On the (un)predictability of a large intragenic fitness landscape

Claudia Banka,b,c,1, Sebastian Matuszewski a,b,c,1, Ryan T. Hietpas d,e, and Jeffrey D. Jensen b,c,2,3

a Instituto Gulbenkian de Ciência, 2780–156 Oeiras, Portugal; b School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; c Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland; d Eli Lilly and Company, Indianapolis, IN 46225; and e Department of Biochemistry & Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605

Edited by Andrew G. Clark, Cornell University, Ithaca, NY, and approved October 11, 2016 (received for review August 2, 2016)

The study of fitness landscapes, which aims at mapping genotypes to fitness, is receiving ever-increasing attention. Novel experimental approaches combined with next-generation sequencing (NGS) methods enable accurate and extensive studies of the fitness effects of mutations, allowing us to test theoretical predictions and improve our understanding of the shape of the true underlying fitness landscape and its implications for the predictability and repeatability of evolution. Here, we present a uniquely large multiallelic fitness landscape comprising 640 engineered mutants that represent all possible combinations of 13 amino acid-changing mutations at 6 sites in the heat-shock protein Hsp90 in Saccharomyces cerevisiae, in salinity. Despite a prevalence of negative epistasis in the landscape, we find that the global fitness peak is reached via four positively epistatic mutations. Combining traditional and extending recently proposed theoretical and statistical approaches, we quantify features of the global multiallelic fitness landscape. Using subsets of the data, we demonstrate that extrapolation beyond a known part of the landscape is difficult owing to both local ruggedness and amino acid-specific epistatic hotspots and that inference is additionally confounded by the nonrandom choice of mutations for experimental fitness landscapes.

Significance

The study of fitness landscapes is fundamentally concerned with understanding the relative roles of stochastic and deterministic processes in adaptive evolution. Here, the authors present a uniquely large and complete multiallelic intragenic fitness landscape of 640 systematically engineered mutations in the heat-shock protein Hsp90 in yeast. Using a combination of traditional and recently proposed theoretical approaches, they study the accessibility of the global fitness peak and the potential for predictability of the fitness landscape topography. They report local ruggedness of the landscape and the existence of epistatic hotspot mutations, which together make extrapolation and hence predictability inherently difficult if mutation-specific information is not considered.

Author contributions: C.B., R.T.H., and J.D.J. designed research; R.T.H. performed research; C.B. and S.M. analyzed data; and C.B., S.M., R.T.H., and J.D.J. wrote the paper.

The authors declare no conflict of interest.

Author contributions: C.B., R.T.H., and J.D.J. designed research; R.T.H. performed research; C.B. and S.M. analyzed data; and C.B., S.M., R.T.H., and J.D.J. wrote the paper.

The authors declare no conflict of interest.

1 C.B. and S.M. contributed equally to this work.
2 Present address: Center for Evolution & Medicine, School of Life Sciences, Arizona State University, Tempe, AZ 85281.
3 To whom correspondence should be addressed. Email: jeffrey.d.jensen@asu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1612676113/-/DCSupplemental.
high salinity. With all possible combinations of 13 mutations of various fitness effects at 6 positions, the presented landscape is not only uniquely large but also distinguishes itself from previously published work regarding several other experimental features—namely, by its systematic and controlled experimental setup using engineered mutations of various selective effects and by considering multiple alleles simultaneously. We begin by describing the landscape and identifying the global peak, which is reached through a highly positively epistatic combination of four mutations. Based on a variety of implemented statistical measures and models, we describe the accessibility of the peak, the pattern of epistasis, and the topography of the landscape. To accommodate our data, we extend several previously used models and statistics to the multiallelic case. Using subsets of the landscape, we discuss the predictive potential of such modeling and the problem of selecting nonrandom mutations when attempting to quantify local landscapes to extrapolate global features.

Results and Discussion

We used the EMPIRIC approach (24, 25) to assess the growth rate of 640 mutants in yeast Hsp90 (Materials and Methods). Based on previous screenings of fitness effects in different environments (26) and on different genetic backgrounds (20), and on expectations of their biophysical role, 13 amino acid-changing point mutations at 6 sites were chosen for the fitness landscape presented here (Fig. 1). The fitness landscape was created by assessing the growth rate associated with each individual mutation on the parental background and all possible mutational combinations. A previously described Monte Carlo Markov chain (MCMC) approach was used to assess fitness and credibility intervals (ref. 27 and Materials and Methods).

![Fig. 1. Individual amino acid substitutions and their effect on the parental background in elevated salinity, obtained from 1,000 samples from the posterior distribution of growth rates. Boxes represent the interquartile range [i.e., the 50% confidence interval (CI)], whiskers extend to the highest/lowest data point within the box ±1.5 times the interquartile range, and circles represent outliers; gray and white background shading alternates by amino acid position. The box below indicates with colored dots which mutations are involved in the focal landscapes discussed throughout the main text: the four mutations leading from the parental type to the global optimum ("opt"), the four individually most beneficial mutations on the parental background ("best"), and the four mutations with the individually lowest growth rates on the parental background ("worst"). (Inset) Parental sequence at positions 582–590 and assessed amino acids by position.](image)

![Fig. 2. (A) Empirical fitness landscape by mutational distance to parental type. Each line represents a single substitution. Vertical lines appear when multiple alleles have been screened at the same position. There is a global pattern of negative epistasis. The three focal landscapes are highlighted. (B) Close-up on the beneficial portion of the landscape. (C) Expected ("Exp.") versus observed ("Obs.") fitness for the focal landscapes, obtained from 1,000 posterior samples. We observe strong positive epistasis in the landscape that contains the global optimum, whereas the other two are dominated by negative epistasis. In A–C, the y axis depicts growth rate as a proxy for fitness.](image)

The Fitness Landscape and Its Global Peak. Fig. 2 presents the resulting fitness landscape, with each mutant represented based on its Hamming distance from the parental genotype and its median estimated growth rate. Lines connect single-step substitutions, with vertical lines occurring when there are multiple mutations at the same position (Fig. 1). With increasing Hamming distance from the parental type, many mutational combinations become strongly deleterious. Thus, we observe strong negative epistasis between the substitutions that, as single steps on the parental background, have small effects. This pattern is consistent with Fisher’s Geometric Model (28) when combinations of individually beneficial or small-effect mutations overshoot the optimum and with classic arguments predicting negative epistasis based on mutational load (29). It is also intuitively comprehensible on the protein level, where the accumulation of too many mutations is likely to destabilize the protein and render it dysfunctional (30).

The global peak of the fitness landscape is located four mutational steps away from the parental type (Fig. 2B), with 98% of posterior samples identifying the peak. The fitness advantage of the global peak reaches nearly 10% over the parental type and is consistent between replicates (Materials and Methods and SI Appendix, Fig. S3.1). Perhaps surprising given the degree of conservation of the studied genomic region (figure S5 in ref. 24), it is important to note that these fitness effects are measured under highly artificial experimental conditions including high salinity, which are unlikely to represent a natural environment of yeast. The effects of the individual mutations comprising the peak in a previous experiment without
added NaCl were $-0.04135$, $-0.01876$, $-0.03816$, and $-0.02115$ for mutations W585L, A587P, N588L, and M589A, respectively, emphasizing the potential cost of adaptation associated with the increased salinity environment (data from ref. 20; see also ref. 26).

Curiously, the global peak is not reached by combining the most beneficial single-step mutations on the parental background but via a highly synergistic combination of one beneficial and three “neutral” mutations (i.e., mutations that are individually indistinguishable from the parental type in terms of growth rate). In fact, each of the five beneficial local optima shows a similar signature of positive epistasis (SI Appendix, Fig. S3.2). Fig. 2C demonstrates that a combination of the four mutations involved in the global peak (opt) predicts only a 4% fitness advantage. Furthermore, even a combination of the four individually most-beneficial, single-step mutations on the parental background in the dataset (best) (considering at most one mutation per position) only predicts a benefit of 6%. Notably, the actual combination of these four mutations on the parental background is highly deleterious and thus exhibits strong negative epistasis. Although negative epistasis between beneficial mutations during adaptation has been reported more frequently, positive epistasis has also been observed occasionally (18, 31), particularly in the context of compensatory evolution. In fact, negative epistasis between beneficial mutations and positive epistasis between neutral mutations has been predicted by de Visser et al. (32). Furthermore, our results support the pattern recently found in the gene underlying the antibiotic resistance enzyme TEM-1 $\beta$-lactamase in Escherichia coli, showing that large-effect mutations interact more strongly than small-effect mutations such that the fitness landscape of large-effect mutations tends to be more rugged than the landscape of small-effect mutations (13) and that mutations that were selected for their combined beneficial effect on a wild-type background tend to interact synergistically, whereas mutations selected for their individual effects interact antagonistically (4, 13).

Adaptive Walks on the Fitness Landscape. Next, we studied the empirical fitness landscape within a framework recently proposed by Draghi and Plotkin (33). Given the empirical landscape, we simulated adaptive walks and studied the accessibility of the sites toward global local optima. In addition, we evaluated the length of adaptive walks starting from any mutant in the landscape, until an optimum is reached. In the strong selection weak mutation limit (34), we can express the resulting dynamics as an absorbing Markov chain, where local optima correspond to the absorbing states, and in which the transition probabilities correspond to the relative fitness increases attainable by the neighboring mutations (Materials and Methods). This approach allowed us to derive analytical solutions for the mean and variance of the number of steps to reach a fitness optimum (SI Appendix, Supporting Information 1: Extended Materials and Methods), and the probability to reach a particular optimum starting from any given mutant in the landscape (Fig. 3 and SI Appendix, Figs. S3.3 and S3.4).

Using this framework, we find that the global optimum can be reached with nonzero probability from almost 95% of starting points in the landscape, and is reached with high probability from a majority of starting points indicating high accessibility of the global optimum (Fig. 3). The picture changes when restricting the analysis to adaptive walks initiating from the parental type (Fig. 3 and SI Appendix, Figs. S3.3 and S3.4). Here, although 73% of all edges and 78% of all vertices are included in an adaptive walk to the global optimum, it is reached with only 26% probability. A local optimum two substitutions away from the parental type (Fig. 3C) is reached with a much higher probability of 47%. Hence, adaptation on the studied landscape is likely to stall at a suboptimal fitness peak. This observation indicates that the local and global landscape pattern may be quite different, an observation that is confirmed and discussed in more detail below (see Predictive Potential of Landscape Statistics). In line with the existence of multiple local fitness peaks, we find that pairs of alleles at different loci show pervasive sign (30%) and reciprocal sign (8%) epistasis (35), whereas the remaining 62% are attributed to magnitude epistasis (i.e., there is no purely additive interaction between alleles; for a discussion of the contribution of experimental error, see SI Appendix, Fig. S3.5).

Epistasis Measures and the Topography of the Fitness Landscape. Next, we considered the global topography of the fitness landscape. Various measures of epistasis and ruggedness have been proposed, most of them correlated and hence capturing similar features of the landscape (10). However, drawing conclusions has proven difficult because the studied landscapes were created according to different criteria. Furthermore, the majority of published complete landscapes are too small to be divided into subsets (but see refs. 10 and 12), preventing tests for the consistency and hence the predictive potential of landscape statistics. The landscape studied here provides us with this opportunity. Moreover, because multiple alleles at the same site are contained within the landscape, we may study whether changes in the shape of the landscape are site- or amino acid-specific.

We computed various landscape statistics (roughness-to-slope ratio, fraction of epistasis, median in the landscape of neutral mutations) (34) for all 360 diallelic landscapes (Materials and Methods) and compared them with predictions from theoretical landscape models [NK (36–38), RMF (39, 40), HoC (41), egg-box landscapes (14); for brief definitions of these terms, see SI Appendix, Supporting Information 2: Overview of Different Fitness Landscape Models Introduced in the Main Text]. Whenever necessary, we provide an analytical extension of the used statistic to the case of multiallelic landscapes (Materials and Methods). To assess consistency and predictive potential, we computed the whole set of statistics for (i) all landscapes in which 1 amino acid was completely removed from the landscape (a cross-validation approach (42), subsequently referred to as the “drop-one” approach), (ii) all possible 360 diallelic sublandscapes, and (iii) for all 1,570 diallelic 4-step landscapes containing the parental genotype, highlighting as special examples the three focal landscapes discussed.

We find that the general topography of the fitness landscape resembles that of a RMF landscape with intermediate ruggedness, which is characterized by a mixture of a random HoC component and an additive component (Fig. 4A and B). Whereas the whole set of landscape statistics supports this topography and our conclusions, the gamma statistics measuring landscape-averaged correlations in fitness effects, recently proposed in ref. 14, proved to be particularly illustrative. We will therefore focus on these in the main text; we refer to SI Appendix for additional results (e.g., measurement uncertainty and adaptive walks).
Predictive Potential of Landscape Statistics. When computed based on the whole landscape and on a drop-one approach, the landscape appears quite homogeneous, and the gamma statistics show relatively little epistasis (Figs. 4B and 5B). At first sight, this result contradicts our earlier statement of strong negative and positive epistasis but can be understood, given the different definitions of the epistasis measures used: above, we have measured epistasis based on the deviation from the multiplicative combination of the single-step fitness effects of mutations on the parental background. Because these effects were small, epistasis was strong in comparison. Conversely, the gamma measure is independent of a reference genotype and captures the fitness decay with a growing number of substitutions as a dominant and quite additive component of the landscape.

Only mutation 588P has a pronounced effect on the global landscape statistics and seems to act as an epistatic hotspot by making a majority of subsequent mutations (of individually small effect) on its background strongly deleterious (clearly visible in Figs. 4B and 5C). This finding can be explained by looking at the biophysical properties of this mutation. In wild-type Hsp90, amino acid 588N is oriented away from solvent and forms hydrogen bond interactions with neighboring amino acids (24). Proline lacks an amide proton, which inhibits hydrogen bond interactions. As a result, substituting 588N with a proline could disrupt hydrogen bond interactions with residues that may be involved in main chain hydrogen bonding and destabilize the protein. In addition, the pyrrolidine ring of proline is extremely rigid and can constrain the main chain, which may restrict the conformation of the residue preceding it in the protein sequence (43).

The variation between inferred landscape topographies increases dramatically for the 360 di-allelic, 6-locus sublandscapes (Fig. 4C). Whereas all sublandscapes are largely compatible with an RMF landscape, the decay of landscape-wide epistasis with mutational distance (as measured by \( \gamma(d) \)) shows a large variance, suggesting large differences in the degree of additivity. Interestingly, various sublandscapes, typically carrying mutation *588P, show a relaxation of epistatic constraint with increasing mutational distance (i.e., increasing \( \gamma(d) \)) that is not captured by any of the proposed theoretical fitness landscape models, suggestive of systematic compensatory interactions (but see the egg-box model for an explicit example featuring nonmonotonicity in \( \gamma(d) \)). The variation in the shape of the fitness (sub)landscapes is also reflected in the corresponding roughness-to-slope ratio (inset of Fig. 4C and D), further emphasizing heterogeneity of the fitness landscape with local epistatic hotspots.

Finally, the 1,570 di-allelic four-locus landscapes containing the parental genotype, although highly correlated genetically, reflect a variety of possible landscape topographies (Fig. 4D), ranging from almost additive to egg-box shapes, accompanied by an extensive range of roughness-to-slope ratios. The three focal landscapes discussed above are not strongly different compared with the overall variation and yet show diverse patterns of epistasis between substitutions (Fig. 5).

Fig. 4. (A) Expected pattern of landscape-wide epistasis measure \( \gamma_A \) (SI Appendix, Eq. S1.15) with mutational distance for theoretical fitness landscapes with 6 loci from ref. 14. (B) Observed decay of \( \gamma_A \) with mutational distance under the drop-one approach is quite homogenous, except when hot spot mutation 588P is removed; 95% CIs are contained within the lines for the global landscape. (C) Observed decay of \( \gamma_A \) for all di-allelic six-locus sublandscapes. Depending on the underlying mutations, \( \gamma_A \) is vastly different, suggesting qualitative differences in the topography of the underlying fitness landscape and in the extent of additivity in the landscape. Three focal landscapes representative of different types and \( \gamma_A \) for the full landscape have been highlighted. (D) Observed decay of \( \gamma_A \) for all di-allelic four-locus sublandscapes containing the parental type, indicative of locally different landscape topographies. Highlighted are the three focal landscapes. (Insets) Histograms of the roughness-to-slope ratio \( r/s \) for the respective subset, with horizontal lines indicating values for highlighted landscapes \( r/s \) for global landscape (blue) is significantly different from HoC expectation, \( p = 0.01 \). Similar to \( \gamma_A \), there is a huge variation in \( r/s \) for subparts of the fitness landscape, especially when considering the four-locus subsets.

Fig. 5. (A) Epistasis measure \( \gamma(A_i, d) = \gamma(A_i \rightarrow \Gamma d) \) (SI Appendix, Eq. S1.8) on the x axis) between any two substitutions, averaged across the entire landscape. The majority of interactions are small to moderate (blue). Parts of the fitness landscape show highly localized and mutation-specific epistasis, ranging from strong magnitude epistasis (white) to sign and reciprocal sign epistasis (yellow). (B) The average epistatic effect \( \gamma_{A_{,d}} \) (SI Appendix, Eq. S1.10) of a mutation occurring on any background is always small. (C) The average epistatic effect \( \gamma_{A_{,A}} \) (SI Appendix, Eq. S1.13) of a background on any new mutation is usually small, except for mutations to *588P, which shows a strong magnitude effect. (D) Locus-specific gamma for the four mutations leading to the global optimum (Top), the four largest single-effect mutations (Middle), and the single-effect mutations with the lowest fitness (Bottom). The optimum landscape exhibits strong sign epistasis between loci 587 and 588 and between loci 588 and 589. Also, the best landscape exhibits pervasive epistasis with sign epistasis between loci 588 and loci 585 and 586, respectively. We observe almost no epistasis in the worst landscape.
Thus, predicting fitness landscapes is difficult indeed. Extrapolation of the landscape, even across only a single mutation, may fail due to the existence of local epistatic hotspot mutations. Although the integration of biophysical properties into landscape models is an important step forward (e.g., ref. 44), we demonstrate that such models need to be mutation-specific. Considering a site-specific model [e.g., BLOSUM matrix (45)] is not sufficient. Newer models such as DeepAlign may provide the opportunity to allow integration of mutation-specific effects via aligning two protein structures based on spatial proximity of equivalent residues, evolutionary relationship, and hydrogen bonding similarity (46).

Conclusion

Originally introduced as a metaphor to describe adaptive evolution, fitness landscapes promise to become a powerful tool in biology to address complex questions regarding the predictability of evolution and the prevalence of epistasis within and between genomic regions. Due to the high-dimensional nature of fitness landscapes, however, the ability to extrapolate will be paramount to progress in this area, and the optimal quantitative and qualitative approaches to achieve this goal are yet to be determined.

Here, we have taken an important step toward addressing this question via the creation and analysis of a landscape comprising 640 engineered mutants of the Hsp90 protein in yeast. The unprecedented size of the fitness landscape, along with the multiallelic nature, allows us to test whether global features could be extrapolated from subsets of the data. Although the global pattern indicates a rather homogeneous landscape, smaller sublandscapes are a poor predictor of the overall global pattern because of “epistatic hotspots.”

In combination, our results highlight the inherent difficulty imposed by the duality of epistasis for predicting evolution. In the absence of epistasis (i.e., in a purely additive landscape), evolution is globally highly predictable because the population will eventually reach the single-fitness optimum, but the path taken is locally entirely unpredictable. Conversely, in the presence of (sign and reciprocal sign) epistasis evolution is globally unpredictable, because there are multiple optima and the probability to reach any one of them depends strongly on the starting genotype. At the same time, evolution may become locally predictable with the population following obligatory adaptive paths that are a direct result of the creation of fitness valleys owing to epistatic interactions.

The empirical fitness landscape studied here appears to be intermediate between these extremes. Although the global peak is within reach from almost any starting point, there is a local optimum that will be reached with appreciable probability, particular when starting from the parental genotype. From a practical standpoint, these results thus highlight the danger inherent to the common practice of constructing fitness landscapes from ascertained mutational combinations. However, this work also suggests that one promising way forward for increasing predictive power will be the utilization of multiple small landscapes used to gather information about the properties of individual mutations, combined with the integration of site-specific biophysical properties.

Materials and Methods

Here, we briefly outline the materials and methods used. A more detailed treatment of the theoretical work is presented in SI Appendix.

Data Generation. Codon substitution libraries consisting of 640 combinations (single up to sextuplet mutants) of 13 previously isolated individual mutants within the 582-to-590 region of yeast Hsp90 were generated from optimized cassette-ligation strategies, as previously described (24) and cloned into the p417GPD plasmid that constitutively expresses Hsp90.

Constitutively expressed libraries of Hsp90 mutation combinations were introduced into the S. cerevisiae shuffto strain DBY288 (can1-100 ade2-1 his3-11,15 leu2-3,12 trp1-1 ura3-1 hsp82: LEU2 HSP82::LEU2 hsp::::pgals) using the lithium acetate method (47). Following transformation, the library was amplified for 12 h at 30 °C under nonselective conditions using galactose (Gal) medium with 100 µg/mL ampicillin (1.7 g/L yeast nitrogen base without amino acids, 5 g/L ammonium sulfate, 0.1 g/L aspartic acid, 0.02 g/L arginine, 0.03 g/L valine, 0.1 g/L glutamic acid, 0.4 g/L serine, 0.2 g/L threonine, 0.03 g/L isoleucine, 0.05 g/L phenylalanine, 0.03 g/L tyrosine, 0.04 g/L adenine hemisulfate, 0.02 g/L methionine, 0.1 g/L leucine, 0.03 g/L lysine, and 0.01 g/L uracil with 1% raffinose and 1% Gal). After amplification, the library culture was transferred to selective medium similar to Gal medium but raffinose and Gal are replaced with 2% (wt/vol) dextrose. The culture was grown for 8 h at 30 °C to allow shutoff of the wild-type copy of Hsp90 and then shifted to selective medium containing 0.5 M NaCl for 12 generations. Samples were taken at specific time points and stored at −80 °C.

Yeast lysis, DNA isolation, and preparation for Illumina sequencing were performed as previously described (25). Sequencing was performed by Elim Biopharmaceuticals and produced ≥30 million reads of 99% confidence at each read position based on PHRED scoring (48, 49). Analysis of sequencing data were performed as previously described (26).

Estimation of Growth Rates. Individual growth rates were estimated according to the approach described in ref. 20 using a Bayesian MCMC approach proposed in ref. 27. Nucleotide sequences coding for the same amino acid according to the approach described in ref. 20 using a Bayesian MCMC approach proposed in ref. 27. Nucleotide sequences coding for the same amino acid were interpreted as replicates with equal growth rates. The resulting MCMC output consisted of 10,000 posterior estimates for each amino acid mutation corresponding to an average effective sample size of 7,419 (minimum 725). Convergence was assessed using the Hellinger distance approach (50) combined with visual inspection of the resulting trace files.

Adaptive Walks. In the strong selection weak mutation limit (51), adaptation can be modeled as a Markov process only consisting of subsequent fitness-increasing, one-step substitutions that continue until an optimum is reached (so-called adaptive walks). This process is characterized by an absorbing Markov chain with a total of n different states (i.e., mutants), consisting of k absorbing (i.e., optima) and n − k transient states (i.e., nonoptima). Defining w(g) as the fitness of genotype g, and g0 as the genotype carrying a mutant allele at loci j, the selection coefficient is denoted by $s_j(g) = w(g_j) - w(g)$, such that the transition probabilities p$_{g_j|g_j'}$ for going from any mutant g to any mutant g' are given by the fixation probability (52, 53) normalized by the sum over all adaptive, one-mutant neighbors of the current genotype g. If g is a (local) optimum, p$_{g|g} = 1$. Putting the transition matrix P (54) in its canonical form and computing the fundamental matrix then allows to determine the expectation and the variance in the number of steps before reaching any optimum and to calculate the probability to reach optimum g when starting from genotype g' (55). Robustness of the results and the influence of specific mutations were assessed by deleting the corresponding columns and rows in P (i.e., by essentially treating the corresponding mutation as unobserved) and recalculating and comparing all statistics to those obtained from the full dataset.

Correlation of Fitness Effects of Mutations. Strength and type of epistasis were assessed by calculating the correlation of fitness effects of mutations γ (14), which quantifies how the selective effect of a focal mutation is altered when put onto a different genetic background, averaged over all genotypes of the fitness landscape. Extending recent theory (14), we calculated the matrix of epistatic effects between different pairs of alleles (A, B) and (A, B) termed $A_j A_j' B_j - A_j A_j' B_j'$. The vector of epistatic effects between a specific pair of alleles (A, B) on all other pairs of alleles $A_i A_i' B_j - A_i A_i' B_j'$ (SI Appendix, Eq. S1,5) the vector of epistatic effects between pairs of alleles on a specific allele pair (A, B) termed $A_j A_j' B_j - A_j A_j' B_j'$ (SI Appendix, Eq. S1,5) and the decay of correlation of fitness effects (56) (SI Appendix, Eq. S1,5) with Hamming distance d averaged over all genotypes g of the fitness landscape.

Fraction of Epistasis. Following refs. 35 and 56, we quantified whether specific pairs of alleles between two loci interact epistatically and, if so, whether these display magnitude epistasis [i.e., fitness effects are nonadditive (57), but fitness increases with the number of mutations], sign epistasis (i.e., one of the two mutations considered has an opposite effect in both backgrounds) or reciprocal sign epistasis (i.e., if both mutations show sign epistasis). In particular, we calculated the type of epistatic interaction between mutations $g_{ij}$ and $g_{ij'}$ (with $i \neq j$) with respect to a given reference genotype.
type} over the entire fitness landscape. There was no epistatic interaction if $\left| s(g_i) - s(g_j) \right| = 10^{-4}$, magnitude epistasis if $s(g_i) - s(g_j) > 0$ and $s(g_i) - s(g_j) \geq 0$, reciprocal sign epistasis if $s(g_i) - s(g_j) < 0$ and $s(g_i) - s(g_j) < 0$, and sign epistasis in all other cases (58).

**Roughness-to-Slope Ratio.** Following ref. 11, we calculated the roughness-to-slope ratio $\rho$ by fitting the fitness landscape to a multidimensional linear model using the least-squares method. The slope of the linear model corresponds to the average additive fitness effect (10, 23), whereas the roughness is given by the variance of the residuals. Generally, the better the linear model fits, the smaller the variance in residuals such that the roughness-to-slope ratio approaches 0 in a perfectly additive model. Conversely, a very rugged fitness landscape would have a large residual variance and, thus, a very large roughness-to-slope ratio (as in the HoC model). In addition, we calculated a test statistic $Z_2$, by randomly shuffling fitness values in the sequence space to evaluate the statistical significance of the obtained roughness-to-slope ratio from the dataset $\hat{P}_{\text{MCMC}}$ generated by $Z_2 = \frac{\hat{RSL} - R_{\text{obs}}}{\sqrt{V_{\text{MCMC}}/n}}$.

**ACKNOWLEDGMENTS.** We thank Dan Bolon, Brian Charlesworth, Pamela Cote, Inês Fragata, and Hermina Ghenu for helpful comments and discussion. This project was funded by grants from the Swiss National Science Foundation and a European Research Starting Grant (to J.D.J.). Computations were performed at the Vital-IT Center (www.vital-it.ch) for high-performance computing of the Swiss Institute of Bioinformatics.


Supporting Information 1: Extended Materials and Methods

Adaptive walks. Under the strong selection weak mutation (SSWM) limit [34], adaptation follows an absorbing Markovian process characterized by a series of fitness-increasing substitutions along one-mutant neighbours until reaching a fitness optimum (forming a so-called adaptive walk), with a total of \( n \) different states (i.e., mutants), consisting of \( k \) absorbing (i.e., optima) and \( n - k \) transient states (i.e., non-optima).

Defining \( w(g) \) as the fitness of genotype \( g \), and \( g_{[j]} \) as the genotype \( g \) carrying a mutant allele at locus \( j \), the selection coefficient is denoted by

\[
s_j(g) = w(g_{[j]}) - w(g),
\]

such that the transition probabilities for going from any given genotype \( g \) to any mutant genotype \( \tilde{g} \) are given by the transition matrix \( P \)

\[
P = \begin{cases}
0 & \text{if } \bar{D}_{\text{HD}}(g, \tilde{g}) > 1 \text{ or } s_j(g) \leq 0 \text{ with } \tilde{g} = g_{[j]} \\
\frac{1 - \exp[-2s_j(g)]}{\sum_{j \in M(g)} (1 - \exp[2s_i(g)])} & \text{if } \tilde{g} \in M(g) \text{ with } \tilde{g} = g_{[j]} \\
1 & \text{if } M(g) = \emptyset \text{ and } \bar{g} = g,
\end{cases}
\]

where \( \bar{D}_{\text{HD}}(g, \tilde{g}) \) marks the Hamming distance between two genotypes \( g \) and \( \tilde{g} \) (i.e., the sum of pairwise allele differences across all loci), and \( M(g) := \{ j : s_j(g) > 0, \bar{D}_{\text{HD}}(g, g_{[j]}) = 1 \} \) denotes the set of all adaptive, one-mutant neighbours of the current genotype \( g \). Note that when considering more than two alleles per locus, \( g_{[j]} \) (and accordingly \( s_j(g) \)) are not uniquely determined and – apart from the focal genotype \( g \) – depend on the (mutant) allele found at locus \( j \). However, since both entities become explicit once \( \tilde{g} \) is specified (and for keeping notation simple), we prefer to define a single selection coefficient (i.e., eq. S1_1) only, that is used consistently in all our approaches. Furthermore note, that transition probabilities are given by the classical fixation probability of a given genotype [52, 53] renormalized over the fixation probability of all beneficial genotypes in a one-mutant neighbourhood of the current genotype \( g \). [54] recently used an even more general form of \( P \) which also took into account differences in mutation rates and also allowed for the fixation of slightly deleterious and neutral mutations. However, under the assumption that there is no mutational bias (i.e., all single-step mutants are equally likely to occur) and that selection is strong as compared to mutation (i.e., the SSWM regime), both definitions are identical.

The canonical form of \( P \) can then be obtained by permutation, such that

\[
P = \begin{pmatrix}
Q & R \\
0 & I_k
\end{pmatrix},
\]

where \( Q \) is a \((n - k) \times (n - k)\) matrix which contains the transition probabilities between transient states; \( R \) is a \((n - k) \times k\) matrix which gives the transition probabilities from any transient to any absorbing state; \( 0 \) is the \( k \times (n - k) \) zero matrix; and \( I_k \) is the \( k \times k \) identity matrix [55].

Using the above representation, all basic properties of the absorbing Markov chain can be calculated from the fundamental matrix

\[
N = \sum_{i=0}^{\infty} Q^i = (I_{(n-k)} - Q)^{-1}.
\]

In particular, the expected number of steps before absorption (i.e., the expected number of steps on the fitness landscape before reaching any optimum) is given by
E[t] = N1,  \[S1\_5\]

where 1 is a column vector of length \((n - k)\) with all entries being 1, and the \(i^{th}\) entry of \(E[t]\) gives the expected number of steps when starting from state (mutant) \(i\).

Similarly, the variance in the number of steps before being absorbed can be computed as

\[
\text{Var}[t] = (2N - I_i) E[t] - E[t] \odot E[t], \quad \[S1\_6\]
\]

where \(\odot\) denotes the Hadamard product.

Finally, the probability of being absorbed in state \(j\) when starting from transient state \(i\) (i.e., reaching optimum \(j\) when the initial genotype is \(i\)), is given by the \((n - k) \times k\)-matrix

\[
B = NR. \quad \[S1\_7\]
\]

Thus, these methods give an easy and computationally fast way of quantifying and predicting adaptive walks on fitness landscapes. Furthermore, the robustness of these results and the influence of particular mutants can be assessed by deleting the \(i^{th}\) column and row of \(P\) – i.e., by essentially treating mutant \(i\) as unobserved –, recalculating the above statistics and comparing these to the statistics obtained from the full data set. Similarly, entire mutations (i.e., amino acids) can be left out to assess their relative effect on the fitness landscape and the generality of our analysis (and the statistics used).

**Measuring epistasis.** We applied different metrics for quantifying epistasis over the entire fitness landscape as well as for particular mutations and assessed their consistency in capturing the strength of gene\(\times\)gene interactions. In particular, we follow the definition of epistasis by [57] (originally termed epistacy), as the deviation from additivity when combining two genetic effects which is measured by the difference in log-fitness between the effects of the double mutant and the single mutant relative to the wild-type fitness.

**Correlation of fitness effects of mutations:** \(\gamma\). The first measure has recently been introduced by [14] and is defined as the single-step correlation of fitness effects for mutations between neighbouring genotypes. It quantifies how the selective effect of a focal mutation is altered when it occurs in a different genetic background averaged over all genotypes of the fitness landscape. Geometrically, \(\gamma\) measures the correlation between slopes (with respect to genotype-fitness hypercube) of the same mutation put into different genetic backgrounds. Thus, if the fitness effect of a mutation is independent of its genetic background (i.e., if there is no epistasis), the correlation in slopes will be perfect \((\gamma = 1)\), whereas it will be zero if the fitness slopes of each genotype are independent of the fitnesses of other genotypes (as in the House-of-Cards model; 41). Depending on the scale \(\gamma\) can either be used to quantify the strength of gene\(\times\)gene interactions between specific mutations or as an overall measure for the entire fitness landscape. However, in its original form \(\gamma\) was defined for di-allelic data only and thus needs to be extended by considering pairs of specific alleles at different loci [14].

Let \(A_i = \{1, 2, \ldots, m_i\}\) denote the set of different alleles present at locus \(i\) for all polymorphic loci \(i \in \{1, 2, \ldots, n\}\) such that \(|A_i| \geq 2\) for all loci \(i\). Further, let \(G\) be the set of all genotypes that can be formed by combining all alleles such that the total number of genotype is \(|G| = \prod_{i=1}^{n} m_i\).

Then the matrix of epistatic effects between loci \(i\) and \(j\) carrying alleles \(A_i, B_i \in A_i\) and \(A_j, B_j \in A_j\) is given by

\[
\gamma(A_i, B_i) \rightarrow (A_j, B_j) = \text{Cor}[s_{A_i, B_i} (g), s_{A_j, B_j} (g|A_i, B_i)] = \frac{\sum_g s_j(g) s_j(g|a_{ij})}{\sum_g (s_j(g))^2}, \quad \[S1\_8\]
\]

where \(g := \{x \in G | x_i = A_i\text{ or } x_i = B_i\text{ and } x_j = A_j\text{ or } x_j = B_j\} \subseteq G\) such that the sum is only calculated over the subset of genotypes carrying one of the two focal alleles at each focal locus. Thus, \(\gamma(A_i, B_i) \rightarrow (A_j, B_j)\) is a quadratic matrix of dimension \((\sum_{i=1}^{n} |A_i|(|A_i| - 1))/2\). Note that in the case where \(|A_i| = 2\) for all loci \(i\), we obtain equation 9 in [14].

Likewise, the epistatic effect of a mutation in locus \(i\) with alleles \((A_i, B_i)\) on other loci (and pairs of alleles) can be calculated as

\[
\gamma(A_i, B_i) \rightarrow = \text{Cor}[s(g), s(g|A_i, B_i)] = \frac{\sum_{j \neq i} \sum_{a_j} \sum_g s_j(g) s_j(g|a_{ij})}{\sum_{j \neq i} \sum_{a_j} \sum_g (s_j(g))^2}, \quad \[S1\_9\]
\]

where the summation index \(a_j = \{(A_j, B_j) \mid A_j, B_j \in A_j\text{ and } A_i \neq B_i\}\) is over the set of subsets of size two that can be constructed from all alleles found at locus \(j\). Note that the third summation index \(g\) changes depending on \(a_j\).

An additional summation allows calculation of the epistatic effect of a mutation in locus \(i\) carrying allele \(A_i\) on other loci (and pairs of alleles) can be calculated as
where the last summation is over all different alleles present at locus \( i \).

Then, summing over \( I_i = \{(A_i, B_i) \mid A_i, B_i \in A_i \land A_i \neq B_i\} \), i.e., the elements of the set of subsets of size two that can be constructed from all alleles found at locus \( i \) that contain allele \( A_i \).

Finally, \( \gamma_d \), that is the decay of correlation of fitness effects with Hamming distance \( d \) (i.e., the cumulative epistatic effect of \( d \) mutations averaged over the entire fitness landscape) is calculated as

\[
\gamma_d = \text{Cor}[s(g), s(g_{i_1 \ldots i_d})] = \sum_{g \in \mathcal{G}} s(g) \sum_{g_{i_1 \ldots i_d}} s(g_{i_1 \ldots i_d}) s(g_{i_2 \ldots i_d}) \frac{1}{\sum_{g \in \mathcal{G}} s(g)^2},
\]

where \( \mathcal{G} \) is the set of all genotypes.

**Fraction of epistasis.** The second statistic quantifies whether specific pairs of alleles between two loci interact epistatically, and if so whether they display magnitude epistasis (i.e., fitness effects are non-additive, but fitness increases with the number of mutations), sign epistasis (i.e., one of the two mutations considered has an opposite effect in both backgrounds) or reciprocal sign epistasis (i.e., if both mutations show sign epistasis; [35, 56]; for an implementation see [58]).

Using equation (S1_1) the type of epistatic interaction between a mutation at locus \( i \) and \( j \) (with \( i \neq j \)) when introduced on some reference genotype \( g \), \( \epsilon(g, i, j) \), can formally be given as

\[
\epsilon(g, i, j) = \begin{cases} 
  \text{none} & \text{if } |s_i(g_{[j]}) - s_i(g)| = 0 \\
  \text{magnitude} & \text{if } s_j(g)s_i(g_{[j]}) \geq 0 \text{ and } s_i(g)s_i(g_{[j]}) \geq 0 \\
  \text{reciprocal sign} & \text{if } s_j(g)s_i(g_{[j]}) < 0 \text{ and } s_i(g)s_i(g_{[j]}) < 0 \\
  \text{sign} & \text{else}
\end{cases}.
\]

Note that for numerical purposes we allowed for a very small deviation \( \varepsilon = 10^{-6} \) in the case of no epistatic interaction. For ease of notation, we treat the elements in \( A_i \) as fixed and ordered, and define \( I(g_{[j]}) \) as the index of the allele present at locus \( i \) in genotype \( g \) in \( A_i \) such that the fraction of epistasis over all \( \mathcal{G} \) can be calculated as

\[
\mathcal{E}(x) = c^{-1} \sum_{g} \sum_{k=I(g_{[j]})+1}^{m_i} \sum_{l=k+1}^{m_j} 1(g_{[i]}) \epsilon(g, k, l),
\]

where \( \mathcal{E}(x) \) is the fraction of epistasis.
where $x \in \{\text{none}, \text{magnitude}, \text{sign}, \text{reciprocal sign}\}$, $\mathbb{1}_x(y)$ is the indicator function that is 1 if $x = y$ and 0 otherwise, and 

$$
c = \left( \sum_{y} \sum_{k=1}^{m} \sum_{l=k+1}^{m} 1 \right)
$$

is a normalization constant.
Supporting Information 2: Overview of different fitness landscape models introduced in the main text

In the following, we briefly introduce the various fitness landscape models introduced and discussed in the main text and provide references to selected relevant literature. For an excellent general overview, we refer the reader to a review by [10].

**The additive model.** In the additive model the log-fitness of each genotype is given by the sum over the individual per locus log-fitness effects. Thus, the fitness effect of a specific allele – drawn from a Normal distribution with mean $\mu_a$ and variance $\sigma_a^2$ [14] [see also 39] – is independent of its genetic background (i.e., it is constant across all genetic backgrounds), such that all mutations are non-interacting (i.e., there is no epistasis) and the resulting single-peaked fitness landscape is (maximally) smooth. In particular, the roughness-to-slope ratio is 0 and $E[\gamma_d] = 1$ for the entire range of mutational distances $d$. Note that when fitnesses are not measured on log-scale, the fitness of a genotype is simply the product over the individual fitness effects, which is why this model is sometimes also referred to as the multiplicative model.

**The House-of-cards model.** On the other extreme, in the House-of-cards (HoC) model [41] the fitness of each genotype is an i.i.d. normally distributed random variable with zero mean and variance $\sigma_{\text{HoC}}^2$ resulting in an uncorrelated, maximally rugged fitness landscape that is characterized by multiple fitness maxima and minima [36, 37]. In particular, the fitness effect of an allele entirely depends on its genetic background such that there is complete interaction between all loci (i.e., full epistasis) which is also reflected in the roughness-to-slope ratio and $E[\gamma_d]$ that become infinity and zero (for $d > 0$), respectively.

**The Rough Mount Fuji model.** Introduced by [40], the Rough-Mount-Fuji (RMF) model – named after the eponymous mountain in Japan –, which was initially formulated in the context of protein evolution [see also 39, for a simplified version], interpolates between the former two extremes. The fitness of a genotype is composed by an additive component (parametrized by $\mu_a$ and $\sigma_a^2$; see above) and a HoC component (parametrized by $\sigma_{\text{HoC}}^2$) such that the extent of epistatic interactions ranges between none (additive model) to complete (HoC model) depending on the relative sizes of these three parameters. In particular, when $\sigma_{\text{HoC}}^2 \ll \mu_a^2 + \sigma_a^2$, the RMF model becomes an additive model whereas for $\sigma_{\text{HoC}}^2 \gg \mu_a^2, \sigma_a^2$ it essentially behaves like a HoC model [14]. Accordingly, $0 \leq E[\gamma_d] = \text{const} \leq 1$ for $d > 0$ (see Figure 4a) and the roughness-to-slope ratio ranges from zero to infinity.

**The Kauffman NK model.** Another frequently used fitness landscape model that, as the RMF model, also interpolates between the additive and the HoC model [36, 37] is the Kauffman NK model, where $N$ di-allelic loci interact with $K \in \{0, 1, \ldots, L - 1\}$ randomly assigned other loci. In particular, for $K = 0$ the NK model collapses to an additive model whereas for $K = L - 1$ it approaches the HoC model. Although there are different ways how groups of interacting loci can be chosen [38], properties such as the mean number and height of local optima tend only to be weakly dependent on the exact choice being made. For the NK model, $E[\gamma_d]$ is a non-negative monotonically decreasing function in $d$, and the roughness-to-slope ratio can again range from zero to infinity.

**The eggbox model.** Introduced by [14], the eggbox model is a maximally epistatic, anticorrelated fitness landscape model (i.e., all loci interact with each other up to interactions of order $L$), in which the fitness effect of an allele switches from the highest to lowest value (or vice versa) between genetic backgrounds one step apart. Accordingly, depending on whether two genotypes are separated by an odd or even Hamming distance, their absolute fitness difference is either twice the mean allelic fitness effect or zero. Thus, this model generates an extreme case of reciprocal sign epistasis in which each mutation is either deleterious or compensatory, multiple local optima exist, and $\gamma_d$ accordingly oscillates between -1 and 1.
Supporting Information 3: Supporting Figures

**Fig. S3_1.** (A) Scatterplot demonstrating the correlation of estimated growth rates ($R^2 = 0.944$). Orange dots indicate median growth rates, rectangles indicate the limits of the 95% credibility intervals (in blue if they do not overlap between replicates). (B) In contrast, estimated initial population sizes are much more variable between replicates, demonstrating that initial representation in the mutant library had no effect on the estimated growth rate.
Fig. S3_2. Expected versus observed fitness for all 5 beneficial local optima of the fitness landscape, obtained from 1000 posterior samples.
Fig. S1_3. Graphical illustration of the sub-landscape of all beneficial mutants with respect to the parental type. (A) Each vertex (circle) corresponds to a mutant that is beneficial with respect to the parental type (triangle), colored according to their fitness. Arrows connect mutants that differ by a single substitution, with their direction indicating an increase in fitness. The five local optima in this sub-landscape are indicated as squares. Inset: distribution of average lengths of adaptive walks starting from any type in the full landscape (i.e., absorbing times of the Markov chain). The red line indicates mean path lengths for adaptive walks from the parental genotype. (B)-(F) Paths leading towards the different optima – highlighted by their sequence name – when beginning an adaptive walk at the parental genotype. Inset (B)-(F): Distribution of absorbing probabilities, that is, the probability to reach a specific optimum starting from a given mutant computed for all mutants in the data set. The red line corresponds to the respective probability when starting from the parental sequence. The global optimum (panel E) is in general reached with a very high probability, but there are starting points from which it is poorly accessible. $C(\epsilon, v)$ quantifies the landscape coverage: although 73% of all edges and 78% of all vertices can be included in walks from the parental genotype to the global optimum, it is reached in only 26% of all possible walks. Conversely, 47% of walks will reach the local optimum in panel C, which is lower in fitness, but closer to the parental type in terms of its Hamming distance. This demonstrates that evolution on this landscape is not well predicted by fitness only, and that a population may become stuck in a sub-optimal fitness peak.

Fig. S1_3. Graphical illustration of the sub-landscape of all beneficial mutants with respect to the parental type. (A) Each vertex (circle) corresponds to a mutant that is beneficial with respect to the parental type (triangle), colored according to their fitness. Arrows connect mutants that differ by a single substitution, with their direction indicating an increase in fitness. The five local optima in this sub-landscape are indicated as squares. Inset: distribution of average lengths of adaptive walks starting from any type in the full landscape (i.e., absorbing times of the Markov chain). The red line indicates mean path lengths for adaptive walks from the parental genotype. (B)-(F) Paths leading towards the different optima – highlighted by their sequence name – when beginning an adaptive walk at the parental genotype. Inset (B)-(F): Distribution of absorbing probabilities, that is, the probability to reach a specific optimum starting from a given mutant computed for all mutants in the data set. The red line corresponds to the respective probability when starting from the parental sequence. The global optimum (panel E) is in general reached with a very high probability, but there are starting points from which it is poorly accessible. $C(\epsilon, v)$ quantifies the landscape coverage: although 73% of all edges and 78% of all vertices can be included in walks from the parental genotype to the global optimum, it is reached in only 26% of all possible walks. Conversely, 47% of walks will reach the local optimum in panel C, which is lower in fitness, but closer to the parental type in terms of its Hamming distance. This demonstrates that evolution on this landscape is not well predicted by fitness only, and that a population may become stuck in a sub-optimal fitness peak.
Fig. S3.4. (A) Path probability of reaching a specific local optimum when beginning an adaptive walk at the parental genotype. Results are presented on log scale. Boxes represent the interquartile range (i.e., the 50% C.I.), whiskers extend to the highest/lowest data point within the box ± 1.5 times the interquartile range, and black circles represent close outliers. (B) Path length (i.e., the number of substitutions) to reach a specific local optimum. Inset: Color coding for the different optima with the total number of different paths leading to the respective optimum in brackets. There is a total of 3776 different paths that can be taken from the parental genotype, 66% of which lead to the global optimum (yellow). Most of these, however, require multiple substitutions, such that the probability to reach the global optimum is generally low. In contrast, the local optimum close to the parental genotype (orange) can be reached with only a few substitutions, and thus, with relatively high probability.

Fig. S3.5. (A) The fraction of epistasis (eq. S1_17) calculated for different values of $\varepsilon$, quantifying the error threshold beyond which interactions between different alleles are considered to be epistatic. Expectedly, as $\varepsilon$ increases the number of non-epistatic interactions between alleles increases, which in the majority of cases were previously classified as magnitude epistatic interactions. Likewise, the fraction of sign and reciprocal sign epistasis decreases. The red dashed line shows the mean absolute difference between median growth rates obtained from the two independent replicates calculated over all mutants with median growth rate larger than the mean median growth rate estimated for the substitution F583* (i.e., when substituting a stop codon) between the two replicates. Using this threshold for $\varepsilon$, we find that approximately 10% of all interactions are additive, approximately 50% show magnitude epistasis, whereas the sign and reciprocal sign epistasis remains roughly constant at 30% and 8%, respectively. (B) The mean and the 95% interquantile range (IQR) of the fraction of epistasis calculated over 10,000 posterior samples obtained from the MCMC simulations using the standard threshold $\varepsilon = 10^{-6}$. There is only little to no variation in the fraction of epistasis and the mean value is essentially identical to the proportions reported in the main text, indicating robustness of the inferred fraction of epistasis, and confirming that variation in posterior samples is typically small (i.e., growth rates are estimated with high precision).