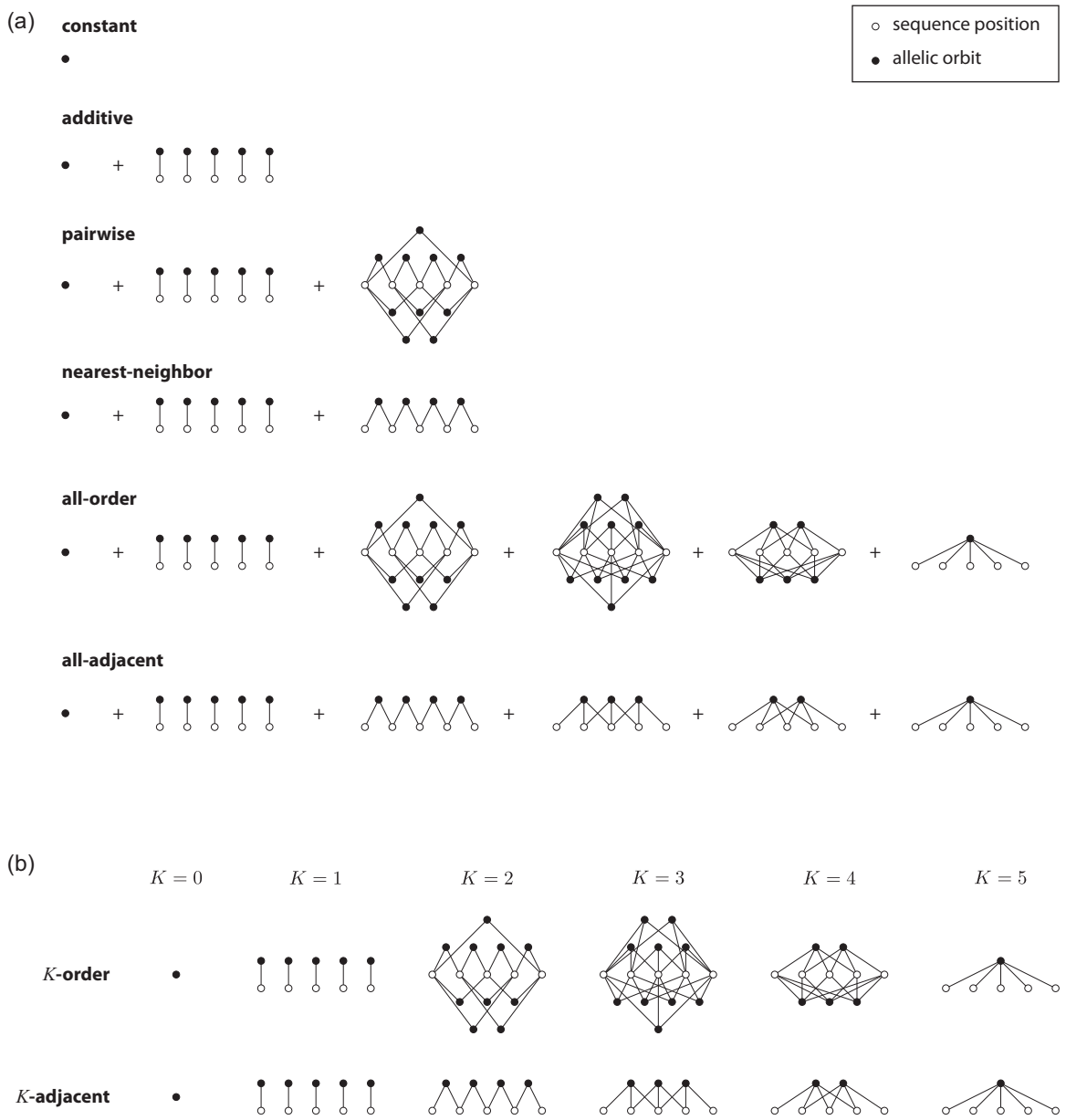

FIG. 3. Structure of generalized one-hot models analyzed in Table I for sequences of length $L = 5$. Open circles represent sequence positions. Closed circles represent allelic orbits, i.e., sets of sequence features that are closed under the action of H_{PSCP}. Edges indicate position indices shared by the features in each allelic orbit. (a) Structure of constant, additive, pairwise, nearest-neighbor, all-order, and all-adjacent models. (b) Structure of K-order models and K-adjacent models for various interaction orders K.

apply the same permutation to characters at every position in a sequence. Another is the group of position permutations, H_{PP}. This group comprises transformations that permute the positions in a sequence without otherwise changing the characters therein. SM Sec. 7.1 shows that Theorem 1 does not hold for either H_{GCP} or H_{PP}. Consequently, one cannot compute distilled embeddings using the irreducible representations of either group.

A third symmetry group is H_{Ham}, which describes combinations of position permutations and position-specific character permutations. H_{Ham} is the largest symmetry group that preserves Hamming distances [31] and includes H_{PSCP}, H_{PP}, and H_{GCP} as subgroups. Theorem 1 does hold for H_{Ham} due the fact that H_{PSCP} is a subgroup (see SM Sec. 7.2). However, the set of models that are equivariant under H_{Ham} is a subset of the models that are equivariant under H_{PSCP} and the irreducible representations of H_{Ham} are more complex than those of H_{PSCP}. H_{PSCP} is therefore more useful than H_{Ham} for analyzing gauge freedoms.

M. Transformation behavior and parameter interpretability

We now return to the connection between parameter transformation behavior and parameter interpretability. Our above discussion of pairwise models suggested that the ability to interpret individual parameters as quantifying intrinsic allelic effects required the presence of gauge freedoms. We now

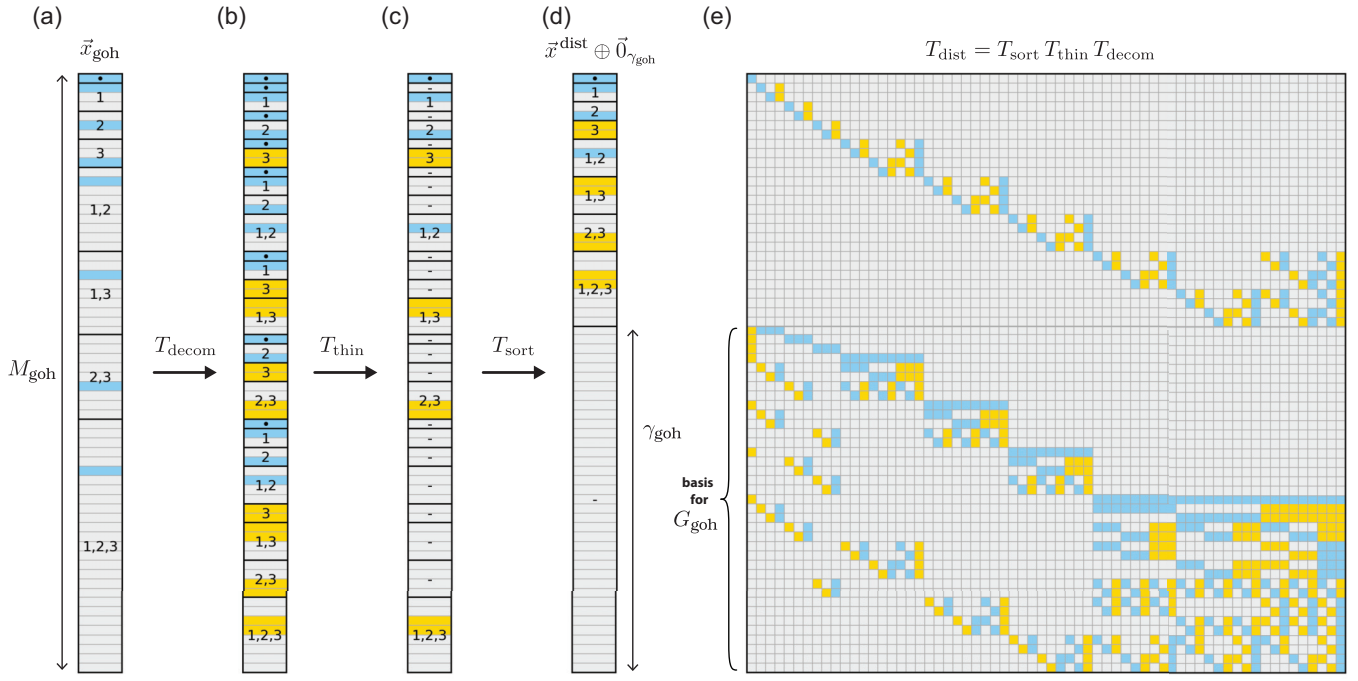

FIG. 4. Embedding distillation for an example generalized one-hot model. (a) Embedding $\vec{x}_{\text{goh}}$ of the $L = 3$ sequence $s = \text{ABC}$ for an all-order one-hot model based on the alphabet $\mathcal{A} = \{\text{A, B, C}\}$. This embedding has degree $M_{\text{goh}} = 64$. (b) Result of multiplication by the decomposition matrix T_{decom}. (c) Result of subsequent multiplication by the thinning matrix T_{thin}. (d) Result of subsequent multiplication by the sorting matrix T_{sort}, which yields $\vec{x}^{\,\text{dist}} \oplus \vec{0}_{\gamma_{\text{goh}}}$ with $\gamma_{\text{goh}} = 37$ being the number of gauge freedoms. In (b)–(d), dots indicate $\vec{x}^{\,\text{triv}}$, dashes indicate zero vectors, and numbers indicate the positions l contributing to each component. (e) Distillation matrix T_{dist} that implements the full distillation procedure in (a)–(d). The last γ_{goh} rows of T_{dist} provide a sparse basis for the gauge space, G_{goh}. In (a)–(e), vector and matrix elements are colored according to their numerical values: blue represents $+1$, yellow represents -1, and gray represents 0.

formalize this observation and conjecture an extension to all linear equivariant models.

Mathematically, we define a generalized allele a to be any subset of $\mathcal{S}$, and say that any sequence $s \in \mathcal{S}$ has allele a if $s \in a$. The corresponding "allelic feature" x_a is defined to be the indicator function on $\mathcal{S}$ for whether a sequence has allele a. An "allelic model" is defined to be a linear sequence-function relationship in which every feature is an allelic feature. In the context of an allelic model, the parameter θ_a that multiplies x_a is said to be an "allelic effect." The parameters of a linear model can therefore be interpreted as allelic effects if and only if every one of the corresponding features is an indicator function for membership in some subset of $\mathcal{S}$.

For an allelic model to have parameters that describe intrinsic allelic effects, the model must be a "permutation model," i.e., the features and parameters of the model must transform under a permutation representation of H_{PSCP}. Requiring an allelic model to be a permutation model puts strong constraints on which sets of alleles it can describe. Because H_{PSCP} permutes sequences, it also permutes alleles. Given a specific allele a, we call the set of alleles created by the action of H_{PSCP} on a an "allelic orbit." It is readily seen that, for an allelic model to be a permutation model, the set of alleles it describes must consist of some number J of complete allelic orbits.

All allelic models that comprise $J \geqslant 2$ allelic orbits have gauge freedoms. To see this, observe that the features in each orbit transform among themselves according to a permutation representation. The features of the full model will therefore transform under a direct sum of J permutation representations. Because every permutation representation contains the trivial representation in its Maschke decomposition, the decomposition of the full model's representation will contain at least J copies of the trivial representation. The model will therefore have at least $J - 1$ gauge freedoms, though additional gauge freedoms can be present as well.

This result is reflected in our above analytic analysis of generalized one-hot models. All generalized one-hot models are allelic permutation models (though the converse is not true; see SM Sec. 9.1), and each allelic orbit of a generalized one-hot model corresponds to a position set A_j in Eq. (34). The lower-bound on the number of gauge freedoms identified here recapitulates the finding above that generalized one-hot models have no gauge freedoms if and only if $J = 1$.

An allelic permutation model that does not have gauge freedoms must therefore comprise only one allelic orbit. An example of a model with only one allelic orbit is a one-hot model of length $L = 1$, e.g., a model describing the effect of only one nucleotide position in a DNA sequence or one amino acid position in a protein sequence. Are single-orbit allelic models useful in practice? We argue that the answer is essentially no. In SM Sec. 9.1 we show that single-orbit generalized one-hot models cannot describe co-occurring alleles. We regard such models as trivial because the entire reason for quantitatively modeling sequence-function relationships

is to deconvolve the influence of co-occurring alleles. There are single-orbit allelic permutation models that describe co-occurring alleles, but all the examples of these we have analyzed either have gauge freedoms or are mathematically equivalent to generalized one-hot models (see SM Sec. 9.1). Moreover, among models whose embeddings are built from direct sums of Kronecker products of single-position embeddings, the generalized one-hot models have the fewest gauge freedoms (see SM Sec. 9.2). Based on these findings, we conjecture that all nontrivial allelic permutation models (i.e., all models whose parameters describe intrinsic allelic effects) have gauge freedoms.

IV. DISCUSSION

Motivated by the connection between gauge freedoms and symmetries in physics, we investigated the relationship between gauge freedoms and symmetries in quantitative models of sequence-function relationships. We found that, for linear models that are equivariant under H_{PSCP} (i.e., the group of PSCP transformations), gauge freedoms arise due to model parameters transforming under redundant irreducible matrix representations. From a conceptual standpoint, this result links the gauge freedoms of models of sequence-function relationships to the transformation behavior of these models under a specific symmetry group of sequence space. From a practical standpoint, this result facilitates the analytic calculation of the number of independent gauge freedoms in a large class of commonly used models. It also enables an embedding distillation algorithm that can efficiently compute a sparse basis for the space of gauge freedoms. This latter capability may prove to be useful particularly when studying models with very large numbers of parameters. Such models are increasingly common, as massively parallel reporter assays, deep mutational scanning experiments, and other multiplex assays of variant effect can now readily measure the activities of hundreds of thousands of sequences in a single experiment (e.g., Refs. [32,33]).

We also investigated the link between parameter transformation behavior and parameter interpretability. In doing so we identified an incompatibility between two different notions of parameter interpretability. In linear models that are equivariant under H_{PSCP}, the ability to interpret individual parameters as quantifying intrinsic allelic effects requires that these parameters transform under a permutation representation of H_{PSCP}. But in many (and possibly in all) nontrivial models, this requirement is incompatible with the ability to interpret the values of individual parameters in the absence of gauge-fixing constraints. Consequently, models that have gauge freedoms can have advantages over equally expressive models that do not have gauge freedoms.

It should be noted that there are indeed useful models that do not have gauge freedoms. One such class of models are the "wild-type" one-hot models, the features of which are limited to those describing mutations away from a specific sequence of interest (e.g., Refs. [33,34]). Note that wild-type models differ in an important way from one-hot models expressed in the wild-type gauge (described in Ref. [29]): the latter models have specific parameters set to zero, whereas the former models lack these parameters entirely. The parameters of wild-type models have a close connection to the quantities that one can actually experimentally measure, i.e., activity differences between alleles. However, these parameters do not transform under a permutation representation of H_{PSCP} and so do not quantify intrinsic allelic effects. Indeed, wild-type models are quite close in spirit to the representation of E&M explicitly in terms of electric and magnetic fields: while these fields are the directly measurable manifestation of E&M, they transform in complicated ways under changes in observer velocity and so do not provide the theoretical clarity–the *intrinsic* description of E&M–that the electromagnetic four-potential does.

Another class of useful models that do not have gauge freedoms are models whose features represent sequence-dependent physical properties, such as the chemical properties of amino acids [35,36] or the physical shape of the DNA double helix [37,38]. These models are not equivariant, however, as their parameters describe the effects of physical properties of alleles, not the effects of alleles themselves. Notably, both classes of model reflect inductive biases that break H_{PSCP} symmetry.

In classical field theories like E&M, there are specific symmetries that are well-established by experiment and that any mathematical formulation of the theory must be consistent with. This does not, however, mean that the equations of the theory must transform in a simple way under those symmetries. Mathematically formulating physical theories so that the equations themselves manifestly respect the symmetries of the theory generally requires overparameterizing the equations, thereby introducing gauge freedoms. Physicists often find it worthwhile to do this, as having fundamental symmetries be manifestly reflected in one's equations can greatly facilitate the interpretation and application of those equations. Solving such equations, however, requires fixing the gauge—introducing additional constraints that make the solution of the equations unique.

Unlike in physics, there is no experimentally established requirement that models of sequence-function relationships be equivariant under any symmetries of sequence space. The specific mathematical form one uses for such models is subjective, and different models are commonly used in different contexts. Citing the ambiguities caused by gauge freedoms, some have argued for restricting one's choice of model to those that have no gauge freedoms. Nevertheless, models that have gauge freedoms are still common in the literature. We suggest that a major reason for this may be that researchers often prefer to use models that manifestly reflect symmetries of sequence space, and therefore have parameters that are interpretable as intrinsic allelic effects. As we showed, these criteria often (and possibly in all nontrivial cases) require the use of overparameterized models. In this way, the origin of gauge freedoms in models of sequence-function relationships does mirror the origin of gauge freedoms in physical theories.

There is still much to understand about the relationship between models of sequence-function relationships, the symmetries of these models, and how these models can be biologically interpreted. This paper and its companion [29] have only addressed gauge freedoms and symmetries in linear models. Some work has explored the gauge freedoms and

symmetries of nonlinear models of sequence-function relationships [39,40], but important questions remain. The sloppy modes [41,42] present in these models are also important to understand, and to our knowledge these have yet to be systematically investigated. Addressing these problems is becoming increasingly urgent due to the expanding interest in interpretable quantitative models of sequence-function relationships (e.g., Ref. [43]).

See Supplemental Material [47] for full derivations of the mathematical results presented above. Python code implementing the embedding distillation algorithm described the section "Computational analysis of generalized one-hot models" and used for generating Fig. 4 is available at [48].

ACKNOWLEDGMENTS

We thank Peter Koo and Vijay Balasubramanian for helpful discussions. This work was supported by NIH Grant No. R35 GM133613 (A.P. and D.M.M.), NIH Grant No. R35 GM133777 (A.P. and J.B.K.), NIH Grant No. R01 HG011787 (J.B.K.), the Alfred P. Sloan foundation (D.M.M.), and additional funding from the Simons Center for Quantitative Biology at CSHL (D.M.M. and J.B.K.).

DATA AVAILABILITY

No experimental data were created or analyzed in this study.

[1] J. B. Kinney and D. M. McCandlish, Massively parallel assays and quantitative sequence-function relationships, Annu. Rev. Genom. Hum. Genet. **20**, 99 (2019).

[2] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing, Proc. Natl. Acad. Sci. USA **106**, 67 (2009).

[3] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, Protein 3D structure computed from evolutionary sequence variation, PLoS One **6**, e28766 (2011).

[4] M. Ekeberg, C. Lovkvist, Y. Lan, M. Weigt, and E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models, Phys. Rev. E **87**, 012707 (2013).

[5] M. Ekeberg, T. Hartonen, and E. Aurell, Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences, J. Comput. Phys. **276**, 341 (2014).

[6] R. R. Stein, D. S. Marks, and C. Sander, Inferring pairwise interactions from biological data using maximum-entropy probability models, PLoS Comput. Biol. **11**, e1004182 (2015).

[7] J. B. Kinney, G. Tkacik, and C. G. Callan, Precise physical models of protein-DNA interaction from high-throughput data, Proc. Natl. Acad. Sci. USA **104**, 501 (2007).

[8] J. P. Barton, E. D. Leonardis, A. Coucke, and S. Cocco, ACE: adaptive cluster expansion for maximum entropy graphical model inference, Bioinformatics **32**, 3089 (2016).

[9] A. Haldane, W. F. Flynn, P. He, and R. M. Levy, Coevolutionary landscape of kinase family proteins: Sequence probabilities and functional motifs, Biophys. J. **114**, 21 (2018).

[10] S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, and M. Weigt, Inverse statistical physics of protein sequences: A key issues review, Rep. Prog. Phys. **81**, 032601 (2018).

[11] A. Haldane and R. M. Levy, Influence of multiple-sequence-alignment depth on Potts statistical models of protein covariation, Phys. Rev. E **99**, 032405 (2019).

[12] H. T. Rube, C. Rastogi, S. Feng, J. F. Kribelbauer, A. Li, B. Becerra, L. A. N. Melo, B. V. Do, X. Li, H. H. Adam, N. H. Shah, R. S. Mann, and H. J. Bussemaker, Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning, Nat Biotechnol. **40**, 1520 (2022).

[13] S. Zamuner and P. D. L. Rios, Interpretable neural networks based classifiers for categorical inputs, arxiv:2102.03202.

[14] C. Feinauer, B. Meynard-Piganeau, and C. Lucibello, Interpretable pairwise distillations for generative protein sequence models, PLoS Comput. Biol. **18**, e1010219 (2022).

[15] A. Gerardos, N. Dietler, and A.-F. Bitbol, Correlations from structure and phylogeny combine constructively in the inference of protein partners from sequences, PLoS Comput. Biol. **18**, e1010147 (2022).

[16] C. Hsu, H. Nisonoff, C. Fannjiang, and J. Listgarten, Learning protein fitness models from evolutionary and assay-labeled data, Nat. Biotechnol. **40**, 1114 (2022).

[17] C. Feinauer and E. Borgonovo, Mean dimension of generative models for protein sequences, bioRxiv (2022), doi:10.1101/2022.12.12.520028.

[18] E. D. Weinberger, Fourier and Taylor series on fitness landscapes, Biol. Cybern. **65**, 321 (1991).

[19] C. T. Zhang and R. Zhang, Analysis of distribution of bases in the coding sequences by a diagrammatic technique, Nucleic Acids Res. **19**, 6313 (1991).

[20] P. F. Stadler, Spectral landscape theory, in *Evolutionary Dynamics: Exploring the Interplay of Selection, Accident, Neutrality and Function*, edited by J. Crutchfield and P. Schuster (Oxford University Press, Oxford, 2003), pp. 231–271.

[21] G. D. Stormo, Maximally efficient modeling of DNA sequence motifs at all levels of complexity, Genetics **187**, 1219 (2011).

[22] F. J. Poelwijk, V. Krishna, and R. Ranganathan, The context-dependence of mutations: A linkage of formalisms, PLoS Comput. Biol. **12**, e1004771 (2016).

[23] D. H. Brookes, A. Aghazadeh, and J. Listgarten, On the sparsity of fitness functions and implications for learning, Proc. Natl. Acad. Sci. USA **119**, e2109649118 (2022).

[24] A. Tareen, M. Kooshkbaghi, A. Posfai, W. T. Ireland, D. M. McCandlish, and J. B. Kinney, MAVE-NN: Learning genotype-phenotype maps from multiplex assays of variant effect, Genome Biol. **23**, 98 (2022).

[25] J. D. Jackson and L. B. Okun, Historical roots of gauge invariance, Rev. Mod. Phys. **73**, 663 (2001).

[26] J. D. Jackson, *Classical Electrodynamics* (John Wiley & Sons, Hoboken, 1998).

[27] Results in quantum physics, such as the Aharanov-Bohm effect [44,45], suggest a reality to the four-potential beyond what

can be inferred solely from classical E&M, though there are arguments against this interpretation [46].

[28] B. E. Sagan, *The Symmetric Group: Representations, Combinatorial Algorithms, and Symmetric Functions*, 2nd ed., Graduate Texts in Mathematics (Springer, Berlin, 2001).

[29] A. Posfai, J. Zhou, D. M. McCandlish, and J. B. Kinney, Gauge fixing for sequence-function relationships, PLoS Comput Biol **21**, e1012818 (2025).

[30] Our analysis is readily extended to models in which $\vec{x}$ and $\vec{\theta}$ are complex vectors (e.g., Ref. [23]). All the results in the Supplemental Material are, in fact, derived for this more general class of models. Here we restrict our discussion to the reals only to simplify the presentation.

[31] R. Happel and P. F. Stadler, Canonical approximation of fitness landscapes, Complexity **2**, 53 (1996).

[32] J. B. Kinney, A. Murugan, C. G. Callan, and E. C. Cox, Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence, Proc. Natl. Acad. Sci. USA **107**, 9158 (2010).

[33] C. A. Olson, N. C. Wu, and R. Sun, A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain, Curr. Biol. **24**, 2643 (2014).

[34] P. Baeza-Centurion, B. Miñana, J. M. Schmiedel, J. Valcárcel, and B. Lehner, Combinatorial genetics reveals a scaling law for the effects of mutations on splicing, Cell **176**, 549 (2019).

[35] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, AAindex: amino acid index database, progress report 2008, Nucleic Acids Res. **36**, D202 (2007).

[36] K. K. Yang, Z. Wu, C. N. Bedbrook, and F. H. Arnold, Learned protein embeddings for machine learning, Bioinformatics **34**, 2642 (2018).

[37] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, The role of DNA shape in protein-DNA recognition, Nature (London) **461**, 1248 (2009).

[38] L. Yang, T. Zhou, I. Dror, A. Mathelier, W. W. Wasserman, R. Gordân, and R. Rohs, TFBSshape: a motif database for DNA shape features of transcription factor binding sites, Nucleic Acids Res. **42**, D148 (2014).

[39] J. B. Kinney and G. S. Atwal, Parametric inference in the large data limit using maximally informative models, Neural Comput. **26**, 637 (2014).

[40] G. S. Atwal and J. B. Kinney, Learning quantitative sequence–function relationships from massively parallel experiments, J. Stat. Phys. **162**, 1203 (2016).

[41] B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna, Parameter space compression underlies emergent theories and predictive models, Science **342**, 604 (2013).

[42] M. K. Transtrum, B. B. Machta, K. S. Brown, B. C. Daniels, C. R. Myers, and J. P. Sethna, Perspective: Sloppiness and emergent theories in physics, biology, and beyond, J. Chem. Phys. **143**, 010901 (2015).

[43] E. Seitz, D. M. McCandlish, J. B. Kinney, and P. K. Koo, Interpreting *cis*-regulatory mechanisms from genomic deep neural networks using surrogate models, Nat Mach Intell **6**, 701 (2024).

[44] Y. Aharonov and D. Bohm, Significance of electromagnetic potentials in the quantum theory, Phys. Rev. **115**, 485 (1959).

[45] M. Peshkin and A. Tonomura, *The Aharonov-Bohm Effect* (Springer Verlag, Berlin, 2005).

[46] L. Vaidman, Role of potentials in the Aharonov-Bohm effect, Phys. Rev. A **86**, 040101(R) (2012).

[47] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.7.023005 for full derivations of our mathematical results.

[48] Analysis code used to make Fig. 4 is available on GitHub at https://github.com/jbkinney/24_posfai2. A snapshot of this code is available on Zenodo at https://doi.org/10.5281/zenodo.15095911.

Supplemental Material for: Symmetry, gauge freedoms, and the interpretability of sequence-function relationships

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

March 27, 2025

Contents

1 Preliminaries

1.1 Notation

- V denotes feature space, i.e., the vector space in which embeddings and parameters live.
- M denotes the dimension feature space.
- $\vec{x}$ denotes the sequence embedding of a model.
- $\vec{\theta}$ denotes the parameters of a model.
- f denotes a linear model.
- γ denotes the number of gauge freedoms.
- G denotes the space of gauge freedoms.
- Θ denotes a gauge space.
- H denotes a group.
- $\leftrightarrow$ denotes a transposition.
- R denotes a group representation.
- $\simeq$ denotes equivalence.
- $\mathcal{U}$ denotes a finite set.
- α denotes the number of characters in an alphabet.
- c denotes a character in an alphabet.
- L denotes the length of a sequence.

Note: We assume throughout this document that V is complex, whereas Main Text assumes that V is real in order to simplify the presentation. Sec. 5.3 shows that our embedding distillation procedure, derived here assuming complex V, still works when V is restricted to the reals.

1.2 Specific sets

- $\mathcal{I}_\alpha$ denotes $(1, 2, \ldots, \alpha)$, an ordered set comprising the first α positive integers.
- $\mathcal{A}$ denotes the alphabet $(c_1, \ldots, c_\alpha)$, an ordered set comprising α characters. Note that $\mathcal{A}$ is written as an unordered set in main text.
- $\mathcal{S}$ denotes sequence space, i.e., the set of sequences of length L built from the α characters in $\mathcal{A}$.

1.3 Specific groups

- S_α denotes the symmetric group and acts on $\mathcal{I}_\alpha$.
- H_{CP} denotes the group of character permutations and acts on $\mathcal{A}$.
- H_{CP}^l (written H_l in Main Text) denotes the group of single-position character permutations and acts on $\mathcal{S}$.
- H_{PSCP} denotes the group of position-specific character permutations and acts on $\mathcal{S}$.
- H_{PP} denotes the group of position permutations and acts on $\mathcal{S}$.
- H_{Ham} denotes the group of Hamming graph symmetries and acts on $\mathcal{S}$.

1.4 Specific embeddings and representations

- $\vec{x}^{\,\mathrm{triv}}$ denotes the trivial embedding of $\mathcal{I}_\alpha$, $\mathcal{A}$, or $\mathcal{S}$. $R^{\,\mathrm{triv}}$ denotes the trivial representation, under which $\vec{x}^{\,\mathrm{triv}}$ transforms.
- $\vec{x}^{\,\mathrm{ohe}}$ denotes the one-hot embedding of $\mathcal{I}_\alpha$ or $\mathcal{A}$. $R^{\,\mathrm{ohe}}$ denotes the one-hot representation, under which $\vec{x}^{\,\mathrm{ohe}}$ transforms.
- $\vec{x}^{\,\mathrm{sim}}$ denotes the simplex embedding of $\mathcal{I}_\alpha$ or $\mathcal{A}$. $R^{\,\mathrm{sim}}$ denotes the simplex representation, under which $\vec{x}^{\,\mathrm{sim}}$ transforms.
- $\vec{x}^{\,\mathrm{tet}}$ denotes the tetrahedral embedding of $\mathcal{A}$, applicable when $\alpha = 4$. $R^{\,\mathrm{tet}}$ denotes the tetrahedral representation, under which $\vec{x}^{\,\mathrm{tet}}$ transforms.
- $\vec{x}_l^{\,\mathrm{ohe}}$, $l \in \mathcal{I}_L$, denote a single-position one-hot embedding of $\mathcal{S}$. $R_l^{\,\mathrm{ohe}}$ denotes the representation under which $\vec{x}_l^{\,\mathrm{ohe}}$ transforms.
- $\vec{x}_l^{\,\mathrm{sim}}$, $l \in \mathcal{I}_L$, denotes a single-position simplex embedding of $\mathcal{S}$. $R_l^{\,\mathrm{sim}}$ denotes the simplex representation, under which $\vec{x}_l^{\,\mathrm{sim}}$ transforms.

1.5 Definitions

- **co-transformation**: Two equivariant embeddings *co-transform* if they are transformed by the same group representation.
- **embedding**: An *embedding* $\vec{x}$ is a map from a set $\mathcal{U}$ to a vector space $\mathbb{R}^m$.
- **equivariance**:
  - An embedding $\vec{x}$ of a set $\mathcal{U}$ is *equivariant* under a group H iff there is a group representation R such that $\vec{x}(hu) = R(h)\vec{x}(u)$ for all $h \in H$ and all $u \in \mathcal{U}$.
  - A linear model f on a set $\mathcal{U}$ is *equivariant* under a group H iff $f(u, \vec{\theta}) = \vec{\theta}^{\,\dagger}\vec{x}(u)$ for all $u \in \mathcal{U}$, where $\vec{x}$ is an embedding of $\mathcal{U}$ that is equivariant under H, and $\vec{\theta} \in \mathbb{R}^m$ where $m = \dim \vec{x}$.
- **equivalence**:
  - Two embeddings $\vec{x}$ and $\vec{y}$ of a set $\mathcal{U}$ are *equivalent*, denoted $\vec{x} \simeq \vec{y}$, iff there is an invertible matrix A (a similarity transformation) such that $\vec{y}(u) = A\vec{x}(u)$ for all $u \in \mathcal{U}$.
  - Two representations R and R' of a group H are *equivalent*, denoted $R \simeq R'$, iff there is an invertible matrix A (a similarity transformation) such that $R(h) = AR'(h)A^{-1}$ for all $h \in H$.
- **degree**: The degree of a representation R, denoted $\deg R$, is the number of rows (or equivalently the number of columns) of each matrix in the representation.
- **full rank embedding**: An embedding $\vec{x}$ of degree m is *full rank* iff $\operatorname{span} \vec{x} = \mathbb{R}^m$.
- **irreducibility**: An *irreducible* embedding is an equivariant embedding that transforms under an irreducible representation.
- **nonzero embedding**: An embedding $\vec{x}$ of a set $\mathcal{U}$ is *nonzero* iff $\vec{x}(u) \neq 0$ for some $u \in \mathcal{U}$.

- **linear model**: A complex-valued model of a sequence-function relationship is a function of the form $f(s\vec{\theta}) = \vec{\theta}^{\dagger}\vec{x}(s)$, where $\vec{x} : \mathcal{S} \to \mathbb{C}^M$ and $\vec{\theta} \in \mathbb{C}^M$.
- **module**:
  - A *module* is a vector space, together with a group representation that transforms the elements of that vector space.
  - The *module* of an equivariant embedding $\vec{x}$ that has degree m and transforms under the representation R is the module defined by the vector space $\mathbb{R}^m$ and representation R.
- **representation**: A group representation R is a map from a group H to a set of matrices that preserve the multiplicaiton rules of H, i.e., for which $R(h_1)R(h_2) = R(h_1h_2)$ for all $h_1, h_2 \in H$.
- **span of an embedding**: The *span* of an embedding $\vec{x}$ of a set $\mathcal{U}$, denoted $\operatorname{span}\vec{x}$, is the span of the set of vectors $\{\vec{x}(u) : u \in \mathcal{U}\}$.

2 Single-character embeddings and representations

In what follows, we consider a variety of singe-character embeddings and their representations. For the general definitions we assume an alphabet $\mathcal{A} = (c_1, \ldots, c_\alpha)$; for the DNA-specific examples we assume an alphabet $\mathcal{A}_{\text{DNA}} = (\text{A}, \text{C}, \text{G}, \text{T})$.

2.1 The trivial embedding $\vec{x}^{\,\text{triv}}$ and representation $R^{\,\text{triv}}$

The trivial embedding $\vec{x}^{\,\text{triv}}$ of characters in $\mathcal{A}$ has dimension 1 and is given by

$$\vec{x}^{\,\text{triv}}(c) = [1] \quad \text{for all} \quad c \in \mathcal{A}. \tag{1}$$

$\vec{x}^{\,\text{triv}}$ transforms under the trivial representation, $R^{\,\text{triv}}$, of H_{CP}. $R^{\,\text{triv}}$ has degree 1 and is given by

$$R^{\,\text{triv}}(h) = [1] \quad \text{for all} \quad h \in H_{\text{CP}}. \tag{2}$$

2.2 The one-hot embedding $\vec{x}^{\,\text{ohe}}$ and representation $R^{\,\text{ohe}}$

The one-hot embedding, $\vec{x}^{\,\text{ohe}}$, is defined to be an α-dimensional vector having elements

$$[\vec{x}^{\,\text{ohe}}(c)]_i = \begin{cases} 1 & \text{if } c_i = c, \\ 0 & \text{otherwise.} \end{cases} . \tag{3}$$

For example, $\vec{x}^{\,\text{ohe}}$ has dimension 4 for DNA alphabet $\mathcal{A}_{\text{DNA}}$ and is given by

$$\vec{x}^{\,\text{ohe}}(\text{A}) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{x}^{\,\text{ohe}}(\text{C}) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{x}^{\,\text{ohe}}(\text{G}) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{x}^{\,\text{ohe}}(\text{T}) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}. \tag{4}$$

$\vec{x}^{\,\text{ohe}}$ transforms under the one-hot representation, $R^{\,\text{ohe}}$, of H_{CP}. $R^{\,\text{ohe}}$ is isomorphic to what is called the "defining representation" of the symmetric group, but here we use the term "one-hot representation" for consistency. H_{CP} is generated by transpositions, i.e., the exchange of two specific characters. We denote each transposition as $c \leftrightarrow c'$, where c, c' are distinct characters in $\mathcal{A}$. For any transposition $c \leftrightarrow c'$, the corresponding one-hot representation has columns given by

$$[R^{\,\text{ohe}}(c \leftrightarrow c')]_{:,j} = \begin{cases} \vec{x}^{\,\text{ohe}}(c') & \text{if } c_j = c, \\ \vec{x}^{\,\text{ohe}}(c) & \text{if } c_j = c', \\ \vec{x}^{\,\text{ohe}}(c_j) & \text{otherwise.} \end{cases} . \tag{5}$$

In the case of $\mathcal{A}_{\text{DNA}}$, H_{CP} is generated by 6 transpositions. For these transpositions, the one-hot representations are

$$R^{\,\text{ohe}}(\text{A} \leftrightarrow \text{C}) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad R^{\,\text{ohe}}(\text{A} \leftrightarrow \text{G}) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad R^{\,\text{ohe}}(\text{A} \leftrightarrow \text{T}) = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \tag{6}$$

$$R^{\,\text{ohe}}(\text{C} \leftrightarrow \text{G}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad R^{\,\text{ohe}}(\text{C} \leftrightarrow \text{T}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad R^{\,\text{ohe}}(\text{G} \leftrightarrow \text{T}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{7}$$

2.3 The simplex embedding $\vec{x}^{\,\mathrm{sim}}$ and representation $R^{\,\mathrm{sim}}$

The simplex embedding and representation can be formulated in multiple equivalent ways. Here and in Main Text we use one formulation that is particularly amenable to analytic calculations: we define the simplex embedding $\vec{x}^{\,\mathrm{sim}}$ of $\mathcal{A}$ to be an $(\alpha-1)$-dimensional vector having elements

$$[\vec{x}_l^{\,\mathrm{sim}}(s)]_{i-1} = \begin{cases} 1 & \text{if } s_l = c_i \text{ and } i \neq \alpha, \\ -1 & \text{if } s_l = c_\alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

In the case of $\mathcal{A}_{\mathrm{DNA}}$, $\vec{x}^{\,\mathrm{sim}}$ has dimension 3 and is given by

$$\vec{x}^{\,\mathrm{sim}}(\mathrm{A}) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{x}^{\,\mathrm{sim}}(\mathrm{C}) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{x}^{\,\mathrm{sim}}(\mathrm{G}) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \vec{x}^{\,\mathrm{sim}}(\mathrm{T}) = \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix}. \tag{9}$$

$\vec{x}^{\,\mathrm{sim}}$ transforms under the simplex representation, $R^{\,\mathrm{sim}}$, of H_{CP}. $R^{\,\mathrm{sim}}$ is isomorphic to the "standard representation" of the symmetric group, but here we use the term "simplex representation" for consistency. For any transposition $c \leftrightarrow c'$, the corresponding simplex representation has columns given by

$$[R^{\,\mathrm{sim}}(c \leftrightarrow c')]_{:,j} = \begin{cases} \vec{x}^{\,\mathrm{sim}}(c') & \text{if } c_j = c, \\ \vec{x}^{\,\mathrm{sim}}(c) & \text{if } c_j = c', \\ \vec{x}^{\,\mathrm{sim}}(c_j) & \text{otherwise.} \end{cases} . \tag{10}$$

In the case of $\mathcal{A}_{\mathrm{DNA}}$, the representations of the 6 generating elements of H_{CP} are

$$R^{\,\mathrm{sim}}(\mathrm{A}\leftrightarrow\mathrm{C}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R^{\,\mathrm{sim}}(\mathrm{A}\leftrightarrow\mathrm{G}) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad R^{\,\mathrm{sim}}(\mathrm{C}\leftrightarrow\mathrm{G}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \tag{11}$$

$$R^{\,\mathrm{sim}}(\mathrm{A}\leftrightarrow\mathrm{T}) = \begin{bmatrix} -1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \quad R^{\,\mathrm{sim}}(\mathrm{C}\leftrightarrow\mathrm{T}) = \begin{bmatrix} 1 & -1 & 0 \\ 0 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix}, \quad R^{\,\mathrm{sim}}(\mathrm{G}\leftrightarrow\mathrm{T}) = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{bmatrix}. \tag{12}$$

2.4 The tetrahedral embedding $\vec{x}^{\,\mathrm{tet}}$ and representation $R^{\,\mathrm{tet}}$

It is worth noting that, instead of the simplex embedding for DNA characters, some studies (e.g., [3]) have proposed what we call the "tetrahedral embedding", $\vec{x}^{\,\mathrm{tet}}$. The tetrahedral embedding is given by

$$\vec{x}^{\,\mathrm{tet}}(\mathrm{A}) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \vec{x}^{\,\mathrm{tet}}(\mathrm{C}) = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \quad \vec{x}^{\,\mathrm{tet}}(\mathrm{G}) = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}, \quad \vec{x}^{\,\mathrm{tet}}(\mathrm{T}) = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}. \tag{13}$$

$\vec{x}^{\,\mathrm{tet}}$ transforms under the tetrahedral representation, $R^{\,\mathrm{tet}}$, of H_{CP}. In the tetrahedral representation, the 6 generating elements of H_{CP} are

$$R^{\,\mathrm{tet}}(\mathrm{A}\leftrightarrow\mathrm{C}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \quad R^{\,\mathrm{tet}}(\mathrm{A}\leftrightarrow\mathrm{G}) = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \quad R^{\,\mathrm{tet}}(\mathrm{A}\leftrightarrow\mathrm{T}) = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{14}$$

$$R^{\,\mathrm{tet}}(\mathrm{C}\leftrightarrow\mathrm{G}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R^{\,\mathrm{tet}}(\mathrm{C}\leftrightarrow\mathrm{T}) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad R^{\,\mathrm{tet}}(\mathrm{G}\leftrightarrow\mathrm{T}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{15}$$

The tetrahedral embedding is equivalent to the simplex embedding, and the tetrahedral representation is equivalent to the simplex representation:

$$A\vec{x}^{\,\mathrm{sim}} = \vec{x}^{\,\mathrm{tet}} \quad \text{and} \quad AR^{\,\mathrm{sim}}A^{-1} = R^{\,\mathrm{tet}} \quad \text{where} \quad A = \begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix}. \tag{16}$$

2.5 Maschke decompositions of $\vec{x}^{\,\mathrm{ohe}}$ and $R^{\,\mathrm{ohe}}$.

The one-hot representation decomposes, by Maschke's Theorem, into the direct sum of the trivial representation and the simplex representation:

$$R^{\,\mathrm{ohe}} \simeq R^{\,\mathrm{triv}} \oplus R^{\,\mathrm{sim}}. \tag{17}$$

Similarly, the one-hot embedding decomposes into the direct sum of the trivial embedding and the simplex embedding:

$$\vec{x}^{\,\mathrm{ohe}} \simeq \vec{x}^{\,\mathrm{triv}} \oplus \vec{x}^{\,\mathrm{sim}}. \tag{18}$$

In the case of $\mathcal{A}_{\mathrm{DNA}}$, these decompositions are accomplished by the similarity transformation matrix T, in the sense that

$$T\,R^{\,\mathrm{ohe}}\,T^{-1} = \left[\begin{array}{c|ccc} R^{\,\mathrm{triv}} & 0 & 0 & 0 \\ \hline 0 & & & \\ 0 & & R^{\,\mathrm{sim}} & \\ 0 & & & \end{array}\right], \quad T\,\vec{x}^{\,\mathrm{ohe}} = \left[\begin{array}{c} \vec{x}^{\,\mathrm{triv}} \\ \hline \\ \vec{x}^{\,\mathrm{sim}} \\ \\ \end{array}\right], \tag{19}$$

where

$$T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \quad \Rightarrow \quad T^{-1} = \frac{1}{4}\begin{bmatrix} 1 & 3 & -1 & -1 \\ 1 & -1 & 3 & -1 \\ 1 & -1 & -1 & 3 \\ 1 & -1 & -1 & -1 \end{bmatrix}. \tag{20}$$

The Maschke decomposition formulas for $R^{\,\mathrm{ohe}}$ and $\vec{x}^{\,\mathrm{ohe}}$, using $R^{\,\mathrm{tet}}$ and $\vec{x}^{\,\mathrm{tet}}$ in place of $R^{\,\mathrm{sim}}$ and $\vec{x}^{\,\mathrm{sim}}$, are given in terms of the alternative decomposition matrix T_{alt} by,

$$T_{\mathrm{alt}}\,R^{\,\mathrm{ohe}}\,T_{\mathrm{alt}}^{-1} = \left[\begin{array}{c|ccc} R^{\,\mathrm{triv}} & 0 & 0 & 0 \\ \hline 0 & & & \\ 0 & & R^{\,\mathrm{tet}} & \\ 0 & & & \end{array}\right] \quad \text{and} \quad T_{\mathrm{alt}}\,\vec{x}^{\,\mathrm{ohe}} = \left[\begin{array}{c} \vec{x}^{\,\mathrm{triv}} \\ \hline \\ \vec{x}^{\,\mathrm{tet}} \\ \\ \end{array}\right], \tag{21}$$

where

$$T_{\mathrm{alt}} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \quad \Rightarrow \quad T_{\mathrm{alt}}^{-1} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}. \tag{22}$$

Note that T_{alt} is a Haddamard matrix of order 4 [3]. Also note that, unlike the simplex embedding, it is unclear how to generalize the tetrahedral embedding to alphabets that contain arbitrary numbers of characters.

3 Results for general groups

3.1 Nonzero embeddings

Claim 3.1. *Let $\vec{x}$ be an embedding of a set $\mathcal{U}$ that transforms under a representation of a group H that is transitive on $\mathcal{U}$. Then either $\vec{x}(u) = \vec{0}$ all $u \in \mathcal{U}$, or $\vec{x}(u) \neq \vec{0}$ for all $u \in \mathcal{U}$.*

Proof. Assume that $\vec{x}(u) = \vec{0}$ for some $u \in \mathcal{U}$, and choose any other $u' \in \mathcal{U}$. Because H is transitive, there is a group element $h \in \mathcal{H}$ such that $u' = hu$. Letting R denote the representation of H that $\vec{x}$ transforms under, we find

$$\vec{x}(u') = R(h)\vec{x}(u) = R(h)\vec{0} = \vec{0}. \tag{23}$$

We thus find that $\vec{x}(u) = \vec{0}$ for some $u \in \mathcal{U}$ implies that $\vec{x}(u) = \vec{0}$ for all $u \in \mathcal{U}$. This completes the proof. We note, however, that if H is not transitive on $\mathcal{U}$, then Claim 3.1 will generally not hold. □

3.2 Equivariance

Claim 3.2. *If f is a linear model based on the embedding $\vec{x}$ of a set $\mathcal{U}$ into a vector space V, and $\vec{x}$ is an equivariant embedding that transforms under a representation R of a group H, then*

$$f\left(u; \vec{\theta}\,\right) = f\left(hu; R(h)^{-1\dagger}\,\vec{\theta}\,\right) \tag{24}$$

for all $\vec{\theta} \in V$, all $u \in \mathcal{U}$, and all $h \in H$.

Proof.

$$\begin{aligned} f(hu; R(h)^{-1\dagger}\vec{\theta}) &= [R(h)^{-1\dagger}\vec{\theta}]^\dagger \vec{x}(hu) && (25)\\ &= \vec{\theta}^\dagger R(h)^{-1} R(h)\vec{x}(u) && (26)\\ &= \vec{\theta}^\dagger \vec{x}(u) && (27)\\ &= f(u;\vec{\theta}). && (28) \end{aligned}$$

□

3.3 Irreducible embeddings are full rank

Claim 3.3. *Every nonzero equivariant embedding that transforms under an irreducible representation is full rank.*

Proof. Let $\vec{x}$ be a nonzero embedding of a set $\mathcal{U}$ into a vector space V. Further assume that $\vec{x}$ transforms under an irreducible representation R of a group H. Since R transforms vectors in V, V is an irreducible H module. For any $h \in H$,

$$\begin{aligned} R(h)\,\mathrm{span}\,\vec{x} &= \mathrm{span}\,\{R(h)\vec{x}(u) : u \in \mathcal{U}\} && (29)\\ &= \mathrm{span}\,\{\vec{x}(hu) : u \in \mathcal{U}\} && (30)\\ &= \mathrm{span}\,\{\vec{x}(u) : u \in \mathcal{U}\} && (31)\\ &= \mathrm{span}\,\vec{x}. && (32) \end{aligned}$$

Therefore, $\mathrm{span}\,\vec{x}$ is an H-invariant subspace of V. Since V is an irreducible module, the only H-invariant subspaces of V are 0 and V. Because $\vec{x}$ is nonzero, $\mathrm{span}\,\vec{x} \neq 0$. We thus find that $\mathrm{span}\,\vec{x} = V$, proving that $\vec{x}$ is full rank. □

3.4 Direct sums of inequivalent irreducible embeddings have full rank (Theorem 2 of Main Text)

Claim 3.4. *(Theorem 2 of Main Text) For each $k \in \{1, \ldots, K\}$, let $\vec{x}_k$ be a nonzero embedding that transforms under an irreducible representation R_k of a group H. Then the direct sum of all $\vec{x}_k$ is full rank if all R_k are pairwise inequivalent.*

Proof. Let $V = \bigoplus_{k=1}^K V_k$ where V_k is the vector space to which the embedding $\vec{x}_k$ maps. Letting $R = \bigoplus_{k=1}^K R_k$, we see that R is a representation of H over V. Given R, define an H-invariant inner product $\langle\cdot,\cdot\rangle$ over V, i.e. an inner product that satisfies

$$\langle R(h)\vec{v}, R(h)\vec{w}\rangle = \langle \vec{v}, \vec{w}\rangle \quad \text{for all} \quad \vec{v}, \vec{w} \in V \quad \text{and} \quad h \in H. \tag{33}$$

Such an inner product can always be constructed (e.g., see p. 16 of [2]). Next, using the notation $\vec{x}(s) = \bigoplus_{k=1}^K \vec{x}_k(s)$, define the subspace W of V given by

$$W = \{\vec{w} \in V : \langle \vec{w}, \vec{x}(s)\rangle = 0 \ \text{ for all } \ s \in \mathcal{S}\}. \tag{34}$$

W is H-invariant, since for all $\vec{w} \in W$ and $h \in H$,

$$\langle R(h)\vec{w}, \vec{x}(s)\rangle = \left\langle \vec{w}, R(h)^{-1}\vec{x}(s)\right\rangle = \left\langle \vec{w}, \vec{x}(h^{-1}s)\right\rangle = 0. \tag{35}$$

We are goingto show that $\dim(W) = 0$.

We start by showing that the projection operator P that projects V orthogonally on to W according to the inner product $\langle\cdot,\cdot\rangle$ is an H-homomorphism. Assume W has dimension n, and let $\{\hat{e}_1, \ldots, \hat{e}_n\}$ be an orthonormal basis of W (i.e., $\langle \hat{e}_i, \hat{e}_j\rangle = \delta_{ij}$ for all $i, j \in \mathcal{I}_n$). We define the projection operator P to be such that

$$P\vec{v} = \sum_{i=1}^n \hat{e}_i \langle \hat{e}_i, \vec{v}\rangle \quad \text{for all} \quad v \in V. \tag{36}$$

Because $\langle\cdot,\cdot\rangle$ is a H-invariant inner product, $\langle R(h)\hat{e}_i, R(h)\hat{e}_j\rangle = \delta_{ij}$ for all $h \in H$, and thus $\{R(h)\hat{e}_i\}_{i=1}^n$ is also an orthonormal basis for W. Consequently, for all $h \in H$ and $\vec{v} \in V$,

$$P\vec{v} = \sum_{i=1}^n R(h)\hat{e}_i \langle R(h)\hat{e}_i, \vec{v}\rangle = R(h)\sum_{i=1}^n \hat{e}_i \left\langle \hat{e}_i, R(h)^{-1}\vec{v}\right\rangle = R(h)PR(h)^{-1}\vec{v}. \tag{37}$$

We therefore see that P commutes with R. P is therefore an H-homomorphism.

By theorem 1.7.8 in [2], there is a set of real scalars $\{c_k\}_{k=1}^K$ such that

$$P = \bigoplus_{k=1}^{K} c_k I_k, \tag{38}$$

where I_k denotes the identity operator on V_k. Because $P\vec{x}(s) = \vec{0}$ for all $s \in \mathcal{S}$,

$$0 = \langle P\vec{x}(s), P\vec{x}(s)\rangle = \sum_{k=1}^{K} \langle c_k\vec{x}_k(s), c_k\vec{x}_k(s)\rangle = \sum_{k=1}^{K} ||c_k||^2||\vec{x}_k(s)||^2. \tag{39}$$

Since all $\vec{x}_k$ are nonzero, this is only possible if all $c_k = 0$. We thus find that P projects all vectors to $\vec{0}$, and hence that W has dimension 0. Since W is the orthogonal complement of span $\vec{x}$, $\vec{x}$ has full rank. This completes the proof. □

4 Results for S_α, the symmetric group

The symmetric group, S_α, is the group of permutations among the integers $\mathcal{I}_\alpha$. S_α is generated by transpositions, which exchange two elements of $\mathcal{I}_\alpha$ while leaving all other elements unchanged. In what follows, we use $i \leftrightarrow j$ to denote the transposition that exchanges elements $i, j \in \mathcal{I}_\alpha$.

4.1 The trivial embedding $\vec{x}^{\,\mathrm{triv}}$ and representation $R^{\,\mathrm{triv}}$

The trivial embedding of $\mathcal{I}_\alpha$, $\vec{x}^{\,\mathrm{triv}}$, has degree 1 and is given by

$$\vec{x}^{\,\mathrm{triv}}(i) = [1] \quad \text{for all} \quad i \in I_\alpha. \tag{40}$$

$\vec{x}^{\,\mathrm{triv}}$ transforms under the trivial representation, $R^{\,\mathrm{triv}}$, which also has degree 1 and is given by

$$R^{\,\mathrm{triv}}(h) = [1] \quad \text{for all} \quad h \in S_\alpha. \tag{41}$$

4.2 The simplex embedding $\vec{x}^{\,\mathrm{sim}}$ and representation $R^{\,\mathrm{sim}}$

Here we define an explicit coordinate realization of the simplex embedding $\vec{x}^{\,\mathrm{sim}}$ of $\mathcal{I}_\alpha$, as well as the simplex representation $R^{\,\mathrm{sim}}$ of S_α, under which $\vec{x}^{\,\mathrm{sim}}$ transforms. Both $\vec{x}^{\,\mathrm{sim}}$ and $R^{\,\mathrm{sim}}$ have degree $\alpha - 1$, and we use $k, l \in \mathcal{I}_{\alpha-1}$ to index the elements of $\vec{x}^{\,\mathrm{sim}}$ and $R^{\,\mathrm{sim}}$. We define the coordinates of $\vec{x}^{\,\mathrm{sim}}$ to be

$$[\vec{x}^{\,\mathrm{sim}}(i)]_k = \delta_{ik} \quad \text{for} \quad i \in \mathcal{I}_{\alpha-1}, \tag{42}$$

$$[\vec{x}^{\,\mathrm{sim}}(\alpha)]_k = -1. \tag{43}$$

Correspondingly, we define the coordinate realization of $R^{\,\mathrm{sim}}$ on transpositions to be

$$[R^{\,\mathrm{sim}}(i \leftrightarrow j)]_{kl} = \delta_{kl} - \delta_{ik}\delta_{il} - \delta_{jk}\delta_{jl} + \delta_{ik}\delta_{jl} + \delta_{jk}\delta_{il}, \tag{44}$$

$$[R^{\,\mathrm{sim}}(i \leftrightarrow \alpha)]_{kl} = \delta_{kl} - \delta_{kl}\delta_{ik} - \delta_{il}. \tag{45}$$

One can readily verify that these definitions work by showing that $\vec{x}^{\,\mathrm{sim}}$ does in fact transform according to $R^{\,\mathrm{sim}}$ for all transpositions in S_α, and thereby showing that $R^{\,\mathrm{sim}}$ is indeed a valid representation of S_α. The simplex embedding and representation in Sec. 2.3 is an example of this formulation for $\alpha = 4$. Finally, we note that $\vec{x}^{\,\mathrm{sim}}$ has the useful property that

$$\sum_{i=1}^{\alpha} \vec{x}^{\,\mathrm{sim}}(i) = \vec{0}. \tag{46}$$

4.3 $\vec{x}^{\,\mathrm{triv}}$ and $\vec{x}^{\,\mathrm{sim}}$ are the only inequivalent irreducible embeddings that transform under S_α

Claim 4.1. *If a nonzero embedding of $\mathcal{I}_\alpha$ transforms under an irreducible representation of S_α, then it is equivalent to either the trivial embedding or the simplex embedding.*

Proof. Assume that $\vec{x}$ is a nonzero embedding of $\mathcal{I}_\alpha$, that $\vec{x}$ has degree m, and that $\vec{x}$ transforms under an irreducible representation R of $\mathcal{S}_\alpha$.

First consider $\alpha = 2$. S_2 has only two inequivalent irreducible representations: the trivial representation and the simplex representation. The trivial representation supports the embedding $\vec{x}^{\,\mathrm{triv}}$, and the simplex representation supports the simplex embedding $\vec{x}^{\,\mathrm{sim}}$. This proves the claim for $\alpha = 2$.

Now consider $\alpha \geq 3$. Select $i, j, k \in \mathcal{I}_\alpha$, where $i \neq j$, $i \neq k$, $j \neq k$, and assume that $\vec{x}(i) = \vec{x}(j) \neq \vec{x}(k)$. The transposition $j \leftrightarrow k$ exchanges j and k without affecting i. Therefore, the representation of $R(j \leftrightarrow k)$ must exchange the vectors $\vec{x}(j)$ and $\vec{x}(k)$ while leaving the vector $\vec{x}(i)$ invariant. But this is impossible because $\vec{x}(i) = \vec{x}(j)$. Consequently, $\vec{x}(i)$ must be either the same for all $i \in \mathcal{I}_\alpha$, or must be different for all $i \in \mathcal{I}_\alpha$. If $\vec{x}(i)$ is the same for all i, then $\vec{x} \simeq \vec{x}^{\,\mathrm{triv}}$. In the rest of this proof we assume that $\vec{x}(i)$ is different for all $i \in \mathcal{I}_\alpha$. We will show that this implies $\vec{x} \simeq \vec{x}^{\,\mathrm{sim}}$.

Claim 3.3 shows that, because $\vec{x}$ is irreducible, $\vec{x}$ is also full rank. We can therefore choose a subset $\{n_1, \ldots, n_m\} \subseteq \mathcal{I}_\alpha$ such that $\{\vec{x}(n_1), \ldots, \vec{x}(n_m)\}$ is a basis for $\mathbb{R}^m$. For simplicity, we assume that $n_i = i$ for all $i \in \mathcal{I}_m$, so that $\mathcal{B} = \{\vec{x}(i) : i \in \mathcal{I}_m\}$ is the basis; we revisit this assumption at the end of the proof. In the basis $\mathcal{B}$, the elements of $\vec{x}$ are

$$[\vec{x}(i)]_k = \delta_{ik} \tag{47}$$

for all $i, k \in \mathcal{I}_m$.

Now choose $a, b, j \in \mathcal{I}_m$ such that $a \neq b$, $a \neq j$, $b \neq j$. We can compute the elements of $R(a \leftrightarrow b)$ in basis $\mathcal{B}$ by noting that

$$[\vec{x}(a)]_i = \sum_{k=1}^{m} [R(a \leftrightarrow b)]_{ik} [\vec{x}(b)]_k = [R(a \leftrightarrow b)]_{ib}, \tag{48}$$

$$[\vec{x}(b)]_i = \sum_{k=1}^{m} [R(a \leftrightarrow b)]_{ik} [\vec{x}(a)]_k = [R(a \leftrightarrow b)]_{ia}, \tag{49}$$

$$[\vec{x}(j)]_i = \sum_{k=1}^{m} [R(a \leftrightarrow b)]_{ik} [\vec{x}(j)]_k = [R(a \leftrightarrow b)]_{ij} \quad \text{for} \quad j \neq a, b. \tag{50}$$

and therefore that,

$$[R(a \leftrightarrow b)]_{ia} = \delta_{ib}, \tag{51}$$

$$[R(a \leftrightarrow b)]_{ib} = \delta_{ia}, \tag{52}$$

$$[R(a \leftrightarrow b)]_{ij} = \delta_{ij} \quad \text{for} \quad j \neq a, b. \tag{53}$$

Combining Eq. 51, Eq. 52, and Eq. 53 into a single expression gives that, for any $a, b \in I_m$ such that $a \neq b$, and any $i, j \in I_m$,

$$[R(a \leftrightarrow b)]_{ij} = \delta_{ij}(1 - \delta_{aj})(1 - \delta_{bj}) + \delta_{aj}\delta_{bi} + \delta_{bj}\delta_{ai}, \tag{54}$$

$$= \delta_{ij} - \delta_{ai}\delta_{aj} - \delta_{bi}\delta_{bj} + \delta_{ai}\delta_{bj} + \delta_{bi}\delta_{aj}. \tag{55}$$

Now choose $c \in \mathcal{I}_\alpha \setminus \mathcal{I}_m$ and $a, j \in \mathcal{I}_m$ such that $a \neq j$. Because $\mathcal{B}$ is a basis, there are coefficients r_i^c, $i \in \mathcal{I}_m$, such that

$$\vec{x}(c) = \sum_{i=1}^{m} r_i^c \vec{x}(i) \tag{56}$$

Consequently, the elements of $\vec{x}(c)$ in the basis $\mathcal{B}$ are,

$$[\vec{x}(c)]_i = r_i^c. \tag{57}$$

We compute the matrix elements of $R(a \leftrightarrow c)$ by noting that

$$[\vec{x}(a)]_i = \sum_{k=1}^{m} [R(a \leftrightarrow c)]_{ik} [\vec{x}(c)]_k = \sum_{k=1}^{m} [R(a \leftrightarrow c)]_{ik} r_k^c, \tag{58}$$

$$[\vec{x}(c)]_i = \sum_{k=1}^{m} [R(a \leftrightarrow c)]_{ik} [\vec{x}(a)]_k = [R(a \leftrightarrow c)]_{ia}, \tag{59}$$

$$[\vec{x}(j)]_i = \sum_{k=1}^{m} [R(a \leftrightarrow c)]_{ik} [\vec{x}(j)]_k = [R(a \leftrightarrow c)]_{ij}. \tag{60}$$

From Eq. 47, Eq. 57, Eq. 59, and Eq. 60 we see that, for all $i, k \in \mathcal{I}_m$,

$$[R(a \leftrightarrow c)]_{ik} = \delta_{ik}(1 - \delta_{ak}) + r_i^c \delta_{ak} \tag{61}$$

are the elements of $R(a \leftrightarrow c)$. Plugging this into Eq. 58 then gives a constraint on the r_i^c coefficients:

$$\delta_{ia} = \sum_{k=1}^{m} [\delta_{ik}(1 - \delta_{ak}) + r_i^c \delta_{ak}] r_k^c = r_i^c (1 - \delta_{ai}) + r_a^c r_i^c. \tag{62}$$

Considering different values for $i \in \mathcal{I}_\alpha$ in Eq. 62 then gives two specific constraints on r_i^c:

$$(r_a^c)^2 = 1 \text{ for } i = a \quad \text{and} \quad r_i^c (1 + r_a^c) = 0 \text{ for } i \neq a. \tag{63}$$

The left side of Eq. 63 gives $r_i^c = \pm 1$. If $r_a^c = 1$, the right side of Eq. 63 gives $r_i^c = 0$ for $i \neq a$. But this would mean that $\vec{x}(c) = \vec{x}(a)$, which violates the requirement that all $\vec{x}(i)$ are distinct. We thus find that $r_i^c = -1$. Therefore, for all $i, k \in \mathcal{I}_m$,

$$[\vec{x}(c)]_i = -1, \quad \text{and} \quad [R(a \leftrightarrow c)]_{ik} = \delta_{ik}(1 - \delta_{ak}) - \delta_{ak}. \tag{64}$$

In particular, this means that there is only a single possible $\vec{x}(c)$ for all $c \in \mathcal{I}_\alpha \setminus \mathcal{I}_m$. Since all $\vec{x}(c)$ must be distinct, there can only be one possible value for $c \geq m$, i.e. $c = \alpha = m + 1$. We therefore find that for $i, j, k, l \in I_{\alpha-1}$, $i \neq j$,

$$\begin{aligned}
[\vec{x}(i)]_k &= \delta_{ik} && (65)\\
[\vec{x}(\alpha)]_k &= -1 && (66)\\
[R(i \leftrightarrow j)]_{kl} &= \delta_{kl} - \delta_{ik}\delta_{il} - \delta_{jk}\delta_{jl} + \delta_{ik}\delta_{jl} + \delta_{jk}\delta_{il}. && (67)\\
[R(i \leftrightarrow \alpha)]_{kl} &= \delta_{kl} - \delta_{ik}\delta_{kl} - \delta_{il}. && (68)
\end{aligned}$$

We therefore see that $\vec{x} = \vec{x}^{\,\text{sim}}$ and $R = R^{\,\text{sim}}$. Finally, we revisit our simplifying assumption that $n_i = i$ for all $i \in \mathcal{I}_m$. It is readily seen that removing this assumption leads to

$$\vec{x} = T\vec{x}^{\,\text{sim}} \quad \text{and} \quad R = TR^{\,\text{sim}}T^{-1} \tag{69}$$

where T is the similarity transformation that maps each $\vec{x}(i)$ to $\vec{x}(n_i)$, i.e.,

$$T = R(g) \quad \text{where} \quad g = \begin{pmatrix} 1 & 2 & \cdots & m \\ n_1 & n_2 & \cdots & n_m \end{pmatrix}. \tag{70}$$

Consequently, $\vec{x} \simeq \vec{x}^{\,\text{sim}}$ and $R \simeq R^{\,\text{sim}}$. This completes the proof. □

4.4 Irreducible embeddings that co-transform under S_α are proportional

Claim 4.2. *Nonzero embeddings that transform under the same irreducible representation of the symmetric group are equal up to a constant of proportionality.*

Proof. Let $\vec{x}$ and $\vec{y}$ be two embeddings of $\mathcal{I}_\alpha$ that transform according to the same irreducible representation R of S_α. First we construct a matrix A such that

$$\vec{y}(i) = A\vec{x}(i) \quad \text{for all} \quad i \in \mathcal{I}_\alpha. \tag{71}$$

From Claim 4.1, $\vec{x}$ and $\vec{y}$ must transform under the trivial representation or a representation equivalent to the simplex representation. If $\vec{x}$ and $\vec{y}$ transform under the trivial representation, then both $\vec{x}(i)$ and $\vec{y}(i)$ are nonzero scalars that do not depend on i, and Eq. 71 is satisfied by setting $A = \vec{y}/\vec{x}$. If $\vec{x}$ and $\vec{y}$ transform under a representation equivalent to the simplex representation, then both $\vec{x}$ and $\vec{y}$ have degree $m = \alpha - 1$. Define the $m \times m$ matrices $X = [\vec{x}(1)\ \vec{x}(2)\ \ldots\ \vec{x}(m)]$ and $Y = [\vec{y}(1)\ \vec{y}(2)\ \ldots\ \vec{y}(m)]$. Next define $A = YX^{-1}$; note that X^{-1} exists because the columns of X are linearly independent. By construction, $\vec{y}(i) = A\vec{x}(i)$ for $i \in I_{\alpha-1}$. Moreover $\sum_{i=1}^{\alpha} \vec{x}(i) = \sum_{i=1}^{\alpha} \vec{y}(i) = \vec{0}$ from Eq. 46. Therefore,

$$\vec{y}(\alpha) = -\sum_{i=1}^{\alpha-1} \vec{y}(i) = -\sum_{i=1}^{\alpha-1} A\vec{x}(i) = A\vec{x}(\alpha) \tag{72}$$

as well. This proves that Eq. 71 is satisfied when $\vec{x}$ and $\vec{y}$ transform under the simplex representation.

Next we show that A commutes with R. For all $h \in \mathcal{S}_\alpha$ and all $i \in \mathcal{I}_\alpha$,

$$AR(h)\vec{x}(i) = A\vec{x}(hi) = \vec{y}(hi) = R(h)\vec{y}(i) = R(h)A\vec{x}(i). \tag{73}$$

Because $\vec{x}$ has full rank, this requires

$$AR(h) = R(h)A \tag{74}$$

for all $h \in S_\alpha$, thereby showing that A and R commute.

Finally, we show that $A = cI$ for some $c \in \mathbb{R}$. This is a consequence of Shur's Lemma (e.g., see Corollary 1.6.6 of ref. [2]), which tells us that any matrix that commutes with an irreducible representation is either invertible or the zero matrix. Now define $B = A - cI$ where c is an eigenvalue of A. B commutes with R, since

$$BR(h) = (A - cI)R(h) = R(h)(A - cI) = R(h)B. \tag{75}$$

But B cannot be invertible because at least one of its eigenvalues is zero. B must therefore be the zero matrix, implying that $A = cI$. This shows that $\vec{y} = c\vec{x}$, completing the proof. □

5 Results for H_{PSCP}, the group of position-specific character permutations

5.1 Irreducible embeddings that co-transform under H_{PSCP} are proportional (Theorem 1 of Main Text)

Claim 5.1. *(Theorem 1 of Main Text) Nonzero sequence embeddings that transform under the same irreducible representation of H_{PSCP} are equal up to a constant of proportionality.*

Proof. As stated in Main Text, the group of position-specific character permutations, H_{PSCP}, is given by the product

$$H_{\text{PSCP}} = H_{\text{CP}}^1 \times \cdots \times H_{\text{CP}}^L, \tag{76}$$

where H_{CP}^l is the group of permutations among the α possible characters at position l in a sequence. Therefore, every element $h \in H_{\text{PSCP}}$ can be expressed as the product

$$h = \prod_{l=1}^{L} h_l, \quad h_l \in H_{\text{CP}}^l. \tag{77}$$

It is known that if we consider representations over an algebraically closed field like $\mathbb{C}$, then all irreducible representations of a product group can be expressed as tensor products of the irreducible representations of the individual group factors (e.g., Theorem 1.11.3 of ref. [2]). Therefore, any irreducible representation R of H_{PSCP} can be expressed as

$$R(h) = \bigotimes_{l=1}^{L} R_l(h_l), \tag{78}$$

where each R_l is an irreducible representation of H_{CP}^l, and where, for each $h \in H_{\text{PSCP}}$, $h_l \in H_{\text{CP}}^l$ is the corresponding factor in Eq. 77.

Now let $\{\vec{x}_q\}_{q=1}^{Q}$ denote a set of embeddings that transform under R. Eq. 78 implies that each $\vec{x}_q$ must be expressible as a tensor product of site-specific embeddings as

$$\vec{x}_q = \bigotimes_{l=1}^{L} \vec{x}_q^l, \tag{79}$$

where each $\vec{x}_q^l$ is a site-specific embedding that transforms under H_{CP}^l. Because H_{CP}^l is isomorphic to S_α, each R_l can be thought of as a representation of $\mathcal{S}_\alpha$, and each $\vec{x}_q^l$ can be thought of as an embedding of $\mathcal{I}_\alpha$. Claim 4.2 says that any two nonzero embeddings of $\mathcal{I}_\alpha$ that transform under the same irreducible representation S_α must be scalar multiples of one another. There must therefore exist, for every $l \in \mathcal{I}_L$, a nonzero embedding $\vec{x}^l$ that transforms under R_l and that satisfies

$$\vec{x}_q^l = c_q^l \vec{x}^l \tag{80}$$

for all $q \in \mathcal{I}_Q$. We therefore find that

$$\vec{x}_q = c_q \vec{x}, \quad \text{where} \quad c_q = \prod_{l=1}^{L} c_q^l \quad \text{and} \quad \vec{x} = \bigotimes_{l=1}^{L} \vec{x}^l. \tag{81}$$

This completes the proof. □

Going back to Eq. 22 of Main Text, we find as a consequence of Claim 5.1 that

$$\bigoplus_{k=1}^{K} \bigoplus_{q=1}^{Q_k} \vec{x}_{kq} = \bigoplus_{k=1}^{K} \bigoplus_{q=1}^{Q_k} c_{qk} \vec{x}_k = \left[\bigoplus_{k=1}^{K} \bigoplus_{q=1}^{Q_k} c_{qk} I_{M_k \times M_k} \right] \left[\bigoplus_{k=1}^{K} \bigoplus_{q=1}^{Q_k} \vec{x}_k \right] \simeq \bigoplus_{k=1}^{K} \bigoplus_{q=1}^{Q_k} \vec{x}_k = \bigoplus_{k=1}^{K} Q_k \vec{x}_k, \tag{82}$$

where $I_{M_k \times M_k}$ denotes the identity matrix. Note that the penultimate step follows from the fact that multiplication by the matrix $\bigoplus_{k=1}^{K} \bigoplus_{q=1}^{Q_k} c_{qk} I_{M_k \times M_k}$ is a similarity transformation. Eq. 22 of Main Text results.

5.2 Generalized distillations of H_{PSCP}-equivariant embeddings

Here we remove the assumption in the Main Text that every irreducible embedding $\vec{x}_{kq}$ in Eq. 16 is nonzero. Removing this assumption and retracing the logic, it is straight-forward to see that only two changes to the Main Text equations are required. First, Eq. 12 of Main Text becomes

$$\vec{x}^{\,\text{dist}} = \bigoplus_{k\in\mathcal{K}} \vec{x}_k, \tag{83}$$

where $\mathcal{K} \subseteq \mathcal{I}_K$ is the set of k such that $\vec{x}_{qk} \neq \vec{0}$ for at least one value of q, and where $\vec{x}_k$ is chosen to be any one of the nonzero embeddings $\vec{x}_{qk}$. Second, the statement that the number of gauge freedoms equals the sum of the degrees of all redundant irreducible representations still holds, but the term "redundant" requires clarification. For every $k \in \mathcal{I}_K$, there are Q_k copies of the representation R_k in the Maschke decomposition (Eq. 15 of Main Text). When $k \in \mathcal{K}$, $Q_k - 1$ of these copies of R_k are redundant, because all but one of the Q_k embeddings $\vec{x}_{kq}$ can be zeroed out by similarity transformations. When $k \in \mathcal{I}_K \setminus \mathcal{K}$, however, all Q_k of these copies of R_k are redundant because all of the $\vec{x}_{kq}$ are already zero. Eq. 22 of Main Text therefore becomes

$$\gamma \;=\; \dim \vec{x} - \dim \vec{x}^{\,\text{dist}} \;=\; \sum_{k\in\mathcal{K}} (Q_k - 1) \deg R_k \;\; + \sum_{k\in\mathcal{I}_K\setminus\mathcal{K}} Q_k \deg R_k. \tag{84}$$

5.3 Restriction to the reals

Claim 5.2. *If an embedding $\vec{x}$ is real then there is a corresponding real distillation matrix T_{dist}.*

Proof. Given an embedding $\vec{x}$, the embedding distillation procedure yields a distillation matrix T_{dist} such that

$$T_{\text{dist}}\, \vec{x} = \vec{x}^{\,\text{dist}} \oplus \vec{0}_\gamma. \tag{85}$$

If $\vec{x}$ is restricted to the reals, and we construct $\vec{x}^{\,\text{dist}}$ using the choices of $\vec{x}^{\,\text{triv}}$ and $\vec{x}^{\,\text{sim}}$ described in this work (both of which are real), then taking one half the sum of Eq. 86 and its complex conjugate yields

$$\Re(T_{\text{dist}})\vec{x} = \vec{x}^{\,\text{dist}} \oplus \vec{0}_\gamma. \tag{86}$$

where $\Re(T_{\text{dist}})$ denotes the real part of T_{dist}. We therefore find that $\Re(T_{\text{dist}})$ is an equivalent real distillation matrix. In particular, the last γ rows of T_{dist} provide a real basis for the space of gauge freedoms G. We conclude that our embedding distillation procedure works when $\vec{x}$ and $\vec{\theta}$ are restricted to the reals. □

6 Analytic results for specific generalized one-hot models (Table 1 of Main Text)

We now compute the number of parameters and gauge freedoms for a variety of generalized one-hot models. To avoid conflicts in notation, we use N and n in place of K and k to parameterize orders of interactions. Note that N here does not refer the number of sequences in $\mathcal{S}$.

6.1 N-order model

Here we compute the number of parameters and gauge freedoms for the N-order model. We begin by computing the dimension of the model embedding $\vec{x}_{N\text{-order}}$. The position sets that define $\vec{x}_{N\text{-order}}$ are

$$\{A_j\}_{j=1}^{J} = \{U : U \subseteq I_L, |U| = N\}. \tag{87}$$

From this we see that there are $J = \binom{L}{N}$ distinct sets A_j, and that $|A_j| = N$ for every $j \in \mathcal{I}_J$. By Main Text Eq. 30, the dimension of $\vec{x}_{N\text{-order}}$ is therefore

$$\dim \vec{x}_{N\text{-order}} \;=\; \sum_{j=1}^{J} \alpha^{|A_j|} \tag{88}$$

$$\;=\; \binom{L}{N} \alpha^N. \tag{89}$$

Next we compute the dimension of the distilled embedding $\dim \vec{x}^{\text{dist}}_{N\text{-order}}$. The position sets that define $\vec{x}^{\text{dist}}_{N\text{-order}}$ are

$$\begin{aligned}
\{B_k\}_{k=1}^{K} &= \{V : V \subseteq A_j, j \in \mathcal{I}_J\} && (90)\\
&= \{V : V \subseteq U, U \subseteq I_L, |U| = N\} && (91)\\
&= \{V : V \subseteq I_L, |V| \leq N\} && (92)\\
&= \bigcup_{n=0}^{N} \{V : V \subseteq I_L, |V| = n\}, && (93)
\end{aligned}$$

where the union is over disjoint sets. From this we see that, for each $n \in \{0, \ldots, N\}$, there are $\binom{L}{n}$ sets B_k such that $|B_k| = n$. By Main Text Eq. 32, the dimension of $\vec{x}^{\text{dist}}_{N\text{-order}}$ is therefore

$$\begin{aligned}
\dim \vec{x}^{\text{dist}}_{N\text{-order}} &= \sum_{k=1}^{K} (\alpha - 1)^{|B_k|} && (94)\\
&= \sum_{n=0}^{N} \binom{L}{n} (\alpha - 1)^n. && (95)
\end{aligned}$$

The number of parameters, $M_{N\text{-order}}$, and the number of gauge freedoms, $\gamma_{N\text{-order}}$, are therefore given by

$$\begin{aligned}
M_{N\text{-order}} &= \dim \vec{x}_{N\text{-order}} && (96)\\
&= \binom{L}{N} \alpha^N, && (97)\\
\gamma_{N\text{-order}} &= \dim \vec{x}_{N\text{-order}} - \dim \vec{x}^{\text{dist}}_{N\text{-order}} && (98)\\
&= \binom{L}{N} \alpha^N - \sum_{n=0}^{N} \binom{L}{n} (\alpha - 1)^n. && (99)
\end{aligned}$$

6.2 Hierarchical N-order models

We define the hierarchical N-order model to be the sum of n-order models taken over all $n \in \{0, 1, \ldots, N\}$. The embedding of the hierarchical N-order model is

$$\vec{x}_{\leq N\text{-order}} = \bigoplus_{n=0}^{N} \vec{x}_{n\text{-order}}. \tag{100}$$

The number of parameters of the hierarchical N-order model is therefore

$$\begin{aligned}
M_{\leq N\text{-order}} &= \sum_{n=0}^{N} M_{n\text{-order}} && (101)\\
&= \sum_{n=0}^{N} \binom{L}{n} \alpha^n. && (102)
\end{aligned}$$

Since every parameter of the models of order $N-1$ and lower is redundant with parameters of the model of order N, the number of gauge freedoms of the hierarchical N-order model is

$$\begin{aligned}
\gamma_{\leq N\text{-order}} &= \dim \vec{x}_{\leq N\text{-order}} - \dim \vec{x}^{\text{dist}}_{\leq N\text{-order}} && (103)\\
&= M_{\leq N\text{-order}} - \dim \vec{x}^{\text{dist}}_{N\text{-order}} && (104)\\
&= \sum_{n=0}^{N} \binom{L}{n} \alpha^n - \sum_{n=0}^{N} \binom{L}{n} (\alpha - 1)^n && (105)\\
&= \sum_{n=0}^{N} \binom{L}{n} [\alpha^n - (\alpha - 1)^n]. && (106)
\end{aligned}$$

Additive models, pairwise models, and all-order models are special cases of the hierarchical N-order model.

6.2.1 $N = 1$: Additive model

The additive model is a hierarchical 1-order model. Setting $N = 1$, Eq. 102 gives

$$\begin{aligned} M_{\text{additive}} &= M_{\leq 1\text{-order}} && (107)\\ &= \binom{L}{0} + \binom{L}{1}\alpha && (108)\\ &= 1 + L\alpha, && (109) \end{aligned}$$

and Eq. 106 gives,

$$\begin{aligned} \gamma_{\text{additive}} &= \gamma_{\leq 1\text{-order}}, && (110)\\ &= \binom{L}{0}(1-1) + \binom{L}{1}(\alpha - (\alpha - 1)), && (111)\\ &= L. && (112) \end{aligned}$$

6.2.2 $N = 2$: Pairwise model

The pairwise model is a hierarchical 2-order model. Setting $N = 2$, Eq. 102 gives

$$\begin{aligned} M_{\text{pairwise}} &= M_{\leq 2\text{-order}} && (113)\\ &= \binom{L}{0} + \binom{L}{1}\alpha + \binom{L}{2}\alpha^2 && (114)\\ &= 1 + L\alpha + \binom{L}{2}\alpha^2, && (115) \end{aligned}$$

and Eq. 106 gives

$$\begin{aligned} \gamma_{\text{pairwise}} &= \gamma_{\leq 2\text{-order}} && (116)\\ &= \binom{L}{0}(1-1) + \binom{L}{1}(\alpha - (\alpha - 1)) + \binom{L}{2}(\alpha^2 - (\alpha - 1)^2) && (117)\\ &= L + \binom{L}{2}(2\alpha - 1). && (118) \end{aligned}$$

6.2.3 $N = L$: All-order model

The all-order model is a hierarchical L-order model. Setting $N = L$, Eq. 102 gives

$$\begin{aligned} M_{\text{all-order}} &= M_{\leq L\text{-order}} && (119)\\ &= \sum_{n=0}^{L} \binom{L}{n}\alpha^n && (120)\\ &= (\alpha + 1)^L, && (121) \end{aligned}$$

and Eq. 106 gives,

$$\begin{aligned} \gamma_{\text{all-order}} &= \gamma_{\leq L\text{-order}}, && (122)\\ &= \sum_{n=0}^{L} \binom{L}{n} \left[\alpha^n - (\alpha - 1)^n\right], && (123)\\ &= (\alpha + 1)^L - \alpha^L. && (124) \end{aligned}$$

6.3 N-adjacent model

Here we compute the number of parameters and gauge freedoms for the N-adjacent model. To facilitate computations, we define $\operatorname{span} U = \max(U) - \min(U) + 1$ for any nonempty set $U \subseteq \mathcal{I}_l$, as well as $\operatorname{span} \emptyset = 0$. We also define the 0-adjacent model to be the constant model, i.e., the model that corresponds to the embedding $\vec{x}_{\text{constant}} = \vec{x}^{\,\text{triv}}$, which has $M_{\text{constant}} = 1$ parameters and $\gamma_{\text{constant}} = 0$ gauge freedoms.

For $N \geq 1$, we begin by computing the degree of the embedding $\vec{x}_{N\text{-adjacent}}$. The position sets that define $\vec{x}_{N\text{-adjacent}}$ are

$$\{A_j\}_{j=1}^{J} = \{U : U \subseteq I_L,\ |U| = N,\ \text{span}\,(U) = N\} \tag{125}$$

$$= \{U_l : l \in \mathcal{I}_{L-N+1}\},\ \text{ where }\ U_l = \{l, \ldots, l+N-1\}. \tag{126}$$

From this we see that there are $J = L - N + 1$ distinct sets A_j, and that $|A_j| = N$ for every $j \in \mathcal{I}_J$. By Main Text Eq. 30, the dimension of $\vec{x}_{N\text{-adjacent}}$ is therefore

$$\dim \vec{x}_{N\text{-adjacent}} = \sum_{j=1}^{J} \alpha^{|A_j|} \tag{127}$$

$$= (L - N + 1)\alpha^N. \tag{128}$$

Next we compute the dimension of the distilled embedding $\vec{x}^{\,\text{dist}}_{N\text{-adjacent}}$. The position sets that define $\vec{x}^{\,\text{dist}}_{N\text{-adjacent}}$ are

$$\{B_k\}_{k=1}^{K} = \{V : V \subseteq A_j,\ j \in \mathcal{I}_J\} \tag{129}$$

$$= \{V : V \subseteq U,\ U \subseteq I_L,\ |U| = N,\ \text{span}\,(U) = N\} \tag{130}$$

$$= \{V : V \subseteq I_L,\ |V| \leq N,\ \text{span}\,(V) \leq N\} \tag{131}$$

$$= \bigcup_{p=0}^{N} \bigcup_{n=0}^{p} \{V : V \subseteq I_L,\ |V| = n,\ \text{span}\,(V) = p\}, \tag{132}$$

where the unions are over disjoint sets, and we have used the fact that $|V| \leq \text{span}\,(V)$ for all $V \subseteq \mathcal{I}_L$. By Main Text Eq. 32, the dimension of $\vec{x}^{\,\text{dist}}$ is therefore

$$\dim \vec{x}^{\,\text{dist}}_{N\text{-adjacent}} = \sum_{k=1}^{K} (\alpha - 1)^{|B_k|} \tag{133}$$

$$= \sum_{p=0}^{N} \sum_{n=0}^{p} z(n,p)(\alpha - 1)^n, \tag{134}$$

where $z(n,p)$ is the number of subsets of $\mathcal{I}_L$ that have n elements and span p.

- For $n = 0$, only $V = \emptyset$ has span $p = 0$. Therefore $z(0,0) = 1$, and $z(0,p) = 0$ for all $p \neq 0$.
- For $n = 1$, there are L subsets: $U = \{l\}$ for every $l \in \mathcal{I}_L$. Each of these subsets has span $p = 1$. Therefore $z(1,1) = L$ and $z(1,p) = 0$ for all $p \neq 1$.
- For every $n \in \{2, \ldots, N\}$, there is at least one subset U having span p for every $p \in \{2, \ldots, N\}$. Given the endpoints of U, there are $\binom{p-2}{n-2}$ ways to choose the elements of U. And given p, there are $L - p + 1$ possible choices for the endpoints of U. Therefore, $z(n,p) = (L - p + 1)\binom{p-2}{n-2}$.

Continuing from Eq. 134,

$$\dim \vec{x}^{\,\text{dist}}_{N\text{-adjacent}} = 1 + L(\alpha - 1) + \sum_{p=2}^{N} \sum_{n=2}^{N} \binom{p-2}{n-2} (L - p + 1)(\alpha - 1)^n \tag{135}$$

$$= 1 + L(\alpha - 1) + (\alpha - 1)^2 \sum_{p=2}^{N} (L - p + 1)\alpha^{p-2} \tag{136}$$

$$= 1 + L(\alpha - 1) + (\alpha - 1)^2 (L - 1) \sum_{m=0}^{N-2} \alpha^m - (\alpha - 1)^2 \sum_{m=0}^{N-2} m\alpha^m. \tag{137}$$

Setting $M = N - 2$, and using

$$\sum_{m=0}^{M} \alpha^m = \frac{\alpha^{M+1} - 1}{\alpha - 1}, \tag{138}$$

and

$$\sum_{m=0}^{M} m\alpha^m = \alpha \frac{d}{d\alpha} \sum_{m=0}^{M} \alpha^m = \frac{\alpha}{(\alpha - 1)^2} \left[1 - \alpha^{M+1} + (M+1)(\alpha - 1)\alpha^M\right], \tag{139}$$

we get (after a remarkable simplification),

$$\dim \vec{x}^{\,\text{dist}}_{N\text{-adjacent}} = (L-N+1)\alpha^N - (L-N)\alpha^{N-1}. \tag{140}$$

The number of parameters, $M_{N\text{-adjacent}}$, and the number of gauge freedoms, $\gamma_{N\text{-adjacent}}$, is (for $N \geq 1$) therefore given by

$$\begin{aligned} M_{N\text{-adjacent}} &= \dim \vec{x}_{N\text{-adjacent}}, && (141)\\ &= (L-N+1)\alpha^N, && (142)\\ \gamma_{N\text{-adjacent}} &= \dim \vec{x}_{N\text{-adjacent}} - \dim \vec{x}^{\,\text{dist}}_{N\text{-adjacent}}, && (143)\\ &= (L-N)\alpha^{N-1}. && (144) \end{aligned}$$

Note the curious fact that the number of gauge freedoms for the N-adjacent model is equal to the number of parameters for the $(N-1)$-adjacent model, i.e.

$$\gamma_{N\text{-adjacent}} = M_{(N-1)\text{-adjacent}}. \tag{145}$$

6.4 Hierarchical N-adjacent models

We define the hierarchical N-adjacent model to be the sum of n-adjacent terms taken over all $n \in \{0, 1, \ldots, N\}$. The embedding of the hierarchical N-adjacent model is

$$\vec{x}_{\leq N\text{-adjacent}} = \bigoplus_{n=0}^{N} \vec{x}_{n\text{-adjacent}}. \tag{146}$$

The number of parameters of the hierarchical N-adjacent model is therefore,

$$\begin{aligned} M_{\leq N\text{-adjacent}} &= \sum_{n=0}^{N} M_{n\text{-adjacent}}, && (147)\\ &= 1 + \sum_{n=1}^{N} (L-n+1)\alpha^n, && (148)\\ &= 1 + \frac{\alpha}{(\alpha-1)^2}\left[(L-N+1)\alpha^{N+1} - (L-N)\alpha^N - (L+1)\alpha + L\right]. && (149) \end{aligned}$$

Since every parameter of the models of order $N-1$ and lower is redundant with parameters of the model of order N, the number of gauge freedoms of the hierarchical N-adjacent model,

$$\begin{aligned} \gamma_{\leq N\text{-adjacant}} &= \dim \vec{x}_{\leq N\text{-adjacent}} - \dim \vec{x}^{\,\text{dist}}_{\leq N\text{-adjacent}} && (150)\\ &= M_{\leq N\text{-adjacent}} - \dim \vec{x}^{\,\text{dist}}_{N\text{-adjacent}} && (151)\\ &= 1 + \frac{\alpha}{(\alpha-1)^2}\left[(L-N+1)\alpha^{N+1} - (L-N)\alpha^N - (L+1)\alpha + L\right] && (152)\\ &\quad -(L-N+1)\alpha^N + (L-N)\alpha^{N-1}, && (153)\\ &= (L-N)\alpha^{N-1} + 1 + \frac{\alpha}{(\alpha-1)^2}\left[(L-N+2)\alpha^N - (L-N+1)\alpha^{N-1} - (L+1)\alpha + L\right]. && (154) \end{aligned}$$

Additive models, nearest-neighbor models, and all-order models are special cases of the hierarchical N-adjacent model.

6.4.1 $N = 1$: Additive model

The additive model is a hierarchical 1-adjacent odel. Setting $N = 1$, Eq. 149 gives

$$\begin{aligned} M_{\text{additive}} &= M_{\leq 1\text{-adjacent}} && (155)\\ &= 1 + \frac{\alpha}{(\alpha-1)^2}\left[L\alpha^2 - (L-1)\alpha - (L+1)\alpha + L\right] && (156)\\ &= 1 + L\alpha, && (157) \end{aligned}$$

and Eq. 154 gives

$$\begin{aligned} \gamma_{\text{additive}} &= \gamma_{\leq 1\text{-adjacent}} && (158)\\ &= (L-1) + 1 + \frac{\alpha}{(\alpha-1)^2}\left[(L+1)\alpha - L - (L+1)\alpha + L\right] && (159)\\ &= L. && (160) \end{aligned}$$

6.4.2 $N = 2$: Nearest-neighbor model

The nearest-neighbor model is a hierarchical 2-adjacent model. Setting $N = 2$, Eq. 149 gives

$$\begin{aligned} M_{\text{neighbor}} &= M_{\leq 2\text{-adjacent}} && (161)\\ &= 1 + \frac{\alpha}{(\alpha-1)^2}\left[(L-1)\alpha^3 - (L-2)\alpha^2 - (L+1)\alpha + L\right] && (162)\\ &= 1 + L\alpha + (L-1)\alpha^2, && (163)\end{aligned}$$

and Eq. 154 gives

$$\begin{aligned} \gamma_{\text{neighbor}} &= \gamma_{\leq 2\text{-adjacent}} && (164)\\ &= (L-2)\alpha + 1 + \frac{\alpha}{(\alpha-1)}\left[L\alpha^2 - (L-1)\alpha - (L+1)\alpha + L\right] && (165)\\ &= 1 + 2(L-1)\alpha. && (166)\end{aligned}$$

6.4.3 $N = L$: All-adjacent model

The all-order adjacent model is a hierarchical L-adjacent model. Setting $N = L$, Eq. 149 gives

$$\begin{aligned} M_{\text{all-adjacent}} &= M_{\leq L\text{-adjacent}} && (167)\\ &= 1 + \frac{\alpha}{(\alpha-1)^2}\left[\alpha^{L+1} - (L+1)\alpha + L\right], && (168)\end{aligned}$$

and Eq. 154 gives,

$$\begin{aligned} \gamma_{\text{all-adjacent}} &= \gamma_{\leq L\text{-adjacent}} && (169)\\ &= 1 + \frac{\alpha}{(\alpha-1)^2}\left[2\alpha^{L} - \alpha^{L-1} - (L+1)\alpha + L\right]. && (170)\end{aligned}$$

7 Results for other symmetry groups: H_{GCP}, H_{PP}, and H_{PSCP}

7.1 Embedding distillation does not work for H_{GCP} or H_{PP}

Embedding distillation does not work for H_{GCP} (the group of global character permutations) or for H_{PP} (the group of position permutations). The reason is that Claim 5.1 (Theorem 1 of Main Text), does not hold for either of these groups: two embeddings that transform under the same representation are not necessarily proportional (or, in fact, related by any similarity transformation). Consider, for example, the following one-dimensional embeddings of a length L sequence:

$$\vec{x}_1(s) = \begin{cases} [1] \text{ if } s_l = s'_l \text{ for all } l, l' \in \mathcal{I}_L \\ [0] \text{ otherwise} \end{cases}, \quad (171)$$

$$\vec{x}_2(s) = \begin{cases} [0] \text{ if } s_l = s'_l \text{ for all } l, l' \in \mathcal{I}_L \\ [1] \text{ otherwise} \end{cases}. \quad (172)$$

Both $\vec{x}_1$ and $\vec{x}_2$ are invariant under H_{GCP} and H_{PP}, and therefore transform under the trivial representation of both groups. However, $\vec{x}_1$ and $\vec{x}_2$ are not related by any similarity transformation, as no similarity transformation can map a zero vector to a nonzero vector, thus showing that Claim 5.1 does not hold for $H_{\text{GCP}} or H_{\text{PP}}$. Consequently, distillation cannot be used to determine the gauge freedoms of either embedding, and the number of gauge freedoms cannot be computed by summing the degrees of the redundant irreducible representations under which these embeddings transform.

To see explicitly how embedding distillation fails for H_{GCP} and for H_{PP}, consider the additive embedding:

$$\vec{x}_{\text{additive}} = \vec{x}^{\,\text{triv}} \oplus \bigoplus_{l=1}^{L} \vec{x}_l^{\,\text{ohe}}. \quad (173)$$

Under H_{GCP}, each $\vec{x}_l^{\,\text{ohe}}$ transforms under the same representation: $R^{\text{ohe}}_{(\alpha)}$, the one-hot representation of S_α (we indicate the degree of the representation in the subscript for clarity). Using $R^{\text{ohe}}_{(\alpha)} \simeq R^{\text{triv}} \oplus R^{\text{sim}}_{(\alpha-1)}$, where $R^{\text{sim}}_{(\alpha-1)}$ is the simplex representation of S_α, we find that $\vec{x}_{\text{additive}}$ transforms under H_{GCP} according to the representation

$$\begin{aligned} R^{\text{additive}}_{\text{GCP}} &= R^{\text{triv}} \oplus [I_L \otimes R^{\text{ohe}}_{(\alpha)}], && (174)\\ &\simeq R^{\text{triv}} \oplus L R^{\text{ohe}}_{(\alpha)}, && (175)\\ &\simeq R^{\text{triv}} \oplus L[R^{\text{triv}} \oplus R^{\text{sim}}_{(\alpha-1)}], && (176)\\ &\simeq [L+1]R^{\text{triv}} \oplus L R^{\text{sim}}_{(\alpha-1)}, && (177)\\ &\simeq R^{\text{dist}}_{\text{GCP}} \oplus R^{\text{redun}}_{\text{GCP}}, && (178)\end{aligned}$$

where the distilled representation is

$$R^{\text{dist}}_{\text{GCP}} = R^{\text{triv}} \oplus R^{\text{sim}}_{(\alpha-1)}, \tag{179}$$

and the redundant irreducible representations are collected into

$$R^{\text{redun}}_{\text{GCP}} = LR^{\text{triv}} \oplus (L-1)R^{\text{sim}}_{(\alpha-1)}. \tag{180}$$

We thus see that

$$\deg R^{\text{redun}}_{\text{GCP}} = \alpha L - \alpha + 1. \tag{181}$$

This does not match the number of gauge freedoms, $\gamma_{\text{additive}} = L$. The reason is that each copy of $R^{\text{sim}}_{(\alpha-1)}$ transforms a simplex embedding of $\mathcal{A}$, one for each position in $\mathcal{I}_L$, and these embeddings are not equivalent. Similarly, H_{PP} acts on $\vec{x}_{\text{additive}}$ via the representation

$$\begin{aligned} R^{\text{additive}}_{\text{PP}} &= R^{\text{triv}} \oplus [R^{\text{ohe}}_{(L)} \otimes I_\alpha], && (182)\\ &= R^{\text{triv}} \oplus \alpha R^{\text{ohe}}_{(L)}, && (183)\\ &\simeq R^{\text{triv}} \oplus \alpha[R^{\text{triv}} \oplus R^{\text{sim}}_{(L-1)}], && (184)\\ &\simeq [1+\alpha]R^{\text{triv}} \oplus \alpha R^{\text{sim}}_{(L-1)}, && (185)\\ &\simeq R^{\text{dist}}_{\text{PP}} \oplus R^{\text{redun}}_{\text{PP}}, && (186) \end{aligned}$$

where the distilled representation is

$$R^{\text{dist}} = R^{\text{triv}} \oplus R^{\text{sim}}_{(L-1)}, \tag{187}$$

and the redundant irreducible representations are collected into

$$R^{\text{redun}}_{\text{PP}} = \alpha R^{\text{triv}} \oplus (\alpha-1)R^{\text{sim}}_{(L-1)}. \tag{188}$$

We thus see that

$$\deg R^{\text{redun}}_{\text{PP}} = \alpha L - L + 1. \tag{189}$$

This does not match the number of gauge freedoms $\gamma_{\text{additive}} = L$. The reason is that each copy of $R^{\text{sim}}_{(L-1)}$ transforms a simplex embedding of $\mathcal{I}_L$, one for each character in $\mathcal{A}$, and these embeddings are not equivalent.

7.2 Embedding distillation does work for H_{Ham}

Our embedding distillation procedure does work for the symmetry group of the Hamming graph. This group, which we denote by H_{Ham}, is given by the semidirect product

$$H_{\text{Ham}} = H_{\text{PSCP}} \rtimes H_{\text{PP}}. \tag{190}$$

From this definition we see that H_{PSCP} is a normal subgroup of H_{Ham}.

Claim 7.1. *Nonzero sequence embeddings that transform under the same irreducible representation of H_{Ham} are equivalent.*

Proof. Assume $\vec{x}$ and $\vec{y}$ are two nonzero embeddings that transform under the same irreducible representation R of H_{Ham}. Let R' denote R restricted to H_{PSCP}. By Maschke's theorem, there is a similarity transformation matrix T_{dist} such that

$$R' = T\left[\bigoplus_{k=1}^{K} R'_k\right]T^{-1}, \tag{191}$$

where, for each $k = 1, \ldots, K$, R'_k is an irreducible representation of H_{PSCP}. Consequently,

$$\vec{x} = T\left[\bigoplus_{k=1}^{K} \vec{x}_k\right], \quad \text{and} \quad \vec{y} = T\left[\bigoplus_{k=1}^{K} \vec{y}_k\right], \tag{192}$$

where, for each $k \in \mathcal{I}_K$, both $\vec{x}_k$ and $\vec{y}_k$ transform under R'_k. We now show that all $\vec{x}_k$ and $\vec{y}_k$ are nonzero. Let $m_k = \deg R'_k$, $\mathcal{K} = \left\{k : k \in \mathcal{I}_K,\ \vec{x}_k = \vec{0}\right\}$, and define the vector space

$$V = T^{-1\dagger}\left[\bigoplus_{k=1}^{K} \begin{cases} \mathbb{R}^{m_k} & \text{if } k \in \mathcal{K}, \\ \left\{\vec{0}\right\} & \text{otherwise} \end{cases}\right]. \tag{193}$$

It is readily see that V is orthogonal to span $\vec{x}$, i.e., $\vec{v}^\dagger \vec{x}(s) = 0$ for all $v \in V$ and all $s \in \mathcal{S}$. But we know from Claim 3.3 that, because $\vec{x}$ transforms under an irreducible representation of H_{Ham}, span $\vec{x}$ has full rank. Consequently, $\mathcal{K} = \emptyset$, and thus all $\vec{x}_k$ are nonzero. Similarly, all $\vec{y}_k$ are nonzero. Since both $\vec{x}_k$ and $\vec{y}_k$ are nonzero embeddings that transform under the same irreducible representation of H_{PSCP}, Claim 5.1 tells us that $\vec{y}_k = c_k \vec{x}_k$ for some nonzero scalar $c_k \in \mathbb{R}$. Therefore,

$$\vec{y} = TCT^{-1}\vec{x} \quad \text{where} \quad C = \bigoplus_{k=1}^{K} c_k I_{m_k}. \tag{194}$$

This shows that $\vec{x}$ and $\vec{y}$ are related by a similarity transformation, thereby completing the proof. □

8 Distillation algorithm

We now describe the embedding distillation algorithm discussed in the section "Computational analysis of models" of Main Text and illustrated in main text Fig. 4. Python code implementing this algorithm is available at https://github.com/jbkinney/23_posfai.

8.1 Overview of the algorithm

Let $\vec{x}$ be an equivariant one-hot embedding of sequences as defined in Main Text. Specifically, define

$$\vec{x} = \bigoplus_{j=1}^{J} \bigotimes_{l \in A_j} \vec{x}_l^{\,\text{ohe}}, \tag{195}$$

where each A_j, $j \in \mathcal{I}_J$, denotes an ordered set of sequence positions, and $A = (A_j)_{j=1}^{J}$ denotes an ordered set of position subsets the sets of positions used to construct the embedding. Distillation consists of multiplying $\vec{x}$ by a "distillation matrix", T_{dist}, such that

$$T_{\text{dist}}\, \vec{x} = \vec{x}^{\,\text{dist}} \oplus \vec{0}_\gamma, \tag{196}$$

where $\vec{x}^{\,\text{dist}}$ is the full-rank distilled embedding, $\vec{0}_\gamma$ is a vector of γ zeros, and γ is the number of gauge freedoms of $\vec{x}$. As in Main Text, we construct T_{dist} by expressing it as

$$T_{\text{dist}} = T_{\text{sort}}\, T_{\text{thin}}\, T_{\text{decom}}. \tag{197}$$

We now describe the operations each of the three factors above perform:

1. T_{decom} decomposes each Kronecker product of $\vec{x}_l^{\,\text{ohe}}$ into a direct sum of Kronecker products of $\vec{x}_l^{\,\text{sim}}$. Specifically, multiplying Eq. 195 by T_{decom} yields,

$$T_{\text{decom}}\, \vec{x} \quad = \quad \bigoplus_{p=1}^{P} \vec{x}_{k_p}, \tag{198}$$

where each $\vec{x}_k$ is an irreducible representation given by

$$\vec{x}_k = \bigotimes_{l \in B_k} \vec{x}_l^{\,\text{sim}}, \tag{199}$$

each B_k is a distinct ordered set of positions (in which positions are listed in increasing order), $B = (B_k)_{k=1}^{K}$ is an ordered set comprising all unique subsets of positions found within all the A_j, and $(k_p)_{p=1}^{P}$ is an ordered set of (generally non-unique) indices within $\mathcal{I}_K$.

2. T_{thin} thins the resulting embedding by zeroing-out all redundant copies of each irreducible embedding. Specifically, multiplying Eq. 198 by T_{thin} yields

$$T_{\text{thin}}\, T_{\text{decom}}\, \vec{x} = \bigoplus_{p=1}^{P} \begin{cases} \vec{x}_{k_p} & \text{if } k_p \neq k_{p'} \text{ for any } p' < p \\ \vec{0}_{m_{k_p}} & \text{otherwise} \end{cases}, \tag{200}$$

where each $m_k = (\alpha - 1)^{|B_k|}$ is the degree of $\vec{x}_k$.

3. T_{sort} sorts the embeddings in the direct sum so that all nonzero embeddings come first. Specifically, multiplying Eq. 200 by T_{sort} yields

$$T_{\text{sort}}\, T_{\text{thin}}\, T_{\text{decom}}\, \vec{x} = \vec{x}^{\,\text{dist}} \oplus \vec{0}_{\gamma}, \tag{201}$$

where

$$\vec{x}^{\,\text{dist}} = \bigoplus_{k=1}^{K} \vec{x}_k. \tag{202}$$

We now show how to compute each of these three matrix factors. We find that the resulting distillation matrix T_{dist} is sparse (because T_{decom}, T_{thin}, and T_{sort} are sparse) and that all nonzero entries of T_{dist} are either $+1$ or -1. We also show how to compute the inverse of each factor without having to invert any matrices, thereby allowing the rapid computation of

$$T_{\text{dist}}^{-1} = T_{\text{decom}}^{-1}\, T_{\text{thin}}^{-1}\, T_{\text{sort}}^{-1}. \tag{203}$$

Note: In what follows we index vector and matrix elements starting from zero. This differs from the one-indexing used in Main Text and elsewhere in the supplement. We adopt this indexing change here to aid coding efforts.

8.2 Computation of the decomposition matrix T_{decom}.

We build the decomposition matrix, T_{decom}, as a direct sum of individual Nth order decomposition matrices, $T_{\text{decom}}^{(N)}$. Specifically,

$$T_{\text{decom}} = \bigoplus_{j=1}^{J} T_{\text{decom}}^{(N_j)}, \tag{204}$$

where $N_j = |A_j|$ for all $j \in \mathcal{I}_J$. Given N we define $T_{\text{decom}}^{(N)}$ so that

$$T_{\text{decom}}^{(N)} \left[\bigotimes_{n=1}^{N} \vec{x}_{l_n}^{\,\text{ohe}} \right] = \bigoplus_{B \subseteq A} \left[\bigotimes_{l \in B} \vec{x}_{l}^{\,\text{sim}} \right] \tag{205}$$

for any subset of positions A that has N elements. We computationally construct $T_{\text{decom}}^{(N)}$ by recursion using

$$T_{\text{decom}}^{(N)} = T_{\text{perm}}^{(N)} \left[T_{\text{decom}}^{(N-1)} \otimes T \right] \tag{206}$$

where T is the single-position decomposition matrix. Using this recursion relation requires defining T and $T_{\text{perm}}^{(N)}$:

- T is an $\alpha \times \alpha$ matrix having elements

$$[T]_{ij} = \begin{cases} 1 & \text{if } i = 0, \\ -1 & \text{if } 0 < i < \alpha,\ j = \alpha - 1, \\ 1 & \text{if } 0 < i < \alpha,\ j = i - 1, \\ 0 & \text{otherwise,} \end{cases} \tag{207}$$

  for $i, j \in \{0, \ldots, \alpha - 1\}$. Note that T matches Eq. 20 when $\alpha = 4$.

- $T_{\text{perm}}^{(N)}$ is an $\alpha^N \times \alpha^N$ permutation matrix having elements

$$\left[T_{\text{perm}}^{(N)} \right]_{ij} = \begin{cases} 1 & \text{if } 0 \leq i < \alpha^{N-1} \text{ and } \ j = \alpha i, \\ 1 & \text{if } \alpha^{N-1} \leq i < \alpha^N \text{ and } j = 1 + (i - \alpha^{N-1}) + \lfloor (i - \alpha^{N-1})/(\alpha - 1) \rfloor, \\ 0 & \text{otherwise.} \end{cases} \tag{208}$$

  for $i, j \in \left\{0, \ldots, \alpha^N - 1\right\}$. T_{perm} needed because Kronecker products do not distribute over direct sums. Specifically, for a general vector $\vec{v}$ having α^N elements,

$$\vec{v} \otimes \left(\vec{x}^{\,\text{triv}} \oplus \vec{x}_{l}^{\,\text{sim}} \right) \neq \left(\vec{v} \otimes \vec{x}^{\,\text{triv}} \right) \oplus \left(\vec{v} \otimes \vec{x}_{l}^{\,\text{sim}} \right). \tag{209}$$

  The matrix $T_{\text{perm}}^{(N)}$ fixes this inequality, i.e.,

$$T_{\text{perm}} \left[\vec{v} \otimes \left(\vec{x}^{\,\text{triv}} \oplus \vec{x}_{l}^{\,\text{sim}} \right) \right] = \left(\vec{v} \otimes \vec{x}^{\,\text{triv}} \right) \oplus \left(\vec{v} \otimes \vec{x}_{l}^{\,\text{sim}} \right). \tag{210}$$

Similarly, we build the inverse of the decomposition matrix, T_{decom}^{-1}, as a direct sum of N'th order inverse decomposition matrices $\left(T_{\text{decom}}^{(N)}\right)^{-1}$, which are computed by recursion using

$$\left(T_{\text{decom}}^{(N)}\right)^{-1} = \left[\left(T_{\text{decom}}^{(N-1)}\right)^{-1} \otimes T^{-1}\right] \left(T_{\text{perm}}^{(N)}\right)^{-1}, \tag{211}$$

with

$$\left[T^{-1}\right]_{ij} = \frac{1}{\alpha} \times \begin{cases} 1 & \text{if } j = 0, \\ (\alpha - 1) & \text{if } 0 < j < \alpha,\ i = j - 1, \\ -1 & \text{otherwise,} \end{cases} \tag{212}$$

for $i, j \in \{0, \ldots, \alpha - 1\}$. Note that T^{-1} matches Eq. 20 when $\alpha = 4$. Also note that, because $T_{\text{perm}}^{(N)}$ is a permutation matrix,

$$\left(T_{\text{perm}}^{(N)}\right)^{-1} = \left(T_{\text{perm}}^{(N)}\right)^{\top}. \tag{213}$$

8.3 Computation of the thinning matrix T_{thin}

For each $p \in \mathcal{I}_P$, define the index offset function

$$\text{offset}(p) = \sum_{p' < p} m_{k_{p'}}, \tag{214}$$

and the first occurence function

$$\text{first}(p) = \min\left\{p' : k'_p = k_p\right\}. \tag{215}$$

The thinning matrix, T_{thin}, is an $M \times M$ having elements

$$\left[T_{\text{thin}}\right]_{ij} = \begin{cases} 1 & \text{if } i = j, \\ -1 & \text{if } i \neq j,\ i = \text{offset}(p) + h, \text{ and } j = \text{offset}(\text{first}(p)) + h \text{ for some } p \in \mathcal{I}_P,\ h \in \left\{0, \ldots, m_{k_p} - 1\right\}, \\ 0 & \text{otherwise,} \end{cases} \tag{216}$$

for $i, j \in \{0, \ldots, M - 1\}$. One can readily verify that the inverse of the thinning matrix thus has elements

$$\left[T_{\text{thin}}^{-1}\right]_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 1 & \text{if } i \neq j,\ i = \text{offset}(p) + h, \text{ and } j = \text{offset}(\text{first}(p)) + h \text{ for some } p \in \mathcal{I}_P,\ h \in \left\{0, \ldots, m_{k_p} - 1\right\}. \\ 0 & \text{otherwise.} \end{cases} \tag{217}$$

8.4 Computation of the sorting matrix T_{sort}

We define the sorted index offset function as

$$\text{sortedoffset}(p) \quad = \quad \begin{cases} \text{offest}_1(p) & \text{if } p = \text{first}(p), \\ \text{offest}_2(p) & \text{if } p \neq \text{first}(p). \end{cases} \tag{218}$$

where

$$\text{offset}_1(p) \quad = \quad \sum_{\substack{p' < p: \\ p' = \text{first}(p')}} m_{k_{p'}}, \tag{219}$$

$$\text{offset}_2(p) \quad = \quad \sum_{\substack{p' \leq P: \\ p' = \text{first}(p')}} m_{k_{p'}} + \sum_{\substack{p' < p: \\ p' \neq \text{first}(p')}} m_{k_{p'}}. \tag{220}$$

The sorting matrix, T_{sort}, is an $M \times M$ matrix having elements

$$\left[T_{\text{sort}}\right]_{ij} = \begin{cases} 1 & \text{if } i = \text{sortedoffset}(p) + h \text{ and } j = \text{offset}(p) + h \text{ for some } p \in \mathcal{I}_P \text{ and } h \in \mathcal{I}_{m_{k_p}} - 1, \\ 0 & \text{otherwise,} \end{cases} \tag{221}$$

for $i, j \in \{0, \ldots, M - 1\}$. It is readily see that T_{sort} is a permutation matrix. Consequently, the inverse of T_{sort} is given by its transpose, i.e.,

$$T_{\text{sort}}^{-1} = T_{\text{sort}}^{\top}. \tag{222}$$

8.5 Computation of the projection matrix P

Let Θ denote the gauge space spanned by the first $M-\gamma$ columns of $T^{\dagger}_{\text{dist}}$. It is readily seen that the resulting gauge-fixing projection matrix for Θ is given by

$$P = T^{\dagger}_{\text{dist}}\Big|_{M-\gamma} T^{-1\dagger}_{\text{dist}}, \tag{223}$$

where $|_{M-\gamma}$ denotes that the last γ columns of a matrix have been set to zero. We note, however, that the gauge space Θ is not one of the parametric gauges discussed in our companion paper [1].

9 Observations motivating the conjecture

Based on the observations in the next two subsections, we conjecture that all allelic permutation models either (1) do not describe co-occurring alleles, (2) are equivalent to models that do not describe co-occurring alleles, or (3) have gauge freedoms.

9.1 Single-orbit allelic models

Claim 9.1. *Single-orbit generalized one-hot models cannot describe co-occurring alleles.*

Proof. Every nontrivial orbit of a generalized one-hot model is defined by a set of generalized one-hot features,

$$\mathcal{O} = \{x^{c_1\ldots c_K}_{l_1\ldots l_K} : c_1, \ldots, c_K \in \mathcal{A}\}, \tag{224}$$

for some specified set of positions $\{l_1, \ldots, l_K\} \subseteq \mathcal{I}_L$, where each feature in the orbit $\mathcal{O}$ is given by

$$x^{c_1\ldots c_K}_{l_1\ldots l_K}(s) = \begin{cases} 1 & \text{if } c_k = s_{l_k} \text{ for } k = 1, \ldots, K, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for all} \quad s \in \mathcal{S}. \tag{225}$$

For any $s \in \mathcal{S}$, choosing $c_k = s_{l_k}$ for all $k = 1, \ldots, K$ yields a feature $x^{c_1\ldots c_K}_{l_1\ldots l_K} \in \mathcal{O}$ that is equal to 1 on s. Moreover, no other feature in $\mathcal{O}$ is nonzero when evaluated on s because this would require that s have a different character at one of the positions l_k. Therefore, for every sequence $s \in \mathcal{S}$, there is exactly one feature in the orbit $\mathcal{O}$ that is nonzero on s. This proves the claim. □

There are, however, are single-orbit allelic models that are not generalized one-hot models that describe co-occurring alleles. One example is the model based on the "two-hot" DNA embedding, in which each feature tests for the presence of one of two possible DNA bases:

$$\vec{x}^{\,\text{two}} = \begin{bmatrix} x^{\{\text{A,C}\}} \\ x^{\{\text{A,G}\}} \\ x^{\{\text{A,T}\}} \\ x^{\{\text{C,G}\}} \\ x^{\{\text{C,T}\}} \\ x^{\{\text{G,T}\}} \end{bmatrix} \quad \text{where} \quad x^{\{c_1,c_2\}}(c) = \begin{cases} 1 & \text{if } c \in \{c_1, c_2\}, \\ 0 & \text{otherwise} \end{cases} \quad \text{for all} \quad c \in \mathcal{A}_{\text{DNA}} \quad \text{and all} \quad \{c_1, c_2\} \subset \mathcal{A}_{\text{DNA}}. \tag{226}$$

This embedding comprises six allelic features. Since each DNA character is matched by three of the features, the embedding describes co-occurring alleles. This embedding is not equivalent to any generalized one-hot embedding, as it has the wrong dimension. In fact, the observation that $\vec{x}^{\,\text{two}}$ has six dimensions, transforms under a permutation representation, and is non-constant implies that it must transform under the representation

$$R^{\text{two}} \simeq 3R^{\text{triv}} \oplus R^{\text{sim}}. \tag{227}$$

The embedding $\vec{x}^{\,\text{two}}$ therefore has two gauge freedoms, which are readily seen to result from the three affine constraints $x^{\{\text{A,C}\}} + x^{\{\text{G,T}\}} = 1$, $x^{\{\text{A,G}\}} + x^{\{\text{C,T}\}} = 1$, and $x^{\{\text{A,T}\}} + x^{\{\text{C,G}\}} = 1$.

Another example of a single-orbit allelic embedding that describes co-occurring alleles is the "three-hot" DNA embedding, in which each feature tests for the presence of one of three possible DNA bases:

$$\vec{x}^{\,\text{three}} = \begin{bmatrix} x^{\{\text{A,C,G}\}} \\ x^{\{\text{A,C,T}\}} \\ x^{\{\text{A,G,T}\}} \\ x^{\{\text{C,G,T}\}} \end{bmatrix} \quad \text{where} \quad x^{\{c_1,c_2,c_3\}}(c) = \begin{cases} 1 & \text{if } c \in \{c_1, c_2, c_3\}, \\ 0 & \text{otherwise} \end{cases} \quad \text{for all} \quad c \in \mathcal{A}_{\text{DNA}} \quad \text{and all} \quad \{c_1, c_2, c_3\} \subset \mathcal{A}_{\text{DNA}}. \tag{228}$$

This embedding comprises four allelic features. Since each DNA character is matched by three of the features, the embedding describes co-occurring alleles. The fact that $\vec{x}^{\,\text{three}}$ has four dimensions, transforms under a permutation representation, and is non-constant implies that it must transform under the representation

$$R^{\text{two}} \simeq R^{\text{triv}} \oplus R^{\text{sim}} \simeq R^{\text{ohe}}. \tag{229}$$

The embedding $\vec{x}^{\,\text{three}}$ therefore has no gauge freedoms. Rather, it is equivalent to the one-hot embedding $\vec{x}^{\,\text{ohe}}$, i.e., equal up to a similarity transformation. Indeed, one can readily verify that

$$\vec{x}^{\,\text{three}} = A\vec{x}^{\,\text{ohe}}, \quad \text{where} \quad A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}. \tag{230}$$

9.2 Models defined by direct sums of Kronecker products of single-position embeddings

Let us consider an embedding of the form

$$\vec{x}(s) = \bigoplus \vec{x}_{l_1 \ldots l_k}, \quad \text{where} \quad \vec{x}_{l_1 \ldots l_k} := \vec{x}_{l_1}(s) \otimes \ldots \otimes \vec{x}_{l_k}(s), \tag{231}$$

the direct sum goes over an arbitrarily chosen set of position sets $\{l_1, \ldots, l_k\}$, and each $\vec{x}_{l_j} = (x^{i_1}_{l_1}, \ldots, x^{i_k}_{l_k})$ is an arbitrary single-position sequence embedding, where each $x^{i_j}_{l_j}$ is a function that only depends on sequence s at position l_j. We assume that none of the single-position features are constant over all characters, i.e. if x^i_l denotes the ith feature in the single-position embedding $\vec{x}_l$ then $x^i_l(c)$, $c \in \mathcal{A}$, is not a constant function. We note that the constant feature can be part of embedding $\vec{x}$, it corresponds to the empty set in the direct sum over position sets in (231).

Let γ denote the number of gauge freedoms of the model that uses embedding $\vec{x}$, and let γ_{goh} denote the number of gauge freedoms of the generalized one-hot model that uses the embedding of the same direct sum/Kronecker product structure as $\vec{x}$, given in (231), but with $\vec{x}_{l_i} = \vec{x}^{\,\text{ohe}}_{l_i}$ for all l_i.

Claim 9.2. *If a model built from embedding $\vec{x}$ given in formula (231) transforms under a permutation representation of H_{PSCP}, then the model has at least as many gauge freedoms as the corresponding one-hot model, i.e.*

$$\gamma \geq \gamma_{\text{goh}}. \tag{232}$$

Proof. Let us assume we have an embedding $\vec{x}$ that satisfies the assumptions above, and let R be the permutation representation under which it transforms. For each h, $R(h)$ is a permutation matrix, and therefore induces a permutation on the set of features which we denote by h'. With this notation, the equivariance-defining equation becomes

$$\vec{x}(h(s)) = h'(\vec{x}(s)), \quad s \in \mathcal{S}, h \in H_{\text{PSCP}}. \tag{233}$$

We first show that, for each $h \in H_{\text{PSCP}}$, h' does not mix features corresponding to different position sets, i.e. h' maps each feature vector $\vec{x}_{l_1 \ldots l_k}$ onto itself. First, if the embedding contains the constant feature x_0, we have $h'(x_0(s)) = x_0(h(s))$ =constant for all s, therefore x_0 is mapped to itself. (We note that we assumed that there is at most one constant feature in the embedding $\vec{x}$.). Next, let us consider how h' acts on single-position features. Let x^i_l denote the ith coordinate of the singe-position embedding $\vec{x}_l$. If x^i_l is mapped to a possibly higher order feature $x^{i_1 \ldots i_k}_{l_1 \ldots l_k}$ ($k \geq 1$), then using the product form $h = h_1 \cdots h_L$, we have

$$x^i_l(h(s)) = x^{i_1 \ldots i_k}_{l_1 \ldots l_k}(s) = x^{i_1}_{l_1}(s) \cdot \ldots \cdot x^{i_k}_{l_k}(s). \tag{234}$$

We see that the left-hand side function only depends on a sequence's character at position l. Therefore, for any position $l_j \neq l$, if we change s at l_j, the left-hand side of the equation above does not change, while the right hand side does. Note that here we use the assumption that none of the single-position features $x^{i_j}_{l_j}(s)$ are constant functions of s. Therefore, we see that, to avoid contradiction, we need to have $k = 1$ and $l_1 = l$, that is, any single-position feature vector $\vec{x}_l(s)$ gets mapped onto itself by h'. Finally, consider how h' acts on higher order features. By the definition of higher order feature vectors as tensor products of single-position feature vectors, and using the product form $h = h_1 \cdots h_L$, we obtain

$$x^{i_1 \ldots i_k}_{l_1 \ldots l_k}(h(s)) = \prod_{j=1}^{k} x^{i_j}_{l_j}(h(s)) = \prod_{j=1}^{k} x^{i_j}_{l_j}(h_{l_j}(s)) = \prod_{j=1}^{k} x^{i'_j}_{l_j}(s) = x^{i'_1 \ldots i'_k}_{l_1 \ldots l_k}(s), \tag{235}$$

where $x^{i'_{l_j}}_{l_j} := h'_{l_j}(x^{i_{l_j}}_{l_j})$. We thus find that $\vec{x}_{l_1 \ldots l_k}$ also gets mapped onto itself by h'.

We conclude that for each $h \in H$, the permutation matrix $R(h)$ is block-diagonal with blocks $R_{l_1 \ldots l_k}(h)$ corresponding to the position sets $l_1 < \ldots < l_k$ in formula (231). Therefore

$$\vec{x}(h(s)) = \bigoplus_{l_1 < \ldots < l_k} \vec{x}_{l_1 \ldots l_k}(h(s)) = \bigoplus_{l_1 < \ldots < l_k} R_{l_1 \ldots l_k}(h) \vec{x}_{l_1 \ldots l_k}(s). \tag{236}$$

Considering a term in the right hand side expression above, we saw in equation (235) that how $R_{l_1 \ldots l_k}(h)$ acts on $\vec{x}_{l_1 \ldots l_k}$ is completely determined by how it acts on single-site feature vectors, specifically, $R_{l_1 \ldots l_k}(h) = R_{l_1}(h) \otimes \ldots \otimes R_{l_k}(h)$, where $R_{l_j}(h)$ is the block of $R(h)$ that acts on the single-position feature vector $\vec{x}_{l_j}$. In summary, we found that

$$R = \bigoplus R_{l_1} \otimes \ldots \otimes R_{l_k}, \tag{237}$$

where each R_{l_i} is permutation representation of S_α.

Let us look at the Maschke decomposition of a factor R_{l_j} in the expression above. A permutation representation is reducible, in particular, its Maschke decomposition always contains the trivial representation. From Claim 4.2, we further know that the Maschke decomposition of R_{l_j} can only contain copies of the trivial representation R^{triv} and copies of the simplex representation R^{sim} with degree $\alpha - 1$. But if it were to contain copies of only the trivial representation, then $\vec{x}_{l_j}$ would be the trivial embedding which would contradict our assumption that none of the first order features are constant over all characters. This means that the Maschke decomposition of R_{l_j} also has to contain a copy of the simplex representation. Therefore the Maschke decomposition of R_{l_j} looks like

$$R_{l_j} \simeq Q^{\text{triv}} R^{\text{triv}} \oplus Q^{\text{sim}} R^{\text{sim}}, \tag{238}$$

where $Q^{\text{triv}} \geq 1$ and $Q^{\text{sim}} \geq 1$.

Since $Q^{\text{triv}} = 1$ and $Q^{\text{sim}} = 1$ in the Maschke decomposition of the corresponding one-hot model, we see that whenever a coordinate of $\vec{x}$ is zeroed out in the embedding distillation procedure, so is the corresponding coordinate of $\vec{x}^{\text{ohe}}$, and therefore, $\gamma \geq \gamma_{\text{goh}}$. □

References

[1] 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.

[2] Sagan BE. (2001) *The Symmetric Group: Representations, Combinatorial Algorithms, and Symmetric Functions. 2nd Ed.* Springer, New York.

[3] Stormo GD. (2011) Maximally efficient modeling of DNA sequence motifs at all levels of complexity. *Genetics*, 187(4):1219–1224. https://doi.org/10.1534/genetics.110.126052.