An extension of the Walsh-Hadamard transform to calculate and model epistasis in genetic landscapes of arbitrary shape and complexity

Andre J. Faure[1*], Ben Lehner[1,2,3,4], Verónica Miró Pina[1], Claudia Serrano Colome[1], and Donate Weghorn[1,2]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[3]Institució Catalana de Recerca i estudis Avançats (ICREA), Barcelona, Spain
[4]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK
[*]Corresponding author email: andre.faure@crg.eu (AJF)

Abstract

Accurate models describing the relationship between genotype and phenotype are necessary in order to understand and predict how mutations to biological sequences affect the fitness and evolution of living organisms. The apparent abundance of epistasis (genetic interactions), both between and within genes, complicates this task and how to build mechanistic models that incorporate epistatic coefficients (genetic interaction terms) is an open question. The Walsh-Hadamard transform represents a rigorous computational framework for calculating and modeling epistatic interactions at the level of individual genotypic values (known as genetical, biological or physiological epistasis), and can therefore be used to address fundamental questions related to sequence-to-function encodings. However, one of its main limitations is that it can only accommodate two alleles (amino acid or nucleotide states) per sequence position. In this paper we provide an extension of the Walsh-Hadamard transform that allows the calculation and modeling of background-averaged epistasis (also known as ensemble epistasis) in genetic landscapes with an arbitrary number of states per position (20 for amino acids, 4 for nucleotides, etc.). We also provide a recursive formula for the inverse matrix and then derive formulae to directly extract any element of either matrix without having to rely on the computationally intensive task of constructing or inverting large matrices. Finally, we demonstrate the utility of our theory by using it to model epistasis within both simulated and empirical multiallelic fitness landscapes, revealing that both pairwise and higher-order genetic interactions are enriched between physically interacting positions.

Author Summary

An important question in genetics is how the effects of mutations combine to alter phenotypes. Genetic interactions (epistasis) describe non-additive effects of pairs of mutations, but can also involve higher-order (three- and four-way etc.) combinations. Quantifying higher-order interactions is experimentally very challenging requiring a large number of measurements. Techniques based on deep mutational scanning (DMS) represent valuable sources of data to study epistasis. However, the best way to extract the relevant pairwise and higher-order epistatic coefficients (genetic interaction terms) from this data for the task of phenotypic prediction remains an unresolved problem. The Walsh-Hadamard transform represents a rigorous computational framework for calculating and modeling epistatic interactions at the level of individual genotypic values. Critically, this formalism currently only allows for two alleles (amino acid or nucleotide states) per sequence position, hampering applications in more biologically realistic scenarios. Here we present an extension of the Walsh-Hadamard transform that overcomes this limitation and demonstrate the utility of our theory by using it to model epistasis within both simulated and empirical multiallelic genetic landscapes.

Introduction

A fundamental challenge in biology is to understand and predict how changes (or mutations) to biological sequences (DNA, RNA, proteins) affect their molecular function and ultimately the phenotype of living organisms. The phenomenon of 'epistasis' (genetic interactions) – broadly defined as the dependence of mutational effects on the genetic context in which they occur [1, 2, 3] – is widespread in biological systems, yet knowledge of the underlying mechanisms remains limited. Defining the extent of epistasis and better understanding of its origins has relevance in fields ranging from genetic prediction, molecular evolution, infectious disease and cancer drug development, to biomolecular structure determination and protein engineering [3].

Evolutionarily related sequences, natural genetic variation within populations, and more recently results of techniques such as deep mutational scanning (DMS) [4] – also known as massively parallel reporter assays (MPRAs) and multiplex assays of variant effect (MAVEs) – represent valuable sources of data to study epistasis [5, 1]. In particular, DMS enables the systematic measurement of mutational effects across entire combinatorially complete genetic landscapes [5, 6, 7, 8, 9, 10, 11, 12, 13]. Importantly, the typical use of engineered genotypes, haploid individuals and near-identical environmental (laboratory) conditions in these experiments allows population genetic considerations – such as dominance, variable allele frequencies and linkage disequilibrium – to be ignored [14]. In other words, measurements obtained from deep mutational scanning and related methods permit the modeling of epistasis in the mechanistic sense (sequence-to-function encoding) rather than in the evolutionary sense, i.e. based on the dynamics of population genotype frequencies. Nevertheless, precisely how to extract the most biologically relevant pairwise and higher-order epistatic coefficients (genetic interaction terms) from this type of data is an unresolved problem.

Quantitative definitions of epistasis vary among fields, but it has been argued that one particular formulation termed 'background-averaged' epistasis, also known as 'ensemble' epistasis [1, 12], may provide the

most useful information on the epistatic structure of biological systems [2]. The underlying rationale is that by averaging the effects of mutations across many different genetic backgrounds (contexts), the method is robust to local idiosyncrasies in the relationship between genotype and phenotype. It has been previously pointed out that the definition of background-averaged epistasis is conceptually similar to that of 'statistical epistasis' attributed to Fisher, but instead of measuring the average effect of allele substitutions against the population average genetic background i.e. averaging over all genotypes present in a given population (taking into account their individual frequencies), the approach instead averages over all possible genotypes (assuming equal genotype weights) [1, 2].

The current mathematical formalism of background-averaged epistasis is based on the Walsh-Hadamard transform [2]. Interestingly, although widely used in physics and engineering, the Walsh-Hadamard transform was first applied to non-biological fitness landscapes in the field of genetic algorithms (GA) [15], subsequently being proposed as the basis of a framework for the computation of higher-order epistasis in empirical settings [16]. However, the Walsh-Hadamard transform can only accommodate two alleles (amino acid or nucleotide states) per sequence position, with no extension to multialleleic landscapes (cardinality greater than two) yet made, as confirmed by multiple recent reports [2, 17, 18, 19]. Alternative implementations for multiallelic landscapes either rely on 'one-hot encoding' elements of larger alphabets as biallelic sequences – requiring the manipulation of prohibitively large Walsh-Hadamard matrices – or constructing graph Fourier bases [18, 20], which is mathematically complex and provides no straightforward way to interpret epistatic coefficients. The result is that the application of background-averaged epistasis has been severely limited and its properties remain largely unexplored in more biologically realistic scenarios.

In this work we provide an extension of the Walsh-Hadamard transform that allows the calculation and modeling of background-averaged epistasis in genetic landscapes with an arbitrary number of states (20 for amino acids, 4 for nucleotides, etc.). We also provide a recursive formula for the inverse matrix, which is required to infer epistatic coefficients using regression. Furthermore, we derive convenient formulae to directly extract any element of either matrix without having to rely on the computationally intensive task of constructing or inverting large matrices. Lastly, we apply these formulae to the analysis of both simulated and empirical multiallelic DMS datasets, demonstrating that sparse models inferred from the background-averaged representation (embedding) of the underlying genetic landscape more regularly include epistatic terms corresponding to direct physical interactions.

Results

Extension of the Walsh-Hadamard transform to multiallelic landscapes

In this work, a genotype sequence is represented as a one-dimensional ordering of monomers, each of which can take on s possible states (or alleles), for example $s = 4$ for nucleotide sequences or $s = 20$ for amino acid sequences. Without loss of generality, the s states can be labelled $0, 1, 2, \ldots, s-1$, where 0 denotes the wild-type allele. We are going to consider genotype sequences of length $n \in \mathbb{N}$, i.e. sequences taking values

in $\mathcal{S}^n$, where $\mathcal{S} := \{0, 1, \dots, s-1\}$.

Each genotype $\vec{i} \in \mathcal{S}^n$ is associated with its phenotype $y_{\vec{i}}$. Note that here we use the term 'phenotype' as shorthand for 'molecular phenotype score' from a quantitative laboratory assay (DMS) reporting on a molecular function for each genotype of interest. In quantitative genetics terminology this might be referred to as 'genotypic value' because environmental deviation is negligible due to the controlled nature of the experiments, but our subject here is the macromolecule not an individual from a population [14]. In the context of empirical genotype-phenotype landscapes, the phenotypic effect of a genotype $\vec{i}$ is typically measured with respect to the wild-type, i.e. it is given by $y_{\vec{i}} - y_{(0,\dots,0)}$.

It is important to emphasize that in what follows we implicitly restrict ourselves to the haploid reference base, because our primary goal is the modeling of sequence-to-function encodings for *individual* genotype sequences – for the ultimate purpose of understanding and engineering macromolecules – not the modeling of sequence evolution or quantification of sources of phenotypic variance in populations.

If the phenotypic effects of individual mutations were independent, they would be additive, meaning that the phenotypic effect of $\vec{i} = (i_1, \dots, i_n)$ would be the sum of the phenotypic effects of the single mutants $(i_1, 0, \dots, 0), \dots, (0, \dots, 0, i_n)$. The epistatic coefficient quantifies how much the observed phenotypic effect of $\vec{i}$ deviates from this assumption. In the case of background-averaged epistasis, we quantify the interactions between a set of mutations by averaging over all possible genotypes for the remaining positions in the sequence. For example, if $n = 3$ and $s = 2$, the pairwise epistatic coefficient involving the mutations at positions 2 and 3 is calculated by averaging over all states (backgrounds) for the remaining positions, in this case given by the two states of the first position ($*$ denotes the positions at which the averaging is performed), i.e.

$$\begin{aligned}
\varepsilon_{(*,1,1)} &= \frac{1}{2}\Big(\big[(y_{(1,1,1)} - y_{(1,0,0)}) - (y_{(1,1,0)} - y_{(1,0,0)}) - (y_{(1,0,1)} - y_{(1,0,0)})\big] + \\
&\qquad \big[(y_{(0,1,1)} - y_{(0,0,0)}) - (y_{(0,1,0)} - y_{(0,0,0)}) - (y_{(0,0,1)} - y_{(0,0,0)})\big]\Big) \\
&= \frac{1}{2}\Big(\big[y_{(1,1,1)} - y_{(1,1,0)} - y_{(1,0,1)} + y_{(1,0,0)}\big] + \big[y_{(0,1,1)} - y_{(0,1,0)} - y_{(0,0,1)} + y_{(0,0,0)}\big]\Big).
\end{aligned} \tag{1}$$

More generally, in [2] it is shown that for $s = 2$ and any sequence length n, phenotypic effects can be decomposed into background-averaged epistatic coefficients with

$$\bar{\varepsilon}_n = \hat{V}_n \cdot \hat{H}_n \cdot \bar{y}_n, \tag{2}$$

where $\bar{y}_n$ is the vector $(y_{\vec{i}}, \vec{i} \in [0,1]^n)$, $\bar{\varepsilon}_n$ is the vector $(\varepsilon_{\vec{j}}, j \in [*,1]^n)$ and $\hat{H}_n$ and $\hat{V}_n$ are $2^n \times 2^n$ matrices defined recursively as follows:

$$\hat{H}_{n+1} = \begin{pmatrix} \hat{H}_n & \hat{H}_n \\ \hat{H}_n & -\hat{H}_n \end{pmatrix} \qquad \hat{H}_0 = 1, \tag{3}$$

$$\hat{V}_{n+1} = \begin{pmatrix} \frac{1}{2}\hat{V}_n & 0 \\ 0 & -\hat{V}_n \end{pmatrix} \qquad \hat{V}_0 = 1. \tag{4}$$

The matrix $\hat{H}$ is known as the Walsh-Hadamard transform [21, 22] and $\hat{V}$ is a diagonal weighting (or normalisation) matrix to correct the sign and account for averaging over different numbers of backgrounds as a function of epistatic order [2].

In this work, we provide an extension of this theory to describe background-averaged epistasis for sequences with an arbitrary number of states s. Before writing a general formula, we consider the simplest possible multi-state (multiallelic) landscape i.e. a sequence of length $n = 1$ with $s = 3$,

$$\begin{pmatrix} \varepsilon_{(*)} \\ \varepsilon_{(1)} \\ \varepsilon_{(2)} \end{pmatrix} = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} y_{(0)} \\ y_{(1)} \\ y_{(2)} \end{pmatrix} := V_1 \cdot H_1 \cdot \bar{y}_1. \tag{5}$$

Consistent with the definition of background-averaged epistasis for biallelic landscapes [2], the zeroth-order epistatic coefficient $\varepsilon_{(*)}$ is the mean phenotypic value across all genotypes and the first-order epistatic coefficients $\varepsilon_{(1)}$ and $\varepsilon_{(2)}$ are simply the respective individual phenotypic effects of genotypes $y_{(1)}$ and $y_{(2)}$ with respect to the wild-type. However, the key feature of H_1 for multiallelic landscapes – and where it departs from the canonical Walsh-Hadamard transform – is the introduction of zero elements to exclude phenotypes that are irrelevant for the calculation of a given epistatic coefficient. In other words, these phenotypes are excluded because they correspond neither to relevant intermediate genotypes nor alternative genetic backgrounds.

If we now consider a sequence of length $n = 2$ with $s = 3$, then the H_2 and V_2 matrices become 9×9 ($s^n \times s^n$) and can be constructed from recurring to the case $n = 1$ above, giving

$$\begin{pmatrix} \varepsilon_{(*,*)} \\ \varepsilon_{(*,1)} \\ \varepsilon_{(*,2)} \\ \varepsilon_{(1,*)} \\ \varepsilon_{(1,1)} \\ \varepsilon_{(1,2)} \\ \varepsilon_{(2,*)} \\ \varepsilon_{(2,1)} \\ \varepsilon_{(2,2)} \end{pmatrix} = \begin{pmatrix} \frac{1}{9} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} y_{(0,0)} \\ y_{(0,1)} \\ y_{(0,2)} \\ y_{(1,0)} \\ y_{(1,1)} \\ y_{(1,2)} \\ y_{(2,0)} \\ y_{(2,1)} \\ y_{(2,2)} \end{pmatrix}$$

$$:= \quad V_2 \cdot H_2 \cdot \bar{y}_2,$$

where the colors highlight the block structure of the matrices. In V_2, the red square corresponds to $\frac{1}{s}V_1$ and

the light red squares to $-V_1$. In H_2, the gray squares correspond to H_1 and the blue squares to $-H_1$. In Table 1 we show the results of background-averaged epistatic coefficients calculated by applying the above formula to an empirical multiallelic landscape with $n = 2$ and $s = 3$ [6].

Nucleic acid sequence	**Base $s = 3$ representation**	**Phenotypic effect** $\bar{y}_2$	**Epistatic term** $\bar{\varepsilon}_2 = V_2 \cdot H_2 \cdot \bar{y}_2$
GC	(0,0)	0 (±0.01)	-0.17 (±0.02)
GA	(0,1)	-0.14 (±0.05)	-0.21 (±0.06)
GT	(0,2)	-0.07 (±0.02)	0.02 (±0.02)
AC	(1,0)	-0.13 (±0.02)	-0.24 (±0.05)
AA	(1,1)	-0.8 (±0.15)	-0.53 (±0.16)
AT	(1,2)	-0.01 (±0.03)	0.19 (±0.04)
TC	(2,0)	-0.19 (±0.02)	-0.05 (±0.03)
TA	(2,1)	0 (±0.04)	0.33 (±0.07)
TT	(2,2)	-0.18 (±0.03)	0.08 (±0.04)

Table 1: Interaction terms based on background-averaged epistasis ($\bar{\varepsilon}$) for an empirical multiallelic genotype-phenotype landscape consisting of all combinations of two mutations each at positions 6 and 66 in the tRNA-Arg(CCU) [6], i.e. $n = 2$ and $s = 3$. The first two columns indicate nucleic acid sequences and their base 3 representations. Here the 'GC' reference (wild-type) genotype corresponds to that of *S. cerevisiae*, denoted by $(0, 0)$. The second two columns show the measured phenotypic effects and corresponding background-averaged epistatic coefficients together with their empirical errors and their propagated errors. See Results for a regression analysis of the entire dataset and Supplementary Methods for a derivation of the error propagation.

More generally, for any value of s, when $n = 1$,

$$\begin{pmatrix} \varepsilon_{(*)} \\ \varepsilon_{(1)} \\ \varepsilon_{(2)} \\ \vdots \\ \varepsilon_{(s-1)} \end{pmatrix} = \begin{pmatrix} 1/s & 0 & 0 & \dots & 0 \\ 0 & -1 & 0 & \dots & 0 \\ 0 & 0 & -1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & \dots & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} y_{(0)} \\ y_{(1)} \\ y_{(2)} \\ \vdots \\ y_{(s-1)} \end{pmatrix} := V_1 \cdot H_1 \cdot \bar{y}_1, \tag{6}$$

where $\varepsilon_{(*)}$ corresponds to averaging phenotypes over all possible genotypes and the remaining coefficients simply correspond to the phenotypic effects of each mutation.

For $n = 2$, we have to consider different combinations of mutations in both positions. In this case, the phenotypes can be written as

$$y_{(0,0)}, y_{(0,1)}, \dots, y_{(0,(s-1))}, y_{(1,0)}, \dots, y_{(1,(s-1))}, \dots, y_{((s-1),0)}, \dots, y_{((s-1),(s-1))} \,.$$

A natural ordering of the phenotypes is given by interpreting genotype $\vec{i}$ as the base s representation of an integer (see Table 1). From this, we can see how the first s genotypes correspond to combining the wild-type allele at the first position with a state from the case $n = 1$, i.e. to genotypes that can be written $0 \frown \vec{i} := (0, \vec{i})$, with $\vec{i} \in \mathcal{S}^1$. The next s genotypes correspond to the first mutated allele at the first position combined with

all the genotypes of $n = 1$, i.e. $1 \frown \vec{i}, \vec{i} \in \mathcal{S}^1$, and so on. Therefore, we can write the matrices H and V following a block structure. In the case $n = 2$ and any given s, we would then have

$$H_2 = \begin{pmatrix} H_1 & H_1 & H_1 & \dots & H_1 \\ H_1 & -H_1 & 0 & \dots & 0 \\ H_1 & 0 & -H_1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ H_1 & 0 & \dots & 0 & -H_1 \end{pmatrix}, \tag{7}$$

where the number of H_1 blocks corresponds to the number of states of the first position, so s. Moreover, each of these blocks must be normalized to yield the corresponding background-averaged epistatic terms. Therefore V_2 can also be expressed as a function of V_1 as follows:

$$V_2 = \begin{pmatrix} \frac{1}{s}V_1 & 0 & \dots & 0 \\ 0 & -V_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -V_1 \end{pmatrix}. \tag{8}$$

Given these two matrices, we can write the background-averaged epistatic coefficients for the case of $n = 2$ and s different states per position as $\bar{\varepsilon}_2 = V_2 \cdot H_2 \cdot \bar{y}_2$. More generally, the decomposition of phenotypic effects into background-averaged epistatic coefficients is given by

$$\bar{\varepsilon}_n = V_n \cdot H_n \cdot \bar{y}_n, \tag{9}$$

where H_n and V_n can be defined recursively as

$$H_{n+1} = \begin{pmatrix} H_n & H_n & H_n & \dots & H_n \\ H_n & -H_n & 0 & \dots & 0 \\ H_n & 0 & -H_n & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ H_n & 0 & \dots & 0 & -H_n \end{pmatrix} \qquad H_0 = 1 \qquad \text{and} \qquad H_n \text{ is } s^n \times s^n, \tag{10}$$

$$V_{n+1} = \begin{pmatrix} \frac{1}{s}V_n & 0 & 0 & \dots & 0 \\ 0 & -V_n & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & -V_n \end{pmatrix} \qquad V_0 = 1 \qquad \text{and} \qquad V_n \text{ is } s^n \times s^n. \tag{11}$$

Recursive inverse matrix

Equation (9) defines the vector of epistatic coefficients, $\bar{\varepsilon}_n$, as a function of the vector of phenotypes, $\bar{y}_n$, which in general is the quantity that is measured experimentally. However, usually phenotypic measurements are only available for a subset of genotypes. An alternative is therefore to estimate the epistatic coefficients $\bar{\varepsilon}_n$ by regression,

$$\bar{y}_n = H_n^{-1} \cdot V_n^{-1} \cdot \bar{\varepsilon}_n \,, \tag{12}$$

where the product $H_n^{-1} \cdot V_n^{-1}$ represents a matrix of sequence features. This is analogous to the more widely used one-hot encoding strategy, which implicitly relies on a 'background-relative' (or 'biochemical') view of epistasis when regressing to full order [2]. We discuss other advantages of estimating background-averaged epistatic coefficients using regression at the end of this manuscript.

Since V_n is a diagonal matrix, its inverse is also a diagonal matrix whose elements are the inverse of the elements of V_n.

The inverse of H_n is the matrix A_n which can be defined recursively as

$$A_{n+1} = \frac{1}{s}\begin{pmatrix} A_n & A_n & A_n & \dots & A_n \\ A_n & (1-s)A_n & A_n & \dots & A_n \\ A_n & A_n & (1-s)A_n & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & A_n \\ A_n & A_n & \dots & A_n & (1-s)A_n \end{pmatrix} \qquad A_0 = 1 \qquad \text{and} \qquad A_n \text{ is } s^n \times s^n. \tag{13}$$

See Proposition 1 in S1 Text for a proof of this result. This is the most efficient method to determine the full matrix A_n (see Results).

Formulae to obtain elements of the matrices

When regressing phenotypes on genotypes, a common goal is to determine whether epistatic coefficients up to the rth order (where $r < n$) are sufficient to describe the complexity of the biological system. Furthermore, as mentioned above, some fraction of phenotype values within combinatorially complete genetic landscapes are typically unavailable, representing missing data. Restricting the epistatic order and missing phenotypes respectively correspond to omitting rows and columns from H_n (and vice versa from A_n). Formulae to directly obtain elements of the matrices in equations (9) and (12) would therefore be convenient.

In order to write the matrix element $(H_n)_{ij}$, we need to compare the genotype sequences $\vec{i}, \vec{j} \in \mathcal{S}^n$,

$$\vec{i} = (i_1, i_2, \dots, i_n)$$
$$\vec{j} = (j_1, j_2, \dots, j_n)\,,$$

where $\vec{i}$ denotes the ith element in $\mathcal{S}^n$, $\mathcal{S} = \{0, 1, \dots, s-1\}$, and the elements of $\mathcal{S}^n$ are ordered by the base s representation of integers. For instance, for any value of n, we will denote the wild-type state with index $i = 1$ and write $\vec{i} = \vec{1} = (0, \dots, 0)$. The element denoted with index $i = 2$ would be $\vec{i} = \vec{2} = (0, \dots, 0, 1)$ and so on.

The elements of H_n can be written as

$$(H_n)_{ij} = \begin{cases} (-1)^{(E_n)_{ij}} & \text{if } (M_n)_{ij} = n \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where M and E are $s^n \times s^n$ matrices whose elements are

$$(E_n)_{ij} = \sum_{\substack{k=1 \\ i_k \cdot j_k > 0}}^{n} \delta_{i_k j_k} \tag{15}$$

$$(M_n)_{ij} = \sum_{\substack{k=1 \\ i_k \cdot j_k > 0}}^{n} \delta_{i_k j_k} + \sum_{\substack{k=1 \\ i_k \cdot j_k = 0}}^{n} 1 = (E_n)_{ij} + \sum_{\substack{k=1 \\ i_k \cdot j_k = 0}}^{n} 1, \tag{16}$$

where δ_{ij} denotes the Kronecker delta of i, j, which is equal to 1 when $i = j$ and 0 if $i \neq j$. In words, $(E_n)_{ij}$ counts the number of positions at which the genotype sequences $\vec{i}$ and $\vec{j}$ carry the same mutated allele and $(M_n)_{ij}$ is equal to $(E_n)_{ij}$ plus the number of positions where $\vec{i}$ or $\vec{j}$ carry the wild-type allele. See Proposition 2 in S1 Text for a proof of this result.

Furthermore, the elements of A_n can be written as

$$(A_n)_{ij} = \frac{1}{s^n}(1-s)^{(E_n)_{ij}}, \tag{17}$$

where E_n is defined as in (15). See Proposition 3 in S1 Text for a proof of this result.

Finally, the matrices V_n and V_n^{-1} are diagonal matrices whose diagonal elements can be written as

$$(V_n)_{ii} = (-1)^{n - W_n(\vec{i})} \frac{1}{s^{W_n(\vec{i})}} \tag{18}$$

and

$$(V_n^{-1})_{ii} = (-1)^{n - W_n(\vec{i})} s^{W_n(\vec{i})}, \tag{19}$$

where

$$W_n(\vec{i}) := \sum_{k=1}^{n} w_k, \text{ with } w_k := \delta_{i_k 0}$$

and $\vec{i}$ again denotes the ith element in $\mathcal{S}^n$ when ordered by the base s representation of integers. In words, $w_k = 1$ if the genotype sequence $\vec{i}$ carries the wild-type allele at position k and $W_n(\vec{i})$ counts the number of

positions in $\vec{i}$ carrying the wild-type allele. We prove this result in Proposition 4 in S1 Text.

Generalization to different numbers of states per position

We can generalize the formulae described in the previous subsection further by considering that each position can have different numbers of states. In this case, we can denote s_k the number of possible states at position k. For $n = 1$, this corresponds to exactly the same matrix as in the previous case but with $s = s_1$, which is the number of possible states in this position. For $n = 2$, the matrix changes because now the new position can have a different number of possible states, s_2. Following the recursive definition of H_n, we can construct H_2 by repeating H_1 s_2 times, with the structure stated in (10). Therefore, we have

$$H_2 = \underbrace{\begin{pmatrix} \overbrace{H_1}^{s_1} & H_1 & H_1 & \dots & H_1 \\ H_1 & -H_1 & 0 & \dots & 0 \\ H_1 & 0 & -H_1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ H_1 & 0 & \dots & 0 & -H_1 \end{pmatrix}}_{s_2 \text{ blocks of size } s_1 \Longrightarrow s_2 s_1}. \tag{20}$$

So the structure is exactly the same but the size of the matrix for each n varies according to the number of possible states of the new position. The definition of H_n is the same as in (10) but the dimensions of the matrix are $\prod_{k=1}^{n} s_k \times \prod_{k=1}^{n} s_k$. Similarly, the inverse matrix A_{n+1} can be written recursively as

$$A_{n+1} = \frac{1}{s_{n+1}} \begin{pmatrix} A_n & A_n & A_n & \dots & A_n \\ A_n & (1-s_{n+1})A_n & A_n & \dots & A_n \\ A_n & A_n & (1-s_{n+1})A_n & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & A_n \\ A_n & A_n & \dots & A_n & (1-s_{n+1})A_n \end{pmatrix}, \quad A_0 = 1 \quad \text{and} \quad A_n \text{ is } \prod_{k=1}^{n} s_k \times \prod_{k=1}^{n} s_k. \tag{21}$$

The matrix A_n defined in (21) is the inverse of the matrix H_n in the general case where each position can have a different number of states.

In this general case, the elements of H_n and A_n can be written as

$$(H_n)_{ij} = \begin{cases} (-1)^{(E_n)_{ij}} & \text{if } (M_n)_{ij} = n \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

$$(A_n)_{ij} = \frac{\prod_{k=1}^{n} \left(1 - s_k\right)^{e_k}}{\prod_{k=1}^{n} s_k}, \tag{23}$$

where E_n and M_n are defined as in (15) and $e_k = \begin{cases} 1 & \text{if } i_k = j_k \neq 1 \\ 0 & \text{otherwise} \end{cases}$.

The matrices V_n and V_n^{-1} are diagonal matrices whose diagonal elements can be written as

$$(V_n)_{ii} = (-1)^{n-W_n(\vec{i})} \prod_{k=1}^{n} \left(\frac{1}{s_k}\right)^{w_k} \tag{24}$$

and

$$(V_n^{-1})_{ii} = (-1)^{n-W_n(\vec{i})} \prod_{k=1}^{n} {s_k}^{w_k}, \tag{25}$$

where

$$W_n(\vec{i}) := \sum_{k=1}^{n} w_k, \text{ with } w_k := \delta_{i_k 0}.$$

We prove the results in this subsection in Propositions 5, 6 and 7 in S1 Text.

The above formulae permit the calculation and modeling of background-averaged epistasis in arbitrarily-shaped genetic landscapes, i.e. with any number of alleles (states) per position, as well as the direct construction of sub-matrices for regression to any desired epistatic order and/or in the presence of missing data. In the following subsections we report benchmarking results comparing the performance of alternative methods to obtain H_n and A_n, as well as results from the application of our theory extension to an empirical multiallelic genotype-phenotype landscape.

Benchmarking

Fig 1a-d provides a visualization of the matrices H_n and A_n for different values of n and s, clearly showing a self-similar pattern in all cases due to their recursive nature.

In this paper, we provide different methods to construct $A_n = H_n^{-1}$. First, H_n can be numerically inverted using standard matrix inversion algorithms (here we use the `linalg.inv` function from the SciPy library in Python), referred to as "Recursive H_n inverse" in Fig 1e,f. Alternatively, the recursive definition of the inverse given by equation (13) can be used, which we refer to as "Recursive A_n". As can be seen in Fig 1e, this method is faster than numerically inverting H_n.

Finally, we also provide a convenient formula for extracting specific individual elements of A_n (Proposition 3), referred to as "All elements A_n" in Fig 1e,f. This method is more computationally intensive than the previously described methods, due to the formula relying on the computation of $(E_n)_{ij}$, which equates to counting the number of sequence positions that are identically mutated in vectors $\vec{i}$ and $\vec{j}$, each of size n. However, in situations where subsets of elements (or sub-matrices) – rather than full matrices – are desired, Proposition 3 provides a method that can be faster and more memory efficient (see "10 elements A_n" in Fig 1e,f).

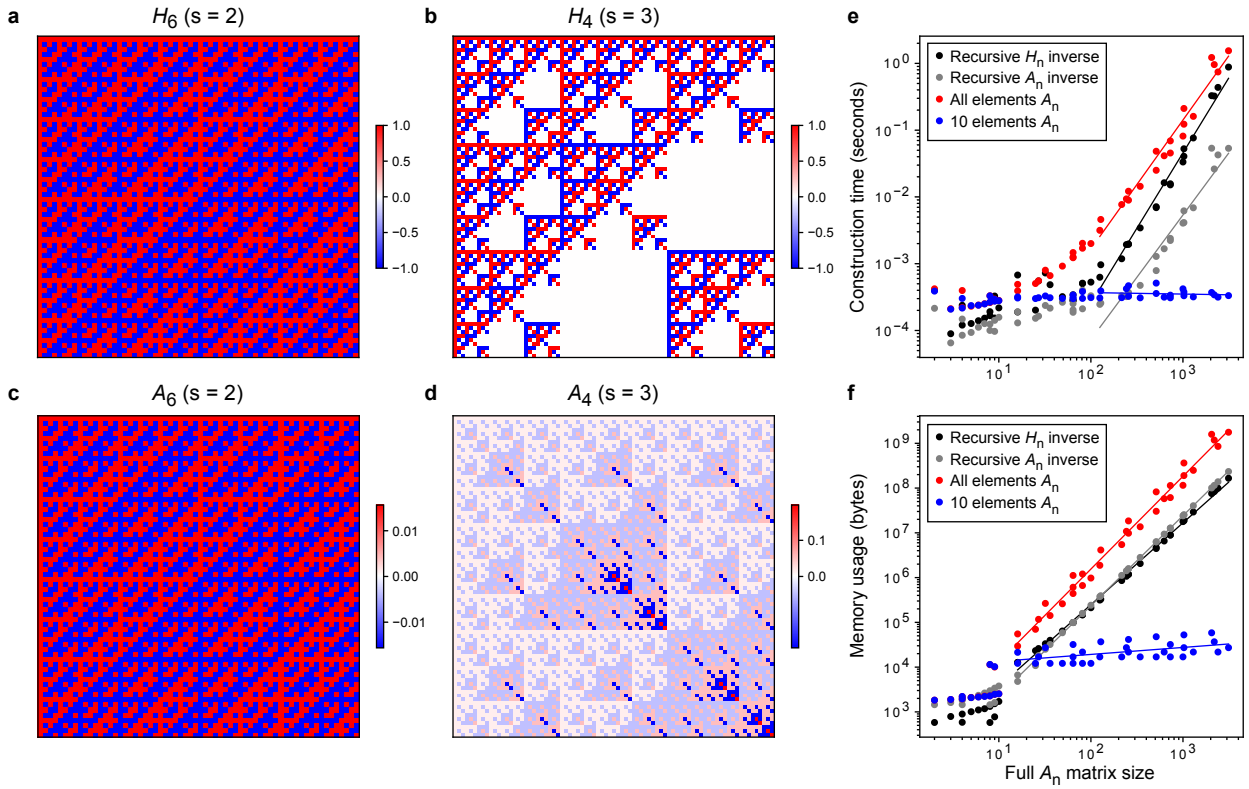

Figure 1: Benchmarking results and heat map representations of matrices corresponding to the binary (biallelic) and multi-state (multiallelic) extension of the Walsh-Hadamard transform, and their corresponding inverses. **a**, H_6 Walsh-Hadamard transform (s=2). **b**, H_4 multi-state extension of the Walsh-Hadamard transform for $s = 3$. **c**, A_6 Inverse Walsh-Hadamard transform. **d**, A_4 multi-state extension of the inverse Walsh-Hadamard transform for $s = 3$. **e**, Computational time on a MacBook Pro (13-inch, 2017, 2.3GHz dual-core Intel Core i5) for extracting elements of A_n matrices of various dimensions and numbers of states (alleles) per position ($s \in [2, 10]$). Comparisons are shown between numerically inverting the recursively constructed H_n (using `scipy.linalg.inv`), i.e. "Recursive H_n inverse", using the recursive formula for A_n, using the formula to extract all elements of A_n and extracting 10 random elements of A_n (see legend). The mean across 10 replicates is depicted. Linear regression lines were fit to data from matrices with at least 100 elements. **f**, Similar to **e** but indicating memory usage. Linear regression lines were fit to data from matrices with at least 10 elements.

For example, in the case of a 10-mer DNA sequence, constructing the full inverse transform A_{10} with $s = 4$ would require $> 10^{23}$ bytes (100 million petabytes) of memory in the best-case scenario ("Recursive H_n inverse" in Fig 1f, log-linear extrapolation). Similarly, the full inverse transform for a 4-mer amino acid sequence (A_4 with $s = 20$) would impose a memory footprint $> 10^{20}$ bytes. On the other hand, calculating the subset of elements from these matrices required for the prediction of a single phenotype using epistatic coefficients up to third order (three-way genetic interaction terms) is feasible in both situations using Proposition 3 (3,675 and 29,678 elements; 2.5 GB and 192 GB of memory; 1.8 and 99 seconds, respectively). This memory footprint can easily be diminished further using data chunking, which is a unique benefit of this method.

Application to a simulated multiallelic genotype-phenotype landscape

In order to demonstrate the utility of our extension of the Walsh-Hadamard transform, we used it to model epistasis within a simulated multiallelic genetic landscape. Knowledge of the 'ground truth' epistatic coefficients allowed us to (i) determine whether background-averaged coefficients can be accurately inferred using regression and (ii) investigate the impact of data sparseness on the results i.e. in the context of missing phenotypic values. In most practical applications – particularly in the case of large empirical genotype-phenotype landscapes – measuring the phenotypic effects of all mutation combinations is infeasible. The simulated combinatorially complete landscape consisted of all possible combinations of four alleles at six different positions i.e. a total of $4^6 = 4,098$ genotypes (Fig 2a, left). First-order and a subset of second-order epistatic coefficients were randomly drawn from Gaussian distributions, while all higher-order epistatic coefficients were set to zero (see Methods). We arbitrarily selected three pairs of positions at which mutations genetically interact i.e. ground truth non-zero pairwise epistatic terms (Fig 2a, left). Fitness values for all variants were reconstructed using equation (12) with random noise added to simulate measurement error (see Methods).

We then trained Lasso regression models of the form in equation (12) to predict fitness values from sequence, where the inferred model parameters correspond to background-averaged epistatic coefficients up to third order (S1a Fig; see Methods). To assess the impact of data sparsity on the results, we sub-sampled the simulated phenotype (fitness) values to obtain training dataset sizes ranging from 64% to 1% of all variants in the combinatorial complete landscape. The sparse models generalize well, explaining more than 80% of fitness variance in the held out test data even when the overwhelming majority of variants (96%) are missing (Fig 2b, left; 'Background-averaged models', red). Importantly, the values of non-zero background-averaged epistatic coefficients closely match those of the ground truth simulated coefficients (Pearson's $r = 0.99$, Fig 2c, left). Reducing the size of the training set tends to reduce the number of coefficients retained in sparse models, but the ground truth values of non-zero coefficients are well recovered (Pearson's $r = 0.99$, S2a Fig).

For comparison, we used the same procedure to fit Lasso regression models of the form $\bar{y} = G^{-1} \cdot \bar{\varepsilon}$, where G^{-1} represents a matrix of one-hot encoded sequence features i.e. the presence or absence of a given mutation – or mutation combination (interaction) – with respect to the reference (wild-type) genotype is denoted by a '1' or '0' respectively (S1a Fig, Fig 2b, left; 'One-hot models', blue). The definition of G and its relationship to the biochemical (or background-relative) view of epistasis is explained in [2]. Although one-hot models show

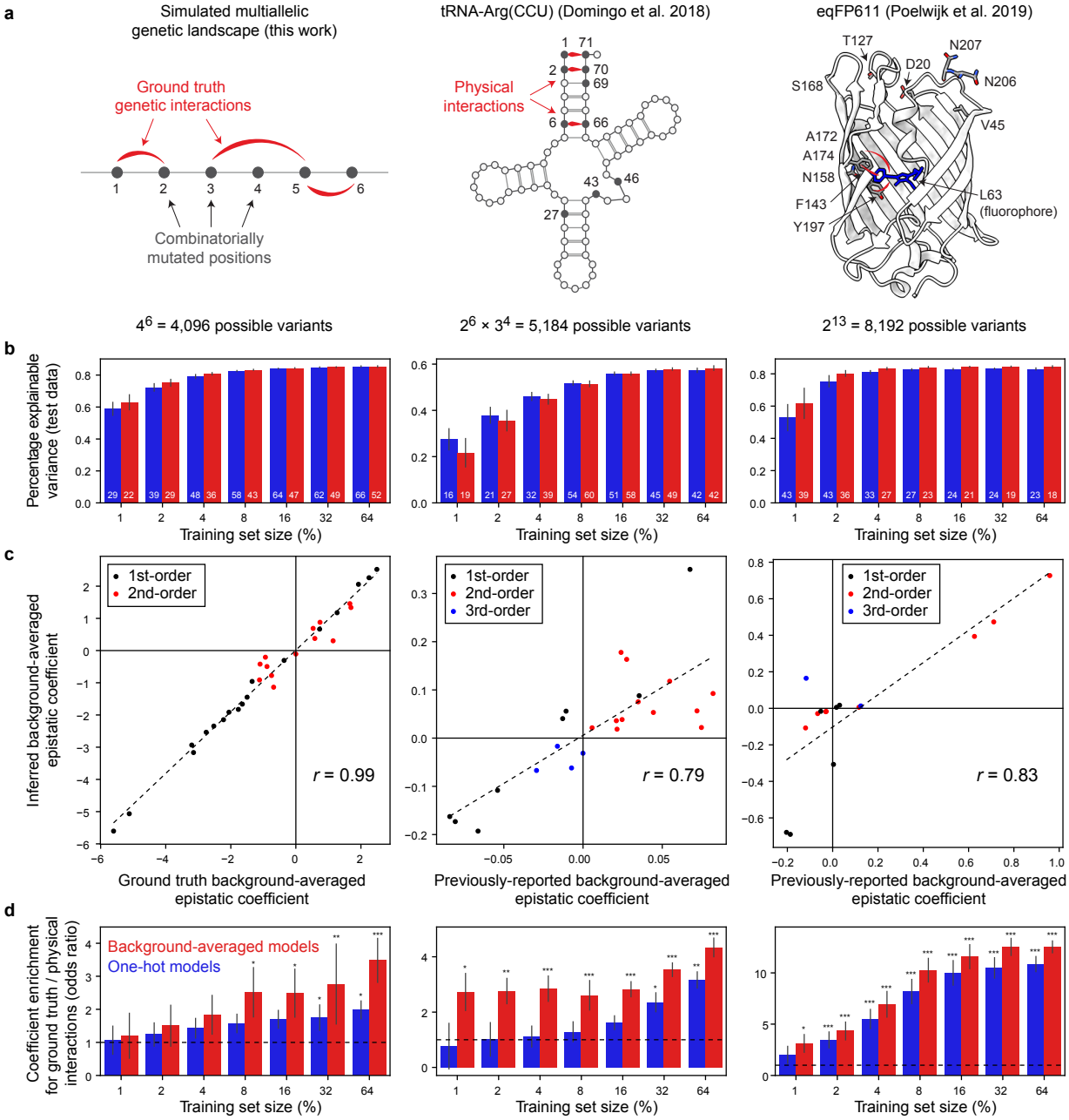

Figure 2: Learning sparse models from simulated and empirical combinatorial fitness landscapes. **a**, Schematic description of combinatorial DMS datasets. Left: simulated multiallelic landscape consisting of all combinations of 4 nucleotides in a DNA 6-mer where ground truth interactions ([1,2], [3,5] and [5,6]) are depicted with red arcs (see Methods). Middle: secondary structure of *S. cerevisiae* tRNA-Arg(CCU) indicating variable positions (closed circles) combinatorially mutated in a DMS experiment as described in [6]. Three Watson–Crick base pairing (WCBP) interactions involving pairs of these positions ([1,71], [2,70] and [6,66]) are indicated. Right: crystal structure of a blue fluorescent variant (TagBFP) of the *Entacmaea quadricolor* protein eqFP611 (PDB: 3M24) and 13 positions (12 shown) that differ in the red fluorescent variant (mKate2) that were mutated in a DMS experiment as described in [7]. Physical interactions between the fluorophore (L63) and three proximal residues ([63,143], [63,174] and [63,197]) are indicated. **b**, Performance of sparse models fit using different proportions of the DMS datasets indicated in panel a. The median number of model coefficients is indicated. Colour scale as in panel d. **c**, Scatter plots comparing sparse model-inferred background-averaged epistatic coefficients (training set size = 64%) to ground truth values (left panel) or previously-reported coefficients (remaining panels) [6, 7]. **d**, Enrichment of direct physical interactions (red arcs in panel a) in non-zero epistatic coefficients. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. Error bars indicate the interquartile range.

similar performance to background-averaged models on simulated test data, they tend to be more complex with larger numbers of coefficients (Fig 2b), including spurious third-order epistatic terms although their ground truth values are zero (S1c Fig).

In summary, these results demonstrate that our theory extensions allow the accurate inference of background-averaged epistatic coefficients in multiallelic genetic landscapes, even in situations when large amounts of phenotypic data are missing.

Application to empirical combinatorial fitness landscapes

We next applied our theory to model epistasis within two different empirical combinatorial fitness landscapes (Fig 2a, middle, right). In the first DMS assay, a budding yeast strain was used in which the single-copy arginine-CCU tRNA (tRNA-Arg(CCU)) gene is conditionally required for growth [6]. A library of variants of this gene was designed to cover all 5,184 ($2^6 \times 3^4$) combinations of the 14 nucleotide substitutions observed in ten positions in post-whole-genome duplication yeast species. The library was transformed into *S. cerevisiae*, expressed under restrictive conditions and the enrichment of each genotype in the culture was quantified by deep sequencing before and after selection. After reprocessing of the raw data, we retained high quality fitness estimates for 3,847 variants (74.2%, see Methods).

Although the findings in [6] were based on the application of background-averaged epistasis theory, the prior limitation of the Walsh-Hadamard transform to only two alleles per sequence position required the authors to adopt an *ad hoc* strategy that involved performing separate analyses on combinatorially complete biallelic sub-landscapes. After imputing fitness values for missing variants we decomposed phenotypes into background-averaged epistatic coefficients using equation (9), which yielded very similar values for significant coefficients of all orders (S2b Fig, see Methods). However – as shown above for simulated data – our theory extensions permit the modeling of background-averaged epistasis in the context of multiallelic landscapes with missing data. We followed the same strategy as described in the previous section, using the tRNA DMS data to train Lasso regression models to predict variant fitness from sequence incorporating epistatic coefficients up to eighth order (Fig 2b, middle, see Methods). The resulting models include many higher-order epistatic coefficients (S1c Fig, 'Background-averaged models') yet exhibit extreme sparsity, with the median number of non-zero coefficients of any order ranging from 19 to 60 i.e. approximately 1% of all possible coefficients of eighth order or less (Fig 2b, middle). Models trained on at least 10% of the training data tend to explain more than 50% of the total explainable variance (Fig 2b, middle) and model-inferred background-averaged epistatic terms correlate well with those reported previously [6] (Pearson's $r = 0.79$, Fig 2c, middle).

To evaluate whether the inferred models report on biologically relevant features of the underlying genetic landscapes, we tested whether sparse model coefficients were more likely to comprise genetic interactions (or modulators thereof) involving known physically interacting positions in the wild-type tRNA secondary structure (Fig 2a, middle). Regardless of data sparsity, background-averaged model coefficients tend to be significantly enriched for physical interactions (Fig 2d, middle). On the other hand, in the case of even mod-

erate sub-sampling of training data (16%), one-hot model coefficients show no such enrichment (Fig 2d, middle). The results of this enrichment analysis closely mirror those obtained using simulated data (Fig 2d, left), suggesting that they are likely to generalise to other empirical genotype-phenotype landscapes.

Finally, we repeated similar analyses using a second DMS dataset comprising fluorescence measurements for all combinations of single substitution mutations at 13 positions ($2^{13} = 8,192$ genotypes) separating two variants of the *Entacmaea quadricolor* protein eqFP611 (Fig 2, right column; see Methods) [7]. Lasso models capture the majority of fitness variance even with high levels of data sparsity (Fig 2b, right) and model-inferred background-averaged coefficients correlate well with those obtained using the decomposition in equation (9) (Pearson's $r = 0.83$, Fig 2c, right). We also find that direct pairwise physical interactions involving the fluorophore (L63, Fig 2a, right) are enriched in model-derived background-averaged epistatic terms (Fig 2d, right), a result that is robust to missing data and recapitulates our observations obtained using ground truth genetic interactions in simulated DMS data.

Discussion

We have provided an extension to the most rigorous computational framework available for describing and modeling empirical genotype-phenotype mappings. Beyond the study of background-averaged epistasis with respect to mutations in the primary sequence, this also permits the inclusion of 'epimutations' (changes in the epigenetic state of DNA), amino acid post-translational modifications or even particular environmental/experimental conditions.

In the simplest application, background-averaged epistatic coefficients (genetic interaction terms) can be directly computed from phenotypic measurements via the decomposition in equation (9). However, estimating epistatic coefficients by regression – as in equation (12) – is a more natural choice in the presence of missing data, when data for multiple related phenotypes is available [23] and/or in the presence of global epistasis [24, 25]. Our mathematical results provide three alternative methods to compute the multi-state (multiallelic) extension of the inverse Walsh-Hadamard transform A_n, one of which allows the direct extraction of specific elements or sub-matrices. In which situations might this capability be desirable?

First, constructing full A_n matrices – particularly by numerical inversion – is impractical for large genetic landscapes. Second, the result of the product $H_n^{-1} \cdot V_n^{-1}$ represents a matrix of sequence features when setting up the inference of epistatic (model) coefficients $\bar{\varepsilon}_n$ from phenotypic measurements $\bar{y}_n$ as a regression task [23, 24, 26, 27]. The ability to construct this feature matrix in batches (of rows) allows computational resource-efficient iteration over large datasets when using frameworks such as TensorFlow or PyTorch.

Third, there are currently no methods to comprehensively map empirical genotype-phenotype landscapes with size greater than the low millions of genotypes. Therefore, assaying landscapes of this size or larger will typically involve experimental measurement of a (random) sub-sample of genotypes, corresponding to distinct rows in A_n. In other words, it is usually unnecessary to construct full A_n matrices when modeling real experimental data. Finally, there is evidence of extreme sparsity in the epistatic architecture of biomolecules

where only a small fraction of theoretically possible genetic interactions are non-zero [7]. The feasibility of sampling very large background-averaged epistatic coefficient spaces may improve methods to infer accurate genotype-phenotype models.

Using results from the analysis of both simulated and empirical multiallelic fitness landscapes, we have shown that sparse regression models relying on a background-averaged definition of epistasis can efficiently capture salient features of the underlying biological system – namely direct physical interactions or ground truth genetic interactions – even in situations of sparse sampling of phenotypes. This behaviour, which we speculate is due to a richer representation of the sequence feature space compared to one-hot models (i.e. higher level of constraint during model fitting; S1a Fig), is particularly desirable in the case of very large genetic landscapes where comprehensive phenotyping is infeasible. However, more work is needed to determine whether this result holds more generally. One difficulty in such comparisons between approaches is the requirement for a set of interactions or landscape features that are known to be critical for biomolecular function. Here we rely on Watson–Crick base pairing interactions and direct physical interactions involved in fluorophore orientation whose respective importance for RNA secondary structure and fluorescence activity are well-established.

More broadly, this work opens the door to investigations of the biological properties of background-averaged epistasis in empirical genetic landscapes of arbitrary shape and complexity. Beyond applications within the field of DMS, we believe our theory extensions have the potential to influence research in evolutionary and synthetic biology including protein engineering. In future it will be important to compare the performance and properties of models relying on this definition of epistasis to those of other recently proposed models that incorporate higher-order genetic interactions for phenotypic prediction [28, 29].

Methods

Simulated multiallelic fitness landscape

The simulated multiallelic genetic landscape (Fig 2a, left) consists of all possible combinations of 4 nucleotides per position in a 6-mer i.e. $4^6 = 4,096$ possible genotypes. Consistent with observations of empirical fitness landscapes where single mutant phenotypes tend to be biased towards negative (detrimental) effects and larger in size than pairwise and higher-order epistatic terms, ground truth first-order epistatic coefficients (single nucleotide substitution effects) were drawn randomly from a normal distribution with a mean of -1 and standard deviation of 2 i.e. $\epsilon = N(\mu, \sigma^2) = N(-1, 4)$. Ground truth non-zero second-order terms (pairwise genetic interactions) were arbitrarily selected between the following pairs of positions: [1,2], [3,5] and [5,6], where their coefficient values were randomly sampled from $\epsilon = N(0, 1)$. All other second- and higher-order coefficients were set to zero. Fitness scores for all variants were calculated using equation (12). We simulated measurement error by adding Gaussian noise to all fitness scores of similar magnitude to the variance of first-order epistatic coefficients i.e. $\sigma_\epsilon^2 = N(0, 4)$. Finally, we also simulated versions of the same multiallelic fitness landscape with both relatively lower and higher measurement noise where $\sigma_\epsilon^2 = N(0, 1)$ and $\sigma_\epsilon^2 = N(0, 9)$, respectively (S1 Fig).

Empirical combinatorial fitness landscapes

Raw sequencing (FASTQ) files obtained from the tRNA-Arg(CCU) DMS experiment in [6] were re-processed with DiMSum v1.3 [30] using default parameters with minor adjustments. We obtained fitness estimates for 5,059 out of a total of 5,184 possible variants (97.6%) in the combinatorially complete genetic landscape. We restricted the data to a high quality subset by requiring fitness estimates in all six biological replicates as well as at least 10 input read counts in all input samples. This resulted in a total of 3,847 retained variants (74.2%) for downstream analysis. For comparisons to previously-reported background-averaged coefficients (Fig 2c, S2b Fig), we used the author-processed data and analysis scripts [6]. Likewise, we used the author-processed data for imputation of missing phenotype values (replaced with the mean fitness at every mutation order) and calculation of background-averaged epistatic coefficients using the decomposition in equation (9) (S2b Fig). For the eqFP611 fluorescent protein DMS experiment, we used the author-processed fitness estimates (brightness scores) [7].

Lasso regression models

We trained Lasso regression models to predict variant fitness estimates from nucleotide or amino acid sequences using the 'scikit-learn' Python package. Training data comprised random subsets of 1, 2, 4, 8, 16, 32 and 64% of retained variants of all mutation orders. All remaining held out variants comprised the 'test' data which was unseen during model training in each case.

To train models inferring background-averaged epistatic coefficients we used feature matrices of the form $H_n^{-1} \cdot V_n^{-1}$ (see equation (12)). For comparison, one-hot encoded matrices of sequence features were used. Linear regression was performed using 10-fold cross validation to determine the optimal value of the L1 regularization parameter λ in the range [0.005, 0.25] ('LassoCV' and 'RepeatedKFold' functions). Final models were fit to all training data. In order to estimate model-related statistics and performance results we fit 100 models to different random subsets of the training data for each model type and training data fraction. In Fig 2, S1 Fig and S2 Fig we plot the median of the indicated measures over all models, where error bars indicate the interquartile range. For model performance estimates in Fig 2b, the maximum explainable variance was calculated by substracting the total estimated technical variance (as reported by DiMSum fitness errors [30]) from the total fitness variance. For the simulated and eqFP611 DMS datasets, the maximum explainable variance was assumed to be 100%. In comparisons between sparse model-derived background-averaged epistatic coefficients and ground truth or previously-reported coefficients (scatter plots in Fig 2d and S1 Fig) we required model-derived terms to be non-zero in $> 80\%$ of models corresponding to random training data subsets.

To test enrichment of physical interactions in Lasso model coefficients we used the following approach: for each model, all position pairs represented in non-zero epistatic coefficients of at least second order were determined. The number of position pairs corresponding to direct physical interactions was counted and an associated enrichment score (odds ratio) and P-value calculated using Fisher's Exact Test. The background set consisted of all position pairs in all possible epistatic coefficients. To test the appropriateness of the

null hypothesis we also repeated enrichment analyses using random models i.e. randomly chosen sets of epistatic coefficients matching the numbers of non-zero coefficients in Lasso models and their distribution over different epistatic orders. The results of this analysis performed using the simulated dataset are shown in S1b Fig, left. Finally, the middle and right panels in S1b Fig assess the impact of measurement noise on the results of this enrichment analysis, revealing that noise level is anti-correlated with enrichment score and significance thereof, as expected.

Acknowledgments

This work was funded by the Spanish Ministry of Science and Innovation (Severo Ochoa Centre of Excellence) and the CERCA Programme / Generalitat de Catalunya. A.J.F. was funded by a Ramón y Cajal fellowship (RYC2021-033375-I) financed by the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033) and the European Union (NextGenerationEU/PRTR). B.L. was funded by European Research Council (ERC) Advanced grant (883742), the Spanish Ministry of Science and Innovation (LCF/PR/ HR21/52410004, EMBL Partnership), the Bettencourt Schueller Foundation, the AXA Research Fund and Agencia de Gestio d'Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322). V.M.P. was funded by the Spanish Ministry of Science and Innovation (PGC2018-100941-A-I00, PID2021-128976NB-I00). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank all members of the Lehner and Weghorn Labs for helpful discussions and suggestions.

References

[1] Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics. 2008;9(11):855-67.

[2] Poelwijk FJ, Krishna V, Ranganathan R. The Context-Dependence of Mutations: A Linkage of Formalisms. PLoS Computational Biology. 2016 Jun;12(6):e1004771.

[3] Domingo J, Baeza-Centurion P, Lehner B. The Causes and Consequences of Genetic Interactions (Epistasis). Annu Rev Genomics Hum Genet. 2019 Aug;20:433-60.

[4] Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014 Aug;11(8):801-7.

[5] de Visser JAG, Krug J. Empirical fitness landscapes and the predictability of evolution. Nature Reviews Genetics. 2014;15(7):480-90.

[6] Domingo J, Diss G, Lehner B. Pairwise and higher-order genetic interactions during the evolution of a tRNA. Nature. 2018 Jun;558(7708):117-21.

[7] Poelwijk FJ, Socolich M, Ranganathan R. Learning the pattern of epistasis linking genotype and phenotype in a protein. Nature communications. 2019 Sep;10(1):1-11.

[8] Baeza-Centurion P, Miñana B, Schmiedel JM, Valcárcel J, Lehner B. Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. Cell. 2019;176(3):549-63.

[9] Pokusaeva VO, Usmanova DR, Putintseva EV, Espinar L, Sarkisyan KS, Mishin AS, et al. An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. PLoS genetics. 2019;15(4):e1008079.

[10] Bendixsen DP, Collet J, Østman B, Hayden EJ. Genotype network intersections promote evolutionary innovation. PLoS biology. 2019;17(5):e3000300.

[11] Soo VW, Swadling JB, Faure AJ, Warnecke T. Fitness landscape of a dynamic RNA structure. PLoS genetics. 2021;17(2):e1009353.

[12] Moulana A, Dupic T, Phillips AM, Chang J, Nieves S, Roffler AA, et al. Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA. 1. Nature Communications. 2022;13(1):1-11.

[13] Rotrattanadumrong R, Yokobayashi Y. Experimental exploration of a ribozyme neutral network using evolutionary algorithm and deep learning. Nature communications. 2022;13(1):1-14.

[14] Lynch M, Walsh B, et al. Genetics and analysis of quantitative traits. vol. 1. Sinauer Sunderland, MA; 1998.

[15] Goldberg DE. Genetic Algorithms and Walsh Functions: Part I, A Genetle Introduction. Complex systems. 1989;3:129-52.

[16] Weinreich DM, Lan Y, Wylie CS, Heckendorn RB. Should evolutionary geneticists worry about higher-order epistasis? Current opinion in genetics & development. 2013;23(6):700-7.

[17] Poelwijk FJ, Ranganathan R. The relation between alignment covariance and background-averaged epistasis. arXiv. 2017;10.48550/ARXIV.1703.10996.

[18] Brookes DH, Aghazadeh A, Listgarten J. On the sparsity of fitness functions and implications for learning. Proc Natl Acad Sci U S A. 2022 Jan;119(1).

[19] Ogbunugafor CB. The mutation effect reaction norm (mu-rn) highlights environmentally dependent mutation effects and epistatic interactions. Evolution. 2022 02;76(s1):37-48.

[20] Weinberger ED. Fourier and Taylor series on fitness landscapes. Biological cybernetics. 1991;65(5):321-30.

[21] Beer T. Walsh transforms. American Journal of Physics. 1981;49(5):466-72.

[22] Stoffer DS. Walsh-Fourier Analysis and its Statistical Applications. Journal of the American Statistical Association. 1991;86(414):461-79.

[23] Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. Nature. 2022;604(7904):175-83.

[24] Tareen A, Kooshkbaghi M, Posfai A, Ireland WT, McCandlish DM, Kinney JB. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. Genome biology. 2022;23(1):1-27.

[25] Otwinowski J, McCandlish DM, Plotkin JB. Inferring the shape of global epistasis. Proceedings of the National Academy of Sciences. 2018;115(32):E7550-8.

[26] Forcier TL, Ayaz A, Gill MS, Jones D, Phillips R, Kinney JB. Measuring cis-regulatory energetics in living cells using allelic manifolds. Elife. 2018;7:e40618.

[27] Kinney JB, Murugan A, Callan Jr CG, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proceedings of the National Academy of Sciences. 2010;107(20):9158-63.

[28] Zhou J, Wong MS, Chen WC, Krainer AR, Kinney JB, McCandlish DM. Higher-order epistasis and phenotypic prediction. Proceedings of the National Academy of Sciences. 2022;119(39):e2204233119.

[29] Zhou J, McCandlish DM. Minimum epistasis interpolation for sequence-function relationships. Nature communications. 2020;11(1):1-14.

[30] Faure AJ, Schmiedel JM, Baeza-Centurion P, Lehner B. DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. Genome Biology. 2020;21(1):1-23.

[31] Horn R, Johnson C. Topics in Matrix Analysis. Cambridge University Press; 1991.

Supporting information captions

S1 Fig. Supplementary figure related to Fig 2b,c.

S2 Fig. Supplementary figure related to Fig 2d.

S1 Text. Supplementary Methods.

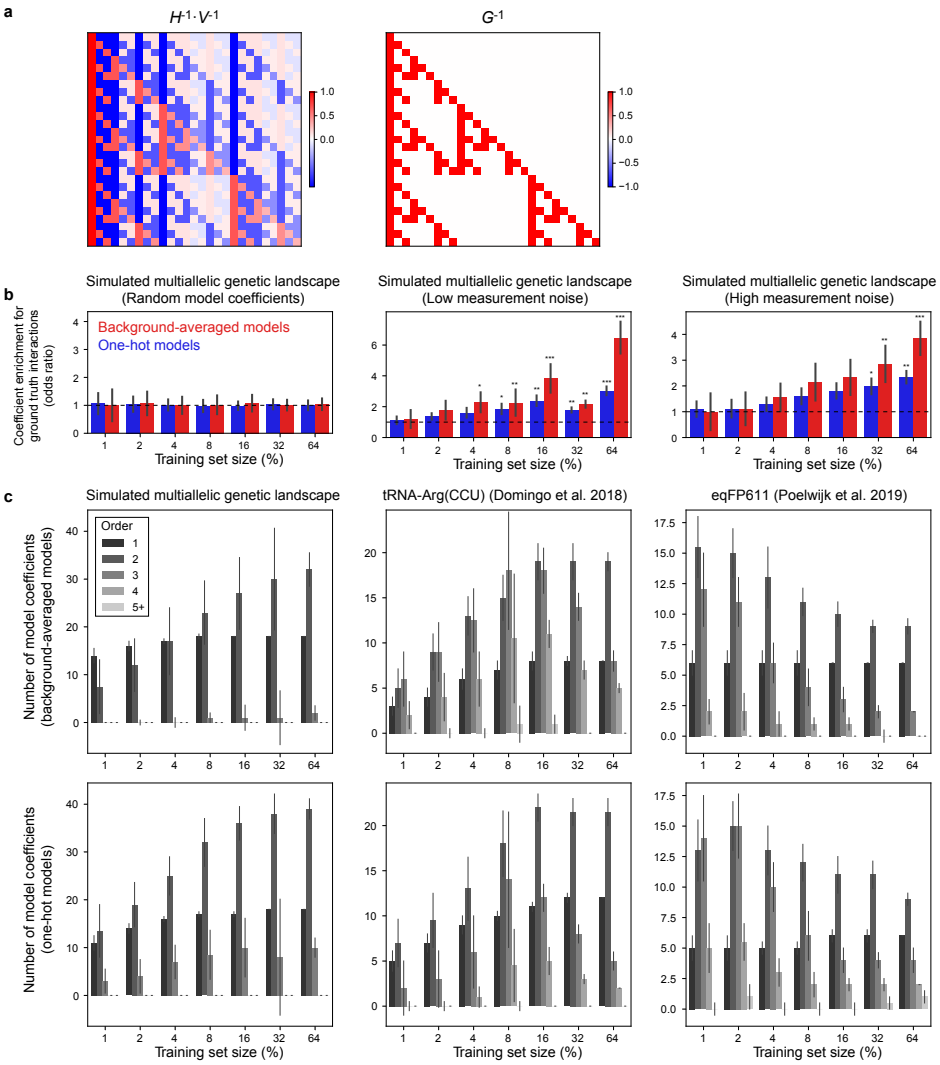

Figure S1: Supplementary figure related to Fig 2b,c. **a**, Cartoon depiction of alternative feature matrices for inferring epistatic coefficients by linear regression. G^{-1} in the right panel indicates the matrix of one-hot encoded sequence features – or embeddings – typically used when fitting models of genotype-phenotype landscapes [2]. The left panel represents the matrix of sequence features used to infer background-averaged epistatic coefficients, as in equation (12). **b**, Enrichment of direct physical interactions (see Fig 2a) in non-zero epistatic coefficients from sparse models of a simulated multiallelic fitness landscape. Left: results for random models with matching numbers of randomly selected epistatic coefficients at all coefficient orders (see Fig 2c, left). Middle, Right: results for sparse models fit to low and high measurement noise versions of the simulated fitness dataset, respectively. **c**, Numbers of non-zero espistatic coefficients of different orders in Lasso regression models inferred using different random fractions of the simulated and empirical combinatorial fitness landscapes indicated.

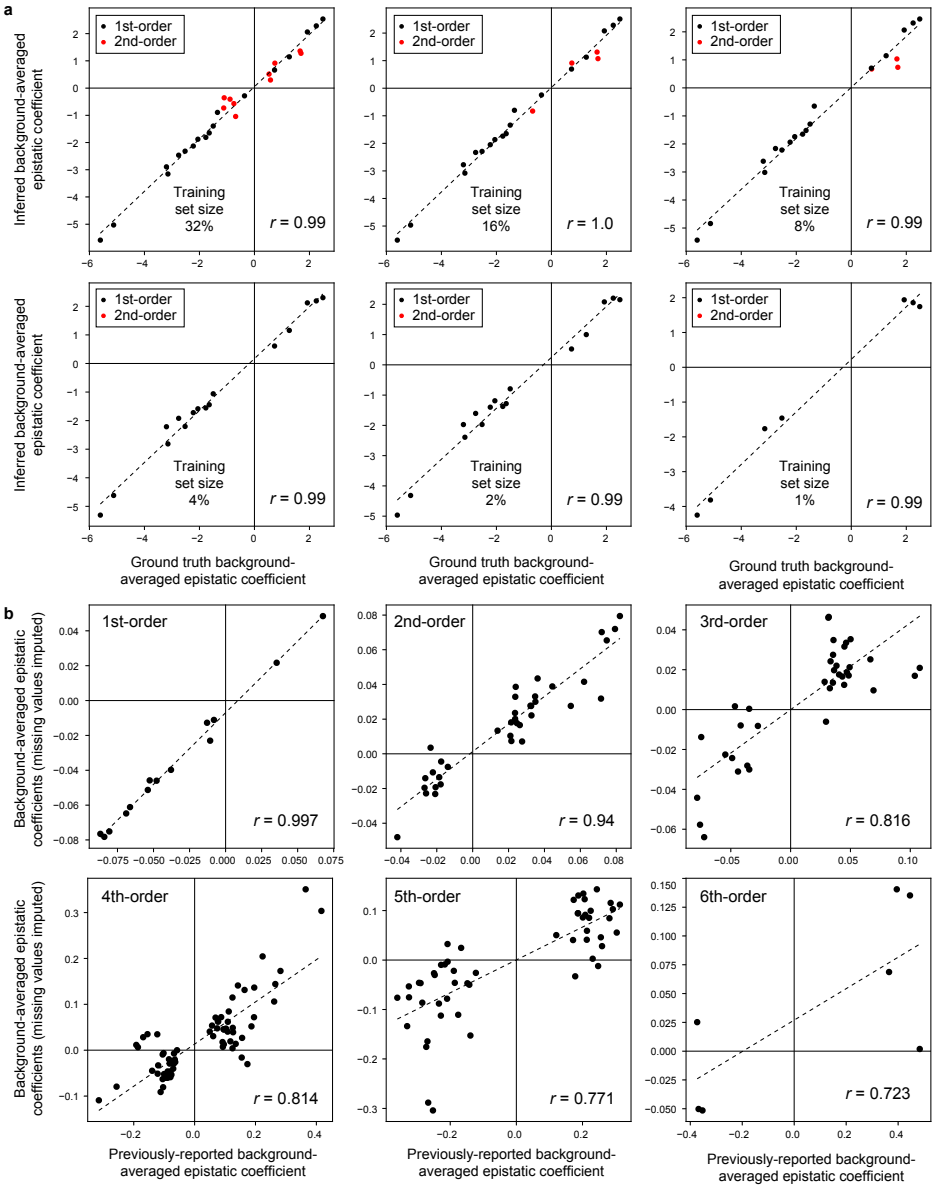

Figure S2: Supplementary figure related to Fig 2d. **a**, Scatter plots comparing sparse model-inferred background-averaged epistatic coefficients of a simulated multiallelic fitness landscape to ground truth values for varying training dataset sizes. **b**, Scatter plots comparing background-averaged epistatic coefficients obtained using the tRNA-Arg(CCU) DMS dataset and equation (9) to previously-reported significant coefficients [6] after imputing fitness values for missing variants (see Methods) and shown separately for coefficient orders 1-6. Only significant coefficients (FDR < 0.05) observed in at least 10 backgrounds are shown [6].

Supporting Information 1 - Supplementary Methods

Here we provide the proofs of the mathematical results shown in the main text.

Proposition 1. *Let us define the matrices A_n recursively as*

$$A_{n+1} = \frac{1}{s}\begin{pmatrix} A_n & A_n & A_n & \dots & A_n \\ A_n & (1-s)A_n & A_n & \dots & A_n \\ A_n & A_n & (1-s)A_n & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & A_n \\ A_n & A_n & \dots & A_n & (1-s)A_n \end{pmatrix} \qquad A_0 = 1 \qquad \textit{and} \qquad A_n \ \textit{is} \ s^n \times s^n. \tag{13}$$

For $n \in \mathbb{N}$, A_n is the inverse of the matrix H_n defined in equation (10).

Proof. Let us prove this by induction. For $n = 1$ we have

$$H_1 = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & \dots & 0 & -1 \end{pmatrix} \tag{26}$$

$$A_1 = \frac{1}{s}\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1-s & 1 & \dots & 1 \\ 1 & 1 & 1-s & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 \\ 1 & 1 & \dots & 1 & 1-s \end{pmatrix}. \tag{27}$$

The rows and columns of these two matrices can be described as follows:

$$(H_1)_{i\cdot} = \begin{cases} (1, 1, \dots, 1) & \text{if } i = 1 \\ (1, 0, \dots, 0, h_i, 0, \dots, 0) & \text{if } i \neq 1, \end{cases}$$

$$(A_1)_{\cdot j} = \frac{1}{s}\begin{cases} (1, 1, \dots, 1)^T & \text{if } j = 1 \\ (1, 1, \dots, 1, a_j, 1, \dots, 1)^T & \text{if } j \neq 1, \end{cases}$$

where $h_i := (H_1)_{ii} = -1 \ \forall i > 1$ and $a_j := (A_1)_{jj} = 1 - s \ \forall j > 1$.

Therefore,

$$(H_1 \cdot A_1)_{ij} = (H_1)_{i\cdot} \cdot (A_1)_{\cdot j} = \frac{1}{s}\begin{cases} s & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$H_1 \cdot A_1 = \frac{1}{s}\begin{pmatrix} s & 0 & \dots & 0 \\ 0 & s & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & s \end{pmatrix} = I_{s\times s},$$

where $I_{s\times s}$ is the identity matrix of size $s \times s$. Since both H_1 and A_1 are symmetric, it is also true that $A_1 \cdot H_1 = I_{s\times s}$. Therefore, A_1 is the inverse of H_1.

Assume that the hypothesis is true for a fixed $n \in \mathbb{N}$. Let us now prove that it is also true for $n+1$. Following the recursive definitions of H_{n+1} and A_{n+1} in equations (10) and (13), we can write the blocks of these matrices as follows:

$$(H_{n+1})_{[i][\cdot]} = \begin{cases} (H_n, H_n, \dots, H_n) & \text{if } i = 1 \\ (H_n, 0, \dots, 0, \tilde{h}_i, 0, \dots, 0) & \text{if } i \neq 1, \end{cases} \tag{28}$$

where $(H_{n+1})_{[i][j]}$ denotes the block at position i, j in H_{n+1} and $\tilde{h}_i := (H_{n+1})_{[i][i]} = -H_n \ \forall i > 1$;

$$(A_{n+1})_{[\cdot][j]} = \frac{1}{s}\begin{cases} (A_n, A_n, \dots, A_n)^T & \text{if } j = 1 \\ (A_n, A_n, \dots, A_n, \tilde{a}_j, A_n, \dots, A_n)^T & \text{if } j \neq 1, \end{cases} \tag{29}$$

where $(A_{n+1})_{[i][j]}$ denotes the block at position i, j in A_{n+1} and $\tilde{a}_j := s(A_{n+1})_{[j][j]} = (1-s)A_n \ \forall j > 1$. We can therefore write the block at position i, j of the product of these matrices as follows:

$$(H_{n+1} \cdot A_{n+1})_{[i][j]} = (H_{n+1})_{[i][\cdot]} \cdot (A_{n+1})_{[\cdot][j]} = \frac{1}{s}\begin{cases} sH_n \cdot A_n & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

According to the induction hypothesis we know that H_n and A_n are inverse matrices i.e. $H_n \cdot A_n = I_{s^n\times s^n}$. Therefore, the blocks on the diagonal are identity matrices and the blocks outside the diagonal are zeros. This means that $H_{n+1} \cdot A_{n+1} = I_{s^{n+1}\times s^{n+1}}$. Similarly, due to the symmetry of the matrices we can also prove that $A_{n+1} \cdot H_{n+1} = I_{s^{n+1}\times s^{n+1}}$.

We can then conclude that $A_n = H_n^{-1}$ for every $n \in \mathbb{N}$. □

The proof above can be greatly simplified with the use of the Kronecker product and its properties as follows:

Proof. It is straightforward to observe that the recursive notation of H_{n+1} can be expressed as a Kronecker product between H_1 and H_n. With this, we can use that the inverse of the Kronecker product is the product of the inverses of the individual matrices, if and only if, the individual matrices are invertible [31]. In our case, it is easy to see that the latter is true because we have shown that H_1 is invertible (we computed its inverse in the previous proof), and all the other matrices are compute as a Kronecker product that derive uniquely from H_1 (H_2 is invertible because $H_2 = H_1 \otimes H_1$, H_3 is invertible because $H_3 = H_1 \otimes H_2$, and so forth. Therefore,

$$A_{n+1} = (H_{n+1})^{-1} = (H_1 \otimes H_n)^{-1} = H_1^{-1} \otimes H_n^{-1} = A_1 \otimes A_n$$

Now, we only need to compute the inverse of H_1 as we did at the beginning of the original proof, and we obtain the recurrent definition of A_{n+1} (13) by applying the Kronecker product between A_1 as in (27) and A_n. □

Proposition 2. *The elements of H_n can be written as*

$$(H_n)_{ij} = \begin{cases} (-1)^{(E_n)_{ij}} & \text{if } (M_n)_{ij} = n \\ 0 & \text{otherwise,} \end{cases}$$

where M and E are $s^n \times s^n$ matrices whose elements are

$$(E_n)_{ij} = \sum_{\substack{k=1 \\ i_k \cdot j_k > 0}}^{n} \delta_{i_k j_k} \tag{15}$$

$$(M_n)_{ij} = \sum_{\substack{k=1 \\ i_k \cdot j_k > 0}}^{n} \delta_{i_k j_k} + \sum_{\substack{k=1 \\ i_k \cdot j_k = 0}}^{n} 1 = (E_n)_{ij} + \sum_{\substack{k=1 \\ i_k \cdot j_k = 0}}^{n} 1,$$

where δ_{ij} denotes the Kronecker delta of i, j.

Proof. Let us prove the formula by induction. For $n = 1$ and any given s, H_n is given by equation (26). Therefore, we can write $(E_1)_{ij}$ and $(M_1)_{ij}$ as follows:

$$(M_1)_{ij} = \begin{cases} 1 & \text{if } i = 1 \text{ or } j = 1 \\ 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \neq 1 \end{cases}$$

$$(E_1)_{ij} = \begin{cases} 1 & \text{if } i = j \neq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

Therefore, since $n = 1$, $(M_1)_{ij} = n = 1$ only when either $i = 1$, $j = 1$ or $i = j$. In the rest of the cases $(M_1)_{ij} \neq n = 1$ and, according to the formula, the elements of the matrix will be 0. Now, for the cases where $(M_1)_{ij} = n = 1$, we need to check the value of $(E_1)_{ij}$. We can see how $(E_1)_{ij} = 1$ only when $i = j$ and they are different from 1. This means that all the elements of the diagonal of H_1, except the first one, will be $(-1)^1 = -1$ and the first row and first columns will have $(-1)^0 = 1$. The rest of the elements correspond to $(M_1)_{ij} \neq n = 1$ so they will be filled with zeros. Putting all this together, we find the expression as H_1 from equation (26).

Assume now that the expression is true for a fixed $n \in \mathbb{N}$ and let us prove it for $n+1$. In this case, the matrix of H_{n+1} is defined by blocks (see equation (10)). We first define the indices $P \in \{1, \dots, s\}$ and $Q \in \{1, \dots, s\}$ for row and column blocks, respectively. The first matrix block of H_{n+1} corresponds to $P = 1$ and $Q = 1$. The corresponding blocks of the matrices M_{n+1} and E_{n+1}, which are necessary for the derivation of the block in H_{n+1}, are computed by comparing the genotype sequences $\vec{p}, \vec{q} \in \mathcal{S}^{n+1}$ for which $\vec{p} = (0, i_1, \dots, i_n) := 0 \frown \vec{i}$ for $\vec{i} \in \mathcal{S}^n$ and $\vec{q} = 0 \frown \vec{j}, \vec{j} \in \mathcal{S}^n$. More generally, the block in the Pth position with respect to the rows and Q^{th} position with respect to the columns can be obtained by comparing the genotype sequences $\vec{p}, \vec{q} \in \mathcal{S}^{n+1}$ for which $\vec{p} = (P-1) \frown \vec{i}, \vec{i} \in \mathcal{S}^n$ and $\vec{q} = (Q-1) \frown \vec{j}, \vec{j} \in \mathcal{S}^n$. See below for a visual description of the notation.

From these observations, it can easily be deduced that for any $\vec{p} = (P-1) \frown \vec{i}, \vec{q} = (Q-1) \frown \vec{j}$ with $\vec{i}, \vec{j} \in \mathcal{S}^n$, $P-1, Q-1 \in \mathcal{S}$,

$$(M_{n+1})_{pq} = \begin{cases} (M_n)_{ij} + 1 & \text{if } P = 1 \text{ or } Q = 1 \\ (M_n)_{ij} + 1 & \text{if } P = Q \\ (M_n)_{ij} & \text{if } P \neq Q \neq 1 \end{cases} \tag{31}$$

$$(E_{n+1})_{pq} = \begin{cases} (E_n)_{ij} + 1 & \text{if } P = Q \neq 1 \\ (E_n)_{ij} & \text{otherwise,} \end{cases} \tag{32}$$

where $i = p - s^n(P-1)$, $j = q - s^n(Q-1)$, $P = \lceil p/s^n \rceil$ and $Q = \lceil q/s^n \rceil$, with $\lceil . \rceil$ denoting the ceiling function. A visual description of the notation of the block structure of the matrices is given by

$$
X_{n+1} \quad = \quad \text{row block } P\left\{\begin{array}{c}\\ \\ \\ \\ \\ \end{array}\right. \begin{array}{c}
\begin{array}{ccccc} & & \overbrace{\qquad\qquad}^{\text{column block } Q} & & \end{array} \\
\begin{pmatrix}
X_n^{11} & \cdots & X_n^{1Q} & \cdots & X_n^{1s} \\
\vdots & \ddots & \vdots & & \vdots \\
X_n^{P1} & \cdots & X_n^{PQ} & \cdots & X_n^{Ps} \\
\vdots & & \vdots & \ddots & \vdots \\
X_n^{s1} & \cdots & X_n^{sQ} & \cdots & X_n^{ss}
\end{pmatrix} \\
\begin{array}{ccccc} & & \uparrow & & \\ & & \text{column } q & & \end{array}
\end{array} \begin{array}{l} \\ \\ \leftarrow \text{row } p \\ \\ \\ \end{array}
$$

Here X_{n+1} denotes any generic matrix following the structure of the matrices in equations (10), (11), (13), (31) and (32).

Now, similar to the case $n = 1$, we have that $(M_{n+1})_{pq} = n + 1$ only when $(M_n)_{ij} = n$ and either $P = 1$, $Q = 1$ or $P = Q$, which corresponds to the newly added state being $p_1 = P - 1 = 0$, $q_1 = Q - 1 = 0$ or $p_1 = q_1$. In the rest of the cases $(M_{n+1})_{pq} \neq n + 1$ and the elements of the matrix H_{n+1} will be 0. Now, for the cases where $(M_{n+1})_{pq} = n+1$, we will have that $(E_{n+1})_{pq}$ has either the same value of the corresponding entry in E_n or it will be increased by 1. This means, that when $P = Q \neq 1$, the sign of the entry in H_{n+1} will be inverted. Otherwise, the sign of the element of the matrix stays the same. With this we prove the formula for H_{n+1}, since we have shown how we can find the same block structure as in equation (10). By induction, we can conclude that the formula holds for every $n \in \mathbb{N}$. □

Proposition 3. *The elements of A_n can be written as*

$$(A_n)_{ij} = \frac{1}{s^n}(1 - s)^{(E_n)_{ij}}, \tag{17}$$

where E_n is defined as in equation (15).

Proof. Let us prove the formula by induction.

For any given s, A_1 is defined in equation 27. Its diagonal elements are equal to $(1 - s)/s$ for $i > 1$, $1/s$ for $i = 1$ and its off-diagonal elements are equal to $1/s$. It can easily be observed from equation (30) that the RHS of equation (17) is equal to $(1 - s)/s$ for $i = j \neq 1$ and to $1/s$ otherwise, so equation (17) is true for $n = 1$.

Assume now that equation (17) is true for a fixed $n \in \mathbb{N}$ and let us prove it for $n + 1$. We use the recursive definition of A_{n+1} in equation (13) and the block representation of the genotype sequences as in the proof of (2).

Let us start with the first block in the diagonal of A_{n+1}, i.e. $P = 1$ and $Q = 1$, where the entries of the corresponding block in E_{n+1} are derived from comparisons of pairs of genotype sequences of the form $\vec{p} = 0 \frown \vec{i}, \vec{q} = 0 \frown \vec{j}$ for $\vec{i}, \vec{j} \in \mathcal{S}^n$. From equation (13), this block is equal to $\frac{1}{s}A_n$, so writing $i = p \bmod s^n$ and $j = q \bmod s^n$, we have

$$(A_{n+1})_{pq} = \frac{1}{s}(A_n)_{ij} = \frac{1}{s^{n+1}}(1-s)^{(E_n)_{ij}}.$$

From equation (32), $(E_{n+1})_{pq} = (E_n)_{ij}$, which yields the desired result.

Now let us consider the elements in the other diagonal blocks of A_{n+1}, where the entries correspond to pairs of genotype sequences of the form $\vec{p} = (P-1) \frown \vec{i}, \vec{q} = (Q-1) \frown \vec{j}$ with $P = Q \neq 1$ and $\vec{i}, \vec{j} \in \mathcal{S}^n$. From equation (13), this block is equal to $\frac{1}{s}(1-s)A_n$, i.e.

$$(A_{n+1})_{pq} = \frac{1}{s}(1-s)(A_n)_{ij} = \frac{1}{s^{n+1}}(1-s)^{(E_n)_{ij}+1}.$$

From equation (32), $(E_{n+1})_{pq} = (E_n)_{ij} + 1$, which yields the desired result.

Finally, let us consider the elements in the off-diagonal blocks of A_{n+1}, where the entries correspond to pairs of genotype sequences of the form $\vec{p} = (P-1) \frown \vec{i}, \vec{q} = (Q-1) \frown \vec{j}$ with $P \neq Q$ and $\vec{i}, \vec{j} \in \mathcal{S}^n$. From equation (13), this block is equal to $\frac{1}{s}A_n$, so we have

$$(A_{n+1})_{pq} = \frac{1}{s}(A_n)_{ij} = \frac{1}{s^{n+1}}(1-s)^{(E_n)_{ij}}.$$

From equation (32), $(E_{n+1})_{pq} = (E_n)_{ij}$, which completes the proof. □

Proposition 4. *The matrices V_n and V_n^{-1} are diagonal matrices whose diagonal elements can be written as*

$$(V_n)_{ii} = (-1)^{n-W_n(\vec{i})}\frac{1}{s^{W_n(\vec{i})}} \tag{18}$$

and

$$(V_n^{-1})_{ii} = (-1)^{n-W_n(\vec{i})}s^{W_n(\vec{i})}, \tag{19}$$

where

$$W_n(\vec{i}) := \sum_{k=1}^{n} w_k, \text{ with } w_k := \delta_{i_k 0}$$

and $\vec{i}$ again denotes the ith element in $\mathcal{S}^n$ when ordered by the base s representation of integers.

Proof. Let us prove equation (18) by induction, equation (19) follows directly.

One can easily check from equation (11) that the formula holds for $n = 1$. Let us now assume that equation (18) is true for a fixed $n \in \mathbb{N}$ and let us prove it for $n+1$. We use the recursive definition of V_{n+1} in equation (11). Let us consider the element $(V_{n+1})_{pp}$. If $\vec{p} = 0 \frown \vec{i}$, this corresponds to the first block of V_{n+1}, i.e. $P = 1$, where the elements are multiplied by $1/s$ and $W_{n+1}(\vec{p}) = W_n(\vec{i}) + 1$, so

$$(V_{n+1})_{pp} = \frac{1}{s}(V_n)_{ii} = (-1)^{n-W_n(\vec{i})}\frac{1}{s^{W_n(\vec{i})+1}} = (-1)^{n+1-W_{n+1}(\vec{p})}\frac{1}{s^{W_{n+1}(\vec{p})}}.$$

Similarly, if $\vec{p} = (P-1) \frown \vec{i}$ and $P > 1$, i.e. for the other diagonal blocks, from the recursive formula the elements are multiplied by -1 and $W_{n+1}(\vec{p}) = W_n(\vec{i})$, so by writing again $i = p - s^n(P-1)$, where $P = \lceil p/s^n \rceil$, we have

$$(V_{n+1})_{pp} = -(V_n)_{ii} = (-1)^{1+n-W_n(\vec{i})} \frac{1}{s^{W_n(\vec{i})}} = (-1)^{n+1-W_{n+1}(\vec{p})} \frac{1}{s^{W_{n+1}(\vec{p})}},$$

which completes the proof. □

Proposition 5. *The matrix A_n defined in equation (21) is the inverse of the matrix H_n in the general case where each position can have a different number of states.*

Proof. Let us prove by induction that $H_n \cdot A_n = I$ where I is the identity matrix of the corresponding size. Since H_n and A_n are symmetric, this would imply that $A_n \cdot H_n = I$ as well, and therefore, $A_n = H_n^{-1}$.

The case $n = 1$ corresponds exactly to the case $n = 1$ of the proof of Proposition 1 by setting $s = s_1$. Therefore, $H_1 A_1 = I_{s_1 \times s_1}$, and A_1 is the inverse of H_1.

Now, assume the hypothesis is true for a fixed $n \in \mathbb{N}$ and let us prove that this is also true for $n+1$. We can write the rows and columns of the matrices H_{n+1} and A_{n+1} as equation (28) and equation (29), respectively. The only difference is that we need to replace s by s_{n+1} and the size of the matrices is different. Following exactly the same derivation as in Proposition 1 we can conclude that $H_{n+1} A_{n+1} = I$ and this proves by induction that $A_n = H_n^{-1}$. □

Proposition 6. *In this general case, the elements of H_n and A_n can be written as*

$$(H_n)_{ij} = \begin{cases} (-1)^{(E_n)_{ij}} & \text{if } (M_n)_{ij} = n \\ 0 & \text{otherwise} \end{cases}$$

$$(A_n)_{ij} = \frac{\prod_{k=1}^{n} \left(1 - s_k\right)^{e_k}}{\prod_{k=1}^{n} s_k},$$

where E_n and M_n are defined as in equation (15) and $e_k = \begin{cases} 1 & \text{if } i_k = j_k \neq 1 \\ 0 & \text{otherwise} \end{cases}$.

Proof. The proof follows directly from the proofs of Propositions 2 and 3. The only difference in the induction step is that s is replaced by s_{n+1}. □

Proposition 7. *The matrices V_n and V_n^{-1} are diagonal matrices whose diagonal elements can be written as*

$$(V_n)_{ii} = (-1)^{n-W_n(\vec{i})} \prod_{k=1}^{n} \left(\frac{1}{s_k}\right)^{w_k}$$

and

$$(V_n^{-1})_{ii} = (-1)^{n-W_n(\vec{i})} \prod_{k=1}^{n} {s_k}^{w_k},$$

where

$$W_n(\vec{i}) := \sum_{k=1}^{n} w_k, \text{ with } w_k := \delta_{i_k 0}.$$

Proof. The proof follows the same steps as the proof of Proposition 4. The only difference in the induction step is that s is replaced by s_{n+1}. □

Proposition 8. *Assuming that the uncertainty or empirical error of each component of $\bar{y}_n$ is independent of the error of the other components, we can propagate the error to the estimation of the coefficients $\bar{\epsilon}_n$ as follows:*

$$\bar{\sigma}_{\epsilon_n} = \sqrt{(V_n H_n) \circ (V_n H_n)(\bar{\sigma}_{y_n} \circ \bar{\sigma}_{y_n})} \tag{33}$$

where $\bar{\sigma}_{y_n}$ and $\bar{\sigma}_{\epsilon_n}$ denote the vector of uncertainties of the phenotypes and the epistatic coefficients, respectively, and $\circ$ denotes the element-wise product.

Proof. According to equation (9) we can compute the epistatic coefficients as

$$\bar{\epsilon}_n = V_n H_n \bar{y}_n$$

This means that each ϵ is a linear combination of the elements of $\bar{y}_n$, and the coefficients of the linear combination are defined by the elements of $V_n H_n$. Writing the formula as a system of equations and using that the error propagation of a linear function is just the square root of the sum of the squares of each coefficient times the variance of each variable, we obtain:

$$\begin{aligned}
\sigma_{\epsilon_1} &= \sqrt{([V_n H_n]_{11})^2 \sigma_{y_1}^2 + \cdots + ([V_n H_n]_{1N})^2 \sigma_{y_N}^2} \\
&\vdots \\
\sigma_{\epsilon_N} &= \sqrt{([V_n H_n]_{N1})^2 \sigma_{y_1}^2 + \cdots + ([V_n H_n]_{NN})^2 \sigma_{y_N}^2}
\end{aligned}$$

which can be written with matrix notation as

$$\bar{\sigma}_{\epsilon_n} = \sqrt{(V_n H_n) \circ (V_n H_n)(\bar{\sigma}_{y_n} \circ \bar{\sigma}_{y_n})}$$

□

This equation has already been used previously in articles regarding standard $s = 2$ epistasis [2], and it also applies here for any arbitrary n and s with our definition of V_n and H_n.

For the example given in the main text regarding tRNA-ARG(CCU) with $n = 2$ and $s = 3$, we can see that applying the formula above we obtain:

$$
\begin{aligned}
\sigma_{\epsilon_{(*,*)}} &= \sqrt{\frac{\sum_k^9 \sigma_{y_k}^2}{9}} \\
\sigma_{\epsilon_{(*,1)}} &= \sqrt{\frac{\sigma_{y_1}^2 + \sigma_{y_2}^2 + \sigma_{y_4}^2 + \sigma_{y_5}^2 + \sigma_{y_7}^2 + \sigma_{y_8}^2}{3}} \\
\sigma_{\epsilon_{(1,1)}} &= \sqrt{\sigma_{y_1}^2 + \sigma_{y_2}^2 + \sigma_{y_4}^2 + \sigma_{y_5}^2}
\end{aligned}
$$

where the other terms have the same shape as the corresponding same order term above, but with different phenotypes involved. We can see that if we assume equal uncertainties the error of the epistasis coefficients increase with the order of the interaction.