Supplementary Tables:

Table S1 Pearson correlations between site-specific substitution rates and structural properties.

	1qhw			2ppn			1pek		
	#sig	Mean r (std)	P-value	#sig	Mean r (std)	P-value	#sig	Mean r (std)	P-value
RSA vs post mean ω^h									
Real protein	Yes	0.394	1.3e-12	No	0.065	0.508	Yes	0.526	4.3e-21
C-SI	0	0.002 (0.057)	0.503	0	0.011 (0.097)	0.461	1	-0.008 (0.065)	0.469
S-SI	50	0.274 (0.039)	0.000	40	0.313 (0.069)	0.014	43	0.222 (0.052)	0.011
S-SD	47	0.232 (0.044)	0.001	21	0.253 (0.077)	0.042	47	0.260 (0.055)	0.001
WCN vs post mean ω^h									
Real protein	Yes	-0.352	3.5e-10	No	0.031	0.750	Yes	-0.431	6.2e-14
C-SI	0	0.005 (0.054)	0.522	0	-0.009 (0.095)	0.476	0	0.008 (0.059)	0.519
S-SI	50	-0.250 (0.043)	0.000	24	-0.255 (0.065)	0.030	39	-0.196 (0.049)	0.012
S-SD	49	-0.244 (0.044)	0.001	16	-0.225 (0.076)	0.074	38	-0.211 (0.062)	0.025
RSA vs dN^h/dS^h									
C-SI	1	0.006 (0.057)	0.506	0	-0.003 (0.098)	0.471	1	0.011 (0.053)	0.578
S-SI	50	0.407 (0.023)	0.000	50	0.395 (0.037)	0.000	43	0.220 (0.044)	0.003
S-SD	50	0.437 (0.021)	0.000	50	0.458 (0.035)	0.000	50	0.340 (0.043)	0.000
WCN vs dN^h/dS^h									
C-SI	0	0.006 (0.059)	0.485	0	0.007 (0.095)	0.482	1	-0.017 (0.062)	0.483
S-SI	50	-0.373 (0.019)	0.000	46	-0.322 (0.043)	0.002	32	-0.184 (0.047)	0.019
S-SD	50	-0.422 (0.023)	0.000	50	-0.408 (0.033)	0.000	47	-0.262 (0.049)	0.001

sig: Number of trials out of 50 which showed a significant correlation
Mean r (std): Pearson correlation coefficient average over the 50 trials (standard deviation)
P-value: The P-value averaged over the 50 trials
C-SI: C-series site-independence generative model
S-SI: Stability site-independence generative model
S-SD: Stability site-dependence generative model
RSA: Relative solvent accessibility
WCN: Weighed contact number
dN^h/dS^h: expected site-specific substitution rates
post mean ω^h: inferred based on the best fitting M-series model

Table S2 Likelihood ratio test summaries for natural protein alignments.

	1QHW		**2PPN**		**1PEK**	
	sig	MLE	sig	MLE	sig	MLE
M0	--	ω_1 = 0.077	--	ω_1 = 0.025	--	ω_1 = 0.061
M0 vs **M3(k=2)**	yes	ω_1 = 0.007 ω_2 = 0.298 p_1 = 0.715	yes	ω_1 = 0.003 ω_2 = 0.086 p_1 = 0.671	yes	ω_1 = 0.021 ω_2 = 0.235 p_1 = 0.644
M3(k=2) vs **M3(k=3)**	yes	ω_1 = 0.003 ω_2 = 0.187 ω_3 = 0.721 p_1 = 0.661 p_2 = 0.276	no	ω_1 = 0.003 ω_2 = 0.003 ω_3 = 0.086 p_1 = 0.271 p_2 = 0.400	yes	ω_1 = 0.001 ω_2 = 0.070 ω_3 = 0.373 p_1 = 0.344 p_2 = 0.445
M3(k=2) vs **CLM3**	yes	ω_1 = 0.001 ω_2 = 0.516 p_1 = 0.76 δ = 0.15	no	ω_1 = 0.002 ω_2 = 0.146 p_1 = 0.755 δ = 0.003	yes	ω_1 = 0.032 ω_2 = 50.0 p_1 = 0.979 δ = 0.125
BUSTED(ω_3 <1) vs **BUSTED**	no		no		yes	ω_1 = 0.006 ω_2 = 0.703 ω_3 = 50.0 p_1 = 0.811 p_2 = 0.162

Note: The reported maximum likelihood estimates are based on the bolded model for each the model contrasts.

Table S3 Likelihood ratio test summaries from the fifty alignments generated under the specified generative model (C-SI, S-SI, or S-SD) with protein-specific simulation parameters (1QHW, 2PPN, or 1PEK).

	1QHW		**2PPN**		**1PEK**	
	# sig	Mean MLE sig trials (all trials)	# sig	Mean MLE sig trials (all trials)	# sig	Mean MLE sig trials (all trials)
C-SI **M0**	--	ω_1 = 0.607 (0.607)	--	ω_1 = 0.649 (0.649)	--	ω_1 = 0.516 (0.516)
M0 vs **M3(k=2)**	50	ω_1 = 0.363 (0.363) ω_2 = 1.104 (1.104) p_1 = 0.581 (0.581)	50	ω_1 = 0.316 (0.316) ω_2 = 0.958 (0.958) p_1 = 0.482 (0.482)	50	ω_1 = 0.228 (0.228) ω_2 = 0.652 (0.652) p_1 = 0.447 (0.447)
M3(k=2) vs **M3(k=3)**	6	ω_1 = 0.158 (0.203) ω_2 = 0.664 (0.719) ω_3 = 2.158 (1.770) p_1 = 0.218 (0.286) p_2 = 0.710 (0.553)	1	ω_1 = 0.355 (0.230) ω_2 = 0.920 (0.666) ω_3 = 3.946 (1.385) p_1 = 0.416 (0.297) p_2 = 0.548 (0.416)	28	ω_1 = 0.121 (0.145) ω_2 = 0.407 (0.442) ω_3 = 1.119 (2.278) p_1 = 0.161 (0.211) p_2 = 0.630 (0.588)
M3(k=2) vs **CLM3**	7	ω_1 = 0.162 (0.268) ω_2 = 1.023 (0.983) p_1 = 0.413 (0.440) δ = 0.385 (0.176)	17	ω_1 = 0.205 (0.321) ω_2 = 1.088 (5.458) p_1 = 0.421 (0.579) δ = 0.215 (0.267)	33	ω_1 = 0.219 (0.232) ω_2 = 2.284 (2.743) p_1 = 0.446 (0.449) δ = 0.196 (0.159)
BUSTED(ω_3 <1) vs **BUSTED**	0		0		0	
S-SI **M0**	--	ω_1 = 0.178 (0.178)	--	ω_1 = 0.086 (0.086)	--	ω_1 = 0.066 (0.066)
M0 vs **M3(k=2)**	50	ω_1 = 0.040 (0.040) ω_2 = 0.455 (0.455) p_1 = 0.617 (0.617)	50	ω_1 = 0.008 (0.008) ω_2 = 0.224 (0.224) p_1 = 0.626 (0.626)	50	ω_1 = 0.006 (0.006) ω_2 = 0.136 (0.136) p_1 = 0.477 (0.477)
M3(k=2) vs **M3(k=3)**	21	ω_1 = 0.011 (0.017) ω_2 = 0.230 (0.238) ω_3 = 0.825 (0.773) p_1 = 0.405 (0.439) p_2 = 0.447 (0.387)	7	ω_1 = 0.002 (0.003) ω_2 = 0.142 (0.115) ω_3 = 0.656 (0.384) p_1 = 0.516 (0.499) p_2 = 0.398 (0.318)	39	ω_1 = 0.003 (0.003) ω_2 = 0.092 (0.083) ω_3 = 0.281 (0.252) p_1 = 0.409 (0.403) p_2 = 0.423 (0.382)
M3(k=2) vs **CLM3**	10	ω_1 = 0.026 (0.028) ω_2 = 0.480 (0.449) p_1 = 0.621 (0.589) δ = 0.105 (0.062)	14	ω_1 = 0.003 (0.006) ω_2 = 0.342 (0.290) p_1 = 0.680 (0.649) δ = 0.106 (0.046)	22	ω_1 = 0.002 (0.004) ω_2 = 0.179 (0.181) p_1 = 0.502 (0.489) δ = 0.050 (0.031)
BUSTED(ω_3 <1) vs **BUSTED**	0		0		0	
S-SD **M0**	--	ω_1 = 0.198 (0.198)	--	ω_1 = 0.173 (0.173)	--	ω_1 = 0.169 (0.169)
M0 vs **M3(k=2)**	50	ω_1 = 0.070 (0.070) ω_2 = 0.514 (0.514) p_1 = 0.654 (0.654)	50	ω_1 = 0.043 (0.043) ω_2 = 0.325 (0.325) p_1 = 0.536 (0.536)	50	ω_1 = 0.065 (0.065) ω_2 = 0.273 (0.273) p_1 = 0.429 (0.429)
M3(k=2) vs **M3(k=3)**	15	ω_1 = 0.031 (0.028) ω_2 = 0.283 (0.225) ω_3 = 1.024 (0.794) p_1 = 0.416 (0.361) p_2 = 0.463 (0.455)	16	ω_1 = 0.006 (0.015) ω_2 = 0.160 (0.171) ω_3 = 0.690 (0.534) p_1 = 0.278 (0.318) p_2 = 0.588 (0.487)	43	ω_1 = 0.022 (0.024) ω_2 = 0.151 (0.154) ω_3 = 0.406 (2.349) p_1 = 0.170 (0.178) p_2 = 0.564 (0.562)
M3(k=2) vs **CLM3**	25	ω_1 = 0.048 (0.052) ω_2 = 0.550 (0.520) p_1 = 0.644 (0.635) δ = 0.216 (0.148)	35	ω_1 = 0.019 (0.024) ω_2 = 0.450 (0.424) p_1 = 0.594 (0.587) δ = 0.234 (0.182)	50	ω_1 = 0.033 (0.033) ω_2 = 0.314 (0.314) p_1 = 0.363 (0.363) δ = 0.106 (0.106)
BUSTED(ω_3 <1) vs **BUSTED**	0		1	ω_1 = 0.000 ω_2 = 0.371 ω_3 = 34.413 p_1 = 0.431 p_2 = 0.562	0	

Note: The reported maximum likelihood estimates are based on the bolded model for each the model contrasts.

Table S4 Comparison of alignments simulated under C-series site-independence (C-SI), stability-informed site-independence (S-SI), and stability-informed site-dependence (S-SD) models. Reported are the Bonferroni corrected P-values. The distribution of $\omega_1$and ω_2 values are estimated from M3 (k=2).

	1QHW	**2PPN**	**1PEK**
	P-Value	P-Value	P-Value
$\boldsymbol{\omega_1}$			
C-SI v S-SI	1.027802e-52	1.096157e-32	1.243729e-59
C-SI v S-SD	1.527041e-48	2.101497e-28	5.969164e-45
S-SI vs S-SD	3.032326e-18	2.843385e-20	2.174462e-43
$\boldsymbol{\omega_2}$			
C-SI v S-SI	1.111643e-40	2.137810e-32	3.091145e-62
C-SI v S-SD	1.873076e-36	8.048157e-27	2.739571e-48
S-SI vs S-SD	5.490179e-05	6.074310e-11	3.890470e-43
Proportion of conserved sites (1 amino acid)			
C-SI v S-SI	1.971864e-79	2.983453e-85	3.033707e-80
C-SI v S-SD	4.872843e-79	4.481974e-53	1.415823e-57
S-SI vs S-SD	1.024121e-22	2.867718e-47	1.389450e-68
Proportion of sites with $\geq$ 5 amino acids			
C-SI v S-SI	3.633243e-54	1.923767e-56	9.462538e-101
C-SI v S-SD	8.800839e-46	1.000679e-40	9.806678e-62
S-SI vs S-SD	7.711144e-14	1.206305e-20	1.167348e-68

Table S5 Correlation coefficients between RSA/WCN and two measures of epistasis sensitivity, std(dN^h/dS^h) or std(log(dN^h/dS^h)).

	1QHW r (p-value)	2PPN r (p-value)	1PEK r (p-value)
RSA vs std(dN^h/dS^h)	0.34 (1.00e-09)	0.39 (3.84e-05)	0.32 (4.7e-08)
WCN vs std(dN^h/dS^h)	-0.38 (1.03e-11)	-0.41 (7.64e-06)	-0.22 (2.1e-04)
RSA vs std(log(dN^h/dS^h))	-0.20 (0.0004)	-0.36 (0.0001)	-0.21 (0.003)
WCN vs std(log(dN^h/dS^h))	0.20 (0.0006)	0.30 (0.002)	0.15 (0.01)

Table S6 Pearson correlations between expected site-specific substitution rates (dN^h/dS^h) and inferred site-specific rates (post mean ω^h).

	1qhw			2ppn			1pek		
	#sig	Mean r (std)	P-value	#sig	Mean r (std)	P-value	#sig	Mean r (std)	P-value
dN^h/dS^h vs post mean ω^h									
C-SI	49	0.237 (0.044)	0.001	28	0.296 (0.117)	0.046	49	0.394 (0.078)	0.001
S-SI	50	0.681 (0.038)	0.000	50	0.749 (0.043)	0.000	50	0.791 (0.040)	0.000
S-SD	50	0.669 (0.035)	0.000	50	0.698 (0.046)	0.000	50	0.736 (0.049)	0.000

sig: Number of trials out of 50 which showed a significant correlation
Mean r (std): Pearson correlation coefficient average over the 50 trials (standard deviation)
P-value: The P-value averaged over the 50 trials
C-SI: C-series site-independence generative model
S-SI: Stability site-independence generative model
S-SD: Stability site-dependence generative model
dN^h/dS^h: expected site-specific substitution rates
post mean ω^h: inferred based on the best fitting M-series model

Table S7 NCBI Accession numbers for DNA sequences used for the three natural protein alignments (1QHW, 2PPN, 1PEK).

1QHW	2PPN	1PEK
NP_001075457.1	M80199.1	XM_003713956.1
ELK28734.1	U09386.1	XM_016786789.1
DAA35014.1	NM_204330.1	AM412313.1
ELR59971.1	NM_001252190.3	X14688.1
XP_020749961.1	AF483488.1	XM_018327475.1
DAA35015.1	BT021075.1	XM_006967830.1
NM_001256558.1	KY474593.1	EF362571.1
XP_021499432.1	KY474591.1	AF104385.1
XM_022526904.1	KY474590.1	AJ427459.1
XM_008048505.2	BC059689.1	HM635906.2
M76110.1	BT075719.1	XM_014693831.1
CR457078.1	NM_001139669.1	M73795.1
NM_001284443.1	NM_001103022.1	
XP_005078573.1	BT082974.1	

Table S8 Algorithm for finding sequence with high fitness values.

1. Start at random amino acid sequence S
2. Calculate the site-specific fitness landscape at all sites $\langle \mathrm{F}^1(S), \ldots, \mathrm{F}^n(S) \rangle$
3. If a single step uphill move is possible, then randomly choose the next substitution from the set of single amino acid changes that will increase fitness
4. If no uphill move is possible (i.e. local maximum), then randomly choose 20 sites and substitute them to the fittest amino acid at that site
5. Repeat 2-4 until fitness is greater than 0.99

Table S9 PDB codes for protein structures used to calculate average free energy, $\bar{E}(S)$, and standard deviation $\Delta E^2(S)$, for a sequence in the unfolded configurations.

1cnz	1gyh	1jix	1m4l	1nsz	1pby	1t5j	1wer
1dmh	1hz4	1jj2	1mkf	1o4s	1pfk	1t5o	1wkr
1e19	1i4w	1jkm	1moq	1o7j	1qo0	1to6	1woh
1ek6	1iom	1jl5	1mty	1o88	1qop	1uby	2bbv
1esd	1ir6	1jub	1n00	1oc7	1rkd	1umd	2mas
1ga6	1jfb	1kwf	1nbf	1odm	1sbp	1v6s	3sil
1gwu	1jil	1l5o	1nd6	1ojj	1svm	1wch	

Supplementary Figures:

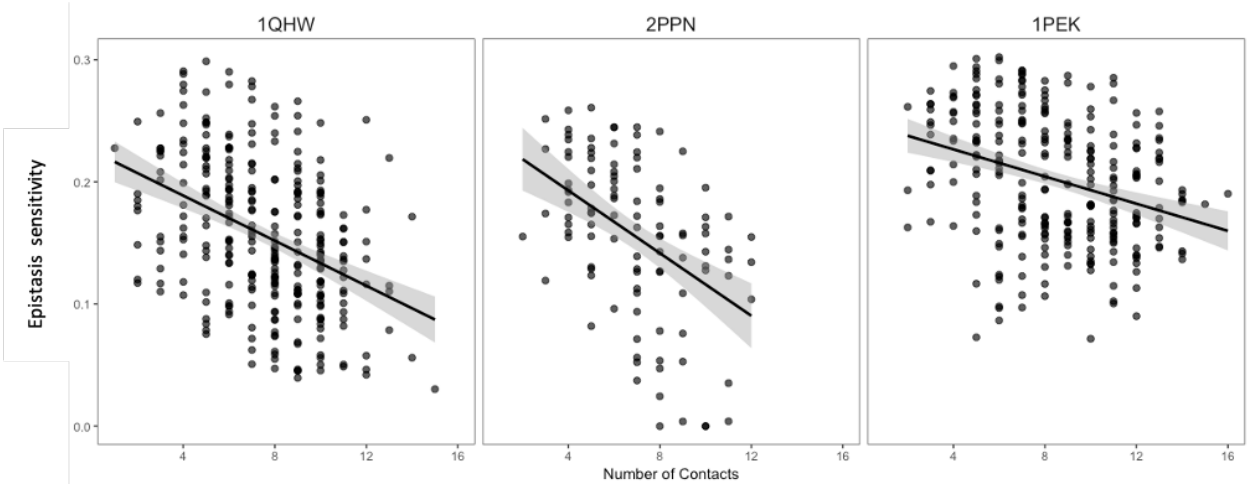


Figure S1. Relationship between epistatic sensitivity and number of contacts. Epistasis sensitivity is measured as the standard deviation in expected rates (dN^h/dS^h) across all 50xN background sequences, where N is the number of taxa = 14, 14, and 12 for the 1QHW, 2PPN, and 1PEK proteins respectively. The lines represent a linear regression and the shaded area the 95% confidence interval.

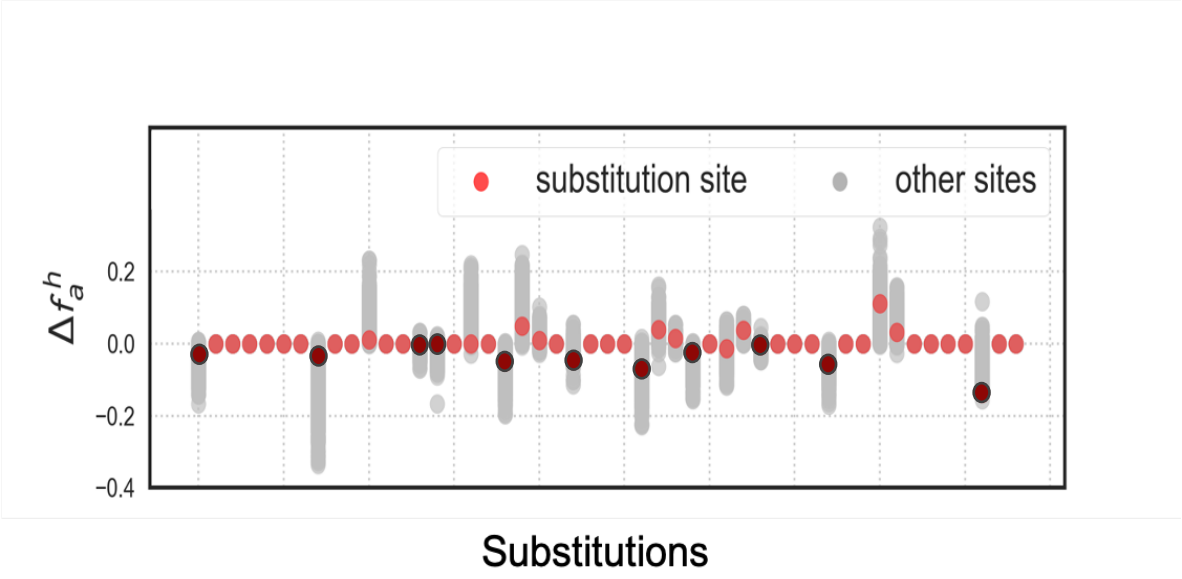
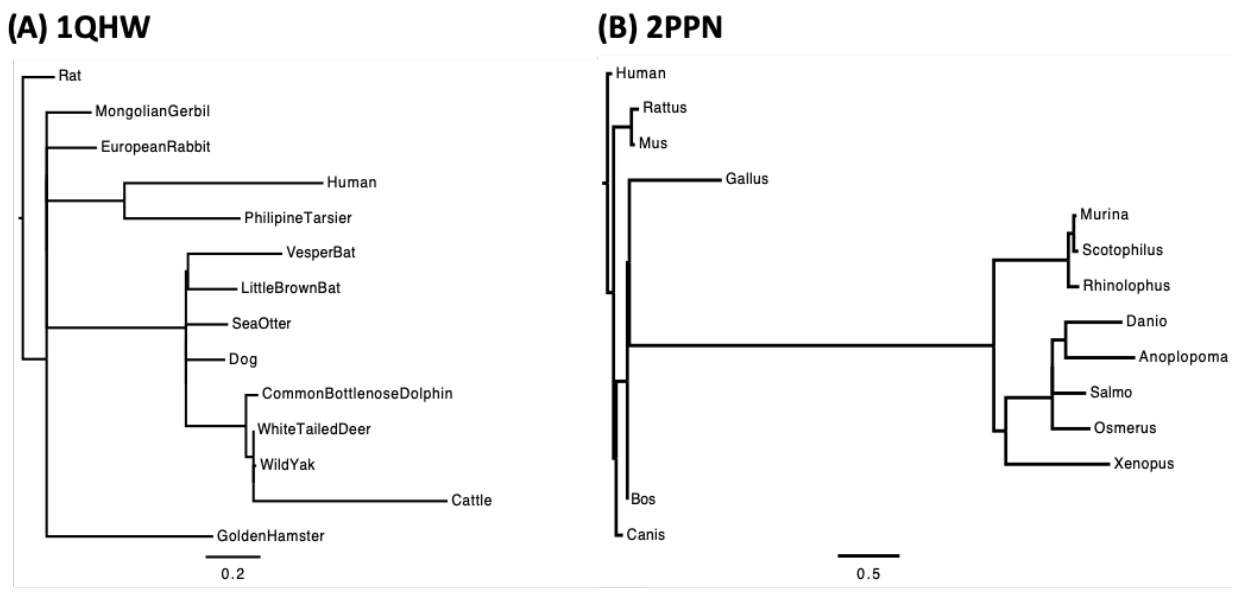


Figure S2. Epistatic dynamics in the evolution of the 1QHW protein. Changes in the fitness of the resident amino acids (Δf_a^h) at all (300) sites in the 1QHW protein. Red dots represent the change in fitness at the substitution site, and therefore a change in the resident amino acid. Dark red dots represent a deleterious substitution. Grey dots represent the change in the fitness of the resident amino acid at all other (299) sites.



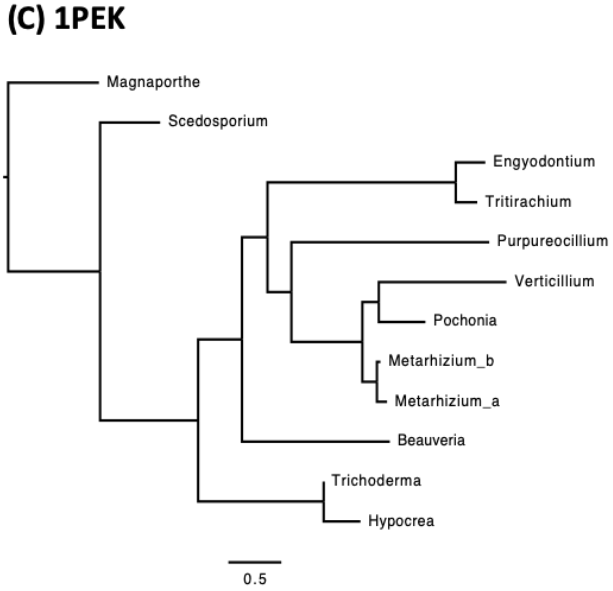


Figure S3 Phylogenetic trees for the three (1QHW, 2PPN, 1PEK) natural protein alignments analysed in this study. The topologies were inferred using IQ-Tree. Branch lengths, measured as the expected number of single nucleotide substitution per codon site, were inferred from codon model M3($k = 3$).

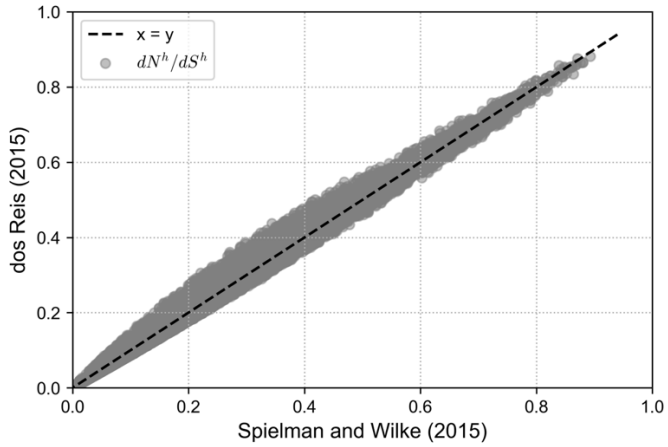


Figure S4 Correlation between the different ways of calculating site-specific substitution rates, dN^h/dS^h. The dos Reis (2015) formulation measures the nonsynonymous mutation rate in reference to the neutral stationary frequencies $\pi^{(0)}$ based on mutational biases only. The Spielman and Wilke (2015) formulation measures the nonsynonymous mutation rate in reference to the site-specific stationary frequencies π^h in the presence of selection pressure and mutational biases. See Jones *et al.* 2017 for further discussion on the interpretation of both formulations. The Pearson correlation coefficient between both formulations was 0.99, P-value = 0.00.

Supplementary Discussion:

Accessing sample size

Given the enormity of sequence space, it is unclear that any sampling, no matter how extensive, could characterize the entire fitness landscape. Since the evolution of natural proteins billions of years ago, even natural proteins have not adequately sampled their respective sequence space and are evolving on a small, localized portion of sequence space. To understand how epistasis influences protein evolution concerning rates of substitution, we made sure to consider comparisons between rate estimates from the S-SD and S-SI models in the same "local-neighbourhood" of sequence space. This avoided the difficulty of comparing behaviour in different regions of sequence space.

Specifically, to calculate the expected rate at a site, we approximate the rate as the average over the extant sequences observed in each S-SD simulated alignment. The extant sequences provide a sample of the "local neighbourhood". To address the robustness of our results to a more extensive sampling of sequences in the local space, we compared the expected rate dN_i^h/dS_i^h considering all extant sequences from each alignment i (as described in equation 5 in main text) to the rate dN_{ij}^h/dS_{ij}^h calculated by leaving out the jth sequence. We calculated the bias and mean squared error (MSE) as

$$bias_i^h = \frac{1}{N}\sum_{j=1}^{N} dN_{ij}^h/dS_{ij}^h - dN_i^h/dS_i^h \quad \text{(S1)}$$

$$MSE_i^h = \frac{1}{N}\sum_{j=1}^{N}(dN_{ij}^h/dS_{ij}^h - dN_i^h/dS_i^h)^2 \quad \text{(S2)}$$

where N is the number of taxa per alignment. The average bias and average MSE in the site-specific rates over all trials are plotted in figure S5, where $bias^h = \frac{1}{50}\sum_{i=1}^{50} bias_i^h$ and $MSE^h = \frac{1}{50}\sum_{i=1}^{50} MSE_i^h$ are the average bias and MSE per rate at a site over all simulated alignments.

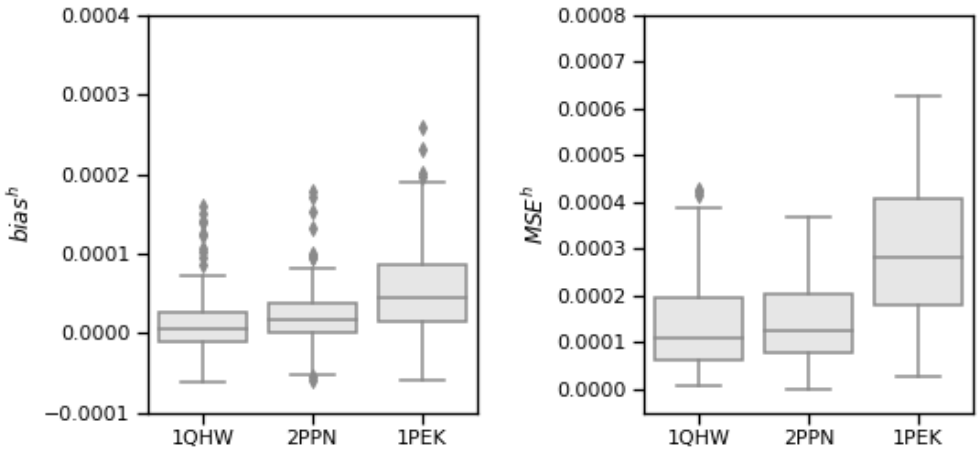


Figure S5. Assessing the average bias and average mean squared error in expected site-specific rates (dN^h/dS^h).

The distributions of the average bias and average MSE, for all three proteins, suggest that calculating expected rates based on the extant sequences does not systematically bias estimates, and has little impact on the expected rate values. However, note that the bias and MSE are slightly higher for the 1PEK simulations. This is likely due to two reasons: (1) the 1PEK tree is deeper than the 1QHW and 2PPN trees (tree length = 13.88, 4.93, 8.04 for 1PEK, 1QHW, and 2PPN) which means that the "local neighbourhood" is larger for the 1PEK simulations, and (2) the number of taxa in the 1PEK alignment (n= 12) is smaller than the number of sequences in the 1QHW and 2PPN alignments (n = 14). The larger "local neighbourhood" in conjunction with the smaller sampling likely lead to the increase in bias and MSE observed. Importantly, however, the bias and MSE are nonetheless very small which suggests that calculating rates as the average over the extant sequences has minimal consequences on rate expectations.