















RESEARCH ARTICLE

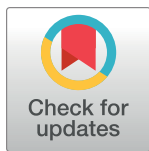
An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape

Victoria O. Pokusaeva¹ , Dinara R. Usmanova² , Ekaterina V. Putintseva¹ , Lorena Espinar^{3,4} , Karen S. Sarkisyan^{1,5,6} , Alexander S. Mishin⁵ , Natalya S. Bogatyreva^{3,4,7} , Dmitry N. Ivankov^{1,7} , Arseniy V. Akopyan¹ , Sergey Ya. Avvakumov¹ , Inna S. Povolotskaya⁸ , Guillaume J. Filion^{3,4} , Lucas B. Carey^{4,9*} , Fyodor A. Kondrashov^{1*} 

1 Institute of Science and Technology Austria, Am Campus 1, Klosterneuburg, Austria, **2** Department of Systems Biology, Columbia University, New York, NY, United States of America, **3** Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), 88 Dr. Aiguader, Barcelona, Spain, **4** Universitat Pompeu Fabra (UPF), Barcelona, Spain, **5** Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia, **6** Medical Research Council London Institute of Medical Sciences, Imperial College London, London, United Kingdom, **7** Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, Pushchino, Moscow region, Russia, **8** Veltischev Research and Clinical Institute for Pediatrics of the Pirogov Russian National Research Medical University, Moscow, Russia, **9** Center for Quantitative Biology and Peking-Tsinghua Joint Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China

 These authors contributed equally to this work.

* lucas.carey@gmail.com (LBC); fyodor.kondrashov@ist.ac.at (FAK)



 OPEN ACCESS

Citation: Pokusaeva VO, Usmanova DR, Putintseva EV, Espinar L, Sarkisyan KS, Mishin AS, et al. (2019) An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS Genet* 15(4): e1008079. <https://doi.org/10.1371/journal.pgen.1008079>

Editor: Dmitri A. Petrov, Stanford University, UNITED STATES

Received: August 21, 2018

Accepted: March 11, 2019

Published: April 10, 2019

Copyright: © 2019 Pokusaeva et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The raw and processed data have been submitted to the NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE99990. A virtual machine containing a running version of the data processing pipeline is available as a Docker image <https://hub.docker.com/r/gui11aume/epi/>. The scripts to reproduce the figures are on Github at <https://github.com/Lcarey/HIS3InterspeciesEpistasis>.

Abstract

Characterizing the fitness landscape, a representation of fitness for a large set of genotypes, is key to understanding how genetic information is interpreted to create functional organisms. Here we determined the evolutionarily-relevant segment of the fitness landscape of His3, a gene coding for an enzyme in the histidine synthesis pathway, focusing on combinations of amino acid states found at orthologous sites of extant species. Just 15% of amino acids found in yeast His3 orthologues were always neutral while the impact on fitness of the remaining 85% depended on the genetic background. Furthermore, at 67% of sites, amino acid replacements were under sign epistasis, having both strongly positive and negative effect in different genetic backgrounds. 46% of sites were under reciprocal sign epistasis. The fitness impact of amino acid replacements was influenced by only a few genetic backgrounds but involved interaction of multiple sites, shaping a rugged fitness landscape in which many of the shortest paths between highly fit genotypes are inaccessible.

Author summary

An intuitive understanding of protein evolution dictates that, with the exception of adaptive substitutions, amino acid states should be freely exchangeable between the same gene from different species. However, the extent to which this assertion holds true has not been tested in a controlled experiment. Here, we show that whether an amino acid state can be

Funding: The work was supported by HHMI International Early Career Scientist Program (55007424), the MINECO (BFU2012-31329, BFU2012-37168, BFU2015-68351-P and BFU2015-68723-P), Spanish Ministry of Economy and Competitiveness Centro de Excelencia Severo Ochoa 2013-2017 grant (SEV-2012-0208), the Unidad de Excelencia María de Maeztu funded by the MINECO (MDM-2014-0370), Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat AGAUR program (2014 SGR 0974), the CERCA Programme of the Generalitat de Catalunya, Russian Foundation for Basic Research grant (18-04-01173), the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie programme (665385) and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013, ERC grant agreement 335980_EinME and Synergy Grant 609989). KSS was supported by EMBO long-term fellowship (ALTF 107-2016). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

exchanged between orthologues depends on other amino acid states in the same protein. Furthermore, we show that the mode of interaction of amino acid states is multidimensional. Assuming that amino acid replacements influence the protein in several independent ways substantially improves our ability to predict the effect of an amino acid state in a protein sequence that has not been observed in nature.

Introduction

Predicting function and fitness of organisms from their genotypes is the ultimate goal of many fields in biology, from medical genetics to systems biology to the study of evolution [1–5]. Among the conceptual frameworks for understanding the genotype-to-phenotype connection is the fitness landscape, which assigns a fitness (phenotype) to every possible genotype (sequence) of a gene or genome under consideration [4,6]. The recognition of the importance of the fitness landscape stimulated the development of a variety of theoretical approaches to describe it, including its general shape and epistatic interactions between alleles, a key property which determines the complexity of the fitness landscape (see [4] and references within). Before the advent of next-generation sequencing, experimental assays of the fitness landscape were few and could not address the issue at the sequence level. Recently, large-scale experimental assays described the shape of the fitness landscape a few mutations away from a local fitness peak (see [7–10] and references within). Also, some assays involving a smaller number of genotypes considered combinations of mutations with established functional [11–17] or evolutionary [18–25] significance.

Empirical evidence of the nature of large-scale fitness landscapes mostly comes from the study of genotypes incorporating random mutations [4,7–10], the majority of which are deleterious [7–10,26]. Thus, our present knowledge of fitness landscapes is primarily driven by the study of deleterious mutations and their interactions, although local adaptive trajectories have also been considered [2,4,16,27–29]. Deleterious mutations were found to engage in synergistic epistasis, whereby the joint effect of multiple mutations was stronger than the sum of their individual effects [4,7–10,16]. Furthermore, sign epistasis among random mutations was mostly rare [5,7–10,16,30], although some of these conclusions differ from study to study (e.g. see [4,30]).

Unfortunately, there are fundamental limitations to assaying the fitness landscape on a large or macroevolutionary level with random mutation libraries. The number of genotypes underlying the fitness landscape is the combinatorial set of all amino acids across the length of the protein [4,6]. For example, for the 220 amino acid protein coded by the His3 gene in *Saccharomyces cerevisiae*, the fitness landscape is a 220-dimensional genotype space with 20^{220} different possible sequences. Such immense spaces are both computationally and experimentally intractable. Fortunately, it may not be necessary to survey all genotypes to study the evolutionary-relevant section of the fitness landscape. Because the vast majority of mutations in protein sequences are deleterious [26], a randomly sampled protein sequence is non-functional [31,32].

Here we propose an evolutionary approach for assaying fitness landscapes on a macroevolutionary scale in a high-throughput manner that avoids the random sampling of mostly non-functional sequences. The functionally and evolutionarily relevant section of the fitness landscape can be represented by the combination of extant amino acid states, those found in extant species. This approach applied previously on a limited scale [18–25] mitigates the problem of exploring a prohibitively large fitness landscape while highlighting the relationships between

evolutionarily-relevant genotypes (Fig 1A). Crucially, substitutions that have been fixed in evolution are fundamentally different from random mutations [26], the former are either neutral or beneficial in at least some genetic contexts and represent the driving force of molecular evolution, while the latter are mostly deleterious and are primarily relevant on a microevolutionary scale. Therefore, current empirical data do not shed much light on the impact of interactions between amino acid states that were fixed in the course of evolution by natural selection. Combinations of extant amino acid states represent the area of the sequence space that considers all possible orders in which amino acid substitutions from evolution could have happened allowing one to assay a much wider functionally relevant area of the sequence space than approaches based on random mutagenesis of a single sequence (Fig 1B and 1C).

Results

Estimating fitness of evolutionary-relevant genotypes

We studied His3, a gene coding for imidazoleglycerol-phosphate dehydratase (IGPD, His3p), an enzyme essential for histidine synthesis. In a multiple alignment of His3 orthologues from 21 yeast species we identified 686 extant amino acid states (S1 Supporting Information), which were evenly distributed across the His3p structure (Fig 1D). These 686 states, which represent the end product of ~400 million years of evolution [33] (Fig 1B and 1C) correspond to $\sim 10^{83}$ sequences, even a tiny fraction of which would be too many to survey. Thus, we sectioned His3 into 12 independent segments such that the full combinatorial set of amino acid states that occurred in His3 during yeast evolution comprised 10,000-100,000 genotypes per segment (see Methods and S1A Fig). The 12 segments were of similar length, constrained by the molecular methods employed for library construction (see Methods), and covered a diverse range of secondary structures and functional elements (S1C Fig). For each of the 12 segments of His3 we performed an independent experiment surveying its fitness landscape. For each segment we used degenerate oligonucleotides to construct genotypes consisting of combinations of amino acids present in extant His3 sequences, and determined the fitness conferred by these genotypes by expressing them in a $\Delta his3$ strain of *S. cerevisiae* and measuring the rate of growth (S1B Fig). This way for each segment we assayed the fitness landscape of the genotype space that was traversed over the course of the last ~400 million years of evolution [33].

The segmentation of the His3 protein into 12 segments at first glance appears not to be relevant to understanding the fitness landscape of the entire protein. Indeed, it would have been more informative to survey combinations of extant amino acid states across the entire His3 sequence. Our choice to study independent segments was dictated by our preference for depth over width, in other words, we were more curious to consider combinations of extant amino acid states from distant orthologues rather than consider combinations of extant amino acids states from a few closely related species across the entire gene. Our choice of selecting random segments across the protein is justified in the same way as surveys of individual genes; segments may interact to form a functional protein while genes interact to contribute to the overall organismal fitness. Indeed, the lessons learned from fitness landscapes of random protein segments are likely to be scalable to the level of the entire protein to a greater extent than how fitness landscapes of an entire protein can be scaled to understand the fitness landscape of a genome.

Across 11 experiments, we measured fitness for a total of 4,018,105 genotypes (875,151 unique amino acid sequences) with high accuracy (S2 Fig). Of these, 422,717 consist solely of combinations of extant amino acid states from His3 orthologues, while the remaining genotypes incorporate other amino acid changes (S2 Supporting Information and Methods).

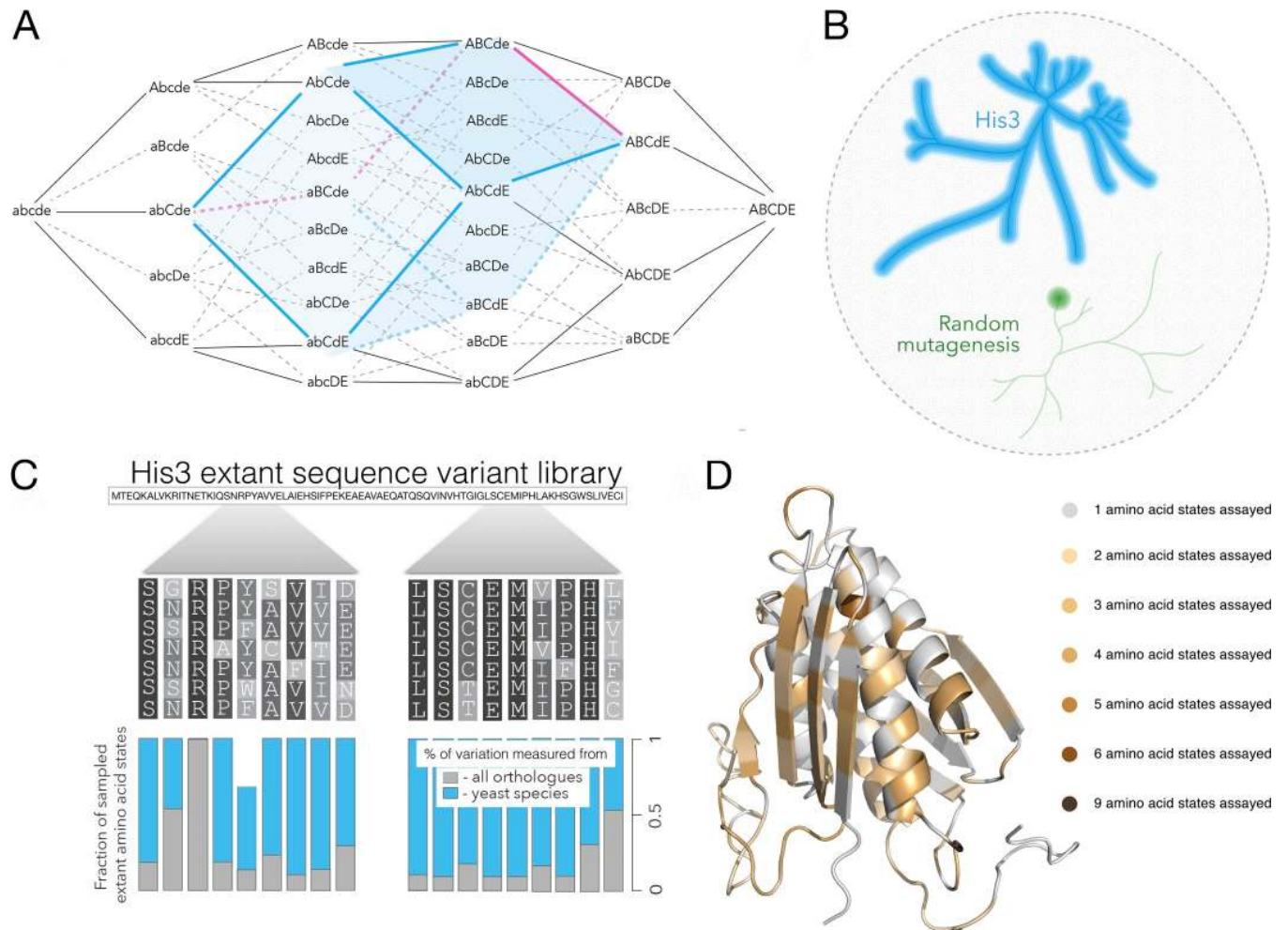


Fig 1. Combinatorial approach to the study of fitness landscapes. A fitness landscape is the representation of fitness for all possible genotypes composed of a specific set of loci. a, Following Fig 1 from Sewall Wright ref. [6] consider the genotype space consisting of 5 loci, each with two allele states (lower and uppercase letters). The entire genotype space is 5-dimensional consisting of 25 genotypes. Given two genotypes found in extant species (abCde and ABCdE in this example), surveying combinations of extant alleles substantially reduces the dimensionality of the genotype space, concomitantly reducing the number of genotypes to assay. The surveyed area (blue cube) considers all combinations of allele substitutions that have occurred in the course of evolution between the two sequences (red line), avoiding the sampling of combinations with less relevance to the evolutionary trajectory (black lines). b, Given the entire multidimensional genotype space (black circle) our approach considers a multidimensional subspace consisting of the combinatorial set of amino acid states from extant species. The blue line represents the yeast phylogeny and the surrounding blue space represents a multidimensional set of combinations of extant amino acids of the sequence under consideration, one His3 gene segment in our study. By contrast, random mutagenesis studies consider only a local segment of the genotype space surrounding a specific genotype (green circle). c, A multiple alignment of orthologous sequences of His3 for segment 2 for which we incorporated almost all extant amino acid states from 21 yeast species (blue bars) and 10-100% extant states from a set of 396 orthologues (grey bars). d, The predicted structure of His3p with amino acid residues that were substituted in our library.

<https://doi.org/10.1371/journal.pgen.1008079.g001>

Throughout the study, we considered the logarithm of fitness to allow for the sum of the impact of individual amino acid replacements on fitness to represent the fitness function. For one segment, 9, the accuracy of our experiment was low, and it was not used in cumulative analyses. For each segment we measured fitness for 60% - 99.8% of all possible genotypes from the combinatorial set of selected extant amino acid states found in 21 yeast species and a smaller fraction of combinations found across all domains of life ([S2 Supporting Information](#)), characterizing the evolutionary relevant fitness landscape ([Fig 1B](#)). For segment 3 for instance, 11 out of 17 amino acid sites had more than one extant amino acid state: L145 = 2, L147 = 2, Q148 = 3, K151 = 2, V152 = 2, D154 = 3, L164 = 3, E165 = 4, A168 = 2, E169 = 4,

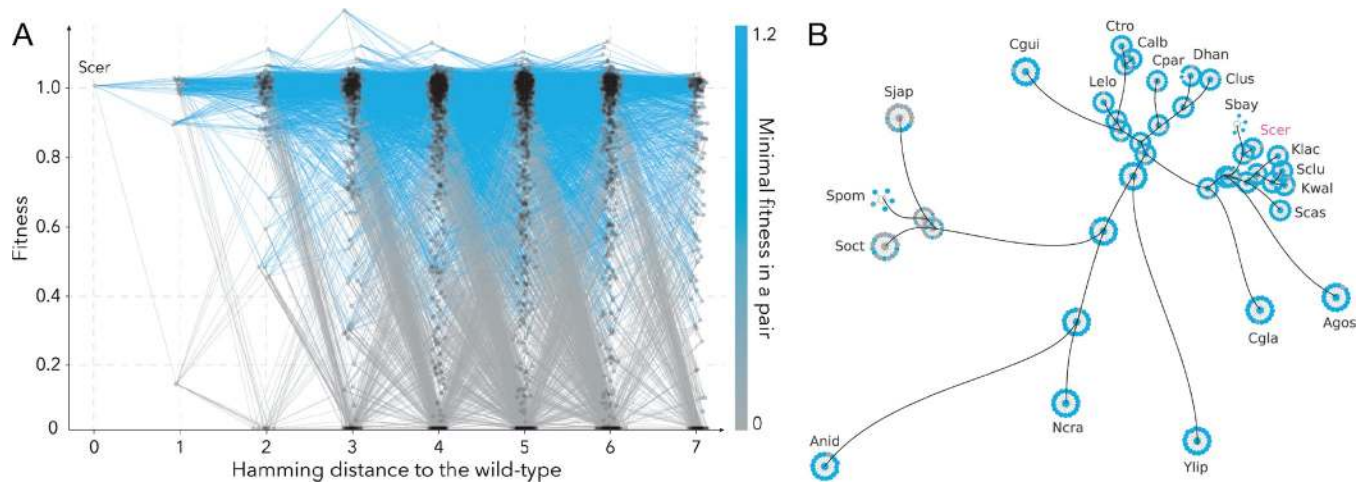


Fig 2. Visual representations of the fitness landscape. a, The fitness landscape for all assayed genotypes in segment 7. Nodes represent unique amino acid sequences with edges connecting those separated by a single amino acid replacements. Colour saturation represents the minimum fitness of the two connected nodes. b, For segment 7, fitness of ancestral and extant nodes and genotypes one amino acid replacement away from the nodes in the background of *S. cerevisiae* gene on the yeast phylogeny (black lines), are shown in colour ranging from grey (lowest fitness) to blue (highest fitness). The abbreviations represent the following species: Scer: *Saccharomyces cerevisiae*, Soct: *Schizosaccharomyces octosporus*, Sbay: *Saccharomyces bayanus*, Cgla: *Candida glabrata*, Scas: *Saccharomyces (Naumovozyma) castellii*, Kwal: *Kluyveromyces waltii*, Klac: *Kluyveromyces lactis*, Sklu: *Saccharomyces (Lachancea) kluyveri*, Agos: *Ashbya gossypii*, Clus: *Clavispora (Candida) lusitanae*, Dhan: *Debaryomyces hansenii (Candida famata)*, Cgui: *Candida (Pichia) guilliermondii*, Ctro: *Candida tropicalis*, Calb: *Candida albicans*, Cpar: *Candida parapsilosis*, Lelo: *Lodderomyces elongisporus*, Ylip: *Yarrowia lipolytica*, Anid: *Aspergillus nidulans*, Ncra: *Neurospora crassa*, Sjap: *Schizosaccharomyces japonicus*, Spom: *Schizosaccharomyces pombe*.

<https://doi.org/10.1371/journal.pgen.1008079.g002>

A170 = 4, with the full yeast combinatorial set consisting of $2^2 \cdot 2^3 \cdot 2^2 \cdot 2^3 \cdot 3^4 \cdot 2^4 \cdot 4^4 = 55,296$ genotypes out of which we determined the fitness for 48,198, or 87% of the possible yeast extant states combinations in our library. Our experimental design not only generated combinations of extant amino acids, but also allowed us to distinguish real variants from less abundant aberrations originating from sequencing errors (S2 Fig). We further reduced genotyping errors by sequencing every variant twice through paired-end sequencing (S2A Fig), and by using an error-correction algorithm (see **Sequencing error rate** section).

Most of the genotypes that can be created by the combination of extant amino acid states are not ancestral His3 protein segments. In a high-dimensional sequence space, the number of actual ancestral sequences is exponentially smaller than the number of combinations of extant states (Fig 1A). Out of the 48,198 extant amino acid state combinations in segment 3 only seven sequences actually match the reconstructed ancestral states of the entire segment. The combinations of extant amino acid states represent the evolutionary-relevant segment of the fitness landscape because they represent all possible alternative evolutionary trajectories in sequence space that consist of the same substitutions that have occurred in evolution of the extant protein sequences under consideration. Furthermore, some of these seven sequences were likely found in a different context of the entire protein, so they do not represent the ancestral states of the entire His3 protein but rather the ancestral states found in segment 3 across different genetic context of the rest of the His3 protein.

A substantial proportion of combinations of extant amino acid states led to genotypes with low fitness (Fig 2, Fig 3A and 3B, S3 Fig), an observation that takes into account the false discovery rate in our data (S2 Supporting Information). This observation could be explained by i) some extant amino acids having a universally deleterious effect, ii) some amino acid states exerting a negative effect on fitness because of intergenic interactions with other *S. cerevisiae* genes, or iii) by epistatic interactions between the extant amino acid states within His3 [34]. We exclude the possibility that some extant amino acid states had a universally lethal effect in

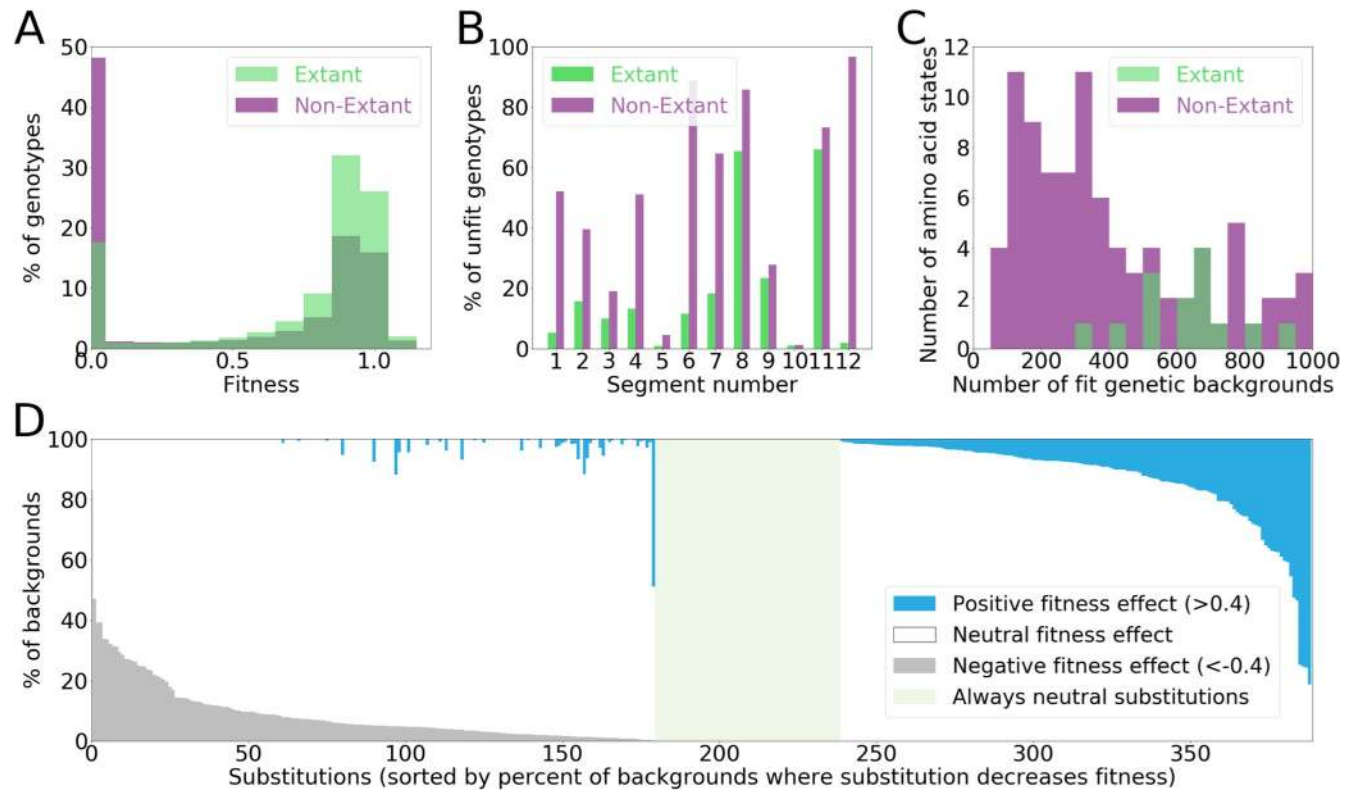


Fig 3. Fitness distributions. a, the distribution of fitness for genotypes composed of combination of extant amino acid states (green) and non-extant amino acid states (purple) at the same positions. b, The fraction of unfit genotypes per segment among genotypes consisting entirely from extant amino acid states (green) and those incorporating non-extant amino acid states (purple). c, Amino acid states can be found in a different number of fit backgrounds. In this figure, we show the number of amino acid states that were found in the lowest number of fit genetic backgrounds. Some non-extant amino acid states were found in just a few genotypes with high fitness. The most infrequent extant amino acid state was found in at least 300 different genotypes that had high fitness. d, The percent of backgrounds in which a specific amino acid replacement is neutral (white), beneficial (dark blue) or deleterious (dark grey). The region marked in green shows amino acid replacements that never have large effects (> 0.4) on fitness. Beneficial and deleterious effects are shown only if the frequency for a given amino acid replacement was higher than the false discovery rate (S2 Supporting Information). Data from segment 9 were excluded for this figure.

<https://doi.org/10.1371/journal.pgen.1008079.g003>

S. cerevisiae background because no extant amino acid states were present only in unfit genetic backgrounds, genotypes conferring a fitness of zero. Indeed, no extant amino acid states had a universally strong deleterious effect, a decline of fitness by more than 0.4 (Fig 3D), and the smallest number of fit backgrounds in which any extant state was found was approximately 300 (Fig 3C). We use the 0.4 threshold because it corresponds to ~1% false discovery rate (see Methods). We exclude the possibility that some extant amino acid states disrupt intergenic interactions because the complete His3 coding sequences from extant species fully complemented a His3 deletion in *S. cerevisiae* (S4C Fig). Thus, the observed genotypes with low fitness can only be explained by epistatic interactions among extant amino acid states within the His3 gene in the same or different segments. Remarkably, 85% (330/389) of replacements between extant amino acid states had substantially different effects on fitness in different backgrounds (Fig 3D). By contrast, only 15% of amino acid replacements are truly neutral, in the sense that they do not exert strong influence on fitness in any genetic background. Three quarters of the universally neutral amino acid replacements were observed in the disordered region of the protein (44/59). Taken together, the His3 fitness landscape across the 11 segments with high accuracy was strongly influenced by epistasis on a macroevolutionary scale, *i.e.* the impact

of an extant amino acid state on fitness often depends on the background in which it occurs [34–37]. An epistatic fitness landscape is rugged in the sense that evolving genotypes must avoid fitness valleys that emerge through deleterious combinations of amino acid states that may also be found in fit genotypes [19,20,37–38]. Characterizing the ruggedness and the mechanisms that determine the underlying epistasis becomes the primary challenge in understanding the fitness landscape of His3.

Unidimensional epistasis of the His3 fitness landscape

The ruggedness of the fitness landscape can be characterized by different measures of complexity of the underlying epistatic interactions. In the simplest case, epistasis may be unidimensional, in the sense that the fitness landscape can be described as a function of an intermediate variable, the fitness potential [40–42]. The fitness landscape is a function from the space of genotypes to fitness. In analogy with a scalar field, we can characterize the ruggedness of this function with standard measures of complexity if genotypes are arranged in a linear space. The simplest case is that of a linear predictor called the fitness potential: $p = c_1x_1 + c_2x_2 + \dots + c_nx_n$, where c_i is a coefficient and x_i is a binary variable that signifies the presence (1) or absence (0) of a given amino acid at a given position. By definition, e^p describes a non-epistatic fitness landscape because the effect of every amino acid replacement is multiplicative, and it depends only on the associated c . Any other function of p leads to epistasis. If the $f(p)$ function is “simple”, meaning that it has a small number of local extrema, such as a bimodal function, the epistasis is called unidimensional [42]. The limitation of simplicity of $f(p)$ is necessary because any function $f_0(x_1, \dots, x_n)$ can be represented by a function $f(p)$ and choosing appropriate coefficients c_1, \dots, c_n in p . Thus, a simple $f(p)$ leads to unidimensional epistasis because the entire genotype space can be reduced to a single dimension [42].

To quantitatively determine how well fitness differences between genotypes can be explained by unidimensional epistasis we used a deep learning approach to estimate the coefficients c for each allele x in the fitness potential and determine the unidimensional function $f(p)$ that best approximated the fitness landscape. We used a dense neural network architecture composed of three layers. Each neuron in the architecture performed a linear transformation of its input and then applied a nonlinear (sigmoid) function. The single neuron of the first layer computes the fitness potential, which is then mapped to a fitness value obtained from $f(p)$, the function of the fitness potential which is found by the three layers of the neural network architecture (see [Methods](#); [Fig 4](#)). For ten protein segments, $f(p)$ converged to the same shape: a threshold function in which organismal fitness remains constant with decreasing fitness potential and then is rapidly reduced to lethal after a certain threshold ([Fig 5A](#)). This analysis, coupled with previous observations of thresholds in fitness landscapes [7,10, 43] suggests that macroevolutionary fitness landscapes may in fact represent instances of truncation selection. Remarkably, no fitness functions showed a defined optimum, implying a lack of stabilizing selection on His3 protein function and that it was maintained at an optimized state throughout its evolution. The ability of the cliff-like threshold fitness function [44] to predict fitness from genotype varied between the His3 segments from near perfect ($r^2 = 0.97$) in segment 7, to relatively poor ($r^2 = 0.44$) in segment 5 ([S5 Fig](#)). Thus, while the fitness landscape of His3 is approximately unidimensional for some segments, it has a higher degree of complexity for others.

Ruggedness and multidimensional epistasis of the His3 fitness landscape

Ruggedness is a general property of fitness landscapes that quantifies the accessible paths of high fitness that connect fit genotypes [45–47]. A path between highly fit genotypes is

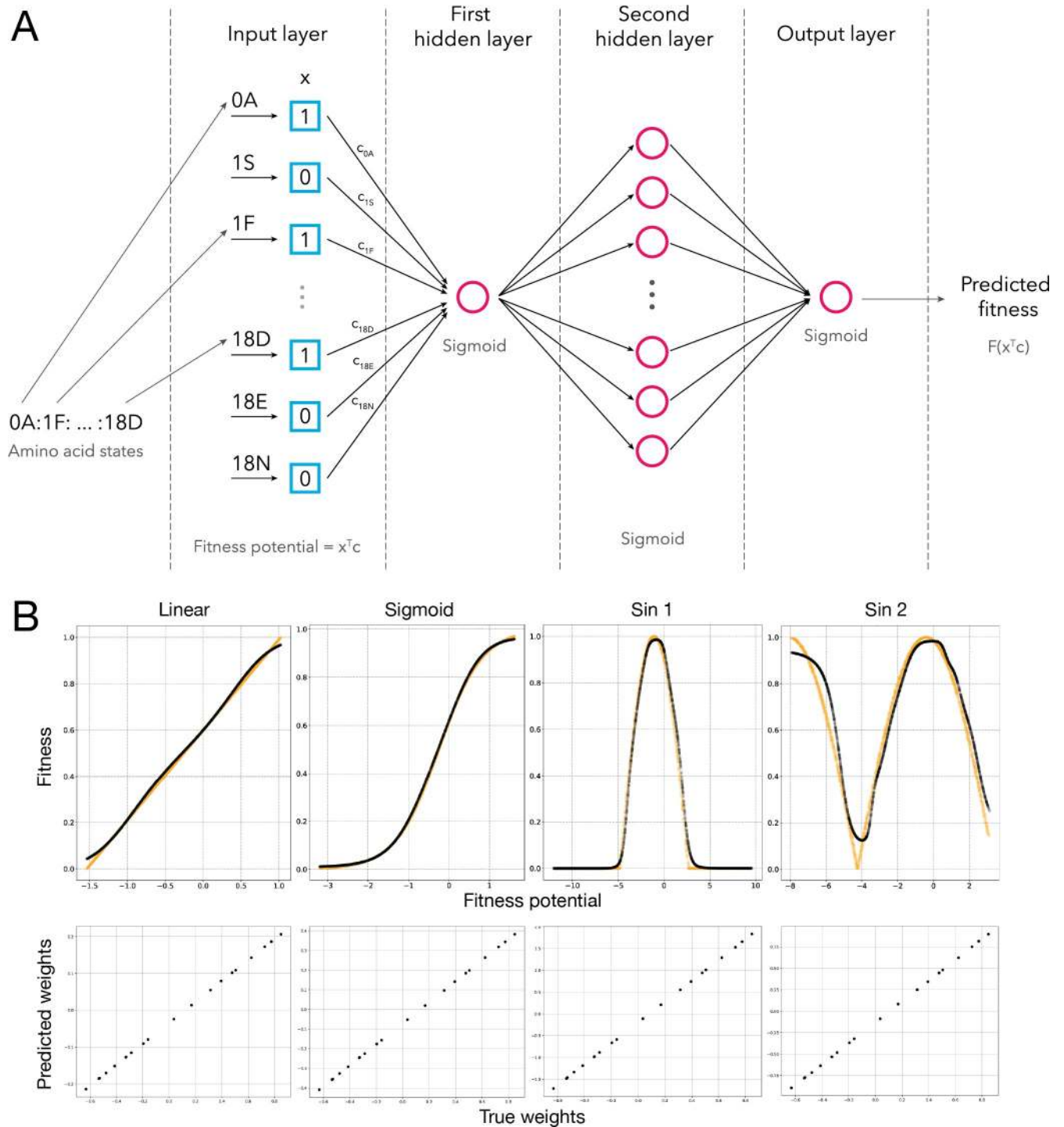


Fig 4. Schematic representation of a deep learning approach able to fit any arbitrary fitness function. a, each genotype was encoded as a binary vector (x). During training, each of the amino acid replacements was assigned a coefficient (c_i), comprising a vector of coefficients (c). The multiplication of these two vectors is the fitness potential of the genotype. After going through three layers, each with a sigmoid activation function, the predicted fitness is obtained. b, The fit of a mock fitness function (yellow) and the fit achieved by our neural network (black). The mock fitness function, was created by generating a set of amino acid states with defined coefficients (effects on fitness potential), which were then combined to generate genotypes across a range of fitness potential values.

<https://doi.org/10.1371/journal.pgen.1008079.g004>

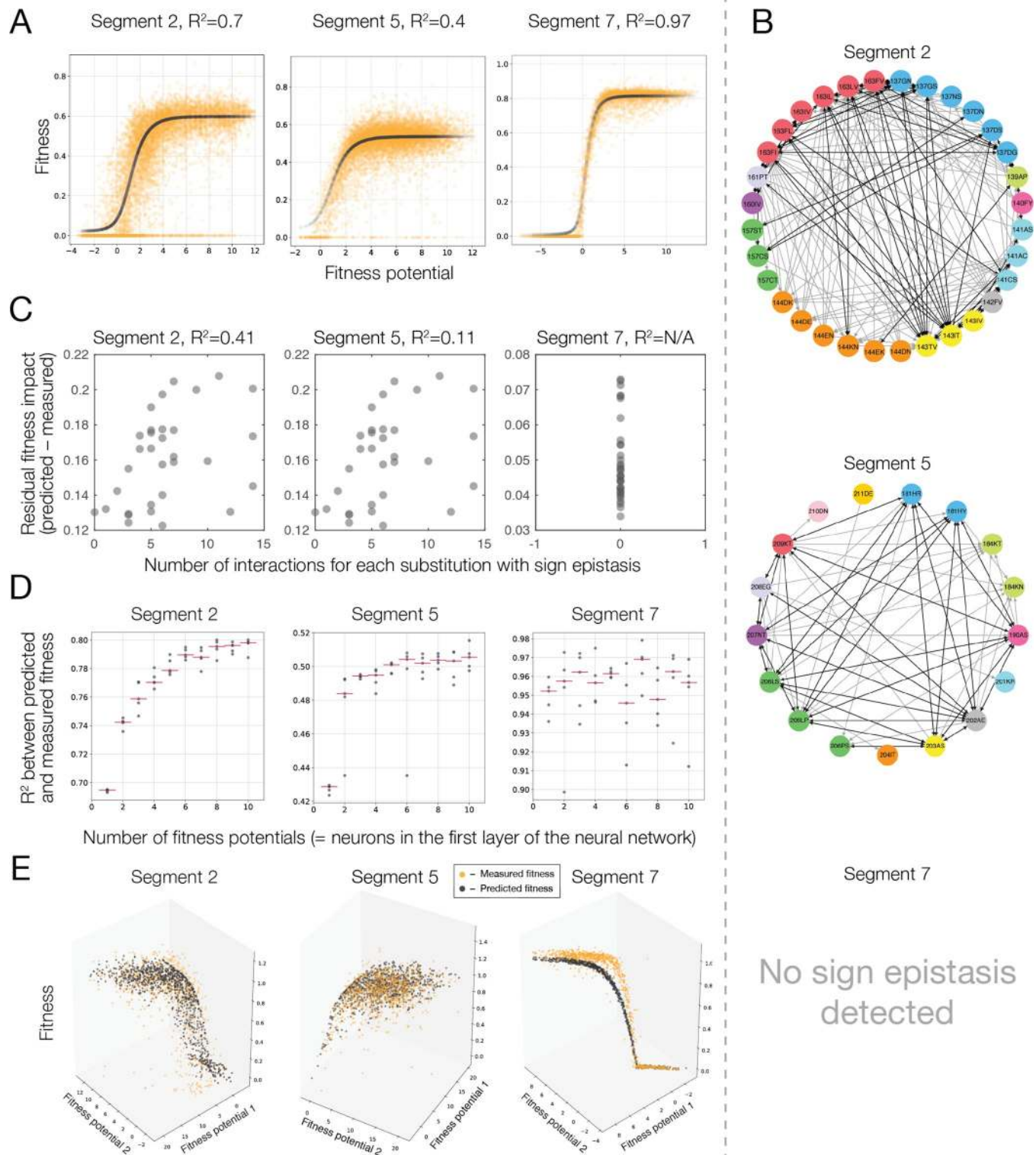


Fig 5. Epistasis and the His3 fitness landscape for segments 2, 5 and 7. a, Fitness as a function of a single fitness potential (black curve, the fitness of individual genotypes is orange). b, A network depiction of sign epistasis between amino acid replacements. Colour coded sites with reciprocal sign epistasis (black lines) and unidirectional interactions (grey arrows) are shown. c, Genotypes containing replacements with a higher number of sign epistatic interactions are less likely to be fit by the threshold function of the fitness potential. d, Increasing the number of neurons in the first layers of the neural network, which is equivalent to increasing the number of underlying fitness potentials, leads to more accurate models for segments with detected sign epistasis. Each dot

corresponds to an independent optimization of model parameters. e, Fitness as a function of two fitness potentials (black dots, measured fitness is depicted in orange).

<https://doi.org/10.1371/journal.pgen.1008079.g005>

inaccessible when one of the intermediate genotypes has low fitness [6,21–23,45,48] (e.g., for genotypes AB and ab, the intermediate are aB and Ab). Such instances also manifest in sign epistasis on the fitness landscape, that the same amino acid replacement may be beneficial or deleterious when occurring in a different genetic background [48,49]. To quantify the ruggedness of the His3 fitness landscape we identified instances of sign epistasis: replacements between extant amino acid states that were strongly beneficial in some backgrounds (increasing fitness by at least 0.4 in absolute fitness) or strongly deleterious (decreasing fitness by at least 0.4 in absolute fitness) in other backgrounds [48]. Some of these instances may be due to miscalled fitness of very few genotypes. Therefore, we considered a pair of extant amino acid states to be under sign epistasis only when sign epistasis was observed in a statistically significant number of different genetic backgrounds (see [Methods](#) section **Quantifying sign epistasis**).

An example of sign epistasis is the C141S replacement in the second segment that had an opposite effect on fitness depending on amino acid at site 143 (I, V or T). The I143T replacement in turn exhibits sign epistasis depending on the amino acid at site 163 (F, I, V or L) ([Fig 6A](#)). These epistatic interactions can be represented by a graph in which nodes represent a pair of extant amino acid states at a specific site and nodes are connected by edges if strong sign epistasis has been detected between them (C141S - I143T - I163F) ([Fig 5B](#)). We found that 86 out of 128 (67%) sites in our library exhibit sign epistasis and 46% (59/128) exhibit reciprocal sign epistasis with 8% (968/11597) of all pairs of sites exhibiting sign epistasis (see [S2 Supporting Information Table 5](#)). Most sites showed a sign epistatic interaction with multiple other sites ([Fig 6C](#), [S6 Fig](#)) demonstrating that, although sign epistasis affects few genotypes, it leads to a fitness landscape that requires the interaction of multiple sites for proper characterization.

Sign epistasis can appear when fitness is described by a unidimensional function of the fitness potential, for example, when the fitness landscape is a unimodal function with an optimum in an intermediate range of the fitness potential [45,49]. However, sign epistasis may also be a sign of multidimensional epistasis, when a unidimensional function of the fitness potential cannot fully describe genotype fitness [42]. Many genotypes were predicted poorly by a unidimensional function of the fitness potential ([S5A Fig](#)). Two lines of evidence suggest

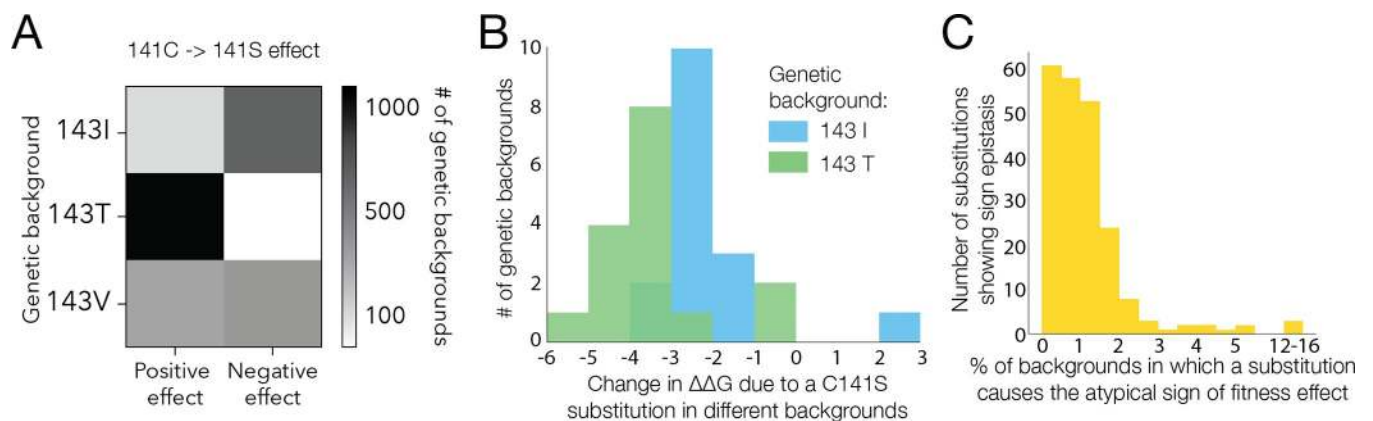


Fig 6. Sign epistasis. a, Amino acid replacement C->S at site 141 in segment 2 more frequently has a positive effect on fitness in the background of T at site 143, a negative effect in the background of 143I and is equally likely to be strongly deleterious or strongly beneficial in the background of 143V. b, Predicted change in folding free energy using Rosetta [64] following a C141S replacement in all genetic backgrounds with an I or T at 143 and that are closer than four mutations away from *S. cerevisiae*. c, the fraction of genotypes in which the amino acid replacement under sign epistasis has the less frequent effect on fitness.

<https://doi.org/10.1371/journal.pgen.1008079.g0062>

that such genotypes reveal the presence of multidimensional epistasis. First, genotypes with a higher number of amino acid replacements influenced by sign epistasis were less well-predicted by a unidimensional fitness function (**Fig 5C and S7B Fig**). Second, we explain a larger fraction of genotypes by using a more complex neural network architecture accommodating multiple fitness potentials instead of one. We found that increasing the number of neurons in the first layer of the neural network architecture, which is equivalent to increasing the number of independent fitness potentials, gradually improves the prediction power of the obtained models for most of the segments (**Fig 5D**). Thus, adding dimensions to the function of fitness potential increases the prediction power of the model. For example, for a two-dimensional case fitness was described by $f_1(p_1, p_2)$ with $p_1 = a_1x_1 + a_2x_2 + \dots + a_nx_n$ and $p_2 = b_1x_1 + b_2x_2 + \dots + b_nx_n$. For several His3 segments, a fitness function with multiple underlying fitness potentials described the fitness landscape more accurately than a simple unidimensional function of a single fitness potential (**S7A Fig**). For instance, for these segments, fitness function of two fitness potentials described the shape with a higher degree of accuracy than a function of a single fitness potential (**Fig 5D and 5E**). By contrast, epistasis in segment 7 is entirely unidimensional (**Fig 5D and 5E and S3 Supporting Information**); we do not see any improvement in the model's predictive power when adding extra dimensions.

Evolutionary trajectories on the His3 fitness landscape

On a smooth fitness landscape, evolution can proceed along any of the evolutionary paths connecting two fit genotypes, as none of the intermediate genotypes confer low fitness (see Box 2 in [50]). Alternatively, the fitness landscape is rugged when it contains non-connected fitness peaks, such that there are no viable paths between some pairs of genotypes that confer high fitness [4,5]. In other words, the presence of deleterious intermediate genotypes between highly fit ones leads to inaccessibility of some evolutionary trajectories between extant or ancestral sequences [6,21–23,48]. The simplest explanation for the substantial ruggedness of the landscape observed in many of the His3 segments lies in the unidimensional threshold fitness function (**Fig 7A**). On a threshold function, a combination of amino acid replacements that are all neutral in some genetic backgrounds can take a genotype beyond the fitness threshold through their additive effect on fitness potential, making some genotypes inaccessible for evolution (**Fig 7A**). Between any two fit genotypes, the fraction of intermediate genotypes that are unfit depends on the fitness potential of the two parental genotypes (**Fig 7B**). Evolution between two fit genotypes with high fitness potential can proceed unhindered because all intermediate genotypes also have high fitness potential and, consequently, high fitness. Conversely, when both fit genotypes are located close to the threshold, many of the intermediate genotypes between them have low fitness and many evolutionary paths between them are inaccessible (**Fig 7C**). Thus, the cliff-like threshold fitness function is the major determinant of the observation that not all paths between two fit genotypes are accessible to evolution (**Fig 7B**). We find that unfit intermediate genotypes are in genetic proximity with each other and are on a limited number of paths; the fraction of inaccessible paths is smaller than if the same number of unfit genotypes were distributed randomly in genotype space (**Fig 7D and 7E**).

The effect of synergistic epistasis dominates the His3 fitness landscape (**Fig 5A**), affecting over 85% of amino acid replacements from our library that occurred in His3 evolution (**Fig 3D**). This synergistic epistasis may reflect the folding free energy of the protein [10,51,52], as evidenced by a weak to modest correlation between the fitness potential and the impact of amino acid replacements on the folding free energy of His3p (**S8 Fig**). Similarly, instances of sign epistasis may also be explained by changes in protein stability; for example, in the 143T background C141S increased fitness and also had a positive effect on stability (**Fig 6B**).

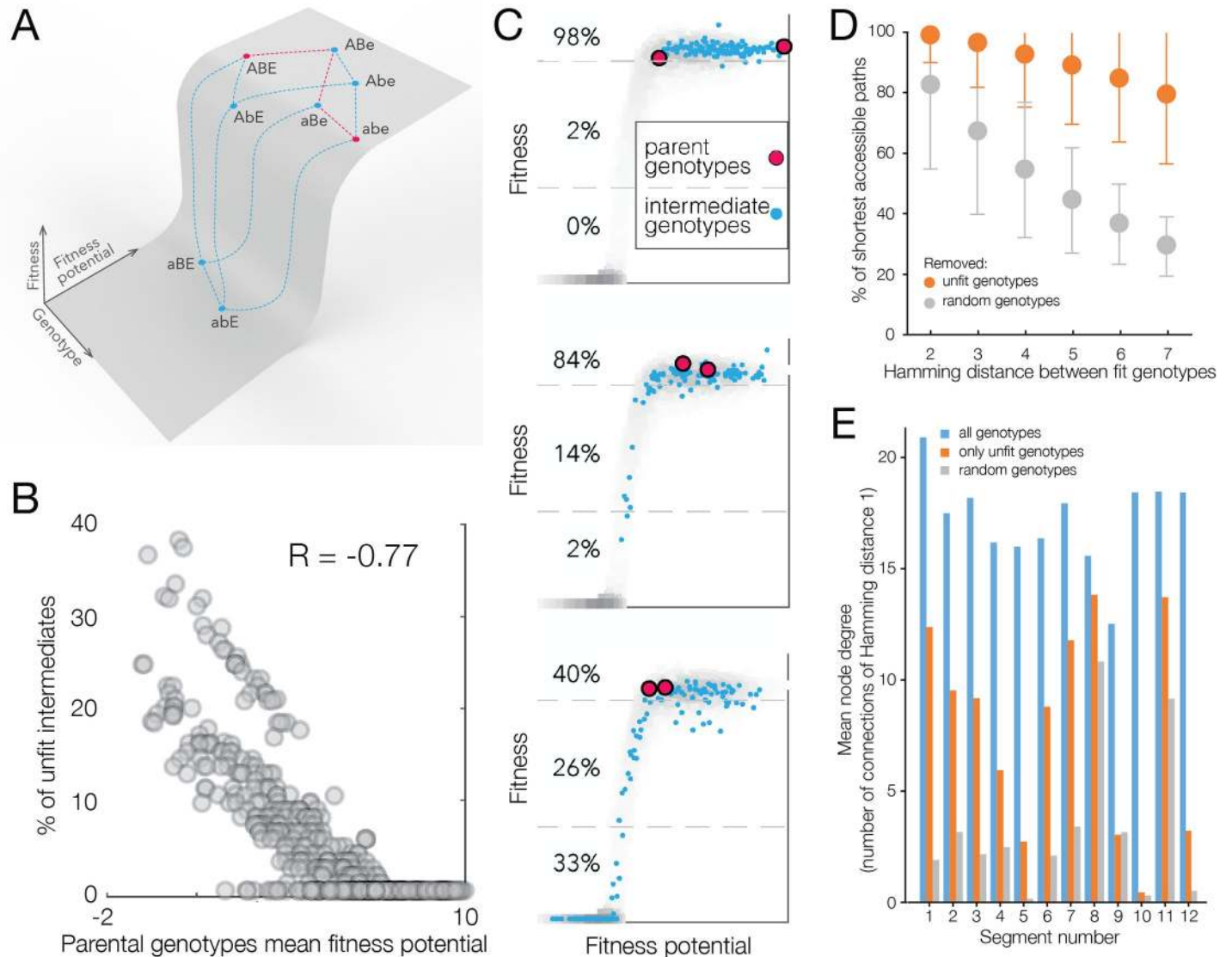


Fig 7. Analysis of evolutionary pathway accessibility. a, A threshold fitness potential function can lead to some paths being inaccessible between two genotypes of high fitness (abe, ABE) if the joint contribution of several alleles to the fitness potential (aBe, aBE) leads to the fitness potential below the threshold. b, The fraction of unfit intermediate genotypes between two fit genotypes as a function of their average fitness potential. c, The grey area represents all genotypes in segment 7. When two fit genotypes (red dots) have high fitness potential, many paths between them will be accessible because many intermediate genotypes will also have high fitness potential and fitness (blue dots). d, Two fit genotypes in our dataset with intermediate genotypes between them. Intermediate genotypes are those that can be found between the two genotypes when making amino acid replacements from one genotype to the other. By this definition, the intermediate genotypes are located on the shortest paths connecting the two genotypes, with shortest paths are those that include only forward replacements. Accessible paths are those that incorporate only fit genotypes. The fraction of accessible shortest paths between two fit genotypes from data (orange) shows a modest decline as a function of Hamming distance between the two genotypes. The fraction of accessible shortest paths between two fit genotypes declines more rapidly when the same number of unfit genotypes as observed in real data are randomly drawn from intermediate genotypes (grey), demonstrating that in real data unfit genotypes are clustered in sequence space. Error bars are standard deviation. e, Genotypes can be represented by a graph with edges connecting genotypes if they can be connected by one amino acid replacement. We calculate the degree of connectivity (number of edges for each node) when all genotypes one replacement away are connected (blue), only unfit (fitness = 0) genotypes are connected (orange) and when unfit genotypes are connected with unfit genotypes selected at random keeping the same number of unfit genotypes as in our data (grey). With the exception of segment 9, in all segments the connectivity of unfit genotypes is higher than random, confirming that unfit genotypes are clustered in sequence space.

<https://doi.org/10.1371/journal.pgen.1008079.g007>

Consistent with protein stability contributing to the observed sign epistasis we find that sites that exhibited reciprocal sign epistasis are close together in the His3p structure (S8 Fig). Only a relatively distant His3 structure to *S. cerevisiae* was used in this analysis, with about 30% divergence, likely reducing the accuracy of the free energy estimates. An additive contribution

to free energy can lead only to a unidimensional fitness function [51], indicating that other non-additive mechanisms, such as catalytic activity or inter-subunit interactions, or non-additive contribution to protein stability must be responsible for the multidimensionality of the His3 fitness landscape.

Inference of inter-segmental epistatic interactions in the His3 gene sequence

Epistasis may be caused by interaction among positions within a segment (intra-segmental epistasis) or by interaction of the segment with the rest of the *S. cerevisiae* His3 sequence (inter-segmental epistasis). In other words, a specific genotype of one segment may be associated with low fitness because of interaction of amino acid states on that segment with amino acid states at other segments. To some extent, the contribution of inter- versus intra-segmental interactions can be decoupled. Given two fit genotypes (e.g., ABC & abc in one His3 segment), any unfit intermediate states (e.g., aBc in the same His3 segment) must be due to intra-segmental epistasis because the rest of the protein remains constant. For each segment, we took as a measurement of intra-segmental epistasis all pairs of fit genotypes and calculated the proportion of unfit intermediate genotypes as a function of the Hamming distance between the two fit genotypes. We then compared this proportion with the total proportion of all unfit genotypes as a function of Hamming distance from *S. cerevisiae*, a measurement that includes both inter- and intra-segmental epistasis. We found three times more inter-segmental than intra-segmental epistasis (S9 Fig), likely because a single segment provides a much smaller target space for interactions than the entire His3 protein. On the other hand, inter-segmental epistasis was just three times more frequent, rather than ten times more frequent, possibly due to higher probability of interaction of amino acid residues that are proximal in primary structure [53]. The proportion of sites under epistatic interactions increased exponentially with Hamming distance (S9 Fig), analogous to Orr's snowball, the accumulation of genetic incompatibilities in the course of speciation [34,54].

Discussion

The concept of the fitness landscape introduced by Sewall Wright (Figs 1 and 2 in [6]) is an indispensable tool for understanding multiple biological phenomena [1–5]. Experimental high-throughput assays of random mutations have begun to unravel some local properties of fitness landscapes [4]. Here, we described a fitness landscape on a macroevolutionary scale by focusing on amino acid states that have been put through the sieve of natural selection. We found that only 15% of amino acid states that were fixed in the evolution of His3 are universally neutral. For the remaining 85%, amino acid replacements had a profound influence on each other's effect on fitness, providing an experimental confirmation that epistasis is one of the defining features of molecular evolution [36]. Substitutions that occur in evolution have properties vastly different from those of random mutations, which are mostly deleterious [26]. Therefore, the way in which combinations of extant amino acid states affect fitness may also be different from that of combinations of random mutations. Unexpectedly, we found that the interaction of extant amino acid states was dominated by synergistic epistasis in a manner similar to that previously found for random mutations [7–10,16]. However, the accumulation of random mutations leads to low fitness much faster than the accumulation of extant amino acid states (compare Fig 2 from [16] and Fig 3B from [10] to Fig 2A).

The experimental data showing that 85% of amino acid states found in extant species confer low fitness in a different genetic background lends strong support to the notion that epistasis is a key factor in protein evolution [34,36]. We showed that the fitness landscape of several segments of the His3 gene cannot be reduced to a single unidimensional forms of epistasis, with a

function of multiple fitness potentials providing a more accurate description of the fitness landscape. By contrast, large-scale fitness landscapes incorporating multiple random mutations away from the wildtype sequence in a constant test environment have not displayed evidence of multidimensional epistasis [8–10,16]; however, it appears to be a more prevalent factor among amino acid replacements that have been subject to positive selection [12–16,19–23,37–39]. We also found that up to 67% of sites with an extant amino acid state were influenced by sign epistasis, resulting in a rugged fitness landscape and a limited number of fitness ridges connecting extant sequences for most His3 segments. However, sign epistasis substantially influenced fitness in only a small number of genotypes (Fig 5D). Overall, the evolutionary-relevant section of the His3 fitness landscape is best described as a fitness ridge, with the crest of the ridge defined by a fitness potential. In some cases, the crest is multidimensional requiring several independent underlying fitness potentials. Evolution can proceed unhindered along the crest (Fig 7A and 7C), however, pathway availability declines rapidly when evolution proceeds close to the edge of the fitness ridge.

Materials and methods

Data access

The raw and processed data have been submitted to the NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE99990. A virtual machine containing a running version of the data processing pipeline is available as a Docker image <https://hub.docker.com/r/gui11aume/epi/>. The scripts to reproduce the figures are on Github at <https://github.com/Lcarey/HIS3InterspeciesEpistasis>.

Study design

The His3 gene was selected for three principal reasons, it is short, conditionally essential and has not been known to be involved in protein-protein interactions. Studying 20^{220} variants of His3 is impossible, thus, we have chosen an approach to survey the fitness landscape in a manner that would elucidate the area most relevant to His3 evolution while managing the technical limitations of our experimental design. We considered amino acid states found in extant species, focusing on yeast species, which translated into a full combinatorial set of $\sim 10^{83}$ unique genotypes. Technically, it is feasible to measure fitness on the order of 100,000 unique genotypes in a single growth experiment. Therefore, we split the His3 gene into 12 independent segments such that the full combinatorial set of extant amino acid states from 21 yeast species in each segment was 10,000 – 100,000 genotypes. We then considered the combinatorial library for each segment in an independent growth experiment, which allowed us to study a tractable section of the sequence space while considering trajectories across a vast part of the space connecting extant species (Fig 1A). We constructed these combinations in 12 plasmid libraries and transformed them into a haploid His3 knockout *S. cerevisiae* strain. Growth rate (fitness) of yeast carrying different mutations in His3 was measured using serial batch culture in the absence of histidine.

We split the His3 gene sequence into segments in a manner agnostic to the structure of the His3 protein (S1A Fig). For technical reasons, a segment consisted of two variable regions with a constant region between them (S1B and S2A Fig). All growth experiments were performed independently for each segment, with the exception of one experiment on a limited group of genotypes from each segment which was done for the normalization of fitness values across different segments (S4 Fig).

As a control, we measured the rate of growth of *S. cerevisiae* whose entire His3 gene sequence came from another distant species. We found that the replacement of an entire gene

sequence of His3 leads to wild-type rates of growth of *S. cerevisiae* even when the His3 sequence comes from very distant yeasts, as far as *S. pombe* (S4 Fig). Therefore, His3 appears to be an independent unit of the fitness landscape and is a good model for the study of fitness landscapes of an isolated gene.

Measuring fitness

Plasmid construction. The His3 open reading frame of *S. cerevisiae* was PCR amplified to include the promoter and transcription terminator regions from 622 base pairs (bp) upstream of the open reading frame (ORF) to 237 bp downstream of the ORF, using primers 126 and 127 (see S2 Supporting Information) from the wild-type prototroph strain FY4. The PCR product was cloned into vector pRS416 using Gibson assembly (NEB, E2611S). The His3 orthologues from other species were amplified from genomic DNA using designed primers (S1 Supporting Information) and were cloned into the vector pRS416_his3, replacing the ORF of *S. cerevisiae* by Gibson assembly (NEB, E2611S). Since the His3 orthologue from *A. nidulans* contains an intron, the whole open reading frame was initially cloned into the vector, and the intron was later removed by PCR-amplifying the whole plasmid without this sequence, followed by recircularization.

Genomic DNA extraction. Genomic DNA from fungi (*Saccharomyces cerevisiae*, *Saccharomyces bayanus*, *Candida glabrata*, *Saccharomyces castellii*, *Kluyveromyces lactis*, *Eremothecium gossypii*, *Debaryomyces hansenii*, *Lodderomyces longosporus*, *Aspergillus nidulans*, *Schizosaccharomyces pombe*, *Candida guilliermondii*, *Saccharomyces kluyveri*, *Kluyveromyces waltii*) was extracted using MasterPure Yeast DNA Purification Kit according to the manufacturer's instructions (Epicentre, MPY80200).

Mutant library construction. Twelve independent mutant libraries, each for different regions of His3 (S2 Supporting Information), were generated based on the results of multiple alignment of 392 His3 orthologues. The alignment was built using the ClustalW alignment feature of the MEGA 6.0 software package [55] and user-corrected.

Mutant libraries were constructed by fusion-PCR, leaving two variable regions separated by a constant region. For each library, two contiguous fragments of His3 were amplified independently, using 1 µg of *S. cerevisiae* (strain FY4) genomic DNA in separate Phusion polymerase reaction mixes (Thermo Fisher Scientific, F530S) in GC buffer. For each PCR, one of the primers was a degenerate oligonucleotide with a constant part at the 5' end required for the fusion-PCR; the other primer was either 126 or 127. The degenerate primer approach led to the integration of non-extant amino acid sequences due to the redundancy of the genetic code. Consider the amino acid Phe in *S. cerevisiae* coded by the codon TTT. When incorporating an extant orthologous state Trp (TGG) two independent T → G nucleotide mutations will be incorporated creating the codons TTG (Leu) and TGT (Cys). If these two amino acids were not found in other species then they would be non-extant and due to the non-random nature of their incorporation in our dataset the non-extant states in this study do not represent a random set of amino acid changes. The cycling conditions for the PCR were 98°C for 30 s; 98°C for 20 s, 60°C for 30s and 72°C for 1 min (25 cycles); and 72°C for 5 min. The products were column-purified (QIAGEN, QIAquick PCR purification kit, 28104), eluted in 50 µl and mixed in equimolar proportion. The fusion-PCR was carried out by diluting 10 µl of the mix to 25 µl of standard Phusion polymerase reaction mix in GC buffer. The cycling conditions of the fusion-PCR were 98°C for 30 s; 98°C for 30 s, 60°C for 2 min and 72°C for 1 min (25 cycles); and 72°C for 5 min. The product of fusion was purified from agarose gel (Qiagen, MinElute Gel Extraction Kit, 28604) and eluted in 10 µl of water. 10 µl of the product was used as a template for additional 5 cycles of PCR reaction in Phusion polymerase reaction mix (Thermo

Fisher Scientific, F530S) in GC buffer, using primers 126 and 127. The cycling conditions were as follows: 98°C for 30 s; 98°C for 20 s, 60°C for 30 s and 72°C for 1 min (5 cycles); and 72°C for 5 min. The product was column-purified (QIAGEN, QIAquick PCR purification kit, 28104), and used as an insert for Gibson assembly.

To create a library of His3 mutants, pRS416 plasmid was amplified using primers 128 and 129. The insert was cloned into the vector using Gibson assembly (NEB, E2611S). Ligated products (200–300 ng/μL) were desalted by drop dialysis using 13 mm diameter, Type-VS Millipore membrane (Merck Millipore, VSWP01300). 20 μL ElectroMAX DH10B competent cells (Invitrogen, 18290015) were electroporated with 3 μL ligated products. 0.01% of the electroporated bacteria were plated on ampicillin-containing medium in order to estimate the complexity of the library; the remaining culture was grown overnight in 100 ml of liquid medium, and the plasmid was extracted the next day. For each library, the maximum number of protein sequences that can be generated was computed. Libraries were generated until to total complexity reached at least 3 times this value.

Yeast transformation and yeast library generation. For each segment, yeast strain LBCY47 (*his3:KanMX_{leu2}Δ0 met15Δ0 ura3Δ0*, derived from BY4741) was transformed with 50 μg of pRS416_His3 mutant library using lithium acetate transformation and plated onto glucose synthetic complete dropout plates lacking uracil. After 40 hours' growth at 30°C, approximately 0.5 million yeast colonies were scraped off the plates, mixed together and washed 2 times with 100 ml of PBS.

Bulk competition. 4×10^9 cells were inoculated into 500 ml of glucose synthetic complete dropout medium lacking uracil with 200 mg/L of G418, and grown at 30°C at 220 RPM for 6–8 h in order to eliminate clones with low fitness irrespective of histidine biosynthesis. Cells were later pelleted and washed with 50 ml of PBS. Approximately 10^{10} cells were inoculated into 1 L of synthetic complete dropout medium lacking histidine, and grown at 30°C at 220 RPM for 168 h with 12 h between bottlenecks: $\sim 10^{10}$ cells were transferred into fresh medium $\sim 10^8$ cells from the culture were kept as sample for the given time point. Bulk competition for each library of mutants was performed in two replicates to account for biological variability.

NGS library preparation. The relative abundance of yeast mutants was measured in 3 samples: 1) the initial population before selection was applied (t0), 2) the population after 12 h of growth in the selective medium (t1), and 3) the final population after 168 h of growth in the selective medium (t14). In order to extract plasmid DNA, 5×10^9 cells from each sample were incubated in 300 μL of zymolyase buffer (1 M sorbitol, 0.1 M sodium acetate, 60mM EDTA (pH 7.0), 2 mg/ml zymolyase, 1% 2-Mercaptoethanol) at 37°C for 3 h. The plasmid DNA was purified from the obtained spheroplasts using QIAprep Spin Miniprep Kit (QIAGEN, 27104) according to the manufacturer's protocol. The obtained DNA was used as a template in a 25 μL of Q5 DNA polymerase reaction mix (NEB, M0491S), using staggered primers for demultiplexing in the following cycling conditions: 98°C for 30s; 98°C for 10s, 60°C for 30s and 72°C for 30s (18 cycles); and 72°C for 2 min. PCR products were purified using Agencourt AM Pure XP beads (Beckman Coulter, A63880), and eluted in 40 μL of TE buffer (pH 8.0). DNA extraction and PCR-amplification were repeated twice for every sample to account for the technical variability.

NGS libraries were prepared from 100 ng of the purified DNA amplicons using Ovation Rapid DR System (Nugen, 0319-32) according to manufacturer's instructions. Each library was visualized on a Bioanalyzer (Agilent Technologies) and quantified by qPCR with a Kapa Library Quantification Kit (Kapa Biosystems, KK4835). Twelve samples were pooled together (accounting for two biological replicates, two technical replicates and three time points) at the final concentration of 4 nM, and sequenced in the same lane. Samples were sequenced as 125-bp paired-end reads on a HiSeq2500 sequencer (Illumina) with v4 sequencing chemistry.

Yeast growth assay. Mutant strains were grown overnight in complete dropout medium lacking uracil. The cultures were diluted to 0.05 OD 600 nm, and grown for 5 h in the same medium. 6 μ L of each culture were transferred into 96-well plates in 125 μ L of complete dropout medium lacking histidine. Growth of the strains was monitored by measuring OD 600 nm every 10 min using Tecan Infinite M1000 PRO microplate reader equipped with an integrated Stacker module.

The growth rate of individual curves was measured as the inverse of the time to grow from OD = 0.135 = $\exp(-2)$ to OD = 0.368 = $\exp(-1)$. If the curve did not reach 0.368, the growth was set to 0. Curves that crossed 0.135 or 0.368 were excluded. The growth rate of a clone was measured as the median of 6 independent growth experiments. We excluded from the analysis clones with discordance between growth in solid and liquid medium, clones that could not be sequenced or that showed evidence of contamination by sequencing, and clones such that the Kullback-Leibler divergence of their read counts compared to all synonymous clones was greater than 0.22. The later criterion ensured that the selected clones were not outliers compared to other variants encoding the same protein.

Growth rates of isolated strains

We isolated 197 strains from all segment libraries of extant amino acid combinations (9–26 strains per segment) and used Sanger sequencing to determine the sequence. For each strain we performed 6 repeats of growth assay and calculated the average growth rate. Fitness values from competition and growth rates are highly correlated ($r = 0.82$, $p = 10^{-48}$). Correlation was significant and greater than 0.6 for all segments except segment 9, where all selected genotypes appeared to be neutral (S4 Fig).

Initial data filtering

The individual sequences of the variants were recovered from pair-end reads with the following steps: the constant region between the two variable regions was identified by inexact matching allowing up to 20% errors using the Seeq library version 1.1.2 (<https://github.com/ezorita/seeq>). The reads are not oriented because the Illumina sequencing adapters were added by ligation, so the constant regions were searched on both reads. Forward and reverse reads were swapped when a match was found on the reverse read. This ensured that all of the sequences are in the same orientation. For multiplexing purposes, the sample identity was encoded in the left and right primers used to PCR-amplify the variants. To demultiplex the reads, we used inexact matching with the candidate primers, allowing up to 20% errors. To merge the reads, the sequence of the reverse reads was reverse complemented and the constant region was searched by inexact matching allowing up to 20% errors. The position of the constant part in each read indicated how they must be stitched together. This approach was faster and less error-prone than using FLASH [56]. In the region of overlap, the consensus sequence was determined by picking the nucleotide with highest quality as indicated in the quality line of the fastq files. If 'N' persisted in the final sequence, the reads were discarded. The PCR primers were trimmed so that all the sequences of the same competition would start and end at the same location.

Reads that did not have the constant region, that could not be oriented or that could not be demultiplexed were discarded. The remaining errors in the reads were corrected by sequence clustering. We used Starcode version 1.0 [57] with default parameters and allowing up to two errors. The corrected reads were translated using the genetic code. Variants encoding the same proteins were not merged; they were kept separate for downstream analyses. A running Docker virtual machine with commented scripts to replay the whole the process is available for download at <https://hub.docker.com/r/gui11aume/epi/>.

DNA sequence variant frequency calculation and data filtering

The total number of reads for 12 segments, 3 time points and 4 replicas are shown in [S2 Supporting Information](#). Genotypes frequencies are defined as the number of reads for a given genotype divided by the total number of reads in that replicate. Mean frequency was calculated over 4 replicas to be used in further analysis. However, to eliminate influence of outliers the median was taken instead of mean if absolute difference between mean and median was greater than the median value. Only genotypes present in both technical replicas of both biological replicas with at least ten reads (summed across all time points) in each of them were kept.

Sequencing error estimate

The length of the segment was designed so that each variable region was read twice by pair-end reads ([S2A Fig](#)). This strategy led to a substantial reduction in the sequencing error rate because mismatches between the two reads were corrected to the nucleotide with the higher quality call.

The raw Illumina sequencing error rate was estimated by measuring the frequency of mismatches between forward and reverse reads. The rationale of this estimate is that each mismatched nucleotide must be a sequencing error for at least one of the reads. The variant calling error rate was estimated by collecting groups of reads that differ by only one nucleotide in the constant region ([S2A Fig](#)). Since the variable regions are identical, such reads come from the same variant and the different nucleotide in the constant region must originate from a calling error (mutations in *S. cerevisiae* are negligible because they occur at a much lower rate). The frequency of such reads was used to approximate the per-nucleotide variant calling error rate. The raw Illumina error rate was computed by custom Python scripts and reads differing by one nucleotide were collected using the “sphere” clustering option of Starcode with a maximum distance of 1 (see the *Data access* section for the code repositories). The results are summarized in [S2 Supporting Information](#).

The design strategy and the low error rate allow us to distinguish variants incorporated in the library from random sequencing errors. Each library contained on the order of 10^5 individual sequence variants, so that each library variant would be found at a frequency several orders of magnitude higher than variants created by sequencing errors. For example, for segment 7, the number of nucleotide variants was 176,879 with 31,815,448 possible single mutants of these variants. The error rate in segment 7 was 0.04% per nucleotide, which translates into 2.4% of reads being miscalled. Thus, the expected frequency of a particular miscalled variant is $0.024 / 31,815,448 \approx 8 \times 10^{-10}$, considerably smaller than the expected frequency of real variants $0.976 / 176,879 \approx 6 \times 10^{-6}$. The estimated frequencies of library variants for all segments of His3 are reported in [S2 Supporting Information](#).

The final variant calling error rate was smaller than the numbers shown in [S2 Supporting Information](#) because errors were further corrected by sequence clustering using Starcode (see *Initial data filtering* section). In line with the rationale above, low frequency erroneous reads are converted to the closest high frequency variant at a maximum Levenshtein distance of 2.

PCR recombination rate estimate

It is known that PCR can create new genotypes by template switching [58]. To test the magnitude of this effect, we took advantage of the two-block design of the variants and estimated the frequency of recombination between the left and the right variable regions. The structure of reads can be represented like this: AAAAAA— — — —BBBBBB. In this example, “A” represents the left part of the variable region of the segment, “—” represents the invariable region and “B” represents the right part of the variable region. Insertions of two or more nucleotides are rare events caused by errors during the library synthesis; so, if the same insertion in a left half of the variant is associated to several variants on the right half, it is likely an occurrence of

template switching (the same holds for insertions in the right half of the variant). For example, if the region with A*AAA*AA-----BBBBBB, where “*” represents a deletion, is found with several different variants of the B segment then such a situation likely represents template switching. Focusing on the initial time point to avert the effect of selection, we counted a total of 11,454 variants with two or more insertions on either the left side or the right side. Among those, 76 had the same insertion as another variant. This means that > 98.6% of the variants were free of template switching. Extrapolating to the rest of the dataset, this means that the “leakage” of reads between variants is substantially lower than the magnitude of the observed epistasis and we can rule out this artefact as a potential explanation for our results. In either case, all errors in our experimental pipeline, including template switching, are taken into account when we calculate the false discovery rate.

Noise estimation

The major factors causing noise in genotype frequency measurements are sampling errors, PCR amplification errors and genetic drift during the competition. For all of these factors, the amount of error depends on the genotype frequency. Therefore, we estimated measurement errors as the function of genotype frequency.

For a given segment, time point and a pair of biological or technical replicas for each genotype we calculated the mean frequency and the squared difference of frequencies from these two replicas. We sorted genotypes by mean frequency and grouped them such that each bin contains 5000 genotypes. We calculated the average frequency and the average squared difference in each bin. Additionally, squared error for frequency 0 was set equal to $\frac{1}{2} \cdot \left(\left(\frac{0.5}{N_i} \right)^2 + \left(\frac{0.5}{N_j} \right)^2 \right)$, where N_i and N_j are total read numbers in replicas i and j . Finally, by linear interpolation we obtained dependencies of squared differences as a function of frequency, $s_{ij}^2(f)$, where i and j are different replicas.

Using squared differences from pairwise comparison of replicas we can estimate variance of mean frequency over four replicas. Let numerate replicas 1, 2, 3, 4 where 1, 2 are technical replicas of the first biological repeat and 3, 4 are the technical replicas of the second biological repeat. Errors coming from the competition (e.g.: genetic drift) are shared for replicas 1, 2 and for replicas 3, 4. Let’s call them Δf_{b_1} and Δf_{b_2} and their variances $\sigma_{b_1}^2$ and $\sigma_{b_2}^2$, respectively. Technical errors of sampling from the population and from PCR are unique for each replica. Let’s call them Δf_{t_i} , $i = 1..4$ and their variances $\sigma_{t_i}^2$, $i = 1..4$, respectively. All variances are function of frequency and when writing $\sigma_{x_i}^2$ we assume $\sigma_{x_i}^2(f)$.

In the introduced notations the mean frequency over 4 replicas is:

$$f = \frac{1}{4} \cdot (f_1 + f_2 + f_3 + f_4) =$$

$$\frac{1}{4} \cdot ((f^* + \Delta f_{b_1} + \Delta f_{t_1}) + (f^* + \Delta f_{b_1} + \Delta f_{t_2}) + (f^* + \Delta f_{b_2} + \Delta f_{t_3}) + (f^* + \Delta f_{b_2} + \Delta f_{t_4})) =$$

$$f^* + \frac{1}{2} \cdot (\Delta f_{b_1} + \Delta f_{b_2}) + \frac{1}{4} \cdot (\Delta f_{t_1} + \Delta f_{t_2} + \Delta f_{t_3} + \Delta f_{t_4}),$$

where f^* is the true frequency. Applying basic properties of variance, the variance of mean frequency:

$$\sigma^2 = \frac{1}{4} \cdot (\sigma_{b_1}^2 + \sigma_{b_2}^2) + \frac{1}{16} \cdot (\sigma_{t_1}^2 + \sigma_{t_2}^2 + \sigma_{t_3}^2 + \sigma_{t_4}^2)$$

To estimate $\sigma_{b_1}^2, \sigma_{b_2}^2, \sigma_{t_1}^2, \sigma_{t_2}^2, \sigma_{t_3}^2, \sigma_{t_4}^2$ we used squared differences from pairwise comparison of replicas calculated above $s_{12}^2, s_{13}^2, s_{14}^2, s_{23}^2, s_{24}^2, s_{34}^2$:

$$\mathbb{E}[s_{12}^2] = \mathbb{E}[(\Delta f_{t_1} - \Delta f_{t_2})^2] = \sigma_{t_1}^2 + \sigma_{t_2}^2$$

$$\mathbb{E}[s_{13}^2] = \mathbb{E}[(\Delta f_{b_1} + \Delta f_{t_1} - \Delta f_{b_2} - \Delta f_{t_3})^2] = \sigma_{b_1}^2 + \sigma_{t_1}^2 + \sigma_{b_2}^2 + \sigma_{t_3}^2$$

$$\mathbb{E}[s_{14}^2] = \mathbb{E}[(\Delta f_{b_1} + \Delta f_{t_1} - \Delta f_{b_2} - \Delta f_{t_4})^2] = \sigma_{b_1}^2 + \sigma_{t_1}^2 + \sigma_{b_2}^2 + \sigma_{t_4}^2$$

$$\mathbb{E}[s_{23}^2] = \mathbb{E}[(\Delta f_{b_1} + \Delta f_{t_2} - \Delta f_{b_2} - \Delta f_{t_3})^2] = \sigma_{b_1}^2 + \sigma_{t_2}^2 + \sigma_{b_2}^2 + \sigma_{t_3}^2$$

$$\mathbb{E}[s_{24}^2] = \mathbb{E}[(\Delta f_{b_1} + \Delta f_{t_2} - \Delta f_{b_2} - \Delta f_{t_4})^2] = \sigma_{b_1}^2 + \sigma_{t_2}^2 + \sigma_{b_2}^2 + \sigma_{t_4}^2$$

$$\mathbb{E}[s_{34}^2] = \mathbb{E}[(\Delta f_{t_3} - \Delta f_{t_4})^2] = \sigma_{t_3}^2 + \sigma_{t_4}^2$$

Therefore, the variance of mean frequency f can be found as:

$$\sigma^2 = \frac{1}{16} \cdot ((s_{13}^2 + s_{14}^2 + s_{23}^2 + s_{24}^2) - (s_{12}^2 + s_{34}^2))$$

Recalling that variance and squared differences are a function of frequency:

$$\sigma^2(f) = \frac{1}{16} \cdot ((s_{13}^2(f) + s_{14}^2(f) + s_{23}^2(f) + s_{24}^2(f)) - (s_{12}^2(f) + s_{34}^2(f)))$$

For each segment and time point we calculated the numerical function $\sigma^2(f)$. Then for each genotype having mean frequency f_x we estimated its variance as $\sigma^2(f_x)$

Merging amino acid genotypes

We merged nucleotide genotypes that corresponded to the same amino acid sequence and summed their frequencies and variances. We filtered out all genotypes x that had any of following patterns:

$f_x^{t_0} = 0, f_x^{t_1} = 0, f_x^{t_2} > 0$ or $f_x^{t_0} = 0, f_x^{t_1} > 0, f_x^{t_2} = 0$ or $f_x^{t_0} > 0, f_x^{t_1} = 0, f_x^{t_2} > 0$. Fraction of such genotypes were $< 0.5\%$ for all segments except S9, for which it was 4.5%

For further analysis, this amino acid dataset was used except when specified.

Fitness estimation

Number of cells in a pool with particular genotype x after time interval t increases exponentially

$$n_x^t = n_x^0 \cdot \text{Exp}[s_x \cdot t],$$

where s_x is absolute fitness. Frequency of genotype x as well depends exponentially on absolute fitness with an additional multiplicative factor:

$$f_x^t = \frac{n_x^t}{N^t} = \frac{n_x^0 \cdot \text{Exp}[s_x \cdot t]}{N^t} = \frac{f_x^0 \cdot \text{Exp}[s_x \cdot t]}{N^t/N^0},$$

where N^t and N^0 are total cell numbers in a pool at time points 0 and t . Factor $\frac{1}{N^t/N^0}$ reflects the total growth of population, it changes with time but is the same for all genotypes. Therefore,

we can rewrite genotype frequency at time t as:

$$f_x^t = f_x^0 \cdot \text{Exp}[(s_x - \overline{s^{0t}}) \cdot t],$$

where $\overline{s^{0t}} = \frac{1}{t} \cdot \text{Log}\left(\frac{N^t}{N^0}\right)$

In the measured dataset for each genotype x we have 3 measurements of frequency $f_x^{t_0}, f_x^{t_1}, f_x^{t_2}$ and their errors $\sigma^2(f_x^{t_0}), \sigma^2(f_x^{t_1}), \sigma^2(f_x^{t_2})$. To estimate genotype fitness, we minimized relative squared errors of exponential fit as function of fitness s_x and initial frequency f_x^0 :

$$(s_x, f_x^0) = \text{argmin}_{s_x, f_x^0} \left(\frac{(f_x^{t_0} - f_x^0)^2}{\sigma^2(f_x^{t_0})} + \frac{(f_x^{t_1} - f_x^0 \cdot \text{Exp}[(s_x - \overline{s^{01}}) \cdot t_1])^2}{\sigma^2(f_x^{t_1})} + \frac{(f_x^{t_2} - f_x^0 \cdot \text{Exp}[(s_x - \overline{s^{02}}) \cdot t_2])^2}{\sigma^2(f_x^{t_2})} \right) \quad (1)$$

This formula contains four parameters common for all genotypes from one segment: $\overline{s^{01}}, \overline{s^{02}}, t_1, t_2$. Further we will perform additional shifting and scaling of fitness values (see next section), therefore, without loss of generality we could set $\overline{s^{01}} = 0$ and $t_1 = 1$. Ideally, t_2/t_1 should equal 14; however, we noticed that this ratio does not hold for many segments and fitted $k = t_2/t_1$ from data instead of using value 14.

To find specific $\overline{s^{02}}$ and k for each segment we selected genotypes with high frequencies at t_0 ($t_0 > 25 \cdot 10^{-6}$) that corresponds to ~ 500 - 1000 reads per technical replicate. Each segment contains 10^3 - 10^4 genotypes that meet this criterion. We minimized eq. (1) for selected genotypes trying all possible combinations of $(\overline{s^{02}}, k)$ from a grid where $\overline{s^{02}} \in [0, 1.2]$ with step 0.01 and $k \in [1, 14]$ with step 0.1 and choose $(\overline{s^{02}}, k)$ which gives minimal (*).

Finally, given $(\overline{s^{02}}, k)$ for each segment we found s_x for each genotype. Errors for fitness values, s_x^{std} , were estimated as standard error of best-fit parameter.

For genotypes with frequencies pattern $f_x^{t_0} > 0, f_x^{t_1} = 0, f_x^{t_2} = 0$ fit of Eq (1) cannot be obtained. Therefore, we defined upper boundary for their fitness value as $s_x^{boundary} = \text{Log}\left(\frac{1}{\max(N_1^{t_1}, N_2^{t_1}, N_3^{t_1}, N_4^{t_1})}\right)$, where $N_i^{t_1}, i = 1..4$ are total read numbers at time point t_1 in i -th replica.

Fitness rescaling

We scaled fitness such that lethal genotypes have fitness 0 and neutral genotypes have fitness 1. We assumed that genotypes with a stop codon or frame shift are lethal. Thus, for each segment we linearly rescaled the fitness distribution so that 95% of genotypes with nonsense mutations have a fitness of 0 and so that the local maximum of the fitness distribution of genotypes with extant amino acids is 1. The scaling around the local maximum led to the shift of fitness values of less than +/- 0.025 in each of the 12 segments compared to the measured wildtype strains and did not affect our results (for scale, we called an amino acid change non-neutral if its effect on fitness was > 0.4). All fitness values that became smaller than 0 were set to 0.

Quality control and comparison of synonymous sequences

We used nucleotide synonymous sequences as an internal control. The error rate for a measurement of fitness of an amino acid sequence depends on the number of synonymous sequences, n , that were used to estimate it. Therefore, we estimated the false discovery rates separately for categories with $n = 1, \dots, 10$ variants. For each amino acid genotype with more than n synonymous variants we merged random combination of n of its nucleotide genotypes and estimated fitness. We then calculated the difference between this fitness and the fitness of the corresponding amino acid sequence. We classified case as “false unfit” if difference was < -0.4 and as “false fit” if difference was > 0.4. The fraction of such cases gives us false discovery rates for genotypes having n synonymous variants. To get total false discovery rates for each segment we averaged

“false unfit” and “false fit” rates for different n with weights equal to the fractions of genotypes in amino acid dataset which have n synonymous variant (S2 Supporting Information). The high correlation between biological replicas (S2 Supporting Information) confirms high accuracy of our high-throughput experiments, with the exception of segment 9.

Analysis of amino acid replacement effects in different backgrounds

For each amino acid replacement, we calculated its fitness effect in different backgrounds. We estimated the fraction of backgrounds in which a replacement exhibits deleterious, beneficial and neutral effects. To get the fraction of backgrounds with neutral effects we utilized the approach of mixture distribution analysis from Sarkisyan *et al.* [10]. We assume that neutral replacements have the same distribution as the fitness effects of synonymous replacements and are caused by measurement noise. We then calculated the fraction of backgrounds in which mutations have a neutral effect as the overlap between the distribution of synonymous mutations and distribution of fitness effects across different backgrounds. In remaining backgrounds, the amino acid replacement was called to have a non-neutral effect. Among them we counted those with strong fitness effects, including deleterious mutations when the fitness effect was < -0.4 and beneficial when the fitness effect was > 0.4 . We concluded that a particular amino acid replacement exhibits a strong deleterious or beneficial effect in some background if the fraction of such backgrounds exceeded the false discovery rate.

Predicting fitness using deep learning

To predict the unidimensional fitness function based on additive contribution of extant amino acid states we used deep learning, a machine learning technique capable of constructing virtually any function, even with a simple neural network architecture [59] (Fig 4B). To convert amino acid sequences into a binary feature matrix we used one-hot encoding strategy, in which each feature (column in the matrix) indicates the presence or absence of a particular amino acid state. For neural network implementation the TensorFlow library was used (www.tensorflow.org/about/bib).

To optimise the accuracy/overfitting ratio, we tested different combinations of neural network architectures and parameters. As a starting point, we selected a number of complex architectures, which describe our data but are prone to overfitting due to their large number of parameters. We then gradually reduced the number of layers and neurons to reduce the overfitting, while empirically controlling for accuracy.

Our final architecture consists of three layers and 22 neurons in total (Fig 4). Each neuron performs a linear transformation of the input and subsequently applies a non-linear sigmoid activation function to the result. The output of the first layer is a single sigmoid of a linear transformation of the feature vector, *i.e.* $\sigma(c_1^T x + b_1)$ where x is the feature vector, c_1 is the vector of coefficients, b_1 is the bias and $\sigma(t) = (1 + e^{-t})^{-1}$. Looking forward, observe that $c_1^T x$ is a fitness potential of the genotype x (see main text). The second layer decompresses the hidden nonlinear representation into 20 sigmoids, the combination of which is further linearly transformed with the only neuron of the third layer and wrapped into another sigmoid function:

$$F(x) = \sigma\left(\sum_{i=1}^{20} c_{3,i} \cdot (c_{2,i} \cdot \sigma(c_1^T x + b_1) + b_{2,i}) + b_3\right)$$

In the formula above, $c_{2,i}$ is the coefficient of the i -th neuron in the n -th layer, and $b_{2,i}$ is the bias of the i -th neuron in the second layer (the biases of the only neurons of the first and third layers are b_1 and b_3 , correspondingly).

The key idea of our approach is that the number of neurons in the first layer of the neural network determines the number of linear combinations of mutations (or fitness potentials) used in order to predict the fitness of the variant. In other words, each neuron in the first layer assigns a single unique weight to every amino acid state in the dataset (Fig 4). Thus, the number of neurons in the first layer of the architecture is the dimensionality of epistasis in the model (*i.e.* one in this case). The fitness potentials are then transformed by a nonlinear phase shift function constructed by the 22 neurons of the neural network.

The simplicity of the architecture minimizes overfitting, which was further prevented by keeping 10% of the data as a test set (in order to see how well the model performs on a fraction of the data it has never seen) and early stopping (training was stopped if the test accuracy did not improve for 10 epochs). The loss function that is being optimised is not convex, which leads to a high probability of getting stuck in different local minima. To ensure reproducibility, each of our models was constructed ten independent times using random train-test splits. The accuracies of the 10 constructed models varied by at most 2%.

Each model was trained for under 100 epochs using mean squared error as the loss function. An adaptive learning rate method proposed by Geoffrey Hinton, RMSProp, was used as the optimiser [60]. This algorithm is a version of a mini-batch stochastic gradient descent, utilising the gradient magnitude of the recent gradients in order to normalise the current ones. All the weights were initialised using Xavier normal initialiser [61].

Paths between pairs of fit genotypes

For analysis in Fig 7D, we first choose two fit “parental” genotypes, one randomly chosen genotype (eg: ABE) and the other parental genotype that is either *S. cerevisiae* wildtype genotype (inter-segmental) or another random fit genotype in the data (intra-segmental) (eg: abe). The two genotypes in this example are Hamming Distance 3 apart ($HD = 3$). We next compute all ($2^{HD} - 2$) intermediate genotypes (eg: AbC, aBc, *et cetera*) and retain the subset that were experimentally measured. We represent the two parental genotypes and all measured intermediate genotypes as an undirected graph in which each genotype is a vertex. All genotypes one amino acid replacement apart are connected by an unweighted edge. The shortest possible path for a given pair of genotypes is of length HD. We find all shortest paths between the two parental genotypes using a breadth-first search. We next remove all vertices (genotypes) that are unfit, and recompute the number of shortest between the two parental genotypes. For example, in Fig 7A, there are six paths of length three if you take into account all genotypes, but only three paths of length three if you take into account only fit genotypes.

Clustering of unfit genotypes in sequence space

For the analysis in Fig 7E, we first represent the two parental genotypes and all measured intermediate genotypes as an undirected graph in which each genotype is a vertex. All genotypes one amino acid replacement apart are connected by an unweighted edge. We can then compute the degree (number of genotypes of distance one) for each vertex (genotype). We do so randomly drawing from all measured genotypes and using only unfit genotypes or using the same number but randomly chosen genotypes. For the randomly chosen genotypes, the value is the average over 1000 runs.

Quantifying sign epistasis

For each amino acid replacement (eg: C → S at position 141), we considered only those that exhibit a large fitness effect (abs. difference > 0.4) comprising a set of amino acid replacements with large effects. For each amino acid replacement, we divided the genetic backgrounds into

two categories: those in which the replacement caused a > 0.4 increase in fitness, and those backgrounds in which the replacement caused > 0.4 decrease in fitness. A single amino acid replacement can cause a large increase in fitness in some backgrounds and a large decrease in others due to two possible reasons: sign epistasis or experimental error. To differentiate the two cases, we identified secondary amino acid replacements that significantly alter the ratio of large increases to large decreases in fitness (Fisher's exact test, Bonferroni corrected p -value < 0.05). We only consider a site to be under sign epistasis if there is a second site that alters the frequency of sign epistasis in a statistically significant manner, *i.e.* more frequently than expected by chance alone.

Ancestral state reconstruction

We reconstructed ancestral amino acid states using maximum likelihood approach implemented in CODEML program of PAML 4 [62].

Structural analysis

Structure prediction. An initial model was obtained with the I-TASSER server [63]. The list of top 10 PDB structural templates picked up by the I-TASSER included high-quality crystal structures of imidazoleglycerol-phosphate dehydratases from *Arabidopsis thaliana* and *Cryptococcus neoformans*. Coordinates of the top-scoring model (C-score = 0.21, estimated TM-score = 0.74 ± 0.11 , estimated RMSD = $5.1 \pm 3.3 \text{ \AA}$) and the predicted normalized B-factor [64] were used for further analysis. The value of the model quality metric (TM-score > 0.5) indicates a model of correct topology. The proteins structurally close to the final model (RMSD $0.6\text{--}1.7 \text{ \AA}$ are PDB IDs 4MU0, 4GQU, 1RHY, 5DNL and 2AE8 from *Arabidopsis thaliana*, *Mycobacterium tuberculosis*, *Cryptococcus neoformans*, *Pyrococcus furiosus* and *Staphylococcus aureus*).

We measured the distribution of distances (in angstroms) between pairs of residues that exhibit strong sign epistasis (S2 Supporting Information, ReallyPositivePair == TRUE), and compared it with the distribution of pairwise distances among residues for which we have sufficient data to be certain that a given pair does not exhibit sign epistasis (S2 Supporting Information, ReallyNegativePair == TRUE).

$\Delta\Delta G$ prediction. Cartesian_ddg application [65] from Rosetta version 2017.08.59291 was used for $\Delta\Delta G$ predictions. Top-scoring I-TASSER model was pre-minimized using the Relax [66] application in dual-space [67] with the flags: -relax:dualspace true; -ex1; -ex2; -use_input_sc; -flip_HNQ; -no_optH false; -relax:min_typedbfgs_armijo_nonmonotone; -nonideal. The best scoring model from 1000 structures was selected. The effect of up to 4 mutations (54,500 genotypes in total) was assessed in Cartesian space with the Talaris_2014 score function, and the -fa_max_dis 9.0 flag. $\Delta\Delta G$ was estimated as a difference of mean score for 3 independent runs for every mutant and the wild-type score.

Supporting information

S1 Fig. Experimental design. **a**, The sequence of the His3 protein from *S. cerevisiae* was separated into 12 independent segments of similar lengths, such that the full combinatorial set of extant amino acid replacements was less than 100,000 possible genotypes. These segments represented different combinations of structural elements of the His3 protein structure. **b**, For each of the 12 segments from His3, we selected extant amino acid states using a multiple alignment of His3 orthologues from 396 species, preferentially incorporating states from 21 yeast species, the variability is shown in segment 3 as an example. Mutant degenerate codon libraries were constructed by fusion PCR of two synthesized variable halves of each segment. These

high-complexity plasmid libraries were transformed into haploid His3 knockout *S. cerevisiae* strain. The growth rate of yeast carrying different extant amino acid state combinations in His3 gene was measured using serial batch culture in the absence of histidine with 12 hours between ~100-fold dilutions. To estimate the fitness of yeast mutants their relative abundance was measured at three points: in the initial population before selection (t_0), in the population after 12 hours of growth in the selective medium (t_1), and in the final population after 168 hours of growth in the selective medium (t_{14}). To assess the fitness of individual mutants the segments from three populations were amplified and sequenced. The relative abundance of each sequence was used as a proxy for abundance of the associated yeast mutant, which in turn determines its fitness. **c**, Secondary structure of His3 mapped to the segments in our experiments.

(TIFF)

S2 Fig. Sequencing strategy and accuracy analysis of His3 segment libraries. **a**, The pair-end reads strategy ensured that all variable regions were read twice. Mismatches between the reads were corrected to the highest quality nucleotide. The differences between the corrected sequence and the expected sequence in the constant region were used to estimate the sequencing error rate. **b**, Histograms of the fraction of nonsense (blue, expected fitness = 1) and extant amino acid state combinations (red). 99.63% of nonsense genotypes have a fitness > 0.4 while 23.46% of extant amino acid combinations have fitness < 0.6. **c**, The distribution of measured fitness of synonymous variants for four different amino-acid genotypes. **d**, The distribution of the difference between fitness of a nucleotide genotype with the mean fitness of all nucleotide genotypes with the corresponding amino acid sequence. We binned the measurements by the frequency of each nucleotide sequence in our data. Only amino acid genotypes with ≥ 10 nucleotide genotypes with measured fitness were used. Bars represent 99% of the distribution. **e**, Correlation of sequencing error rate in each segments, which varied 3-fold, and different measures of our data. None of the correlations were statistically significant, providing evidence that sequencing errors do not substantially contribute to our results. **f**, The measured fitness for all nonsense genotypes binned as a function of genotype frequency at t_0 . **g**, The measured genotype frequencies of 7 spiked-in clones that are synonymous with the wild type *S. cerevisiae* HIS3 gene, and therefore expected to all have nearly equal fitness, in two independent competition experiments. Note that the two y-axes in each replicate are scaled, as the absolute genotype frequencies change due to genotypes with low fitness dropping out.

(TIF)

S3 Fig. Segment-specific fitness distributions for extant and non-extant amino acid states.

a, The fitness distribution for each segment for genotypes consisting only of extant amino acid states (green) or that contain one or more non-extant amino acid states (purple) only at positions with a replacement in the extant library. **b**, The fitness distribution for each segment for genotypes consisting only of extant amino acid states (green) and genotypes with mutations at other positions in that segment (red).

(TIF)

S4 Fig. Growth rate measurement of isolated strains. **a**, Comparison of fitness values from the pooled competition assay with growth rates of isolated strains as measured in a microplate reader. Error bars for growth rates show s.e.m. of 6 replicates. **b**, Pearson correlation coefficients between fitness values from competition and growth rates of isolated strains for each segment. ** signifies p-value < 0.005 (correlation test). **c**, His3p orthologues from different species complement a Δ his3 deletion in *S. cerevisiae*. Growth rates of transformants containing whole HIS3 orthologous genes from other yeast species. Error bars for growth rates show s.e.

m. of ≥ 7 replicates.
(TIF)

S5 Fig. The fitness potential of the His3 fitness landscape. **a**, Fitness potential predicted by the neural network as a function of the measured fitness for all 12 segments. **b**, The correlation between the fitness predicted by the fitness potential and the measured fitness. **c**, Training and test R^2 for each segment for 20-fold cross-validation.
(TIF)

S6 Fig. Sign epistasis dimensionality graphs for all twelve segments. Each node represents a substitution, with multiple replacements at the same site having the same colour. Replacements under reciprocal sign epistasis are indicated by black lines while grey arrows indicate unidirectional sign epistasis.
(TIF)

S7 Fig. Multidimensional description of epistasis in His3 segments. **a**, Increasing the number of neurons in the first layer of the neural network, which is equivalent to increasing the number of underlying fitness potentials, leads to more accurate models for segments with detected sign epistasis. Each dot corresponds to an independent optimization of model parameters. **b**, Number of sign epistatic interactions of certain amino acid replacements against average model prediction power for mutants including these amino acid replacements.
(TIF)

S8 Fig. Protein stability and the fitness potential. **a**, A comparison of correlation coefficients between predicted and measured values across segments. **b,c**, Correlations between the estimated impact of amino acid replacements on folding free energy ($\Delta\Delta G$), fitness potential and fitness. $\Delta\Delta G$ correlates better with fitness potential than with fitness. **d**, Pairs of sites that exhibit sign (connected by a light edge in [S7 Fig](#)) and those that exhibit reciprocal sign epistasis (connected by a dark edge in [S7 Fig](#)) are closer together in the His3p structure than randomly chosen non-connected pairs of positions that exhibit sign epistasis.
(TIF)

S9 Fig. Decoupling inter- and intra-segmental epistasis. **a**, The fraction of unfit genotypes between *S. cerevisiae* and any other genotype consisting of extant amino acid states with high (blue) or any (red) fitness, and genotypes in the latter but not the former category (black) as a function of the Hamming distance between the two boundary genotypes. Points indicate median, the bars and lines indicate 50% of the genotypes and genotypes 2.7 sigmas from the mean, respectively. **b**, The neural network model assigns higher weights to amino acid states that first occur in His3 orthologues farther from *S. cerevisiae*, indicating the presence of intrasegmental interactions. **c**, The fraction of unfit genotypes of all instances when the segment matched an extant or ancestral species as a function of Hamming distance to *S. cerevisiae* sequence.
(TIF)

S1 Supporting Information. Multiple alignment of His3 orthologues.
(FAS)

S2 Supporting Information. A statistical summary of segment libraries and sequencing results.
(XLSX)

S3 Supporting Information. Multidimensional description of epistasis in His3 segments. Fitness as a function of two fitness potentials (black dots, measured fitness is depicted in red).
(GIF)

Acknowledgments

We thank Jochen Hecht of the CRG Genomics Unit and Svetlana Belogorodtseva and Roman Belogorodtsev for the provided technical assistance.

Author Contributions

Conceptualization: Victoria O. Pokusaeva, Dinara R. Usmanova, Ekaterina V. Putintseva, Inna S. Povolotskaya, Guillaume J. Filion, Lucas B. Carey, Fyodor A. Kondrashov.

Data curation: Dinara R. Usmanova, Ekaterina V. Putintseva, Karen S. Sarkisyan, Guillaume J. Filion, Lucas B. Carey, Fyodor A. Kondrashov.

Formal analysis: Victoria O. Pokusaeva, Dinara R. Usmanova, Ekaterina V. Putintseva, Alexander S. Mishin, Natalya S. Bogatyreva, Dmitry N. Ivankov, Arseniy V. Akopyan, Sergey Ya. Avvakumov, Guillaume J. Filion, Lucas B. Carey, Fyodor A. Kondrashov.

Funding acquisition: Guillaume J. Filion, Lucas B. Carey, Fyodor A. Kondrashov.

Investigation: Dinara R. Usmanova, Ekaterina V. Putintseva, Guillaume J. Filion, Lucas B. Carey, Fyodor A. Kondrashov.

Methodology: Victoria O. Pokusaeva, Dinara R. Usmanova, Ekaterina V. Putintseva, Lorena Espinar, Karen S. Sarkisyan, Guillaume J. Filion, Lucas B. Carey, Fyodor A. Kondrashov.

Project administration: Lorena Espinar, Lucas B. Carey, Fyodor A. Kondrashov.

Resources: Lucas B. Carey, Fyodor A. Kondrashov.

Software: Ekaterina V. Putintseva, Guillaume J. Filion.

Supervision: Guillaume J. Filion, Lucas B. Carey, Fyodor A. Kondrashov.

Validation: Guillaume J. Filion, Lucas B. Carey, Fyodor A. Kondrashov.

Visualization: Ekaterina V. Putintseva, Lucas B. Carey, Fyodor A. Kondrashov.

Writing – original draft: Fyodor A. Kondrashov.

Writing – review & editing: Lucas B. Carey, Fyodor A. Kondrashov.

References

1. Dean A. M. & Thornton J. W. Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Rev. Genet.* 8, 675–688 (2007). <https://doi.org/10.1038/nrg2160> PMID: [17703238](https://pubmed.ncbi.nlm.nih.gov/17703238/)
2. Kogenaru M., de Vos M. G. J., & Tans S. J. Revealing evolutionary pathways by fitness landscape reconstruction. *Crit. Rev. Biochem. Mol. Biol.* 44, 169–174 (2009). <https://doi.org/10.1080/10409230903039658> PMID: [19552615](https://pubmed.ncbi.nlm.nih.gov/19552615/)
3. Mackay T. F. C. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Rev. Genet.* 15, 22–33 (2014). <https://doi.org/10.1038/nrg3627> PMID: [24296533](https://pubmed.ncbi.nlm.nih.gov/24296533/)
4. de Visser J. A. G. M. & Krug J. Empirical fitness landscapes and the predictability of evolution. *Nature Rev. Genet.* 15, 480–490 (2014). <https://doi.org/10.1038/nrg3744> PMID: [24913663](https://pubmed.ncbi.nlm.nih.gov/24913663/)
5. de Visser J. A. G. M., Cooper T. F., & Elena S. F. The causes of epistasis. *Proc. Biol. Sci.* 278, 3617–3624 (2011). <https://doi.org/10.1098/rspb.2011.1537> PMID: [21976687](https://pubmed.ncbi.nlm.nih.gov/21976687/)
6. Wright S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. Sixth Int. Congr. Genet.* 1, 356–366 (1932).
7. Olson C. A., Wu N. C. & Sun R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* 24, 2643–2651 (2014). <https://doi.org/10.1016/j.cub.2014.09.072> PMID: [25455030](https://pubmed.ncbi.nlm.nih.gov/25455030/)

8. Puchta O., Cseke B., Czaja H., Tollervey D., Sanguinetti G., and Kudla G. Network of epistatic interactions within a yeast snoRNA. *Science*. 352, 840–844 (2016). <https://doi.org/10.1126/science.aaf0965> PMID: [27080103](https://pubmed.ncbi.nlm.nih.gov/27080103/)
9. Li C. Qian, W. Maclean, C. J. & Zhang, J. The fitness landscape of a tRNA gene. *Science* 352, 837–840 (2016). <https://doi.org/10.1126/science.aae0568> PMID: [27080104](https://pubmed.ncbi.nlm.nih.gov/27080104/)
10. Sarkisyan K.S., et al. Local fitness landscape of the green fluorescent protein. *Nature* 533, 397–401 (2016.) <https://doi.org/10.1038/nature17995> PMID: [27193686](https://pubmed.ncbi.nlm.nih.gov/27193686/)
11. Hietpas R. T., Bank C., Jensen J. D. & Bolon D. N. A. Shifting fitness landscapes in response to altered environments. *Evolution* 67, 3512–3522 (2013). <https://doi.org/10.1111/evo.12207> PMID: [24299404](https://pubmed.ncbi.nlm.nih.gov/24299404/)
12. Parera M. & Martinez M. A. Strong epistatic interactions within a single protein. *Mol. Biol. Evol.* 31, 1546–1553 (2014). <https://doi.org/10.1093/molbev/msu113> PMID: [24682281](https://pubmed.ncbi.nlm.nih.gov/24682281/)
13. Podgornaia A. I. & Laub M. T. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* 347, 673–677 (2015). <https://doi.org/10.1126/science.1257360> PMID: [25657251](https://pubmed.ncbi.nlm.nih.gov/25657251/)
14. Anderson D. W. McKeown, A. N. & Thornton, J. W. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *Elife* 4, e07864 (2015). <https://doi.org/10.7554/eLife.07864> PMID: [26076233](https://pubmed.ncbi.nlm.nih.gov/26076233/)
15. Meini M. R., Tomatis P. E., Weinreich D. M. & Vila A. J. Quantitative description of a protein fitness landscape based on molecular features. *Mol. Biol. Evol.* 32, 1774–1787 (2015). <https://doi.org/10.1093/molbev/msv059> PMID: [25767204](https://pubmed.ncbi.nlm.nih.gov/25767204/)
16. Bank C., Matuszewski S., Hietpas R. T. & Jensen J. D. On the (un)predictability of a large intragenic fitness landscape. *Proc. Natl. Acad. Sci. USA*. 113, 14085–14090 (2016). <https://doi.org/10.1073/pnas.1612676113> PMID: [27864516](https://pubmed.ncbi.nlm.nih.gov/27864516/)
17. Stemmer W. P. Rapid evolution of a protein in vitro by DNA shuffling. *Nature*. 370, 389–391 (1994). <https://doi.org/10.1038/370389a0> PMID: [8047147](https://pubmed.ncbi.nlm.nih.gov/8047147/)
18. Poelwijk F.J., Kiviet D.J., Weinreich D.M., & Tans S.J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* 445, 383–386(2007). <https://doi.org/10.1038/nature05451> PMID: [17251971](https://pubmed.ncbi.nlm.nih.gov/17251971/)
19. Gong L. I. Suchard M. A. & Bloom J.D. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* 2, e00631(2013). <https://doi.org/10.7554/eLife.00631> PMID: [23682315](https://pubmed.ncbi.nlm.nih.gov/23682315/)
20. Gong L. I. & Bloom J.D. Epistatically interacting substitutions are enriched during adaptive protein evolution. *PLoS Genet.* 10, e1004328(2014). <https://doi.org/10.1371/journal.pgen.1004328> PMID: [24811236](https://pubmed.ncbi.nlm.nih.gov/24811236/)
21. Starr T. N. & Thornton J. W. Epistasis in protein evolution. *Protein Sci.* 25, 1204–1218 (2016). <https://doi.org/10.1002/pro.2897> PMID: [26833806](https://pubmed.ncbi.nlm.nih.gov/26833806/)
22. Kumar A. Natarajan C Moriyama H Witt C.C. Weber R.E. Fago A.& Storz J.F. Stability-mediated epistasis restricts accessible mutational pathways in the functional evolution of avian hemoglobin. *Mol. Biol. Evol.* 34, 1240–1251(2017). <https://doi.org/10.1093/molbev/msx085> PMID: [28201714](https://pubmed.ncbi.nlm.nih.gov/28201714/)
23. Tufts D. M., et al. Epistasis constrains mutational pathways of hemoglobin adaptation in high-altitude pikas. *Mol. Biol. Evol.* 32, 287–298 (2015). <https://doi.org/10.1093/molbev/msu311> PMID: [25415962](https://pubmed.ncbi.nlm.nih.gov/25415962/)
24. Stemmer W. P. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc Natl Acad Sci USA*. 91, 10747–10751 (1994). PMID: [7938023](https://pubmed.ncbi.nlm.nih.gov/7938023/)
25. Salinas VH, Ranganathan R. Coevolution-based inference of amino acid interactions underlying protein function. *Elife* 7, pii: e34300 (2018). <https://doi.org/10.7554/eLife.34300> PMID: [30024376](https://pubmed.ncbi.nlm.nih.gov/30024376/)
26. Eyre-Walker A. & Keightley P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8, 610–618 (2007). <https://doi.org/10.1038/nrg2146> PMID: [17637733](https://pubmed.ncbi.nlm.nih.gov/17637733/)
27. Draghi J. A. & Plotkin J. B. Selection biases the prevalence and type of epistasis along adaptive trajectories. *Evolution* 67, 3120–3131 (2013). <https://doi.org/10.1111/evo.12192> PMID: [24151997](https://pubmed.ncbi.nlm.nih.gov/24151997/)
28. Lunzer M., Miller S. P., Felsheim R. & Dean A. M. The biochemical architecture of an ancient adaptive landscape. *Science* 310, 499–501 (2005). <https://doi.org/10.1126/science.1115649> PMID: [16239478](https://pubmed.ncbi.nlm.nih.gov/16239478/)
29. Palmer A. C., Toprak E., Baym M., Kim S., Veres A., Bershtein S. & Kishony R. Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes. *Nat. Commun.* 6, 7385 (2015). <https://doi.org/10.1038/ncomms8385> PMID: [26060115](https://pubmed.ncbi.nlm.nih.gov/26060115/)
30. Lalić J. & Elena S. F. Magnitude and sign epistasis among deleterious mutations in a positive-sense plant RNA virus. *Heredity* 109, 71–77 (2012). <https://doi.org/10.1038/hdy.2012.15> PMID: [22491062](https://pubmed.ncbi.nlm.nih.gov/22491062/)
31. Keefe A. D. & Szostak J. W. Functional proteins from a random-sequence library. *Nature* 410, 715–718 (2001). <https://doi.org/10.1038/35070613> PMID: [11287961](https://pubmed.ncbi.nlm.nih.gov/11287961/)
32. du Plessis L. Leventhal, G.E. & Bonhoeffer, S. How good are statistical models at approximating complex fitness landscapes? *Mol. Biol. Evol.* 33, 2454–2468(2016). <https://doi.org/10.1093/molbev/msw097>

33. Lum P. Y. Edwards S. & Wright R. Molecular, functional and evolutionary characterization of the gene encoding HMG-CoA reductase in the fission yeast, *Schizosaccharomyces pombe*. *Yeast* 12, 1107–1124 (1996). [https://doi.org/10.1002/\(SICI\)1097-0061\(19960915\)12:11%3C1107::AID-YEA992%3E3.0.CO;2-E](https://doi.org/10.1002/(SICI)1097-0061(19960915)12:11%3C1107::AID-YEA992%3E3.0.CO;2-E) PMID: 8896278
34. Kondrashov A.S. Sunyaev S. & Kondrashov F. A. Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. USA* 99, 14878–14883 (2002). <https://doi.org/10.1073/pnas.232565499> PMID: 12403824
35. Kvittek D. J. & Sherlock G. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet.* 7, e1002056 (2011). <https://doi.org/10.1371/journal.pgen.1002056> PMID: 21552329
36. Breen M. S., Kemena C., Vlasov P. K., Notredame C. & Kondrashov F. A. Epistasis as the primary factor in molecular evolution. *Nature* 490, 535–538 (2012). <https://doi.org/10.1038/nature11510> PMID: 23064225
37. Gavrillets S. Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol.* 12, 307–312 (1997). [https://doi.org/10.1016/S0169-5347\(97\)01098-7](https://doi.org/10.1016/S0169-5347(97)01098-7) PMID: 21238086
38. Salverda M. L. M., Dellus E., Gorter F. A., Debets A.J. M., van der Oost J. Hoekstra R. F. Tawfik D. S. & de Visser J.A.G. M. Initial mutations direct alternative pathways of protein evolution. *PLoS Genet.* 7, e1001321(2011). <https://doi.org/10.1371/journal.pgen.1001321> PMID: 21408208
39. Weinreich D. M., Delaney N. F., de Pisto M. A. & Hartl D.L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312, 111–114(2006). <https://doi.org/10.1126/science.1123539> PMID: 16601193
40. Milkman R. Selection differentials and selection coefficients. *Genetics* 88, 391–403 (1978). PMID: 17248802
41. Kimura M. & Crow J. F. Effect of overall phenotypic selection on genetic change at individual loci. *Proc. Natl. Acad. Sci. USA*. 75, 6168–6171 (1978). PMID: 282633
42. Kondrashov F. A. & Kondrashov A. S. Multidimensional epistasis and the disadvantage of sex. *Proc. Natl. Acad. Sci. USA*. 98, 12089–12092 (2001). <https://doi.org/10.1073/pnas.211214298> PMID: 11593020
43. Stiffler M. Hekstra D. & Ranganathan R. Evolvability as a function of purifying selection in TEM-1-lactamase. *Cell* 160, 882–892 (2015). <https://doi.org/10.1016/j.cell.2015.01.035> PMID: 25723163
44. Crow J. F. & Kimura M. Efficiency of truncation selection. *Proc. Natl. Acad. Sci. USA* 76, 396–399 (1979). PMID: 16592610
45. Weinreich D. M. Lan Y. Wylie C. S. & Heckendorn R. B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* 23, 700–707 (2013). <https://doi.org/10.1016/j.gde.2013.10.007> PMID: 24290990
46. Poelwijk F.J. Krishna V. & Ranganathan R. The Context-Dependence of Mutations: A Linkage of Formalisms. *PLoS Comput. Biol.* 12, e1004771 (2016). <https://doi.org/10.1371/journal.pcbi.1004771> PMID: 27337695
47. Sailer Z. R. & Harms M. J. High-order epistasis shapes evolutionary trajectories. *PLoS Comput. Biol.* 13, e1005541 (2017). <https://doi.org/10.1371/journal.pcbi.1005541> PMID: 28505183
48. Weinreich D. M. Watson R. A. & Chao L. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59, 1165–1174 (2005). PMID: 16050094
49. Rokyta D. R. et al. Epistasis between beneficial mutations and the phenotype-to-fitness Map for a ssDNA virus. *PLoS Genet.* 7, e1002075 (2011). <https://doi.org/10.1371/journal.pgen.1002075> PMID: 21655079
50. Kondrashov D. A. & Kondrashov F. A. Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* 31, 24–33 (2015). <https://doi.org/10.1016/j.tig.2014.09.009> PMID: 25438718
51. De Pisto M. A. Weinreich D. M. & Hartl D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Rev. Genet.* 6, 678–687 (2005). <https://doi.org/10.1038/nrg1672> PMID: 16074985
52. Bershtein S., Segal M., Bekerman R., Tokuriki N. & Tawfik D. S. Robustness epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* 444, 929–932 (2006). <https://doi.org/10.1038/nature05385> PMID: 17122770
53. Marks D. S. Colwell L. J. Sheridan R Hopf T. A. Pagnani A. Zecchina R. & Sander C. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6, e28766 (2011). <https://doi.org/10.1371/journal.pone.0028766> PMID: 22163331
54. Orr H. A. The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics* 139, 1805–1813 (1995). PMID: 7789779

55. Tamura K., Stecher G., Peterson D., FilipSKI A. & Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol.* 30, 2725–2729 (2013). <https://doi.org/10.1093/molbev/mst197> PMID: [24132122](https://pubmed.ncbi.nlm.nih.gov/24132122/)
56. Magoč T. & Salzberg S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963 (2011). <https://doi.org/10.1093/bioinformatics/btr507> PMID: [21903629](https://pubmed.ncbi.nlm.nih.gov/21903629/)
57. Zorita E., Cuscó P. & Filion G. J. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* 31, 1913–1919 (2015). <https://doi.org/10.1093/bioinformatics/btv053> PMID: [25638815](https://pubmed.ncbi.nlm.nih.gov/25638815/)
58. Meyerhans A., Vartanian J. P. & Wain-Hobson S. DNA recombination during PCR. *Nucleic acids research* 18, 1687–1691 1990 PMID: [2186361](https://pubmed.ncbi.nlm.nih.gov/2186361/)
59. LeCun Y., Bengio Y. & Hinton G. Deep learning. *Nature*. 521, 436–444 (2015). <https://doi.org/10.1038/nature14539> PMID: [26017442](https://pubmed.ncbi.nlm.nih.gov/26017442/)
60. Ruder, S. An overview of gradient descent optimization algorithms. Preprint at <http://arXiv.org/abs/1609.04747> (2016).
61. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (2010).
62. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591 (2007). <https://doi.org/10.1093/molbev/msm088> PMID: [17483113](https://pubmed.ncbi.nlm.nih.gov/17483113/)
63. Yang J. & Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* 43, W174–W181(2015). <https://doi.org/10.1093/nar/gkv342> PMID: [25883148](https://pubmed.ncbi.nlm.nih.gov/25883148/)
64. Yang J., Wang Y. & Zhang Y. ResQ: An Approach to Unified Estimation of B-Factor and Residue-Specific Error in Protein Structure Prediction. *J. Mol. Biol.* 428, 693–701 (2016). <https://doi.org/10.1016/j.jmb.2015.09.024> PMID: [26437129](https://pubmed.ncbi.nlm.nih.gov/26437129/)
65. Park H., Bradley P., Greisen P. Jr., Liu Y., Mulligan V. K., Kim D.E., Baker D. & DiMaio F. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* 12, 6201–6212 (2016). <https://doi.org/10.1021/acs.jctc.6b00819> PMID: [27766851](https://pubmed.ncbi.nlm.nih.gov/27766851/)
66. Nivón L. G., Moretti R. & Baker D. A Pareto-optimal refinement method for protein design scaffolds. *PLoS One* 8, e59004 (2013). <https://doi.org/10.1371/journal.pone.0059004> PMID: [23565140](https://pubmed.ncbi.nlm.nih.gov/23565140/)
67. Conway P., Tyka M. D., DiMaio F., Konerding D. E. & Baker D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* 23, 47–55 (2014). <https://doi.org/10.1002/pro.2389> PMID: [24265211](https://pubmed.ncbi.nlm.nih.gov/24265211/)