



Article

Selection for protein stability enriches for epistatic interactions

Anna Posfai¹, Juannan Zhou¹, Joshua B. Plotkin², Justin B. Kinney¹ and David M. McCandlish^{1,*}

¹ Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724

² Department of Biology, University of Pennsylvania, Philadelphia, PA, 19104

* Correspondence: mccandlish@cshl.edu; Tel.: 1+516-367-5286

Academic Editor: name

Version June 3, 2018 submitted to Genes

Abstract: A now classical argument for the marginal thermodynamic stability of proteins explains the distribution of observed protein stabilities as a consequence of an entropic pull in protein sequence space. In particular, most sequences that are sufficiently stable to fold will have stabilities near the folding threshold. Here we extend this argument to consider its predictions for epistatic interactions for the effects of mutations on the free energy of folding. Although there is abundant evidence to indicate that the effects of mutations on the free energy of folding are nearly additive and conserved over evolutionary time, we show that these observations are compatible with the hypothesis that a non-additive contribution to the folding free energy is essential for observed proteins to maintain their native structure. In particular through both simulations and analytical results, we show that even very small departures from additivity are sufficient to drive this effect.

Keywords: thermodynamic stability, epistasis, molecular evolution, purifying selection

1. Introduction

The relationship between protein sequence, stability, and function has been a subject of intense investigation for decades. A combination of biophysical and evolutionary models and, more recently, high-throughput mutagenesis experiments have dramatically advanced our understanding of this complex relationship [1–5]. A consensus view has emerged on some aspects of protein functions and evolution—e.g., what accounts for the distribution of thermodynamic stabilities observed in nature. And yet other questions—e.g., whether genetic interactions play a dominant or minor role in protein sequence evolution—remain actively debated, with apparently contradictory empirical and theoretical evidence [1–5].

A nuanced appreciation of the high-dimensional nature of protein sequence space has been essential for resolving questions about protein structure, function, and evolution. The observation that naturally occurring proteins are only marginally, as opposed to maximally, stable was first interpreted as an adaptive feature to permit increased protein flexibility and functionality [6]. But, with some exceptions [7], this view has been largely replaced with a more parsimonious explanation based on the high dimensionality of sequence space: marginal stabilities are observed because, simply, far more sequences are marginally stable than maximally stable [8]. Essential to the development of this explanation was the concept of sequence entropy [9,10] – the idea that the sheer number of protein sequences that map to a given phenotype exerts a strong entropic pull on the distribution of observed phenotypes in viable proteins [11–13]. The field today has mostly settled on a synthetic understanding of how simple biophysical models of energy and folding, along with the structure of sequence space, conspire to explain the distribution of protein stabilities observed in nature [1–5].

33 By contrast to observed stabilities, the role of epistasis in protein evolution and function remains
34 a topic of active debate with unresolved ambiguities. The same bio-physical models that can
35 parsimoniously explain observed distributions of stabilities have been reported to show only a weak
36 context-dependence of mutational effects on stability [14], or, alternatively, reported to show very
37 strong context dependence of mutational effects [13,15,16]. Likewise, experimental studies on the
38 fitness effects of mutations in divergent sequence backgrounds have reportedly very weak epistasis
39 [14], whereas comparative analysis of divergent proteins has implicated an overriding role for epistasis
40 in shaping sequence evolution[17]. How are we to resolve this significant discrepancy about the role
41 of epistasis for protein stability and sequence evolution?

42 In this paper we address this discrepancy by analyzing simple models for the relationship between
43 amino acid sequence and the ΔG of folding. Under selection to maintain a minimum degree of stability,
44 these models predict distributions of folding energies that are roughly consistent with those observed
45 in nature. Moreover, the models predict very weak interactions between pairs of mutations. These
46 predictions are consistent with biophysical measurements of nearly additive mutational effects on
47 stability [18], and with reports of consistent effects over both short [14] and long [19] evolutionary
48 timescales. And yet, at the same time, we show that a non-additive contribution to the folding free
49 energy is essential for allowing proteins to fold stably in our model, for reasons attributable to sequence
50 entropy. These results may help to resolve striking discrepancies in the literature on the importance of
51 epistasis for protein stability and evolution [1–5].

52 2. Methods

53 2.1. Simulations

54 We consider two simple models for the relationship between amino acid sequences and ΔG of
55 folding. All amino acid sequences considered are of length $l = 400$.

56 First we consider a model where epistasis arises due to pair-wise interactions between sites. In
57 this probabilistic model, the additive effect on stability for each possible amino acid in each position
58 in the primary sequence is drawn from a Gaussian distribution with mean μ_{add} and variance σ_{add}^2 .
59 In addition, we allow pair-wise interactions between amino acid sites, and the magnitudes of these
60 interactions are also drawn from a Gaussian distribution with zero mean and variance σ_{epi}^2 . Moreover,
61 these pairwise interactions are specified in such a way that the pairwise interaction terms have no
62 impact on the average effect of any given amino acid substitution, so that the magnitudes of the
63 additive and epistatic effects can be modified independently. That is, the model is equivalent to a
64 “random field model” from the fitness landscape literature[20,21] where the only non-zero terms are
65 the constant, linear, and pairwise interaction terms. See Appendix A for details on the mathematical
66 features and practical implementation of this model.

67 Second, we consider a model where epistasis is modeled as a random deviation from additivity
68 drawn independently for each *genotype*, meaning each sequence of amino acids. This model is similar
69 to the “rough Mount Fuji” model of fitness landscapes [22,23]. In this case we again draw the
70 additive effect of each amino acid in each position from a Gaussian distribution with variance σ_{add}^2 ,
71 but in addition the folding energy of each genotype is perturbed by an independent draw from a
72 zero-mean gaussian with variance σ_{HOC}^2 (where *HOC* denotes “house of cards”, since this component
73 is completely uncorrelated between mutationally adjacent genotypes, similar to the house of cards
74 model of fitness landscapes [24]). Because protein sequence space is too large to store in computer
75 memory, we implement a hashing scheme so that in the simulations these epistatic effects remain
76 consistent for previously observed genotypes, but are drawn anew for genotypes that have not yet
77 been encountered.

78 The simulations of protein sequence evolution under selection are based on a threshold model
79 for thermodynamic stability: proteins with a negative ΔG of folding are deemed viable and all other
80 sequences are deemed inviable. At each step in the simulation, a random position in the protein

81 sequence is chosen and changed to a random alternative amino acid. This new sequence is accepted if
82 it is viable and rejected if it is inviable. These simulations are initialized at the sequence predicted to
83 be most stable based on its additive effects and allowed to equilibrate for 5000 proposed mutations, a
84 time sufficient for the distribution of folding stabilities to become approximately stationary for the
85 conditions consider here. After this relaxation period, the simulations continue for an additional 5000
86 proposed mutations to produce the results shown here. All simulations and all calculations presented
87 were implemented in Mathematica and the corresponding Mathematica notebook is included as
88 supplemental information.

89 The models analyzed here are simpler, and less realistic, than other commonly used models for
90 protein evolution based on force fields [25], contact energies [26], or lattice proteins [27]. However,
91 we employ these models because their simple structure yields to a variety of exact and approximate
92 analytical results, and thus provides a clearer illumination of the theoretical issues involved than the
93 more realistic but less tractable alternatives.

94 3. Results

95 3.1. Epistasis is essential for proper folding of evolved sequences

96 We simulated the evolution of a protein of length 400 under a model where each amino acid at
97 each position makes an additive contribution to the free energy of folding, and where in addition we
98 allow pairwise stability interactions between sites. We imposed truncation selection for spontaneous
99 folding so that only sequences with a negative ΔG of folding are considered viable. The parameters
100 of the simulations were chosen to be roughly consistent with the observed distribution of folding
101 stabilities and mutational effects on stability reported in the literature [e.g. 28–31]

102 Figure 1a shows the distribution of folding energies observed in these simulations after the
103 process was allowed to reach stationarity. The mean of this distribution is only slightly negative,
104 indicating that the evolved proteins are marginally stable, as predicted by theory and observed in
105 nature [8,11,29,31,32]. Examining the effects of single amino acid substitutions for sequences drawn
106 from this distribution (Figure 1b), we observe that the mean is approximately positive 1.133 kcal/mol
107 with standard deviation 1.44 kcal/mol, consistent with empirical observations [28–30], and that the
108 distribution of energetic effects that are fixed over the course of the simulations (Figure 1c) is shifted to
109 have approximately zero mean (0.0009 kcal/mol) and a smaller standard deviation (1.0493 kcal/mol)
110 as observed previously [16,33,34].

111 Interactions between mutations also have a similar magnitude to those observed in previous
112 studies, with double mutants stabilities nearly exactly predicted by the observed additive effects of
113 their constituent single mutations (Figure 1d, $R^2 = 0.99976$). Furthermore, the additive effects of
114 mutations that fix along our simulated evolutionary trajectories remain relatively consistent over time
115 (Figure 1e), with a root mean square change of only 0.5052 kcal/mol at 50% sequence divergence,
116 consistent with the empirical measurements of Risso *et al.* [19], who observed an RMS change of .67
117 kcal/mol among mutations that fix at a similar level of divergence. Because we are working with
118 simulated data, we can also assess the extent to which the observed mutational effects reflect the true
119 additive effects of mutations. Figure 1f shoes that the observed effects of the possible single amino
120 acid substitutions in an evolved background are highly correlated with the true underlying additive
121 effects of these same mutations ($R^2 = 0.887$).

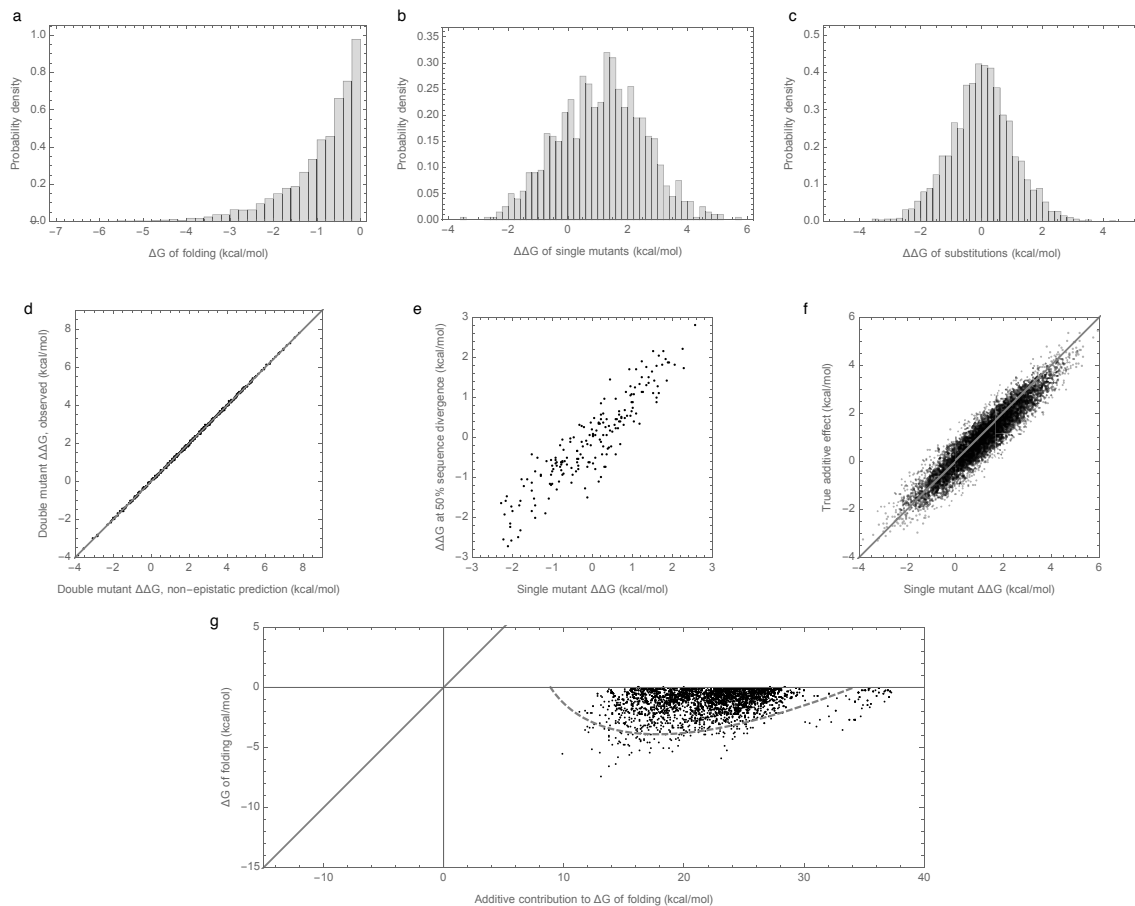


Figure 1. Free energy of folding, stability effects of mutations, and contribution of additive effects to folding stability for populations evolving at stationarity under truncation selection for protein stability with pair-wise energetic interactions: (a) Free energy of folding for evolved sequences. (b) Distribution of stability effects of mutations for evolved sequences. (c) Distribution of stability effects of mutations fixed along simulated trajectories. (d) Stability effects of double mutants for evolved sequences versus predicted stability based on the sum of single mutant effects. 500 random double mutants in a single evolved background shown, $R^2 = 0.99976$. (e) Effects of single mutations that fixed along trajectory in two evolved backgrounds that differ by 50% sequence divergence, $R^2 = 0.8266$. (f) Observed stability effects of mutations are highly correlated with the true average additive effects of the corresponding mutations, $R^2 = 0.887$. (g) Free energy of folding versus additive contribution to folding for evolved sequences. The additive contribution to folding is not a good indicator of the free energy of folding ($R^2 = 0.0672$) and observed sequences cannot fold spontaneously based on the additive contribution alone. The solid curve is derived from our analytical approximations and is predicted to contain 95% of the evolved sequences. Simulations conducted under the pairwise epistasis model with $\mu_{\text{add}} = 1$, $\sigma_{\text{add}}^2 = 1$, $\sigma_{\text{epi}}^2 = 0.0003$.

122 To summarize, our simulations are qualitatively similar to both previous empirical and theoretical
 123 investigations of long-term evolution under selection for protein folding stability, and, on the face of it,
 124 they suggest that epistasis for protein folding stability plays only a minor role. However, when we
 125 actually compute the additive contribution to protein stability observed in our simulations, a very
 126 different picture emerges (Figure 1d). Shockingly, we find that the observed additive contribution to
 127 folding stability is not nearly sufficient to allow spontaneous folding (mean ΔG of folding from the
 128 additive component is 22.45 kcal/mol) so that epistatic interactions are required for folding for all the
 129 sequences observed at stationarity. Furthermore, the additive contribution to folding stability is almost
 130 completely uncorrelated with the actual folding stability ($R^2=0.06$). Thus, epistasis plays an essential

131 role in the simulation results, despite its near absence in the simulated double-mutant data and the
 132 observed conservation of energetic effects at 50% sequence divergence.

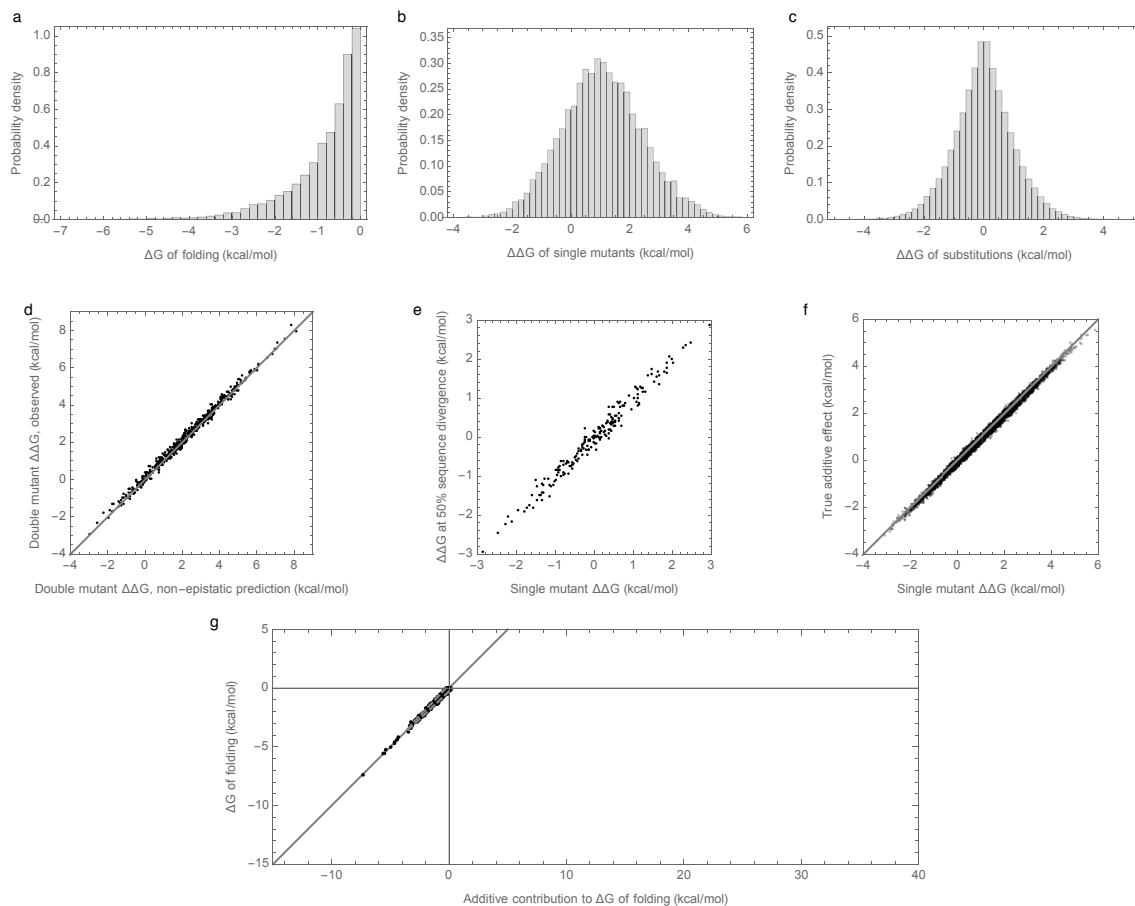


Figure 2. Free energy of folding, stability effects of mutations, and contribution of additive effects to folding stability for populations evolving at stationarity under truncation selection for protein stability under the independent epistatic effects model: (a) Free energy of folding for evolved sequences. (b) Distribution of stability effects of mutations for evolved sequences. (c) Distribution of stability effects of mutations fixed along simulated trajectories. (d) Stability effects of double mutants for evolved sequences versus predicted stability based on the sum of single mutant effects. 500 random double mutants are shown. (e) Effects of single mutations that fixed along trajectory in two evolved backgrounds that differ by 50% sequence divergence, $R^2 = 0.9856$. (f) Observed stability effects of mutations are highly correlated with the true average additive effects of the corresponding mutations. (g) The additive contribution to folding is a good indicator of the free energy of folding ($R^2 = 0.9863$) and 95% of observed sequences can fold spontaneously based on the additive contribution alone. The solid curve is derived from our analytical approximations and is predicted to contain 95% of the evolved sequences. Simulations conducted under the independent random effects model with $\mu_{\text{add}} = 1$, $\sigma_{\text{add}}^2 = 1$, $\sigma_{\text{HOC}}^2 = .01$.

133 3.2. Enrichment for epistasis observed under pairwise, but not independent models

134 In order to better understand the causes of these counter-intuitive results, we considered an
 135 alternative landscape with an identical additive component but with epistasis modeled as random
 136 draw for each sequence (from a zero-mean Gaussian distribution with variance $\sigma_{\text{HOC}}^2 = .01$), Figure 1c.
 137 The results of these simulations are shown in Figure 2. In this case, the distribution of folding
 138 stabilities, distribution of mutational effects, and extent of observed epistasis in pairwise mutations
 139 are qualitatively unchanged from the results observed under the prior, pairwise interaction model

140 (Figure 2a-f vs. Figure 1a-f). However, in this case the paradoxical contribution of epistasis to folding
141 stability is absent, so that the additive contribution to stability is sufficient for spontaneous folding
142 for most evolved sequences, and the observed folding energy is highly correlated with the additive
143 contribution ($R^2=0.9865$, Figure 2d). We therefore conclude that enrichment for epistasis under
144 stabilizing selection occurs with pair-wise epistatic interactions but not with fully random interactions.

145 What explains this difference in behavior between the model with pair-wise epistatic interactions
146 and the model with independent epistatic effects for each sequence? In order to address this question,
147 we conducted a mathematical analysis of the random field model (see Appendix B). What we came
148 to understand was that the amount of epistasis observed in double mutants under the pair-wise
149 interaction model vastly underestimates the total amount of epistasis in the energy landscape. This
150 occurs because making a double mutant only results in changes to relatively few interaction pairs (i.e.
151 those interaction pairs involving the site of either single mutant). However, as additional mutations
152 are added to the sequence, more pairs are perturbed, which unleashes additional epistasis.

153 More precisely, in the mathematical analysis we considered the expected magnitude of the
154 observed epistasis as a function of the number of mutations from an arbitrary focal sequence. That
155 is, we calculated the expected variance in the epistatic contribution among the set of all sequences at
156 a given distance d from this focal sequence. The results are shown in Figure 3 where the variance at
157 $d = 2$ is set to 1, so that the variance is expressed relative to the variance observed in a double mutant
158 analysis. We see that for small d this variance increases roughly linearly, and eventually saturates at
159 almost 50 times the expected variance at $d = 2$. In contrast, the independent random epistasis model
160 is essentially constant at all positive distances. Thus, similar levels of observed epistasis in double
161 mutants make vastly different predictions for the total amount of epistasis under the two models.

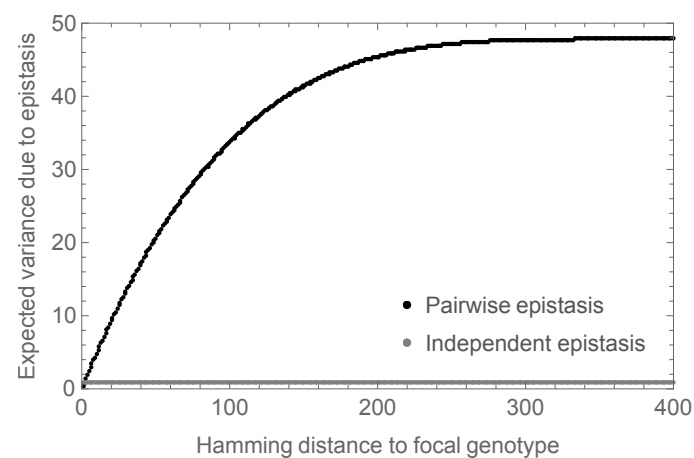


Figure 3. Expected epistatic variance as a function of distance from the focal sequence for amino acid sequences of length $l = 400$. Results for the pair-wise model shown in black, results for the independent epistasis model shown in gray. All variances are normalized relative to the expected variance at $d = 2$ which is set to 1. Notice that epistatic variance at large distances is much larger than epistatic variance at distance $d = 2$ for the pair-wise epistasis model but not for the independent epistasis model.

162 *3.3. Bivariate normal approximation for joint distribution of additive and epistatic contributions captures*
163 *impact of sequence entropy*

164 Our results on the surprising implications of small observed epistatic effects under the pairwise
165 interaction model make the results in Figure 1g appear somewhat more plausible because more
166 epistasis is present in the landscape than is apparent from the double mutants. But this observation
167 still does not provide a definite explanation for the large contribution of epistatic interactions to folding
168 stability.

169 We now provide such an explanation, based on considering the fraction of random sequences
170 that have any given pair of additive and epistatic contributions (see Appendix C for details). In
171 particular, we assume that the distribution of additive contributions to the free energy of folding for
172 random sequences is normally distributed with mean μ_1 and variance σ_1^2 , that the distribution of
173 epistatic contributions is normally distributed with mean 0 and variance σ_2^2 , and that the additive
174 and epistatic contributions are uncorrelated so that the total folding energy of a random sequence
175 $\Delta G = \Delta G_{\text{add}} + \Delta G_{\text{epi}}$ is also normally distributed, with mean $\mu = \mu_1$ and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$.
176 These normal approximations are reasonable considering that ΔG_{add} is calculated by adding up the
177 energy contribution of each site in the sequence, and ΔG_{epi} is calculated by adding up the energy
178 contribution of each pair of sites in the sequence for a large number $l = 400$ of sites. We also note
179 that in the mutation-limited regime depicted in our simulations, the stationary distribution of the
180 simulated random walk will be uniform on the set of genotypes with negative folding energies that
181 are path-connected with our choice of starting genotype [35]. Under the assumption that almost all
182 genotypes with negative folding energies are path-connected, picking a sequence from the stationary
183 distribution is equivalent to picking a sequence from the uniform distribution on sequences with
184 negative folding energies, and so our problem reduces to understanding the distribution of additive
185 folding contributions among all sequences with negative free energies of folding.

186 Under the above approximation, we now consider how—for a typical viable sequence—the
187 additive and epistatic energies jointly produce a negative free energy of folding. The key idea is
188 that there are so many more sequences with positive additive contributions to folding than there are
189 sequences with negative additive contributions that most sequences that fold have a positive additive
190 contribution despite the fact that any particular sequence with a positive additive contribution to the
191 free energy of folding has only a minuscule chance of actually folding.

192 More precisely, let us fix the value of the additive energy at $\Delta G_{\text{add}} = x$, and count the number of
193 sequences, with this given ΔG_{add} , that fold. The number of sequences with $\Delta G_{\text{add}} = x$ is proportional
194 to the probability density for the distribution of ΔG_{add} , $\text{PDF}(\mathcal{N}(\mu_1, \sigma_1^2))(x)$. Adding the epistatic
195 energy to the additive energy, the sequences with $\Delta G_{\text{add}} = x$ that fold are the sequences for which
196 $\Delta G_{\text{epi}} < -x$, i.e. their number is proportional to the cumulative distribution function of ΔG_{epi}
197 evaluated at $-x$, $\text{CDF}(\mathcal{N}(0, \sigma_2^2))(-x)$. Putting the two pieces together, the number of sequences that
198 have $\Delta G_{\text{add}} = x$ and, at the same time fold, is proportional to

$$\text{PDF}(\mathcal{N}(\mu_1, \sigma_1^2))(x) \times \text{CDF}(\mathcal{N}(0, \sigma_2^2))(-x).$$

199 Figure 4 shows this calculation for x values near the viability threshold 0. We see that over
200 this range of folding energies the number of sequences is growing extremely rapidly (Figure 4a) so
201 that, roughly speaking, the number of sequences with a given additive energy increases 10-fold for
202 every additional .45 kcal/mol. Now, Figure 4b shows the fraction of sequences with a given additive
203 contribution that spontaneously fold. This is near 1 for most sequences with a negative contribution,
204 but decreases exponentially for positive additive contributions. The net result (Figure 4c) is that a
205 typical additive contribution for a sequence that folds is often around 22 or 23 kcal/mol. While only a
206 tiny fraction of sequences with additive energies in this range fold (roughly 1 in a million), there are
207 roughly 10 billion times more sequences in this 1 kcal/mol range than there are all sequences that
208 would spontaneously fold based on their additive contribution (i.e. with $\Delta G_{\text{add}} < 0$), so that in the
209 end most sequences that fold have substantially positive additive contributions to the free energy of
210 folding.

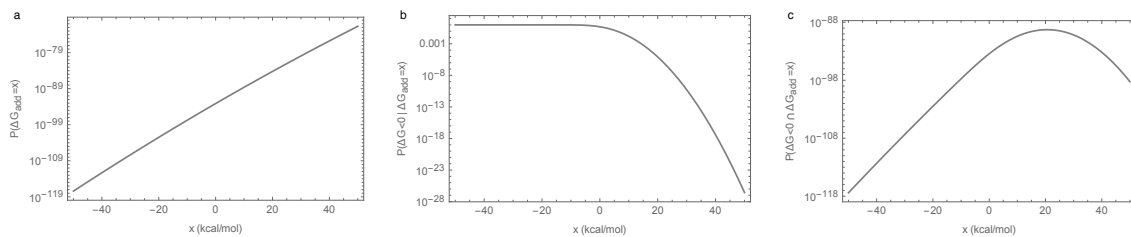


Figure 4. Illustration of the main mechanism behind the essentiality of epistatic interactions for spontaneous folding: (a) Density of random sequences with given additive free energy $P(\Delta G_{\text{add}} = x) = \text{PDF}(\mathcal{N}(\mu_1, \sigma_1^2))(x)$. (b) Fraction of sequences that fold given additive free energy $P(\Delta G < 0 | \Delta G_{\text{add}} = x) = \text{CDF}(\mathcal{N}(0, \sigma_2^2))(-x)$. (c) Density of random sequences that fold and have the given additive free energy $P(\Delta G < 0 \cap \Delta G_{\text{add}} = x)$.

211 The above argument leads to a simple prediction for the joint distribution of the free energy of
 212 folding and additive contribution to folding shown in Figures 1g and 2g: since the joint distribution
 213 for random sequences is bivariate normal, the distribution of observed energies should simply be
 214 this bivariate normal distribution truncated at $\Delta G = 0$ kcal/mol. This approximation is shown in
 215 Figures 1g and 2g by a dashed gray curve that is predicted to contain 95% of the observations, and we
 216 see that this prediction is in reasonable agreement with our simulations.

217 Moreover, under this bivariate normal approximation the average contribution of epistasis to
 218 the mean free energy of folding observed in our simulations can be calculated in a manner exactly
 219 analogous to Galton's classical results on regression to the mean [36], or the difference between the
 220 selection differential and the response to selection in the breeder's equation from quantitative genetics
 221 [37,38]. In particular, we find that the mean additive energy of viable sequences is approximately
 222 $E(\Delta G_{\text{add}} | \Delta G < 0) \approx \mu\sigma_2^2 / (\sigma_1^2 + \sigma_2^2)$ (see Appendix C for details), so that the mean contribution of
 223 epistasis is approximately $-\mu\sigma_2^2 / (\sigma_1^2 + \sigma_2^2)$, or equivalently $-\mu(1 - R^2)$, where R^2 is given by the
 224 squared correlation coefficient of additive and total folding energies taken over all of sequence space.
 225 As a result, even if the mapping from sequence to folding energy is nearly additive, in the sense that R^2
 226 is almost 1, the predicted epistatic contribution to the folding stability can still be substantial provided
 227 that the expected folding energy μ of a random sequence is sufficiently unfavorable.

228 4. Discussion

229 The role of epistasis in long-term protein evolution remains a topic of active debate [1–5]. Here we
 230 have explored a surprising phenomenon where the effects of mutations on the ΔG of folding appear to
 231 combine nearly additively, and nonetheless what little epistasis is present plays a critical role, to the
 232 extent that observed sequences would not be able to fold in the absence of these epistatic interactions.
 233 We showed that this phenomenon occurs in a model where interactions occur between pairs of sites but
 234 not in a model where each sequence differs from its additive prediction by an independent draw from
 235 a normal distribution. The difference between the two models arises because pair-wise interactions
 236 can appear nearly additive in double mutants while still producing a substantial amount of epistasis
 237 over sequence space as a whole. We also present simple analytical approximations that predict the
 238 extent of the epistatic contribution to stability in our simulations. Furthermore, these approximations
 239 suggest that this phenomenon occurs due to sequence entropy: many more sequences can fold due to
 240 a combination of epistatic and additive contributions than can fold based on the additive contributions
 241 to stability alone, and so the epistatic contribution to stability is typically essential when one observes
 242 a random sequence that folds. These results add to a growing literature demonstrating that natural
 243 selection can enrich for epistatic interactions in both adaptive [39–41], and nearly neutral [16] evolution,
 244 such that the mutations that fix during evolution can have a very different pattern of epistasis than
 245 random mutations.

246 Our simulations (Figure 1) recapitulate the known qualitative features of protein evolution under
247 purifying selection for folding stability to a surprising degree, with the exception of matching the
248 observed stability margin, which is smaller in our simulations than for experimentally measured
249 folding energies [29] (free energy of folding is typically -5 to -10 kcal/mol versus -1 kcal/mol in our
250 simulations). However, this extremely small stability margin is a well-known artifact of our decision
251 to model fitness as a step function in stability [8] rather than a more realistic logistic function [11,30],
252 and the fact that our simulations do not include any of the other factors that would tend to increase
253 the stability margin such as selection for mutational robustness [8,32,35] or selection to prevent
254 misfolding due to errors in translation [42]. Nonetheless the simple sequence-to-fitness mapping
255 employed in our simulations allows us to provide a relatively simple and complete theory for the
256 observed phenomenon. Moreover, we emphasize that it is easy to find realistic parameters where
257 the mean additive contribution to stability is far less stable than shown in Figure 1, so we anticipate
258 that the possibility that most sequences fold only due to epistasis would be robust even if sequences
259 experienced a much larger stability margin.

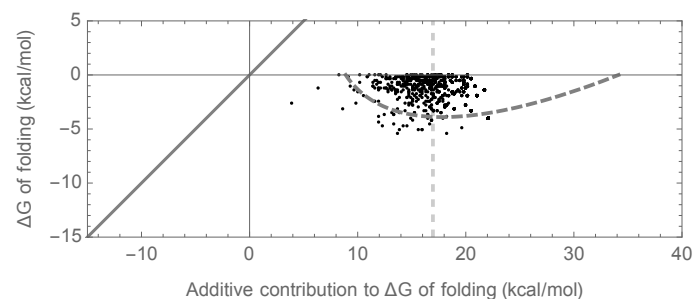


Figure 5. Joint distribution of ΔG of folding and the additive contribution to ΔG of folding for the independent epistasis model with σ_{HOC}^2 chosen so that the bivariate normal approximation matches the bivariate normal approximation shown in Figure 1g. Simulations conducted under the independent epistasis model with $\mu_{\text{add}} = 1$, $\sigma_{\text{add}}^2 = 1$, $\sigma_{\text{HOC}}^2 = 21.6$. Dashed curve shows area predicted to include 95% of sequences at stationarity under the bivariate normal approximation; dashed vertical line shows approximate left-most edge of region where bivariate normal approximation is valid based on a crude percolation theory argument (see text).

260 A different limitation of our results concerns the assumption, in our truncated bivariate normal
261 approximation, that the set of sequences with negative folding energies is mutationally connected,
262 and hence accessible to an evolving population. In particular, the theory breaks down if a large
263 fraction of sequences that fold appear as isolated peaks or small isolated clusters of sequences. Figure 5
264 shows an example of this limitation for the case of the independent model with parameters chosen
265 so that the bivariate normal approximation is identical to the bivariate normal approximation for the
266 pairwise model shown in Figure 1. The figure shows some enrichment for epistasis but not as much as
267 predicted by our bivariate normal approximation. Using the crude percolation-theory argument that
268 the connected network of sequences can extend only up to the additive energy at which each sequence
269 has on average one neighbor that folds due to epistasis [43], we can derive the approximate upper
270 limit of the distribution of additive energies as $-\sigma_{\text{HOC}} \Psi^{-1}(1/(400 \times 19)) = 16.96$, where Ψ^{-1} is the
271 inverse cumulative distribution function of a standard normal distribution. This approximate upper
272 limit is shown by the dashed vertical line in Figure 5. We see that the cloud of observed sequences is
273 primarily to the left of this line, with a notable absence of sequences with substantially more positive
274 additive contributions. This analysis of connectivity of the set of sequences that fold highlights that
275 pairwise interactions have several special features: not only can they appear locally non-epistatic
276 while harboring a substantial amount of epistasis at greater distances, but as long as the individual
277 coefficients remain small they produce energy landscapes that change smoothly over sequence space,

278 producing the enormous connected networks of sequences whose traversal allows the evolution of a
279 sizable epistatic contribution to folding.

280 It is natural to ask whether the essentiality of epistatic interactions for the functionality of evolved
281 sequences is likely to hold in other contexts where additivity is thought to prevail, such as protein-DNA
282 and protein-protein binding [18]. However, this effect is unlikely to occur in most of these cases because
283 the sequences are much shorter and the set of functional sequences makes up a much larger proportion
284 of genotypic space. In particular, it is helpful to consider the z-score of functional sequences relative to
285 random sequences, since the regression to the mean effect observed here is proportional to the absolute
286 value of the z-score. For instance, the average TF binding motif in bacteria has an information content
287 of 23 bits, corresponding to a p-value of 10^{-7} or a z-score of roughly -5, with eukaryotic transcription
288 factors having even smaller information contents and therefore smaller absolute value z-scores [44].
289 In contrast, the z-scores of the spontaneously folding sequences observed in our simulations are on
290 the order of -20, which we would expect to result in a roughly 4-fold larger contribution of epistasis
291 to binding energy at stationarity than for a bacterial transcription factor binding site. Such extreme
292 z-scores are not even possible in short DNA elements, e.g. the most extreme z-score possible in a
293 DNA sequence of length 20 is only -7. Thus, the essentiality of epistatic interactions observed here
294 is likely possible only because protein sequence space is very large compared to other well-studied
295 sequence-function relationships focused on smaller genetic elements.

296 Finally it is important to emphasize that the key question of whether epistatic interactions
297 for protein stability are essential for protein folding in naturally evolved sequences remains open.
298 Our contribution only demonstrates that such an effect is qualitatively consistent with empirical
299 observations on the thermodynamic effects of mutations and the results of prior simulation studies,
300 and suggests that the overall importance of epistasis for stability depends on the precise form of
301 epistasis involved. Intriguingly, the experimental observation that pairwise correlations between
302 site-specific amino acid usages are sometimes necessary for folding [45] provides evidence for both the
303 presence of the low-order epistatic interactions that result in a substantial contribution of epistasis to
304 protein folding and also for the possible essentiality of these interactions. Thus, determining whether
305 epistasis is essential for folding of observed sequences is a key question for the field, from both
306 theoretical and empirical perspectives. Importantly, our analysis shows that most standard designs
307 for examining the extent of epistasis for protein stability cannot adjudicate this question, because
308 they examine how the extent of mutations change at a only single distance from a reference genotype.
309 For instance, the analysis of double mutants considers the change in the effect of a mutation in a
310 sequence at distance 1; and comparison of the effects of mutations on two diverged backgrounds,
311 e.g., [14,19], can only determine the extent of epistasis at that one level of divergence. Rather, the two
312 theories analyzed here differ in how the extent of epistasis changes with distance (e.g. Figure 3 and
313 Appendix B.3). Thus, the critical experiment is to measure how the energetic effects of individual
314 mutations change across several different levels of sequence divergence (c.f. [16]).

315 **Author Contributions:** Conceptualization, A.P., J.B.P, J.B.K., and D.M.M.; Software, A.P., J.Z., and D.M.M.;
316 Formal Analysis A.P., J.Z., J.B.K. and D.M.M.; Investigation A.P. and D.M.M.; Writing - Original Draft Preparation,
317 A.P., J.Z., J.B.P, J.B.K., and D.M.M.; Writing - Review & Editing, A.P., J.B.P. and D.M.M.; Funding Acquisition, J.B.P
318 and J.B.K.

319 **Funding:** J.B.K. and A.P. were supported in part by a grant from the CSHL/Northwell Health alliance. J.B.P.
320 acknowledges support from the David and Lucile Packard Foundation and the U.S. Army Research Office
321 (W911NF-12-R-0012-04).

322 **Acknowledgments:** We thank Ashley Teufel and David Liberles for organizing this special issue of *Genes*.

323 **Conflicts of Interest:** The authors declare no conflict of interest. The founding sponsors had no role in the design
324 of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the
325 decision to publish the results.

326 Appendix A

327 Appendix A.1 Model for folding energy

328 Given an alphabet $\mathcal{A} = \{0, 1, \dots, a-1\}$ and a sequence length l , let \mathcal{S} be the set all possible
 329 strings of length l built from alphabet \mathcal{A} . The free energy of folding $\Delta G(x)$ for each sequence $x \in \mathcal{S}$
 330 is defined as the sum of (1) an additive component that measures the energy contribution of each allele
 331 at each position in the sequence, and (2) an epistatic component that describes the energy contribution
 332 of pairwise interactions among alleles for the pairwise model, or a random draw from a normal
 333 distribution for the independent epistasis model:

$$\Delta G(x) := \Delta G_{\text{add}}(x) + \Delta G_{\text{epi}}(x). \quad (\text{A1})$$

334 To specify each of the terms $\Delta G_{\text{add}}(x)$ and $\Delta G_{\text{epi}}(x)$, we introduce the following notations. Let x_k
 335 denote the k -th letter in the sequence x . Then, for a set of indices $K = \{k_1, \dots, k_{|K|}\}$, let $\beta_{K,\alpha}$ denote the
 336 energetic contribution to the folding energy of x when the substring $x_{k_1}x_{k_2}\dots x_{k_{|K|}}$ is equal to α . Using
 337 this notation, we let:

$$\Delta G_{\text{add}}(x) := \sum_{k=1}^l \beta_{\{k\},x_k} \quad (\text{A2})$$

338 be the additive contribution to the folding energy, and

$$\Delta G_{\text{epi}}(x) := \sum_{k_1=1}^{l-1} \sum_{k_2=k_1+1}^l \beta_{\{k_1,k_2\},x_{k_1}x_{k_2}} \quad (\text{A3})$$

339 be the epistatic contribution for the pairwise model. For the independent epistasis model, we instead
 340 let $\Delta G_{\text{epi}}(x)$ be an independent random draw from a normal distribution with mean 0 and variance
 341 σ_{HOC}^2 .

342 Having described the form of our energy model, we now describe how we choose the $\beta_{K,\alpha}$. Some
 343 care is needed in this choice in order to ensure that $\Delta G_{\text{epi}}(x)$ for the pairwise model is a pure epistatic
 344 contribution, that is, that the average effect of any given point mutation over sequence space is zero.
 345 For the independent epistasis model, no additional steps are needed because the epistatic contribution
 346 is drawn independently for each sequence and each possible point mutation can appear on many
 347 genetic backgrounds, so by the law of large numbers the average epistatic effect of any given point
 348 mutation will be very near zero.

349 We first describe how we choose the $\beta_{K,\alpha}$ for the additive component. For each position k , the
 350 vector of coefficients $\vec{\beta}_{\{k\}} := (\beta_{\{k\},0}, \dots, \beta_{\{k\},a-1})$ specifies the contribution to folding energy of each
 351 allele at the given position. We impose the constraint

$$\frac{1}{a} \sum_{\alpha=0}^{a-1} \beta_{\{k\},\alpha} = \mu_{\text{add}}, \quad (\text{A4})$$

352 so that the mean additive folding energy over all possible sequences is $l\mu_{\text{add}}$, i.e. $\langle \Delta G_{\text{add}}(x) \rangle_x = l\mu_{\text{add}}$.

353 Turning to the epistatic component for the pair-wise model, for each position pair (k_1, k_2) such
 354 that $k_2 < k_1$, the matrix of coefficients $B_{\{k_1,k_2\}} := (\beta_{\{k_1,k_2\},\alpha_1\alpha_2})_{\alpha_1\alpha_2}$ specifies the contribution to folding
 355 energy resulting from the interaction of the alleles at the given positions. We impose the constraints
 356 that all row and all column sums of $B_{\{k_1,k_2\}}$ equal 0, i.e.

$$\sum_{\alpha_1=0}^{a-1} \beta_{\{k_1,k_2\},\alpha_1\alpha_2} = 0, \alpha_2 \in \mathcal{A}, \quad \text{and} \quad \sum_{\alpha_2=0}^{a-1} \beta_{\{k_1,k_2\},\alpha_1\alpha_2} = 0, \alpha_1 \in \mathcal{A}. \quad (\text{A5})$$

357 These constraints ensure that the mean epistatic energy over all possible sequences is 0, i.e.
 358 $\langle \Delta G_{\text{epi}}(x) \rangle_x = 0$. It also follows from conditions (A4) and (A5) that the vectors $\Delta G_{\text{add}} :=$
 359 $(\Delta G_{\text{add}}(x))_{x \in \mathcal{S}}$ and $\Delta G_{\text{epi}} := (\Delta G_{\text{epi}}(x))_{x \in \mathcal{S}}$ are orthogonal, see Appendix B.2.

360 We now turn to our procedure for drawing the additive and pairwise coefficients $\beta_{K,\alpha}$ subject to
 361 the above constraints.

362 Appendix A.2 Choosing the coefficients for the energy function

363 For each position k , we choose $\vec{\beta}_{\{k\}} := (\beta_{\{k\},0}, \dots, \beta_{\{k\},a-1})$ from an a -dimensional normal
 364 distribution that has identical marginals, mean vector $(\mu_{\text{add}}, \dots, \mu_{\text{add}})$, and some covariance matrix
 365 that ensures that constraint (A4) is satisfied. Similarly, for each pair of positions (k_1, k_2) , $k_2 < k_1$, we
 366 choose the elements of $B_{\{k_1, k_2\}} := (\beta_{\{k_1, k_2\}, \alpha_1 \alpha_2})_{\alpha_1 \alpha_2}$ from an a^2 -dimensional normal distribution that
 367 has identical marginals, mean vector $\vec{0}$, and some covariance matrix that ensures that the constraints
 368 in (A5) are satisfied. To draw the coefficients from such distributions we implement the following
 369 procedures.

370 Appendix A.2.1 Choosing the coefficients for the first order terms

371 The subspace of \mathbb{R}^a defined by the constraint in (A4) is a hyperplane specified by the normal
 372 vector $\vec{n} = (1, \dots, 1)$ going through the point $(\mu_{\text{add}}, \dots, \mu_{\text{add}})$. Let \vec{e}_i , $i = 1, \dots, a$, denote the standard
 373 basis in \mathbb{R}^a , i.e. $(\vec{e}_i)_j = 1$ if $j = i$ and 0 otherwise. We obtain a basis spanning \mathbb{R}^a if we take the standard
 374 basis and replace its first vector by \vec{n} :

$$\{\vec{n}, \vec{e}_2, \dots, \vec{e}_a\}.$$

375 By performing Gram-Schmidt orthogonalization on this set of vectors, in the given order, we obtain
 376 an orthogonal basis in \mathbb{R}^a whose first vector is the normal vector \vec{n} . Let \vec{b}_i , $i = 1, \dots, a-1$, denote
 377 the normalized vectors of this basis other than \vec{n} , i.e. the \vec{b}_i vectors are orthonormal and span the
 378 hyperplane specified by constraint (A4). It follows that if $\vec{z} \in \mathbb{R}^{(a-1)}$ is a vector of iid random variables
 379 of normal distribution with mean 0 and some standard deviation σ_{add} , then the vector

$$(\mu_{\text{add}}, \dots, \mu_{\text{add}}) + \sum_{k=1}^{a-1} z_k \vec{b}_k \quad (\text{A6})$$

380 has the desired distribution. In particular, each component $\mu_{\text{add}} + \sum_{k=1}^{a-1} z_k (\vec{b}_k)_i$ has normal distribution
 381 with mean μ_{add} and variance $((\alpha - 1)/\alpha) \sigma_{\text{add}}^2$, i.e.

$$\beta_{\{k\}, \alpha} \sim \mathcal{N}\left(\mu_{\text{add}}, \sqrt{\frac{\alpha - 1}{\alpha}} \sigma_{\text{add}}\right). \quad (\text{A7})$$

382 Appendix A.2.2 Choosing the coefficients for the second order terms

383 For each position pair (k_1, k_2) , $k_2 < k_1$, we choose the elements of $B := (\beta_{\{k_1, k_2\}, \alpha_1 \alpha_2})_{\alpha_1 \alpha_2}$ from an
 384 a^2 -dimensional normal distribution that has identical marginals, mean vector $\vec{0}$, and some covariance
 385 matrix that ensures that the constraints in (A5) are satisfied. In what follows, instead of using the
 386 matrix notation B for the coefficients, we use a vector notation, obtained by concatenating the rows of
 387 B .

388 The subspace of \mathbb{R}^{a^2} defined by the constraints in (A5) is obtained as the intersection of the set of
 389 hyperplanes specified by the normal vectors

$$(\vec{n}_{\text{row}, i})_j = \begin{cases} 1 & \text{if } j \in \{(i-1)a + 1, (i-1)a + 2, \dots, (i-1)a + (a-1)\} \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, 2, \dots, a-1, \quad (\text{A8})$$

$$(\vec{n}_{\text{column},i})_j = \begin{cases} 1 & \text{if } j \in \{i, a+i, 2a+i, \dots, (a-2)a+i\} \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, 2, \dots, a-1, \quad (\text{A9})$$

390 and

$$\vec{n}_j = 1, \quad j \in \{1, 2, \dots, a^2\}. \quad (\text{A10})$$

391 Let $\vec{e}_i, i = 1, \dots, a^2$, denote the standard basis in \mathbb{R}^{a^2} , i.e. $(\vec{e}_i)_j = 1$ if $j = i$ and 0 otherwise. Then,
392 the set of vectors

$$\{\vec{n}\} \cup \{\vec{n}_{\text{row},i} : i = 1, \dots, a-1\} \cup \{\vec{n}_{\text{column},i} : i = 1, \dots, a-1\} \cup \{\vec{e}_i : i = 1, \dots, (a-1)^2\}$$

393 form a basis spanning \mathbb{R}^{a^2} . By performing Gram-Schmidt orthogonalization on this set of vectors, in
394 the given order, we obtain an orthogonal basis in \mathbb{R}^{a^2} whose first $2a-1$ vectors span the orthogonal
395 complement of the subspace defined by the constraints in (A4). Let us denote the normalized version
396 of the rest of the vectors of this basis by $\vec{b}_i, i = 1, \dots, (a-1)^2$. If $\vec{z} \in \mathbb{R}^{(a-1)^2}$ is a vector of iid random
397 variables of normal distribution with mean 0 and some standard deviation $\sigma_{\beta,2}$, then the vector

$$\sum_{k=1}^{(a-1)^2} z_k \vec{b}_k \quad (\text{A11})$$

398 has the desired distribution. In particular, each component $\sum_{k=1}^{(a-1)^2} z_k (\vec{b}_k)_i$ has normal distribution
399 with mean 0 and variance $(\alpha-1)^2/\alpha^2\sigma_{\text{add}}^2$, i.e.

$$\beta_{\{k_1, k_2\}, \alpha_1 \alpha_2} \sim \mathcal{N}\left(0, \frac{\alpha-1}{\alpha} \sigma_{\text{epi}}\right). \quad (\text{A12})$$

400 Appendix B

401 Appendix B.1 Background

402 To derive the expected variance due to epistasis at a given distance from a focal genotypes, as
403 shown in Figure 3, we will first need to introduce some notation.

404 Given an alphabet $\mathcal{A} = \{0, 1, \dots, a-1\}$, let \mathcal{S} be the set all possible sequences (configurations) of
405 length l , with cardinality $|\mathcal{S}| = a^l \equiv N$. We introduce the Hamming distance \mathbf{d} :

$$\mathbf{d} : \mathcal{S} \times \mathcal{S} \rightarrow \{0, 1, \dots, l\}$$

$$\mathbf{d}(x, x') = \mathbf{d}(x', x) = \text{Number of sites where } x \text{ and } x' \text{ differ.}$$

406 Then set of all sequences form the Hamming graph, $\mathcal{G} = (\mathcal{S}, E)$, with set of edges

$$E = \{(x, x') \in \mathcal{S} \times \mathcal{S} \mid \mathbf{d}(x, x') = 1\}$$

407 That is, two sequences x and x' are adjacent on \mathcal{G} , i.e. $x \sim x'$, if and only if $\mathbf{d}(x, x') = 1$.

The graph Laplacian L of \mathcal{G} is a $N \times N$ matrix

$$L(x, x') = \begin{cases} l(a-1) & x = x' \\ -1 & x \sim x' \\ 0 & \text{otherwise} \end{cases}. \quad (\text{A13})$$

408 Applying L to any N -dimensional vector f , we have

$$(Lf)(x) = l(a-1)f(x) - \sum_{x' \sim x} f(x') = \sum_{x' \sim x} f(x) - f(x'), \quad (\text{A14})$$

409 which is the sum of differences between the focal type x and all the adjacent sequences.

410 The graph Laplacian L has $l+1$ distinct eigenvalues ak , $l \geq k \geq 0$, each with multiplicity
 411 $\binom{l}{k}(a-1)^k$. The k -th eigenspace can be interpreted as the space of all energy landscapes of interaction
 412 order k .

413 Appendix B.2 Orthogonality of ΔG_{epi} and ΔG_{add}

414 Before we derive the results shown in Figure 3, we pause to show that with the constraint on the
 415 coefficients β defined previously, ΔG_{add} and ΔG_{epi} are contained in orthogonal eigenspaces therefore
 416 are mutually orthogonal and also orthogonal to the eigenspaces of higher order interactions.

417 Appendix B.2.1 ΔG_{add}

418 Recall the definition:

$$\Delta G_{\text{add}}(x) = \sum_{k=1}^l \beta_{\{k\}, x_k}. \quad (\text{A15})$$

419 We first split the vector ΔG_{add} into its constant and linear components:

$$\Delta G_{\text{lin}} \equiv \Delta G_{\text{add}} - l\mu_{\text{add}}\mathbf{1}, \quad (\text{A16})$$

420 where $\mathbf{1}$ denote the column vector with all one's. For any graph Laplacian, $L\mathbf{1} = 0$, therefore the
 421 constant part of ΔG_{add} is in the null space of L , which is equivalent to the eigenspace associated with
 422 eigenvalue zero. Furthermore, according to equation A14:

$$\begin{aligned} (L\Delta G_{\text{lin}})(x) &= l(a-1)\Delta G_{\text{lin}}(x) - \sum_{x' \sim x} \Delta G_{\text{lin}}(x') \\ &= l(a-1)\Delta G_{\text{lin}}(x) + l^2(a-1)\mu_{\text{add}} - \sum_{x' \sim x} \sum_{k=1}^l \beta_{\{k\}, \{x'_k\}} \\ &= l(a-1)\Delta G_{\text{lin}}(x) + l^2(a-1)\mu_{\text{add}} - \sum_{k=1}^l \sum_{x' \sim x} \beta_{\{k\}, \{x'_k\}} \\ &= l(a-1)\Delta G_{\text{lin}}(x) + l^2(a-1)\mu_{\text{add}} - \sum_{k=1}^l \left((l-1)(a-1)\beta_{\{k\}, \{x_k\}} + \sum_{\alpha \in \mathcal{A} \setminus \{x_k\}} \beta_{\{k\}, \{\alpha\}} \right) \\ &= l(a-1)\Delta G_{\text{lin}}(x) + l^2(a-1)\mu_{\text{add}} - \sum_{k=1}^l \left((l-1)(a-1)\beta_{\{k\}, \{x_k\}} + a\mu_{\text{add}} - \beta_{\{k\}, \{x_k\}} \right) \\ &= l(a-1)\Delta G_{\text{lin}}(x) - \sum_{k=1}^l \left(((l-1)(a-1) - 1)\beta_{\{k\}, \{x_k\}} - (l(a-1) - a)\mu_{\text{add}} \right) \\ &= l(a-1)\Delta G_{\text{lin}}(x) - (l(a-1) - a) \sum_{k=1}^l \left(\beta_{\{k\}, \{x_k\}} - \mu_{\text{add}} \right) \\ &= l(a-1)\Delta G_{\text{lin}}(x) - (l(a-1) - a)\Delta G_{\text{lin}}(x) \\ &= a\Delta G_{\text{lin}}(x) \end{aligned}$$

423 On the fourth line, for each choice of site k , we group the sequences that are adjacent to x into two
 424 groups. The first group consist of $(l-1)(a-1)$ sequences that are identical to the focal x on site k , so
 425 have coefficient $\beta_{\{k\}, \{x_k\}}$. The other group consist of the rest $a-1$ sequences that differ from x at site k .

426 So we need to sum through $\alpha \in \mathcal{A} \setminus \{x_k\}$. Due to the constraint that $\frac{1}{a} \sum_{\alpha=0}^{a-1} \beta_{\{k\},\{\alpha\}} = \mu_{\text{add}}$, this sum
 427 must be equal to $a \mu_{\text{add}} - \beta_{\{k\},\{x_k\}}$. Then on the seventh line we also note that $(l-1)(a-1) - 1 =$
 428 $l(a-1) - a$. Thus, we conclude that ΔG_{lin} is an eigenvector of L with eigenvalue a .

429 To summarize, we have shown that the constant and linear part of ΔG_{add} belong to the zeroth
 430 and first eigenspace of L with eigenvalues 0 and a , respectively. We also note that since $L \Delta G_{\text{add}} =$
 431 $a \Delta G_{\text{lin}} = a \Delta G_{\text{add}} - a \beta_{\emptyset}$, ΔG_{add} is an "elementary" landscape [46,47].

432 Appendix B.2.2 ΔG_{epi}

433 Next, we verify that ΔG_{epi} is an eigenvector of the graph Laplacian L with eigenvalue $2a$. The
 434 spirit is the same as with ΔG_{lin} . First note that $(L \Delta G_{\text{epi}})(x) = l(a-1) \Delta G_{\text{epi}}(x) - \sum_{x' \sim x} \Delta G_{\text{epi}}(x')$.
 435 The second term can be expanded as:

$$\sum_{x' \sim x} \Delta G_{\text{epi}}(x') = \sum_{x' \sim x} \sum_{k_2 < k_1} \beta_{\{k_1, k_2\}, \{x'_{k_1}, x'_{k_2}\}} = \sum_{k_2 < k_1} \sum_{x' \sim x} \beta_{\{k_1, k_2\}, \{x'_{k_1}, x'_{k_2}\}} \quad (\text{A17})$$

$$= \sum_{k_2 < k_1} \left((l-2)(a-1) \beta_{\{k_1, k_2\}, \{x_{k_1}, x_{k_2}\}} + \sum_{\alpha_1 \in \mathcal{A} \setminus \{x_{k_1}\}} \beta_{\{k_1, k_2\}, \{\alpha_1, x_{k_2}\}} \right) \quad (\text{A18})$$

$$+ \sum_{\alpha_2 \in \mathcal{A} \setminus \{x_{k_2}\}} \beta_{\{k_1, k_2\}, \{x_{k_1}, \alpha_2\}} \quad (\text{A19})$$

$$= \sum_{k_2 < k_1} \left((l-2)(a-1) \beta_{\{k_1, k_2\}, \{x_{k_1}, x_{k_2}\}} - 2 \beta_{\{k_1, k_2\}, \{x_{k_1}, x_{k_2}\}} \right) \quad (\text{A20})$$

$$= \sum_{k_2 < k_1} ((l-2)(a-1) - 2) \beta_{\{k_1, k_2\}, \{x_{k_1}, x_{k_2}\}} \quad (\text{A21})$$

$$= ((l-2)(a-1) - 2) \Delta G_{\text{epi}}(x) \quad (\text{A22})$$

436 , where on the third line we divide the sequences that are adjacent to x into three groups. The first
 437 group consist of $(l-2)(a-1)$ sequences that are identical to the focal x on the two sites that are chosen,
 438 so have the same coefficient as x : $\beta_{\{k_1, k_2\}, \{x_{k_1}, x_{k_2}\}}$. The other two groups each consist of sequences that
 439 differ from x at site k_1 or k_2 . Due to the constraint that $\sum_{\alpha \in \mathcal{A}} \beta_{\{k_1, k_2\}, \{\alpha_1, \alpha_2\}} = 0$ for $\alpha = \alpha_1, \alpha_2$, the sum
 440 through the $a-1$ sequences that differ from x on each site must be equal to $-\beta_{\{k_1, k_2\}, \{x_{k_1}, x_{k_2}\}}$

441 Therefore,

$$(L \Delta G_{\text{epi}})(x) = l(a-1) \Delta G_{\text{epi}}(x) - \sum_{x' \sim x} \Delta G_{\text{epi}}(x') \quad (\text{A23})$$

$$= (l(a-1) - (l-2)(a-1) + 2) \Delta G_{\text{epi}}(x) \quad (\text{A24})$$

$$= 2a \Delta G_{\text{epi}}(x). \quad (\text{A25})$$

442 We have shown that ΔG_{lin} and ΔG_{epi} are eigenvectors of the graph Laplacian L and have
 443 eigenvalues a and $2a$, respectively. Therefore they are mutually orthogonal, and furthermore ΔG_{epi}
 444 is orthogonal to ΔG_{add} , since ΔG_{add} is contained in the union of the eigenspaces corresponding to 0
 445 and a , whereas ΔG_{epi} is in the eigenspace corresponding to eigenvalue $2a$. Furthermore, all of ΔG_{add} ,
 446 ΔG_{lin} and ΔG_{epi} are orthogonal to eigenspaces spanned by interactions of order greater than 2.

447 Appendix B.3 Expected variance in energy of distance classes under the random field model

448 We now derive the results shown in Figure 3. For a sequence $x \in \mathcal{S}$, its energy under a general
 449 random field model is [20,21]

$$f(x) = \sum_{k=0}^l \sum_{i=1}^{\binom{l}{k}(a-1)^k} b_{k_i} \phi_{k_i}(x), \quad (\text{A26})$$

450 where $\{\phi_{k_i}\}_{l \geq k \geq 0}$ is the set of orthonormal basis functions that spans \mathbb{R}^S and ϕ_{k_i} is the i -th basis
 451 function of interaction order k . $\{b_{k_i}\}$ are random variables representing the interaction coefficients.
 452 Here we choose $\{\phi_{k_i}\}$ to be eigenvectors of the the graph Laplacian L associated with \mathcal{G} .

453 For simplicity, here we only consider models with interactions of a certain order $k \geq 1$, so

$$f(x) = \sum_i b_{k_i} \phi_{k_i}(x). \quad (\text{A27})$$

454 Since the $\{b_{k_i}\}$ of different orders are statistically independent, to simplify the notation, we derive the
 455 following results for energy landscapes with only k -th order interactions and only consider landscapes
 456 with multiple orders of interaction at the end. In our random field model, the coefficients b_{k_i} 's for a
 457 given k are drawn i.i.d from some distribution with mean 0 and variance σ_k^2 . Without loss of generality,
 458 we set $\sigma_k^2 = 1$. So we have

$$f \sim \mathcal{D}(0, W), \quad (\text{A28})$$

459 where W is the covariance matrix. The distribution \mathcal{D} is not specified here because we are only
 460 concerned with the first and second moment. The covariance matrix W only depends on $\mathbf{d}(x, x')$ and
 461 its entries are given by:

$$\sigma_{f(x)f(x')} = E\left[\sum_i b_{k_i} \phi_{k_i}(x) \sum_j b_{k_j} \phi_{k_j}(x')\right] \quad (\text{A29})$$

$$= \sum_{i,j} E[b_{k_i} b_{k_j}] \phi_{k_i}(x) \phi_{k_j}(x') \quad (\text{A30})$$

$$= \sum_i E[b_{k_i}^2] \phi_{k_i}(x) \phi_{k_i}(x') \quad (\text{A31})$$

$$= \sum_i \phi_{k_i}(x) \phi_{k_i}(x') \quad (\text{A32})$$

$$= \sum_{q=0}^{\min\{k, \mathbf{d}(x, x')\}} a^{-l} (-1)^q (a-1)^{k-q} \binom{\mathbf{d}(x, x')}{q} \binom{l - \mathbf{d}(x, x')}{k-q} \quad (\text{A33})$$

$$\equiv w_k(\mathbf{d}(x, x')) \quad (\text{A34})$$

462 The second to last step is not obvious but is well known in coding theory and some theoretical studies
 463 of fitness landscape and is commonly referred to as the Krawtchouk polynomial [20,48].

464 For an energy landscape f drawn from the above distribution, the sample variance of energies of
 465 all sequences at distance $l \geq d > 0$ to wt is:

$$V(\{f(x)|x \in S(\text{wt}, d)\}) \equiv V(d) = \frac{1}{n} \sum_{x \in S(\text{wt}, d)} (f(x) - \langle f(x') \rangle_{S(\text{wt}, d)})^2 \quad (\text{A35})$$

466 Here, we use $\langle \cdot \rangle_T$ to denote the mean taken over the set T . $S(\text{wt}, d) = \{x \in \mathcal{S} | \mathbf{d}(x, \text{wt}) = d\}$ is the set
 467 of all sequences at distance d to wt and $|S(\text{wt}, d)| = \binom{l}{d} (a-1)^d = n$.

468 Using a well-known expression for the expected sample variance, we can write [49]:

$$E[V(d)] = \frac{n-1}{n} (\overline{\sigma}_d^2 + \sigma_{d,\mu}^2 - \bar{\sigma}_d). \quad (\text{A36})$$

469 The three quantities are:

$$\overline{\sigma}_d^2 = \frac{1}{n} \sum_{x \in S(\text{wt}, d)} \sigma_{f(x)}^2 \quad (\text{A37})$$

$$\sigma_{d, \mu}^2 = \frac{1}{n-1} \sum_{x \in S(\text{wt}, d)} (E[f(x)] - \frac{1}{n} \sum_{x' \in S(\text{wt}, d)} E[f(x')])^2 \quad (\text{A38})$$

$$\overline{\sigma}_d = \langle \sigma_{f(x)f(x')} \rangle_{\{(x, x') \in S^2(\text{wt}, d) | x \neq x'\}} \quad (\text{A39})$$

470 We now turn to deriving each of these three quantities individually.

471 First, we have the mean of variances $\overline{\sigma}_d^2$ for sequences at distance d to wt. By setting $d = 0$ in the
472 Krawchouk polynomial:

$$\sigma_{f(x)}^2 = w_k(0) = a^{-l} \binom{l}{k} (a-1)^k \sigma_k^2 = a^{-l} \binom{l}{k} (a-1)^k, \quad (\text{A40})$$

473 for all $x \in S(\text{wt}, d)$. $\binom{l}{k} (a-1)^k$ is the number of k -th order interactions. And the last step was simplified
474 by setting $\sigma_k^2 = 1$.

475 Next, we have $\sigma_{d, \mu}^2$, which is the variance of the mean energy, $E[f(x)]$, of sequences in $S(\text{wt}, d)$.
476 Since $E[f(x)]$ is constant across all $x \in S(\text{wt}, d)$, we have

$$\sigma_{d, \mu}^2 = 0. \quad (\text{A41})$$

477 Last is the mean covariance between energies of sequences $x, x' \in S(\text{wt}, d)$, $x \neq x'$:

$$\overline{\sigma}_d = \langle \sigma_{f(x)f(x')} \rangle_{\{(x, x') \in S^2(\text{wt}, d) | x \neq x'\}} \quad (\text{A42})$$

$$= \frac{1}{\binom{l}{d} (a-1)^d} \sum_{x \in S(\text{wt}, d)} \langle \sigma_{f(x)f(x')} \rangle_{\{x' | \mathbf{d}(x', \text{wt}) = d, x' \neq x\}} \quad (\text{A43})$$

$$= \langle \sigma_{f(x)f(x')} \rangle_{\{x' | \mathbf{d}(x', \text{wt}) = d, x' \neq x\}} \quad (\text{A44})$$

$$= \frac{1}{\binom{l}{d} (a-1)^d - 1} \sum_{d'=1}^{\min\{l, 2d\}} N(d, d') w_k(d') \quad (\text{A45})$$

478 Due to the assumption that the covariance structure only depends on d , the summand on the second
479 line is the same for all $x \in S(\text{wt}, d)$. Therefore we can arbitrarily pick a sequence x and calculate a
480 weighted sum with weights given by

$$N(d, d') = |\{x' | \mathbf{d}(\text{wt}, x') = d, \mathbf{d}(x, x') = d'\}| = |S(\text{wt}, d) \cap S(x, d')|, \quad (\text{A46})$$

481 which is the number of sequences at distance d to wt and d' to x . Note that $N(d, d')$ is the same for all
482 $x \in S(\text{wt}, d)$ so only depends d and d' . The normalizing factor

$$\sum_{d'=1}^{\min\{l, 2d\}} N(d, d') = \binom{l}{d} (a-1)^d - 1 \quad (\text{A47})$$

483 is the total number of sequences in $S(\text{wt}, d)$ minus the focal sequence x .

484 Our final task is to count $N(d, d')$. First we pick $d \geq s \geq 0$ sites out of d sites on which x and wt
485 differ and set the states of x' on these sites to be the same as wt. The number of choices: $\binom{d}{s}$. Second,
486 since $\mathbf{d}(x', \text{wt}) = d$, we must choose s sites out of the $l-d$ sites where x and wt are identical and set
487 them to be one of the $a-1$ states for x' . The number of choices is $\binom{l-d}{s} (a-1)^s$. Third, now we have x
488 and x' differ on $2s$ sites and since $\mathbf{d}(x, x') = d'$, we need to choose $d' - 2s$ sites for x' out of the $d-s$

Table A1. A simple example for counting $N(d, d')$, with $\mathbf{d}(\text{wt}, x) = \mathbf{d}(\text{wt}, x') = 6$, and $\mathbf{d}(x, x') = 4$.

wt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
x'	0	0	0	0	0	0	0	1	1	1	1	2	2	0	0

489 sites whose states we have not decided yet and set the states of x' to be one of the $a - 2$ states that is
 490 different from both x and wt. The number of choices is $\binom{d-s}{d'-2s}(a-2)^{d'-2s}$

491 Putting these together we have

$$N(d, d') = \sum_{s=0}^{\min\{d, \lfloor \frac{d'}{2} \rfloor\}} \binom{d}{s} \binom{l-d}{s} (a-1)^s \binom{d-s}{d'-2s} (a-2)^{d'-2s}, \quad (\text{A48})$$

492 and so finally we have:

$$\frac{n}{n-1} E[V(d)] = \bar{\sigma}_d^2 + \sigma_{d,\mu}^2 - \bar{\sigma}_d \quad (\text{A49})$$

$$= a^{-l} \binom{l}{k} (a-1)^k - \frac{1}{\binom{l}{d}(a-1)^d - 1} \sum_{d'=1}^{\min\{l, 2d\}} N(d, d') w_k(d') \quad (\text{A50})$$

$$= w_k(0) - \frac{1}{\binom{l}{d}(a-1)^d - 1} \sum_{d'=1}^{\min\{l, 2d\}} N(d, d') w_k(d'). \quad (\text{A51})$$

493 It may be possible to further simplify this expression by plugging in $N(d, d')$ and $w_k(d')$ so that it
 494 is more intelligible. But calculation of $E[V(d)]$ based on this expression is computationally feasible
 495 since both $N(d, d')$ and $w_k(d)$ are easy to calculate.

496 We have derived the results above for energy landscapes with only k -th order interactions and by
 497 setting $\sigma_k^2 = 1$. For an energy landscape with all orders of interactions with Walsh coefficients b_{k_i} 's
 498 drawn i.i.d with mean 0 and variance σ_k^2 for each order k ,

$$E[V(d)] = \frac{1}{1 - \binom{l}{d}^{-1} (a-1)^{-d}} \sum_{k=0}^l \sigma_k^2 \left(w_k(0) - \frac{1}{\binom{l}{d}(a-1)^d - 1} \sum_{d'=1}^{\min\{l, 2d\}} N(d, d') w_k(d') \right). \quad (\text{A52})$$

499 Appendix C

500 Appendix C.1 Bivariate normal approximation

We approximate $(\Delta G_{\text{add}}, \Delta G_{\text{epi}})$ with a bivariate normal distribution with mean vector

$$(\mu_1, 0) := (l\mu_{\text{add}}, 0) \quad (\text{A53})$$

and covariance matrix

$$\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} := \begin{pmatrix} l^{\frac{\alpha-1}{\alpha}} \sigma_{\text{add}}^2 & 0 \\ 0 & \binom{l}{2} \frac{(\alpha-1)^2}{\alpha^2} \sigma_{\text{epi}}^2 \end{pmatrix}, \quad (\text{A54})$$

501 where the parameters were chosen according to the procedure described in Appendix A. Therefore, in
 502 this approximation, the total folding energy $\Delta G = \Delta G_{\text{add}} + \Delta G_{\text{epi}}$ is also normally distributed, with
 503 mean $\mu := \mu_1$ and variance $\sigma^2 := \sigma_1^2 + \sigma_2^2$. Using this normal approximation, we give an analytical
 504 justification for the phenomenon observed in Figure 2, that although the effect of epistasis is small, it is
 505 nonetheless crucial for folding. The analytical formulas we obtain for describing this phenomenon
 506 will also specify the regime of model parameters for which we can expect to see the phenomenon. We

507 shall use two quantities to measure the strength of this phenomenon. For the smallness of the epistatic
 508 effect, we use the measure σ_2^2/σ^2 , the fraction of the variance across all sequences accounted for by
 509 the variance of the epistatic energy. For the importance of the epistatic effect, we use the measure
 510 $E(\Delta G_{\text{add}}|\Delta G < 0)$, the mean of additive energies of viable sequences. If this mean is far above the
 511 viability threshold 0 it indicates that on average epistasis makes a substantial contribution to the ability
 512 of viable sequences to fold.

Now, we analytically approximate the conditional expectation $E(\Delta G_{\text{add}}|\Delta G < 0)$. We use a classical result that is referred to as the regression towards the mean formula for a pair of normally distributed random variables. If (X, Y) has normal distribution with mean (μ_X, μ_Y) and covariance matrix

$$\begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix},$$

then the regression towards the mean formula describes how the means change if we condition on one of the variables being below some cutoff value c :

$$E(X|Y < c) - E(X) = \rho_{XY} \frac{\sigma_X}{\sigma_Y} (E(Y|Y < c) - E(Y)). \quad (\text{A55})$$

Applying this formula to ΔG and ΔG_{add} and the condition that a sequence is viable, i.e. $\Delta G < 0$, we obtain

$$E(\Delta G_{\text{add}}|\Delta G < 0) - \mu_1 = R^2 (E(\Delta G|\Delta G < 0) - \mu), \quad (\text{A56})$$

where

$$R^2 := \frac{\text{Cov}(\Delta G_{\text{add}}, \Delta G)}{\sigma_1 \sigma} \frac{\sigma_1}{\sigma} = \frac{\sigma_1^2}{\sigma^2}$$

is calculated using the fact that $\Delta G = \Delta G_{\text{add}} + \Delta G_{\text{epi}}$ and that the additive and epistatic energies are uncorrelated by (A54). Also by (A54), $\mu = \mu_1$, hence we can express the mean additive folding energy of viable sequences from (A56) as

$$E(\Delta G_{\text{add}}|\Delta G < 0) = (1 - R^2) \mu_1 + R^2 E(\Delta G|\Delta G < 0). \quad (\text{A57})$$

The conditional mean on the right hand side of the equation above can be calculated as

$$E(\Delta G|\Delta G < 0) = \mu - \sigma \frac{\psi\left(-\frac{\mu}{\sigma}\right)}{\Psi\left(-\frac{\mu}{\sigma}\right)}, \quad (\text{A58})$$

513 where ψ and Ψ are the PDF and CDF, respectively, of the standard normal distribution, and as μ
 514 becomes large compared to σ , $E(\Delta G|\Delta G < 0)$ approaches 0.

Returning to (A57), we obtain the estimate

$$E(\Delta G_{\text{add}}|\Delta G < 0) \approx (1 - R^2) \mu_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \mu_1. \quad (\text{A59})$$

515 This means that no matter how small the epistatic effect is, measured by $\sigma_2^2/(\sigma_1^2 + \sigma_2^2)$, if the mean
 516 of the additive energy μ_1 is large enough in comparison, the role of the epistatic energy is crucial for
 517 protein folding.

Plugging in our model parameters as given in (A53) and (A54), and using the approximations $\alpha - 1 \approx \alpha$ and $l - 1 \approx l$, the estimate in (A59) becomes

$$\frac{l^2 \mu_{\text{add}} \sigma_{\text{epi}}^2}{l \sigma_{\text{epi}}^2 + 2 \sigma_{\text{add}}^2}.$$

518 The choice of parameters $\mu_{\text{add}} = 1$, $\sigma_{\text{add}} = 1$, and $\sigma_{\text{epi}} = 0.0003$ then yield 22.59, which is very close to
519 22.45, the mean additive energy of sequences observed in the simulation.

520

- 521 1. Sikosek, T.; Chan, H.S. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The*
522 *Royal Society Interface* **2014**, *11*, 20140419.
- 523 2. Bastolla, U.; Dehouck, Y.; Echave, J. What evolution tells us about protein physics, and protein physics
524 tells us about evolution. *Current opinion in structural biology* **2017**, *42*, 59–66.
- 525 3. Starr, T.N.; Thornton, J.W. Epistasis in protein evolution. *Protein Science* **2016**, *25*, 1204–1218.
- 526 4. Echave, J.; Wilke, C.O. Biophysical models of protein evolution: understanding the patterns of evolutionary
527 sequence divergence. *Annual review of biophysics* **2017**, *46*, 85–103.
- 528 5. Storz, J.F. Compensatory mutations and epistasis for protein function. *Current opinion in structural biology*
529 **2018**, *50*, 18–25.
- 530 6. Li, H.; Tang, C.; Wingreen, N.S. Are protein folds atypical? *Proceedings of the National Academy of Sciences*
531 **1998**, *95*, 4987–4990.
- 532 7. Tokuriki, N.; Tawfik, D.S. Stability effects of mutations and protein evolvability. *Current opinion in structural*
533 *biology* **2009**, *19*, 596–604.
- 534 8. Taverna, D.M.; Goldstein, R.A. Why are proteins marginally stable? *Proteins: Structure, Function, and*
535 *Bioinformatics* **2002**, *46*, 105–109.
- 536 9. Shakhnovich, E.I. Protein design: a perspective from simple tractable models. *Folding and Design* **1998**,
537 *3*, R45–R58.
- 538 10. Govindarajan, S.; Goldstein, R.A. On the thermodynamic hypothesis of protein folding. *Proceedings of the*
539 *National Academy of Sciences* **1998**, *95*, 5545–5549.
- 540 11. Goldstein, R.A. The evolution and evolutionary consequences of marginal thermostability in proteins.
541 *Proteins: Structure, Function, and Bioinformatics* **2011**, *79*, 1396–1407.
- 542 12. Serohijos, A.W.; Shakhnovich, E.I. Merging molecular mechanism and evolution: theory and computation
543 at the interface of biophysics and evolutionary population genetics. *Current opinion in structural biology*
544 **2014**, *26*, 84–91.
- 545 13. Goldstein, R.A.; Pollock, D.D. Sequence entropy of folding and the absolute rate of amino acid substitutions.
546 *Nature ecology & evolution* **2017**, *1*, 1923.
- 547 14. Ashenberg, O.; Gong, L.I.; Bloom, J.D. Mutational effects on stability are largely conserved during protein
548 evolution. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 21071–21076.
- 549 15. Pollock, D.D.; Thiltgen, G.; Goldstein, R.A. Amino acid coevolution induces an evolutionary Stokes shift.
550 *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E1352–9.
- 551 16. Shah, P.; McCandlish, D.M.; Plotkin, J.B. Contingency and entrenchment in protein evolution
552 under purifying selection. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, E3226–E3235,
553 [<http://www.pnas.org/content/112/25/E3226.full.pdf>]. doi:10.1073/pnas.1412933112.
- 554 17. Breen, M.S.; Kemena, C.; Vlasov, P.K.; Notredame, C.; Kondrashov, F.A. Epistasis as the primary factor in
555 molecular evolution. *Nature* **2012**, *490*, 535–538.
- 556 18. Wells, J.A. Additivity of mutational effects in proteins. *Biochemistry* **1990**, *29*, 8509–8517.
- 557 19. Risso, V.A.; Manssour-Triedo, F.; Delgado-Delgado, A.; Arco, R.; Barroso-delJesus, A.; Ingles-Prieto,
558 A.; Godoy-Ruiz, R.; Gavira, J.A.; Gaucher, E.A.; Ibarra-Molero, B.; Sanchez-Ruiz, J.M. Mutational
559 Studies on Resurrected Ancestral Proteins Reveal Conservation of Site-Specific Amino Acid
560 Preferences throughout Evolutionary History. *Mol. Biol. Evol.* **2015**, *32*, 440–455,
561 [<http://mbe.oxfordjournals.org/content/32/2/440.full.pdf+html>]. doi:10.1093/molbev/msu312.
- 562 20. Stadler, P.F.; Happel, R. Random field models for fitness landscapes. *Journal of Mathematical Biology* **1999**,
563 *38*, 435–478. doi:10.1007/s002850050156.
- 564 21. Neidhart, J.; Szendro, I.G.; Krug, J. Exact results for amplitude spectra of fitness landscapes. *Journal of*
565 *Theoretical Biology* **2013**, *332*, 218–227.
- 566 22. Aita, T.; Uchiyama, H.; Inaoka, T.; Nakajima, M.; Kokubo, T.; Husimi, Y. Analysis of a local fitness
567 landscape with a model of the rough Mt. Fuji-type landscape: Application to prolyl endopeptidase and
568 thermolysin. *Biopolymers* **2000**, *54*, 64–79.

- 569 23. Neidhart, J.; Szendro, I.G.; Krug, J. Adaptation in tunably rugged fitness landscapes: the rough Mount Fuji
570 model. *Genetics* **2014**, *198*, 699–721.
- 571 24. Kingman, J. A simple model for the balance between selection and mutation. *Journal of Applied Probability*
572 **1978**, *15*, 1–12.
- 573 25. Guerois, R.; Nielsen, J.E.; Serrano, L. Predicting changes in the stability of proteins and protein complexes:
574 a study of more than 1000 mutations. *Journal of molecular biology* **2002**, *320*, 369–387.
- 575 26. Miyazawa, S.; Jernigan, R.L. Residue–residue potentials with a favorable contact pair term and an
576 unfavorable high packing density term, for simulation and threading. *Journal of molecular biology* **1996**,
577 *256*, 623–644.
- 578 27. Chan, H.; Bornberg-Bauer, E. Perspectives on protein evolution from simple exact models. *Applied*
579 *bioinformatics* **2002**, *1*, 121–144.
- 580 28. Tokuriki, N.; Stricher, F.; Schymkowitz, J.; Serrano, L.; Tawfik, D.S. The stability effects of protein mutations
581 appear to be universally distributed. *Journal of molecular biology* **2007**, *369*, 1318–1332.
- 582 29. Zeldovich, K.B.; Chen, P.; Shakhnovich, E.I. Protein stability imposes limits on organism complexity and
583 speed of molecular evolution. *Proceedings of the National Academy of Sciences* **2007**, *104*, 16152–16157.
- 584 30. Wylie, C.S.; Shakhnovich, E.I. A biophysical protein folding model accounts for most mutational fitness
585 effects in viruses. *Proceedings of the National Academy of Sciences* **2011**, *108*, 9916–9921.
- 586 31. Miyazawa, S. Selection maintaining protein stability at equilibrium. *Journal of theoretical biology* **2016**,
587 *391*, 21–34.
- 588 32. Bloom, J.D.; Raval, A.; Wilke, C.O. Thermodynamics of Neutral Protein Evolution. *Genetics* **2007**, *175*, 255 –
589 266, [<http://www.genetics.org/content/175/1/255.full.pdf+html>]. doi:10.1534/genetics.106.061754.
- 590 33. Serrano, L.; Day, A.G.; Fersht, A.R. Step-wise mutation of barnase to binase: a procedure for engineering
591 increased stability of proteins and an experimental analysis of the evolution of protein stability. *Journal of*
592 *molecular biology* **1993**, *233*, 305–312.
- 593 34. Serohijos, A.W.; Shakhnovich, E.I. Contribution of selection for protein folding stability in shaping the
594 patterns of polymorphisms in coding regions. *Molecular biology and evolution* **2013**, *31*, 165–176.
- 595 35. van Nimwegen, E.; Crutchfield, J.P.; Huynen, M. Neutral evolution of mutational robustness. *Proc. Natl.*
596 *Acad. Sci. U. S. A.* **1999**, *96*, 9716 – 9720.
- 597 36. Galton, F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of*
598 *Great Britain and Ireland* **1886**, *15*, 246–263.
- 599 37. Lynch, M.; Walsh, B.; others. *Genetics and analysis of quantitative traits*; Vol. 1, Sinauer Sunderland, MA,
600 1998.
- 601 38. Kimura, M.; Crow, J.F. Effect of overall phenotypic selection on genetic change at individual loci. *Proceedings*
602 *of the National Academy of Sciences* **1978**, *75*, 6168–6171.
- 603 39. Draghi, J.A.; Plotkin, J.B. Selection biases the prevalence and type of epistasis along adaptive trajectories.
604 *Evolution* **2013**, *67*, 3120–3131.
- 605 40. Greene, D.; Crona, K. The changing geometry of a fitness landscape along an adaptive walk. *PLoS*
606 *computational biology* **2014**, *10*, e1003520.
- 607 41. Blanquart, F.; Achaz, G.; Bataillon, T.; Tenaillon, O. Properties of selected mutations and genotypic
608 landscapes under Fisher’s geometric model. *Evolution* **2014**, *68*, 3537–3554.
- 609 42. Drummond, D.A.; Wilke, C.O. Mistranslation-induced protein misfolding as a dominant constraint on
610 coding-sequence evolution. *Cell* **2008**, *134*, 341–352.
- 611 43. Gavrillets, S.; Gravner, J. Percolation on the fitness hypercube and the evolution of reproductive isolation. *J.*
612 *Theor. Biol.* **1997**, *184*, 51 – 64.
- 613 44. Wunderlich, Z.; Mirny, L.A. Different gene regulation strategies revealed by analysis of binding motifs.
614 *Trends in genetics* **2009**, *25*, 434–440.
- 615 45. Socolich, M.; Lockless, S.W.; Russ, W.P.; Lee, H.; Gardner, K.H.; Ranganathan, R. Evolutionary information
616 for specifying a protein fold. *Nature* **2005**, *437*, 512.
- 617 46. Vassilev, V.K.; Fogarty, T.C.; Miller, J.F. Smoothness, ruggedness and neutrality of fitness landscapes: from
618 theory to application. In *Advances in evolutionary computing*; Springer, 2003; pp. 3–44.
- 619 47. Stadler, P.F. Landscapes and their correlation functions. *Journal of Mathematical Chemistry* **1996**, *20*, 1–45.
- 620 48. Van Lint, J.H. *Introduction to coding theory*; Vol. 86, Springer Science & Business Media, 2012.

621 49. Bradley, R.S. Estimation of bias and variance of measurements made from tomography scans. *Measurement*
622 *Science and Technology* **2016**, *27*. doi:10.1088/0957-0233/27/9/095402.

623 © 2018 by the authors. Submitted to *Genes* for possible open access publication under the terms and conditions
624 of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).