PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Selection pressures on evolution of ribonuclease H explored with rigorous free–energy–based design

Ryan L. Hayes[a,b,1], Charlotte F. Nixon[c], Susan Marqusee[c,d,e], and Charles L. Brooks III[b,f,1]

Edited by José Onuchic, Rice University, Houston, TX; received July 14, 2023; accepted November 22, 2023

Understanding natural protein evolution and designing novel proteins are motivating interest in development of high-throughput methods to explore large sequence spaces. In this work, we demonstrate the application of multisite λ dynamics (MSλD), a rigorous free energy simulation method, and chemical denaturation experiments to quantify evolutionary selection pressure from sequence–stability relationships and to address questions of design. This study examines a mesophilic phylogenetic clade of ribonuclease H (RNase H), furthering its extensive characterization in earlier studies, focusing on *E. coli* RNase H (ecRNH) and a more stable consensus sequence (AncCcons) differing at 15 positions. The stabilities of 32,768 chimeras between these two sequences were computed using the MSλD framework. The most stable and least stable chimeras were predicted and tested along with several other sequences, revealing a designed chimera with approximately the same stability increase as AncCcons, but requiring only half the mutations. Comparing the computed stabilities with experiment for 12 sequences reveals a Pearson correlation of 0.86 and root mean squared error of 1.18 kcal/mol, an unprecedented level of accuracy well beyond less rigorous computational design methods. We then quantified selection pressure using a simple evolutionary model in which sequences are selected according to the Boltzmann factor of their stability. Selection temperatures from 110 to 168 K are estimated in three ways by comparing experimental and computational results to evolutionary models. These estimates indicate selection pressure is high, which has implications for evolutionary dynamics and for the accuracy required for design, and suggests accurate high-throughput computational methods like MSλD may enable more effective protein design.

protein folding | consensus sequence | free energy | selection pressure | ribonuclease H

Significance

Novel computational methods for protein design are combined with experiment to explore the application of rigorous free energy calculations using multisite λ dynamics to a sequence space of 32,768 sequences representing the sequences "between" *E. coli* ribonuclease H (RNase H) and the consensus sequence of an ancestral reconstruction of an RNase H clade. Good agreement between computed and measured stabilities across representatives of this space suggest that multisite λ dynamics can play an important role in protein design and understanding of protein evolutionary landscapes.

There is a growing interest in exploring the large protein sequence spaces that arise from many mutations using high-throughput techniques. These studies are useful for understanding evolutionary history through ancestral sequence reconstruction (1–3), improved protein design (4), knowledge of the distribution of stability effects of mutations (5–8), and design of new functions (9, 10).

The large sequence spaces encountered in ancestral sequence reconstructions and their corresponding extant protein families are useful both for protein design, because mutations have already been screened by evolution (11–13), and for studying protein evolution, because sequences reveal evolutionary pathways and selection pressures (1, 14). Consensus sequences, which include the most common mutations in extant members of a protein family, are a common means of protein design from multiple sequence alignments. Consensus sequences are usually more stable than extant sequences and often even functional (13), but sometimes are less stable (15). Whether this loss in stability is due to a few deleterious mutations, epistatic coupling between mutations (12), or other effects, and how to restore stability are all questions that can be addressed with high-throughput methods.

Studies of proteins resurrected from ancestral sequence reconstructions can also provide valuable insight into evolutionary processes. Such studies not only identify evolutionary pathways but also inform the mechanism by which epistatic coupling between mutations can block other pathways (16, 17). In this work, we focus on bacterial ribonuclease H1 (RNase H) because it is a well-characterized system for which many proteins in the reconstructed ancestral phylogeny have been studied. RNase H studies have identified selective pressures present during evolution: In thermophiles, thermostability is selected, but the mechanism of stabilization fluctuates between residual structure and core contacts (1), and while kinetic stability is initially low, it increases as evolution progresses (14). While these estimates of selection pressure are qualitative, simple quantitative estimates of selection pressure have been made using models derived from multiple sequence

Author affiliations: [a]Department of Chemical and Biomolecular Engineering, University of California, Irvine, CA 92697; [b]Department of Chemistry, University of Michigan, Ann Arbor, MI 48109; [c]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720; [d]California Institute for Quantitative Biosciences, University of California, Berkeley, CA 94720; [e]Department of Chemistry, University of California, Berkeley, CA 94720; and [f]Biophysics Program, University of Michigan, Ann Arbor, MI 48109

Author contributions: R.L.H., C.F.N., S.M., and C.L.B. designed research; R.L.H. and C.F.N. performed research; R.L.H., C.F.N., and S.M. contributed new reagents/analytic tools; R.L.H., C.F.N., S.M., and C.L.B. analyzed data; and R.L.H., C.F.N., S.M., and C.L.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: rhayes1@uci.edu or brookscl@umich.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2312029121/-/DCSupplemental.

Published January 9, 2024.

alignments of protein families (18, 19). These estimates assume the probability of observing a particular sequence is Boltzmann distributed according to its folding free energy

$$P(S) \propto \exp(-\Delta G(S)/kT_s), \quad [1]$$

and the temperature T_s of the distribution quantifies selection pressure for stability; lower selection temperatures imply more stringent selection (20).

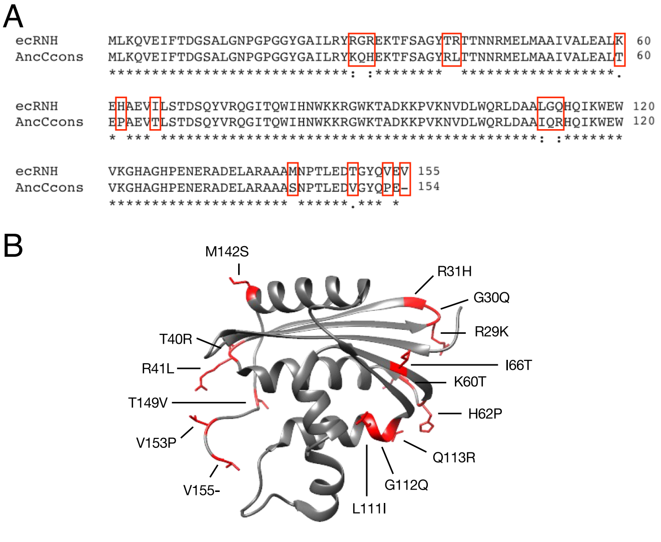
While experimental methods exist and continue to be developed to characterize the large sequence spaces from ancestral sequence reconstruction, design libraries, and random mutagenesis (2, 3, 5–8, 10), the ability to explore these spaces computationally provides a complement useful in informing and directing both the design of new proteins (4, 9) and studies of natural proteins (21, 22). There are many computational tools for predicting the effect of protein mutations and for protein design. For computational efficiency, protein design tools typically estimate the effect of a mutation using several approximations. These approximations include using a single or a few structures rather than a full ensemble and using more approximate interaction potentials. A widely employed design tool is Rosetta (23), which has been used effectively for many different design tasks (24–27) and to predict protein folding or binding free energies (28–31). Several other design methods exist based on similar principles (32–35), and entirely different methods based on deep learning are beginning to appear (36–38). In addition, other tools, such as FoldX (39), are employed to predict mutational folding free energy changes with reasonable accuracy utilizing methods based on physical interactions, statistical observations from databases, or machine learning (28). All of these techniques have shown utility in protein design problems, but make significant approximations in their estimates of changes in free energy due to introduced mutations.

In contrast to these tools, alchemical free energy methods make rigorous free energy estimates using more accurate interaction potentials and full molecular ensembles obtained from molecular dynamics simulations. The $\Delta\Delta G$ for a slow physical process like protein folding is computed by evaluating the difference in free energy of a chemical transformation such as mutation in two physical ensembles. This removes the need for long simulations that sample the slow physical transition between ensembles. These methods have been demonstrated to possess the necessary accuracy to be useful during lead optimization in computer-aided drug design (40–42) and have also shown promise in initial studies of protein folding and binding free energies (43–49).

There are a wide variety of alchemical free energy methods, including free energy perturbation (50), thermodynamic integration (51), nonequilibrium methods (52), enveloping distribution sampling (53, 54), and multisite λ dynamics (MSλD) (55, 56). Most alchemical methods require many simulations to compare two sequences, typically differing by a point mutation, and are not sufficiently scalable to address large protein sequence spaces. MSλD, however, requires only a single simulation to compare combinatorial sequence spaces arising from permutations of many mutations and thus is uniquely well suited for high-throughput studies of large protein sequence spaces and for protein design. MSλD compares sequences in this combinatorial space by sampling chemical degrees of freedom in addition to the spatial degrees of freedom typically sampled by molecular dynamics. Since its inception (55), MSλD has undergone a renaissance that includes technical developments like the generalization from single site to multisite λ dynamics and the introduction of implicit constraints (56, 57), biasing potential replica exchange and adaptive landscape flattening to improve sampling (58, 59), and soft-core interactions and PME electrostatics to improve the accuracy and robustness of results (49, 59, 60). These studies have set the stage for application of MSλD to large protein sequence spaces.

Earlier studies have shown MSλD has comparable accuracy to other alchemical methods (43–45, 47, 49) and comparable or higher accuracy than the approximate approaches used by Rosetta (30, 61) and can accurately predict the effect of up to five simultaneous mutations (49). While these results were notable, obstacles in sequence substitutions of proline, glycine, and charge perturbations, as well as obstacles in sampling large sequence spaces needed to be addressed. Perturbations between *E. coli* RNase H (ecRNH) and the ancestor C clade consensus sequence of RNase H (AncCcons) comprise 15 concurrent mutations, including proline, glycine, and charge-changing mutations (Fig. 1) (15), and spurred development of methods to overcome these obstacles (62, 63). Determining sequence–stability relationships and designing proteins with improved stability within this space of 15 mutations is a challenging task; Fig. 1 shows all mutations are surface mutations, over half are in loops, most are conservative, and all are likely to have small effects because they are essentially evolutionary noise.

The purpose of this study is twofold: First, we focus on determining the magnitude of this evolutionary noise to quantify selection pressure rather than on explaining the small effects of individual mutations; second, we seek to optimize these conservative mutations to design for stability. Because of the small magnitude of effects of individual mutations, it is likely that otherwise successful design methods may fail at this task, even though the 2.1-kcal/mol stability difference between ecRNH and AncCcons reveals there are stability gains to be made by optimizing surface residues. While such gains are modest compared to the stability gains sometimes observed with protein design (67), realizing them would signify the ability to fine tune natural or designed proteins, and optimizing conservative surface mutations is desirable in some contexts (68).



```
ecRNH      MLKQVEIFTDGSALGNPGPGGYGAILRYRGREKTFSAGYTRTTNNRMELMAAIVALEALK  60
AncCcons   MLKQVEIFTDGSALGNPGPGGYGAILRYKQHEKTFSAGYRLTTNNRMELMAAIVALEALT  60
           ****************************: :********  ******************.

ecRNH      EHAEVILSTDSQYVRQGITQWIHNWKKRGWKTADKKPVKNVDLWQRLDAALGQHQIKWEW 120
AncCcons   EPAEVTLSTDSQYVRQGITQWIHNWKKRGWKTADKKPVKNVDLWQRLDAAIQRHQIKWEW 120
           * *** ********************************************: :*******

ecRNH      VKGHAGHPENERADELARAAAMNPTLEDTGYQVEV 155
AncCcons   VKGHAGHPENERADELARAAASNPTLEDVGYQPE- 154
           ********************* ******.*** *
```

Fig. 1. (*A*) Sequence alignment of ecRNH and AncCcons with fully conserved residues denoted with asterisks, strongly similar residues denoted with colons, and weakly similar residues denoted with periods (64). The 15 mutations between the sequences are highlighted with red boxes. (*B*) Chimera rendering of ecRNH (PDB entry 2RN2) with the 15 mutation side chain residues labeled and shown as red ball and stick (65, 66).

Within this sequence space of modest mutations, we identified and tested the most and least stable chimeras to assess MSλD sequence design capacity, along with several other mutants to assess MSλD convergence and Rosetta design capacity. The designed chimera utilized half as many mutations to achieve roughly the same stability improvement over ecRNH as AncCcons. The agreement of MSλD with experiment for 12 sequences was excellent, with a Pearson correlation of 0.86 and a root mean squared error (RMSE) of 1.18 kcal/mol. As a control, the leading protein design algorithm Rosetta struggled with these tasks, designing a chimera with only a quarter of the stability improvement of MSλD, and showing poor correlation with experiment. The selection temperature is estimated globally for the RNase H family as 140 to 168 K by finding the energy scale of a dimensionless fitness score derived from multiple sequence alignments through comparison with experiment or with MSλD. The selection temperature is also estimated locally between the two sequences as 110 to 146 K from the SD predicted by MSλD for the 15 selected point mutations. These estimates indicate relatively stringent selection compared to the accuracy of conventional stability prediction methods and the distribution of possible mutational effects (7), but corroborate previous estimates for selection temperature by other methods (18, 19). The ability of MSλD to predict the effect of up to 15 mutations with kcal/mol level accuracy and provide insight into evolutionary selection pressures in large sequence spaces will be useful in future studies of protein evolution, biophysics, and design.

Results

In order to explore the ability to design in a large sequence space and gain insight into RNase H evolution, MSλD was used to predict the stability of chimeras between ecRNH (sequence 1) and AncCcons (sequence 2), a 2.13-kcal/mol more stable sequence differing by 15 mutations (see Table 1 and Fig. 1 for sequence identities). There are 2^{15}, or 32,768, chimeras for all permutations of mutations, and this is a two orders of magnitude increase over the largest sequence space of 240 sequences (49) or chemical space of 512 ligands (69) previously studied with MSλD. Consequently, preliminary simulations were needed to determine the amount of sampling required to obtain robust results. From 5 independent trials of 100 ns of free energy sampling with MSλD, we identified the most stable single mutants (sequences 3 to 4) and the least stable single mutant (sequence 5), as well as the most stable chimera (sequence 6) and the least stable chimera (sequence 7). These sequences were expressed and experimentally tested for stability, and the results were used to assess convergence for varying levels of sampling with MSλD.

Results from the five independent 100-ns simulations indicated poor convergence due to both statistical noise and insufficient relaxation of the protein structure, as running more independent trials or longer trials improved agreement for sequences 1 to 7 (Fig. 2*A*). While the Pearson correlation of 0.78 between prediction and experiment was satisfactory, the RMSE of 2.05 kcal/mol was larger than that expected from previous studies (49). Sampling these large chemical spaces is facilitated by an adaptive procedure that identifies biases that permit the full space to be explored as efficiently as possible (49, 59, 63). Thus, we reoptimized the biases to better facilitate sampling and ran another 5 trials of 100 ns, which resulted in rather different, but not significantly improved, results due to statistical noise. Running another 15 independent 100-ns trials for both the first and second sets of biases substantially reduced statistical noise, and averaging the results of both sets of 20 trials of 100 ns together gave a Pearson correlation of 0.76 with experiment and an RMSE of 1.42 kcal/mol. Finally, 12 simulations of 400 ns each were run and yielded a Pearson correlation of 0.72 and RMSE of 1.03 kcal/mol for sequences 1 to 7. While the 12 × 400-ns simulations comprised roughly the same amount of sampling as the 40 × 100-ns simulations, the longer duration of the simulations allowed the system, especially the flexible C terminus, to relax more fully, resulting in more accurate predictions.

Having determined that 12 × 400 ns was a sufficient amount of sampling, new predictions for the most stable chimera (sequence 8) and least stable chimera (sequence 9) were made. Experimentally characterizing these sequences revealed significantly more stabilization and destabilization than observed in the previously predicted sequences (6 and 7). Furthermore, sequence 8 was as stable as sequence 2, to within experimental uncertainty, meaning that our methodology enabled design of a sequence as stable as the most stable known sequence, but differing by 8 mutations. Including these sequences and the Rosetta sequences described in the next paragraph, MSλD agreement with experiment improved to a Pearson correlation of 0.86 and an RMSE of 1.18 kcal/mol (Fig. 2*B*).

Since MSλD enabled effective design within this sequence space, for comparison we examined the ability of Rosetta, the leading computational protein design package, to design in this space. Because of its speed, Rosetta can readily handle much

Table 1. Relative folding free energies of various sequences at 25 °C

#	Sequence (S)[*]	Experiment $\Delta\Delta G(S)$[†]	MSλD 5 × 100 ns $\Delta\Delta G(S)$[†]	MSλD 12 × 400 ns $\Delta\Delta G(S)$[†]	Rosetta $\Delta\Delta G(S)$[‡]
1	000000000000000	0.00 ± 0.40	0.00 ± 0.43	0.00 ± 0.44	0.00 ± 0.36
2	111111111111111	−2.13 ± 0.60	−2.60 ± 0.12	−0.97 ± 0.34	−6.91 ± 0.33
3	000010000000000	0.11 ± 0.13	−0.75 ± 0.21	−0.72 ± 0.47	−1.03 ± 0.49
4	000000100000000	−0.81 ± 0.95	−1.06 ± 0.33	−1.15 ± 0.34	0.49 ± 0.34
5	000000010000000	0.29 ± 0.27	1.12 ± 0.39	1.02 ± 0.41	3.69 ± 0.39
6	010110100111011	−1.37 ± 0.32	−5.11 ± 0.26	−2.87 ± 0.31	−2.54 ± 0.44
7	100001011001001	0.38 ± 0.53	4.11 ± 0.56	1.99 ± 0.30	−2.40 ± 0.34
8	000111100110100	−2.00 ± 0.40	−4.85 ± 0.26	−3.97 ± 0.30	−4.70 ± 0.50
9	110001011101101	1.20 ± 0.30	3.05 ± 0.68	3.55 ± 0.15	−8.63 ± 0.30
10	100000000000000	0.57 ± 0.29	2.22 ± 0.48	0.76 ± 0.48	−8.15 ± 0.45
11	110011000100100	1.16 ± 0.32	0.40 ± 0.32	1.27 ± 0.19	−17.36 ± 0.29
12	111101100101000	−0.47 ± 0.63	−1.35 ± 0.18	−0.52 ± 0.24	−11.22 ± 0.44

[*] Sequences indicate either ecRNH (0) or AncCcons (1) at each position listed in Fig. 1.
[†] Experimental and MSλD free energies are in kcal/mol.
[‡] Rosetta free energies are in Rosetta Energy Units, which are roughly 0.2 to 0.35 kcal/mol.

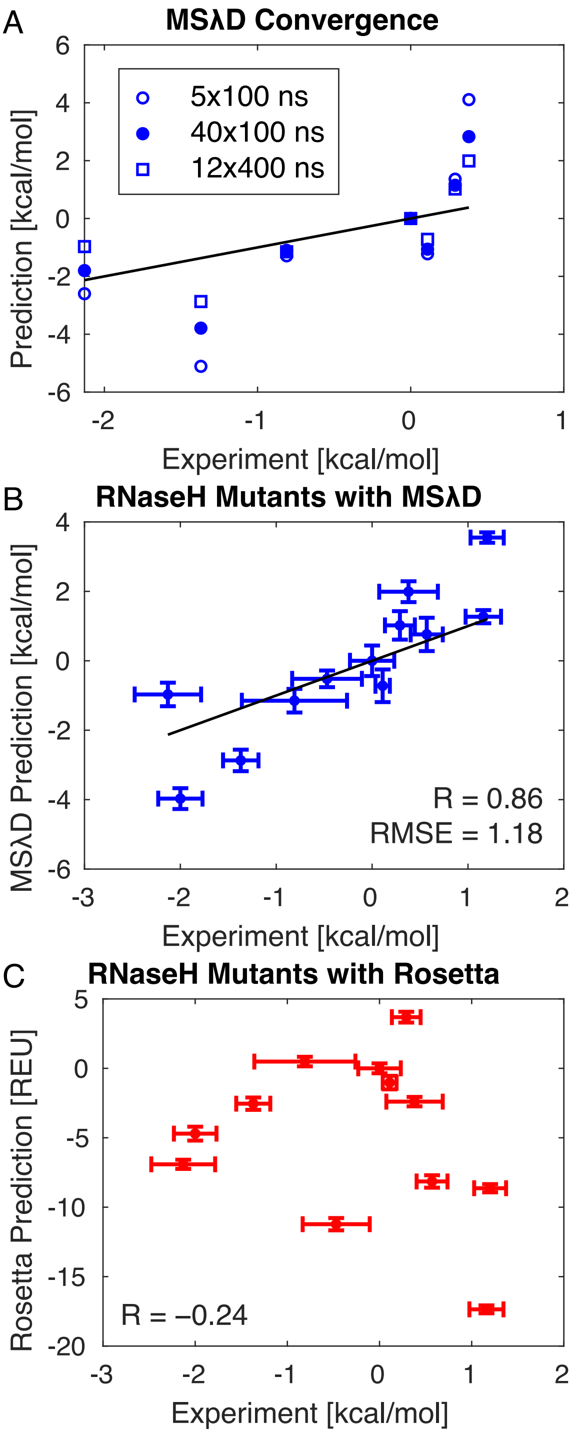


Fig. 2. (*A*) Convergence of MSλD results on the first 7 sequences with increased sampling. Every sequence except sequence 2 approaches the y = x line with increased sampling and longer simulations. (*B*) Agreement of MSλD predictions with experiment for all 12 sequences. The solid line is y = x. (*C*) Agreement of Rosetta predictions with experiment for all 12 sequences.

larger sequence spaces, but for the sake of comparison, it was limited to this sequence space. Rosetta $\Delta\Delta G$ calculations were used to identify the most stable single mutant (sequence 10) and most stable chimera (sequence 11). Notably, half of the stabilization in sequence 11 is attributed to the single mutation in sequence 10, underlining the importance of this mutation to Rosetta. While Rosetta can predict $\Delta\Delta G$, it is more optimized for designing sequences, so Rosetta was also used to design an optimized sequence (12). Experimentally, both the single mutant and the chimera identified by Rosetta $\Delta\Delta G$ calculations as stabilizing turned out to be destabilizing, leading to a negative Pearson correlation with experiment of −0.24 over the full dataset (Fig. 2*C*). The designed sequence 12 showed a small degree of stabilization relative to ecRNH, but this stabilization of 0.47 kcal/mol is only a quarter of the 2.00 kcal/mol stabilization of the MSλD chimera. Since the Rosetta reference energy is parameterized to reproduce natural amino acid abundances (30), and these abundances are affected by factors beyond stability (8), an evolutionary preference correction was computed from the data in ref. 8 (*SI Appendix*, Table S1), but adding this correction to the experimental stability did not improve the agreement between Rosetta and experiment.

The unprecedented accuracy of MSλD simulations provides an opportunity to evaluate selection pressure in different ways from previous studies. A dimensionless fitness score E_{Potts} can be obtained by fitting a Potts model of single site h_i and pairwise interaction J_{ij} terms to the probability distribution P of a multiple sequence alignment.

$$E_{\text{Potts}} = \sum_i h_i(S_i) + \sum_{i<j} J_{ij}(S_i, S_j) \qquad [2]$$

$$P(S) \propto \exp(E_{\text{Potts}}(S)). \qquad [3]$$

If desired, Eq. **3** may be viewed as a Boltzmann factor where the temperature has been absorbed into the fitness score because no energy scale is defined in the purely bioinformatic fitting procedure. Comparison with Eq. **1** reveals the dimensionless fitness score is equal to the stability $\Delta G(S)$ over $-kT_s$ up to an additive constant. Previous works have subsequently determined the selection temperature by comparing fitness scores to experimental stabilities for several sequences (18, 19), by comparing fitness score and coarse-grained force field predictions of the difference between natural and random sequences (18), and by examining the SD of random point mutation effects on fitness score (19). In this work, we again evaluate the selection temperature by comparing the fitness score with experimental results for our 12 sequences, but we also evaluate it in two additional ways utilizing MSλD predictions. First, the fitness score is compared with stability predictions over all 32,768 chimeras. Second, the selection pressure is estimated without reference to the fitness score by evaluating the predicted root mean square effect of the 15 evolutionarily selected mutations and comparing it to a simple model.

We first evaluate selection temperature by comparing the fitness score with the 12 experimental measurements (Fig. 3*A*). The slope of the best-fit line is equal to $-1/kT_s$, but determining the best-fit line depends on the independent variable and the sources of noise. In standard linear regression, the independent variable is placed on the x-axis, and vertical errors δE_{Potts} caused by noise on the y-axis are minimized, but if the x-axis is the dominant source of noise, it is more appropriate to minimize the horizontal errors $\delta\Delta G(S)$, and both lines are shown in Fig. 3A. The ratio between these two slopes (corresponding to T_s = 99 K and 140 K) is the square of the Pearson correlation (−0.840), which is comparable to the correlation of −0.84 previously observed between fitness score and experiment (18). In principle, the selection temperature is used to predict how much the dependent variable of fitness changes in response to the independent variable of stability. Furthermore, the experimental errors are small (around 0.3

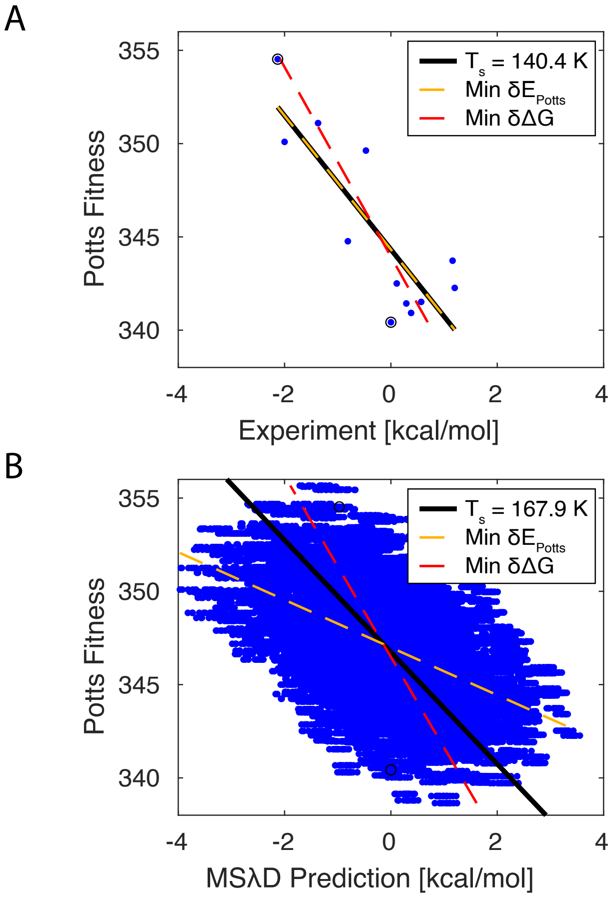


Fig. 3. Selection temperature from fitting the fitness score to (*A*) experimental measurements and (*B*) MSλD predictions. Best-fit lines minimizing y-axis noise (min $\delta E_{\mathrm{Potts}}$) and x-axis noise (min $\delta\Delta G$) are shown in yellow and red, respectively. Best-fit lines accounting for relative noise on both axes are shown in black with the corresponding selection temperature. Open black circles indicate ecRNH and AncCcons.

kcal/mol), so experimental measurements closely match the independent variable of stability. The primary source of noise is the fitness score, both from fitness effects beyond stability, such as function, and from approximations in fitted parameters and functional form of the fitness score. Consequently, the line minimizing vertical errors $\delta E_{\mathrm{Potts}}$ with a slope corresponding to T_s = 140 K is most appropriate.

Before using the computational results to estimate selection pressure, it is necessary to determine the statistics of the computational error, because this noise affects both computational estimates of selection temperature. The error statistics cannot be directly determined from experimental results, because the set of 12 measured sequences was biased by selection for sequences predicted by MSλD to be most or least stable. These sequences are statistically enriched in errors that bias them to appear more or less stable than experiment, respectively (see *SI Appendix* for discussion of overprediction). Instead, the degree of overprediction for the optimal sequences is used to estimate the statistical properties of the error. Sites are assumed to be independent, and both the true effects of point mutations as well as the computational errors are assumed to be Gaussian distributed with SDs of σ_1 and σ_2 (*Materials and Methods*). Simulations reveal that the approximation that sites are independent is reasonable (*SI Appendix*). Expectation values for the predicted and true stabilities of extremal sequences can be calculated analytically under these approximations (*SI Appendix*), and comparison with their measured values implies the SD of point mutations is σ_1 = 0.41 kcal/mol for their true effect and σ_2 = 0.48 kcal/mol for the computational error.

While these parameters were calculated from only the difference between sequences 8 and 9, they are consistent with other measured sequences. Evaluating the SD of the 15 point mutation predictions and SD of the four point mutation measurements (corrected for selection bias) on sequences 3, 4, 5, and 10 reveals the statistics of the data match analytical predictions to within 10%. The model also allows prediction of relevant quantities (see *SI Appendix* for derivations). First, the Pearson correlation of 0.86 observed between MSλD and the 12 experimentally measured sequences is biased relative to the space of all 32,768 sequences due to selection of sequences with large stability effects. The expected correlation with experiment if all sequences were measured is 0.652. Second, the expectation value for the stability of the true most stable sequence is predicted to be −2.80 kcal/mol, though the identity of the most stable sequence remains unknown. In light of this, the ability to design sequence 8 with a stability within 1 kcal/mol of the most stable sequence is quite remarkable.

With a statistical description of computational errors, one can estimate the selection temperature by comparing fitness scores and computational predictions for all 32,768 sequences (Fig. 3*B*). The computational errors in the stability predictions along the x-axis are significantly larger than the experimental errors in Fig. 3*A*, which lowers the Pearson correlation to −0.513 and suggests it is not appropriate to treat either axis as the independent variable in fitting. Instead, both the x-axis and the y-axis are correlated to the independent variable of true stability, and the best-fit line must weigh noise from both axes appropriately. Since switching independent and dependent variables changes the slope by a factor of R^2, the desired slope that minimizes the vertical errors in fitness as a function of true stability can be found by multiplying the slope minimizing the horizontal deviation of the true stability as a function of fitness by the square of the correlation between fitness and true stability. The Pearson correlation between the true stability and the computational prediction over all sequences is analytically predicted to be 0.652. The Pearson correlation (−0.513) between two dependent variables (fitness score and stability prediction) both independently correlated to an independent variable (true stability) is the product of their individual correlations, which suggests a correlation of −0.787 between fitness and the true stability. The slope minimizing horizontal errors of predicted stability as a function of fitness score is independent of noise on the x-axis and also minimizes the horizontal errors of the unknown true stabilities as a function of fitness score. Multiplying this slope of −4.84 $(\mathrm{kcal/mol})^{-1}$ by $(-0.787)^2$ gives the desired slope and corresponds to T_s = 168 K. Notably, this estimate of the selection temperature required quantification of the computational error, but only used it to estimate the correlation between fitness and true stability.

Finally, one can estimate the selection temperature independent of the fitness function by noting that the 15 point mutants studied are all tolerated by evolution, and the SD of selected mutations will be a function of the selection temperature. We determine this relationship for a simple idealized model with completely independent sites and a Gaussian distribution of stability changes available to evolution. *SI Appendix*, Fig. S1 reveals that as long as the SD of available mutations is above a threshold of roughly $1.25kT_s$, the SD of selected mutations between two sequences lies within roughly 15% of a limiting value. If both sequences are selected, this limit is $\sqrt{7/2}kT_s$, whereas if one sequence is selected and the other is the optimal sequence, this limit is $\sqrt{2}kT_s$. The RNase H system lies between these two limits because ecRNH was selected, while AncCcons is an average over the phylogenetic clade with fewer destabilizing mutations. Consequently, measurement of the root mean square effect of mutations results in a range of selection temperatures between these limits. The root mean square effect of the 15 point mutations in this space could be measured experimentally by testing 11 more sequences but would be complicated by the experimental uncertainty of roughly 0.3 kcal/mol that is comparable in magnitude to the small effect of the point mutations. The effect of point mutations is already available from MSλD as 0.61 kcal/mol but is similarly inflated by computational error. Using the fitted value of 0.41 kcal/mol for the SD of true point mutant effects implies a selection temperature range of 110 to 146 K.

Discussion

The results we have presented have biophysical implications for evolution. The selection temperature model, in which extant sequences are Boltzmann distributed by their stability (20), is a simplistic model but quantifies the selection temperature in RNase H as 110 to 168 K. How stringent these selection pressures are depends on how large the corresponding energies of 0.22 to 0.33 kcal/mol are relative to other relevant energies. The energetics of random mutations or of epistatic couplings between mutations that produce barriers between fitness optima (3, 17, 70, 71) are relevant because they define the fitness landscape within which thermal selection allows a protein to evolve. The SD of random mutations is 1.5 kcal/mol in one protein (7), and similar in others (19), which is much larger than the selection pressures we measured, indicating tight selection. In computational protein design, errors in stability predictions are relevant because they play an analogous role to selection temperature by driving the sequence away from fitness optima. The magnitude of these errors depends on the dataset and is obscured by unit differences but is comparable to the magnitude of random mutations, implying natural selection is much more stringent than computational design. This presents a protein design paradox because computational protein design should produce sequences more random and less stable than evolution, but studies have shown (67) computationally designed proteins are often more stable than evolved sequences.

The results presented here underline this paradox by corroborating previous estimates of the selection temperature (18, 19). One notable difference between the present study and previous studies is that this study uses naturally occurring mutations, while previous studies used random point mutations (19) or sequences (18). Random mutations are more likely to occupy regions of sequence space that are not robustly parameterized, leading to greater errors in the fitness score. One study estimated selection temperatures by fitting the fitness score and a coarse-grained energy function (18). Comparison between the fitness score and the coarse-grained model was performed by sampling sequences in the random and evolved basins. Estimates between 60 and 125 K were obtained for most families with an outlier at 220 K (18). Another study estimated the selection temperature from the ratio between the SDs of stability and fitness score for random single nucleotide point mutations (19). Selection temperatures ranged widely from 60 to 280 K, with a median value of 115 K. The SD of stability was calibrated in the PDZ protein family by mutational studies and assumed to be the same in all protein families. However, the SD of stability was also measured for three other families and varied by up to a factor of 1.6 below the assumed constant value, suggesting this is a fairly crude method and that these estimates are on the high side. While the wide variation in selection temperatures observed by these previous studies is surprising, and may be due to underdetermined parameters in the fitness scores, our study provides consistent estimates of selection temperature compared with these two published works.

In this study, the selection temperature is estimated both by fitting experimental results or MSλD predictions to a fitness score and by fitting MSλD predictions to a model of point mutation effects. These two estimates give global and local perspectives on selection pressure within a protein family. The fitness score is derived from a multiple sequence alignment, which requires a few thousand protein sequences to give robust results (72), and averages selection pressures over all sequences in the alignment to give a global estimate of selection pressure. In contrast, fitting to the simple model of mutations between a pair of sequences is local by nature, and the two sequences must be fairly close in sequence space; otherwise, the assumption that sites are independent is suspect. This method can give information about clade-specific differences in selection pressure and provides an alternative means of estimating selection pressure without the fitness score. Looking forward, MSλD calculations and fitness scores can estimate the selection temperature without knowledge of the computational error using the methods from Fig. 3*B* if the correlation between fitness score and true stability is known. Other hybrid approaches are possible: using fully random sequences as in ref. 18, but with MSλD, or using MSλD to verify the assumption in ref. 19 that all protein families have the same distribution of random mutation effects. Thus, the approaches we have presented together with previous methods form a versatile toolbox for quantifying selection pressure.

Protein fitness is a consequence of protein function, but many features modulate protein function. Properties like protein stability, solubility, and degradation indirectly affect function by modulating the amount of protein in functionally competent configurations. Some mutations directly affect function by modulating binding or catalysis, while others affect specificity against other toxic functions. In principle, there is selection pressure for each of these features, but mutations affecting stability are much more common than mutations affecting other features. Consequently, the selection temperature model approximates stability as the sole driver of evolution and other features as noise, which is justified by the strong correlation of roughly 0.8 between stability and fitness score. Fig. 3*B* shows sources of noise can affect the estimation of selection temperature and must be carefully quantified. Other selected features, especially direct modulation of function, may have opposite effects when estimating selection temperature from random mutations and evolutionarily selected mutations. On average, random mutations degrade both stability and other selected features, leading to a correlation between stability and other features that underestimates selection temperature. In contrast, selected mutations often embody a tradeoff between stability and other selected features (73) that overestimates selection temperature. This may explain why our estimates of selection temperature with selected mutations are slightly higher than previous studies using random mutations. However, the rough agreement between estimates shows this effect is small and confirms stability is the primary driver of fitness.

Quantification of the selection temperature suggests new lines of inquiry. First, further study of the selection temperature may help resolve the protein design paradox. The agreement between multiple estimates of the selection temperature strengthens this paradox, even as new datasets (8) become available to study evolutionary selection. Resolving this paradox may give deeper insights into protein design or protein evolution. These insights in turn may foreshadow the benefit that can be expected from more accurate design methods like MSλD or allow one to anticipate in advance whether a particular design method applied to a design target is likely to be successful. Second, the selection temperature quantifies how tightly evolution selects for stability, and in thermostable proteins, it might be expected that selection for stability is stronger. It is therefore an interesting question whether point mutations in the thermophilic T. thermophilus clade might suggest a lower selection temperature with tighter selection than the point mutations we examined in the mesophilic *E. coli* clade. Finally, the large variability in selection temperature over various protein families is surprising, and further quantification of selection temperature in these families may indicate if these differences are robust and provide insight into the molecular causes of differences in selection temperature.

The results presented here also have clear implications for the use of MSλD in determining sequence–stability relationships and exploring design questions in large sequence spaces. As mentioned in the introduction, the small magnitude of individual mutations made this a difficult design task, which was underlined by the results of our attempts to use Rosetta in this context. Rosetta has shown spectacular success designing useful proteins in less constricted sequence spaces (9, 27), so accuracy in optimizing surface residues may seem gratuitous, but such fine tuning is useful in many contexts, including optimizing antibody thermostability (68). Rosetta has previously been shown to yield results in good agreement with experiment for point mutations (with a Pearson correlation of 0.74) (30), and yielded a Pearson correlation of 0.66 for a set of 32 T4 lysozyme buried point mutants we previously studied (49) (*SI Appendix*, Fig. S2). Thus, the poor correlation found with Rosetta in this study likely arises because errors over the large number of concurrent surface mutations compound to overshadow their small effect. In contrast, MSλD was able to predict the effect of up to 15 mutations with kcal/mol level accuracy.

The accuracy observed in the MSλD calculations is impressive and unprecedented and comes at an increased computational cost. The 12 × 400-ns MSλD simulations took 27 d on 12 GTX1080 Ti GPUs, or 8,000 total GPU hours. The Rosetta designs took one to three orders of magnitude less time, depending on the sequence, and only required CPUs (*SI Appendix*). Future MSλD calculations will be more rapid, as a basic lambda dynamics engine (BLaDE) has been recently developed that brings the execution time for these calculations from 27 d to under 4 d (74). Together with hardware advances, BLaDE will enable the accuracy MSλD to be applied routinely to free energy and design calculations in large sequence spaces.

The increased accuracy of MSλD will allow such calculations to carve out a unique niche. Complementary use of MSλD and Rosetta is the most immediate application. Rosetta designs frequently fail, and success rates of 1% are not uncommon (4, 9). This is typically not a problem, as hundreds or thousands of designs are often proposed and tested, but is undesirable for time-consuming assays or ambitious design targets. A hierarchical design strategy, either drawing candidate mutations for MSλD simulations from Rosetta, or screening promising Rosetta designs for defects with MSλD, may raise success rates. In the longer term, MSλD has potential for exploring larger sequence spaces of engineering and biophysical interest. Comparably sized combinatorial experimental libraries of 60,000 to 160,000 sequences have already provided profound insights into evolution and design (3, 70, 75), and larger sequence spaces are well within the reach of MSλD since the computational cost for a desired precision scales like the square of the number of sites considered (63). Furthermore, the ability of MSλD to predict the effect of 15 simultaneous mutations with kcal/mol accuracy is highly relevant for biocatalyst design, where final optimized designs are often 30 to 40 mutations from the starting sequence (76). Looking forward, computational protein design is being revolutionized by deep learning approaches (36–38, 77). Deep learning approaches typically design a single stable structure or interface, and physics-based approaches (78) like MSλD can play a complementary role to deep learning by designing for protein dynamics or kinetics around that structure, designing for or against multiple structures, and by designing for unusual ligands or non-natural amino acids where training data is sparse.

Conclusions

In this work, we applied MSλD predictively to design ribonuclease H variants and to understand the effects of selection on natural sequence variability. These methods are broadly applicable to many protein systems and the evolutionary insights are general. We successfully designed a chimera roughly as stable as AncCcons utilizing only half as many mutations. We obtained unprecedented agreement with experiment for large jumps in sequence space of up to 15 mutations, obtaining a Pearson correlation of 0.86 and an RMSE of 1.18 kcal/mol. The space of 32,768 sequences explored in this work is two orders of magnitude larger than any space previously explored with MSλD and, together with the treatment of prolines and glycines and the high accuracy of our results, opens new frontiers for MSλD to explore large sequence spaces in protein biophysics and design. Finally, our estimate of the selection temperature serves to corroborate previous estimates of the selection temperature and opens new questions about evolution within these large sequence spaces.

Materials and Methods

AncCcons Sequence Design. AncCcons is a designed RNase H sequence consisting of the most frequent amino acid at each position of the multiple sequence alignment from the extant sequences in the AncC clade of the RNase H family (1).

RNase H Expression and Purification. Chimera RNase H gBlock gene fragments were purchased from Integrated DNA Technologies (IDT) and restriction cloned into the pET-27b(+) expression vector. Site-directed mutagenesis was used to generate single-point mutants in the ecRNH background. Sequences were confirmed by Sanger sequencing. RNase H expression and purification were performed as described (79), and the purity and mass were confirmed as a single band by SDS/PAGE and mass spectrometry.

Circular Dichroism Spectroscopy. Circular dichroism (CD) experiments were performed using an Aviv 410 CD spectrometer. Urea melts were performed using samples containing 0.04 mg/mL protein and various urea concentrations in 20 mM sodium acetate and 50 mM potassium chloride, pH 5.5. The samples were equilibrated overnight, and the signal at 222 nm was averaged over one minute with stirring in a 1-cm path length Starna Cells cuvette at 25 °C. Urea melts were obtained in triplicate, and the data were fitted to a two-state model with a linear free-energy extrapolation (80).

MSλD Calculations. MSλD simulations calculate $\Delta\Delta G$ by taking the difference between the ΔG of a mutation in the final ensemble (the folded state), and the ΔG of that same mutation in the initial ensemble (the unfolded state). MSλD simulations of the folded state were performed starting from the crystal structure PDBID: 2RN2 (65) with all three cysteines mutated to alanine. For the folded ensemble, all 15 mutations were simulated simultaneously. The unfolded ensemble was broken up into several short pentapeptides centered on each mutating residue, as done previously (49). In cases where multiple mutating residues would have been present in the same pentapeptide (e.g., R29K, G30Q, and R31H), all mutations were evaluated in a longer peptide fragment with two non-mutating residues on each end (e.g., R27 through K33). The effects of mutations in different peptide fragments were treated as additive.

Obtaining converged estimates of the free energy requires many transitions in alchemical sequence space facilitated by biasing potentials that remove barriers between alchemical endpoints. Adaptive landscape flattening was used to obtain the biasing potentials (59); the recently developed least squares approach (49) was augmented to flatten couplings between sites with intersite ψ, χ, and ω terms as described in ref. 63. After initial flattening runs, 5 parallel 5-ns simulations followed by 5 parallel 20-ns simulations were carried out to further refine the biases.

Previously, simulations included a copy of each side chain at a mutation site with interactions scaled by their respective λ, all bonded to a single copy of the backbone (49). This approach is not viable for glycine and proline mutations because the parameters and even connectivity of the backbone atoms change. An alternative perturbation strategy that includes a copy of each whole residue and breaks proline rings with soft bonds (81, 82) while rigorously avoiding the artifacts described in ref. 83 was designed specifically for this study and has been described in detail elsewhere (62).

Multiple alchemical free energy studies have shown that charge-changing mutations in proteins suffer from lower accuracy (45, 49). When simulations are treated with particle mesh Ewald (PME) electrostatics (60, 84, 85) and the net charge of the system before mutation is neutralized with counterions, an additive correction dominated by the discrete solvent term gives improved results independent of simulation box size (60, 86). Free energy estimates were corrected by the discrete solvent term, which was 0.688 kcal/mol in our simulations.

The sequence space explored in this study is two orders of magnitude larger than any space previously explored with MSλD (41, 49, 69) and required development of a new free energy estimator to analyze results (63). The Potts model estimator, which only includes single site terms h_i and pairwise coupling terms J_{ij}, was borrowed from direct coupling analysis (DCA), which uses it to predict protein contacts from multiple sequence alignments (87, 88). The Potts model is defined as

$$\Delta\Delta G(S) = \sum_i h_i(S_i) + \sum_{i<j} J_{ij}(S_i, S_j), \qquad [4]$$

where the model parameters h_i and J_{ij} are fit to the data, in this case, the λ trajectories, by likelihood optimization. The application of the estimator to MSλD is described in more detail elsewhere (63). While most couplings are small, the 6 couplings larger than 0.2 kcal/mol do make a substantial contribution to the free energy. Furthermore, the predicted couplings are precise; experimental measurement of the couplings is subject to larger errors.

Simulation Details. Simulations were run in the CHARMM molecular dynamics package (89, 90) using DOMDEC GPU acceleration (91) and the CHARMM 36 protein force field (92). Folded simulations were run in roughly 72 Å boxes, and unfolded peptides were run in roughly 40 Å boxes, which allowed a 10 Å margin between the solute and the box edge. Simulation conditions were designed to mimic the 50 mM KCl and 20 mM CH_3COONa and pH = 6.5 experimental conditions. The simulation volumes were solvated with TIP3P water (93) and solvated with 70 mM KCl. Each system was neutralized by adjusting K^+ and Cl^- ions by opposite amounts. Protonation states of titratable residues at pH = 6.5 were determined by using ProPKA (94) in the folded state and their reference pKa in the unfolded state. In the folded state, H114 was the only deprotonated histidine, and D10 and E48 were the only protonated aspartic and glutamic acids. The folded N and C termini were given standard charged caps (NTER and CTER), while peptide N and C termini were given neutral CH_3CO- and -$NHCH_3$ caps (ACE and CT3). Simulations used PME electrostatics with an interpolation order of 6, κ =0.32 Å^{-1}, a cutoff of 10 Å, and roughly one grid point per angstrom (60, 84, 85). Van der Waals interactions utilized force switching with a switch radius of 9 Å and a cutoff of 10 Å (95). Soft core interactions were used as previously described (59), including this time 1–4 nonbonded interactions as well (62).

Rosetta Calculations. Rosetta's cartesian_ddg tool was used to predict $\Delta\Delta G$, similar to previous work (30, 96). The reported free energy changes were the average of 20 independent cartesian_ddg calculations. For each independent calculation, the relax function was run 20 times, and the lowest energy pose was used as an input for cartesian_ddg. We found that running the relax function 20 times for each independent cartesian_ddg prediction was critical. Using the same optimal structure from relax for all 20 cartesian_ddg trials resulted in higher errors relative to experiment and underestimation of uncertainty. Both relax and cartesian_ddg used a cutoff of 9 Å, which has previously been found to improve agreement with experiment when used with the ref2015_cart scoring function. We note Rosetta reference energies are calibrated for 6 Å cutoffs, and using a longer cutoff with these reference energies will overstabilize larger amino acids. Since larger residues are generally less common than would be expected from stability alone (8), use of this longer cutoff likely partially compensates for the fact Rosetta is parameterized to reproduce natural amino acid frequencies rather than stabilities. Future work should give more attention to calibration of the Rosetta reference energy. Only 14 mutations were considered with Rosetta because the C terminal deletion V155- was harder to model; it was treated as V155 for all sequences. (MSλD simulations suggest V155- has a negligible effect). In order to identify the most stable chimera, the effect of all single mutants in the ecRNH background was computed, and the best additive combination of single mutants was chosen as a starting point. From this starting point, the most stabilizing single mutation was made iteratively until no stabilizing mutations remained, evaluating all point mutations in each new sequence background. A total of three mutations were made before convergence.

In addition to designing by computing $\Delta\Delta G$, Rosetta fastdesign was used to design an optimal sequence without explicit calculation of $\Delta\Delta G$. A script provided by Brian Kuhlman and Andrew Leaver-Fay was used to simultaneously optimize side chain identity and rotameric state in a relax-like procedure starting from the ecRNH crystal structure. The standard ref2015 scoring function and cutoff of 6 Å were used. The consensus sequence of 20 design attempts was taken as the optimal designed sequence, and only the R31H site showed any variation between attempts.

Fitness Score. The multiple sequence alignment for the RNase H family (PF00075) was accessed from the Pfam database on 1 February 2022 (97). Any of the 18,650 sequences with more than 20 deletions or more than 20 insertions relative to ecRNH were removed from the multiple sequence alignment, leaving 12,529 sequences. The fitness score was optimized using asymmetric pseudolikelihood maximization (72) and averaging the J_{ij} and J_{ji} coupling terms. The last four mutation sites explored in this study are not present in PF00075 and thus do not affect the fitness score.

Data, Materials, and Software Availability. Potts models used to generate MSλD results are included in the *SI Appendix*. MSλD setup and simulation scripts, Rosetta scripts, and selection temperature scripts are available at https://github.com/RyanLeeHayes/PublicationScripts/blob/main/2023RNaseH.tgz (98). Updated MSλD scripts are available in References 62 & 63, and updated MSλD ALF scripts are available at https://github.com/RyanLeeHayes/ALF (99).

ACKNOWLEDGMENTS. We gratefully acknowledge funding for R.L.H. and C.L.B. from the NIH (GM130587 and GM37554) and the NSF (CHE 1506273) and for C.F.N. and S.M. from the NIH (GM050945 and GM149319). Susan Marqusee is a Chan Zuckerberg Biohub Investigator. We thank Brandon Frenz for his guidance in the use of Rosetta's cartesian_ddg to calculate $\Delta\Delta G$ and design sequences 10 and 11, and Brian Kuhlman and Andrew Leaver-Fay for guidance and sharing protein design scripts used to design sequence 12.

1. K. M. Hart *et al*., Thermodynamic system drift in protein evolution. *PLoS Biol.* **12**, e1001994 (2014).
2. G. K. A. Hochberg, J. W. Thornton, Reconstructing ancient proteins to understand the causes of structure and function. *Annu. Rev. Biophys.* **46**, 247–269 (2017).
3. T. N. Starr, L. K. Picton, J. W. Thornton, Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409–413 (2017).
4. G. J. Rocklin *et al*., Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
5. C. L. Araya *et al*., A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16858–16863 (2012).
6. C. A. Olson, N. C. Wu, R. Sun, A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643–2651 (2014).
7. A. Nisthal, C. Y. Wang, M. L. Ary, S. L. Mayo, Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 16367–16377 (2019).
8. K. Tsuboyama *et al*., Mega-scale experimental analysis of protein folding stability in biology and protein design. *Nature* **620**, 434–444 (2023).
9. A. Chevalier *et al*., Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).
10. H. Renata, Z. J. Wang, F. H. Arnold, Expanding the enzyme universe: Accessing non-natural reactions by mechanism-guided directed evolution. *Angew. Chem. Int. Ed.* **54**, 3351–3367 (2015).
11. B. Steipe, B. Schiller, A. Plückthun, S. Steinbacher, Sequence statistics reliably predict stabilizing mutations in a protein domain. *J. Mol. Biol.* **240**, 188–192 (1994).
12. B. J. Sullivan *et al*., Stabilizing proteins from sequence statistics: The interplay of conservation and correlation in triosephosphate isomerase stability. *J. Mol. Biol.* **420**, 384–399 (2012).
13. M. Sternke, K. W. Tripp, D. Barrick, Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 11275–11284 (2019).
14. S. A. Lim, K. M. Hart, M. J. Harms, S. Marqusee, Evolutionary trend toward kinetic stability in the folding trajectory of RNases H. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 13045–13050 (2016).
15. C. Nixon *et al*., The importance of input sequence set to consensus-derived proteins and their relationship to reconstructed ancestral proteins. bioRxiv (2023). https://www.biorxiv.org/content/10.1101/2023.06.29.547063v1.full.pdf. Accessed 26 September 2023.
16. J. T. Bridgham, E. A. Ortlund, J. W. Thornton, An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461**, 515–519 (2009).
17. T. N. Starr, J. W. Thornton, Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).
18. F. Morcos, N. P. Schafer, R. R. Cheng, J. N. Onuchic, P. G. Wolynes, Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 12408–12413 (2014).
19. S. Miyazawa, Selection originating from protein stability/foldability: Relationships between protein folding free energy, sequence ensemble, and fitness. *J. Theor. Biol.* **433**, 21–38 (2017).
20. V. Pande, A. Grosberg, T. Tanaka, Statistical mechanics of simple models of protein folding and design. *Biophys. J.* **73**, 3192–3210 (1997).
21. R. R. Cheng, F. Morcos, H. Levine, J. N. Onuchic, Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E563–E571 (2014).
22. R. R. Cheng *et al*., Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes. *Mol. Biol. Evol.* **33**, 3054–3064 (2016).
23. A. Leaver-Fay *et al*., "Chapter nineteen—Rosetta3: An object-oriented software suite for the simulation and design of macromolecules" in *Computer Methods, Part C, Methods in Enzymology*, M. L. Johnson, L. Brand, Eds. (Academic Press, 2011), vol. 487, pp. 545–574.
24. B. Kuhlman *et al*., Design of a novel globular protein fold with atomic-level accuracy. *Science* **21**, 1364–1368 (2003).
25. D. Röthlisberger *et al*., Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
26. J. Karanicolas *et al*., A de novo protein binding pair by computational design and directed evolution. *Mol. Cell* **42**, 250–260 (2011).
27. D. A. Silva *et al*., De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186–191 (2019).
28. V. Potapov, M. Cohen, G. Schreiber, Assessing computational methods for predicting protein stability upon mutation: Good on average but not in the details. *Protein Eng. Des. Sel.* **22**, 553–560 (2009).
29. E. H. Kellogg, A. Leaver-Fay, D. Baker, Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Struct. Funct. Bioinf.* **79**, 830–838 (2011).
30. H. Park *et al*., Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).
31. K. A. Barlow *et al*., Flex ddG: Rosetta ensemble-based estimation of changes in protein-protein binding affinity upon mutation. *J. Phys. Chem. B* **122**, 5389–5399 (2018).
32. B. I. Dahiyat, S. L. Mayo, De novo protein design: Fully automated sequence selection. *Science* **278**, 82–87 (1997).
33. D. N. Bolon, S. L. Mayo, Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 14274–14279 (2001).
34. H. K. Privett *et al*., Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 3790–3795 (2012).
35. K. Druart, Z. Palmai, E. Omarjee, T. Simonson, Protein:Ligand binding free energies: A stringent test for computational protein design. *J. Comput. Chem.* **37**, 404–415 (2016).
36. J. Wang *et al*., Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
37. J. Ingraham *et al*., Illuminating protein space with a programmable generative model. bioRxiv (2022). https://www.biorxiv.org/content/10.1101/2022.12.01.518682v1. Accessed 26 September 2023.
38. J. L. Watson *et al*., Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. bioRxiv (2022). https://www.biorxiv.org/content/10.1101/2022.12.09.519842v1. Accessed 26 September 2023.
39. R. Guerois, J. E. Nielsen, L. Serrano, Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387 (2002).
40. L. Wang *et al*., Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **137**, 2695–2703 (2015).
41. E. P. Raman, T. J. Paul, R. L. Hayes, C. L. Brooks III, Automated, accurate, and scalable relative protein-ligand binding free energy calculations using lambda dynamics. *J. Chem. Theory Comput.* **16**, 7895–7914 (2020).
42. C. E. M. Schindler *et al*., Large-scale assessment of binding free energy calculations in active drug discovery projects. *J. Chem. Inf. Model.* **60**, 5457–5474 (2020).
43. D. Seeliger, B. L. de Groot, Protein thermostability calculations using alchemical free energy simulations. *Biophys. J.* **98**, 2309–2316 (2010).
44. V. Gapsys, S. Michielssens, D. Seeliger, B. L. de Groot, Accurate and rigorous prediction of the changes in protein free energies in a large-scale mutation scan. *Angew. Chem.* **55**, 7364–7368 (2016).
45. T. Steinbrecher *et al*., Predicting the effect of amino acid single-point mutations on protein stability: Large-scale validation of MD-based relative free energy calculations. *J. Mol. Biol.* **429**, 948–963 (2017).
46. A. J. Clark *et al*., Free energy perturbation calculation of relative binding free energy between broadly neutralizing antibodies and the gp120 glycoprotein of HIV-1. *J. Mol. Biol.* **429**, 930–947 (2017).
47. J. Duan, D. Lupyan, L. Wang, Improving the accuracy of protein thermostability predictions for single point mutations. *Biophys. J.* **119**, 115–127 (2020).
48. W. Jespers *et al*., QresFEP: An automated protocol for free energy calculations of protein mutations in Q. *J. Chem. Theory Comput.* **15**, 5461–5473 (2019).
49. R. L. Hayes, J. Z. Vilseck, C. L. Brooks III, Approaching protein design with multisite λ dynamics: Accurate and scalable mutational folding free energies in T4 lysozyme. *Protein Sci.* **27**, 1910–1922 (2018).
50. R. W. Zwanzig, High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* **22**, 1420–1426 (1954).
51. T. P. Straatsma, H. J. C. Berendsen, Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *J. Chem. Phys.* **89**, 5876–5886 (1988).
52. M. R. Shirts, E. Bair, G. Hooker, V. S. Pande, Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Phys. Rev. Lett.* **91**, 140601 (2003).
53. C. D. Christ, W. F. van Gunsteren, Enveloping distribution sampling: A method to calculate free energy differences from a single simulation. *J. Chem. Phys.* **126**, 184110 (2007).
54. S. Riniker *et al*., Comparison of enveloping distribution sampling and thermodynamic integration to calculate binding free energies of phenylethanolamine N-methyltransferase inhibitors. *J. Chem. Phys.* **135**, 024105 (2011).
55. X. Kong, C. L. Brooks III, λ-dynamics: A new approach to free energy calculations. *J. Chem. Phys.* **105**, 2414–2423 (1996).
56. J. L. Knight, C. L. Brooks III, Multisite λ dynamics for simulated structure-activity relationship studies. *J. Chem. Theory Comput.* **7**, 2728–2739 (2011).
57. J. L. Knight, C. L. Brooks III, Applying efficient implicit nongeometric constraints in alchemical free energy simulations. *J. Comput. Chem.* **32**, 3423–3432 (2011).
58. K. A. Armacost, G. B. Goh, C. L. Brooks III, Biasing potential replica exchange multisite λ-dynamics for efficient free energy calculations. *J. Chem. Theory Comput.* **11**, 1267–1277 (2015).
59. R. L. Hayes, K. A. Armacost, J. Z. Vilseck, C. L. Brooks III, Adaptive landscape flattening accelerates sampling of alchemical space in multisite λ dynamics. *J. Phys. Chem. B* **121**, 3626–3635 (2017).
60. Y. Huang, W. Chen, J. A. Wallace, J. Shen, All-atom continuous constant pH molecular dynamics with particle mesh Ewald and titratable water. *J. Chem. Theory Comput.* **12**, 5411–5421 (2016).
61. M. Aldeghi, V. Gapsys, B. L. de Groot, Accurate estimation of ligand binding affinity changes upon protein mutation. *ACS Cent. Sci.* **4**, 1708–1718 (2018).
62. R. L. Hayes, C. L. Brooks III, A strategy for proline and glycine mutations to proteins with alchemical free energy calculations. *J. Comput. Chem.* **42**, 1088–1094 (2021).
63. R. L. Hayes, J. Z. Vilseck, C. L. Brooks III, Addressing intersite coupling unlocks large combinatorial chemical spaces for alchemical free energy methods. *J. Chem. Theory Comput.* **18**, 2114–2123 (2022).
64. F. Sievers *et al*., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
65. K. Katayanagi *et al*., Structural details of ribonuclease H from *Escherichia coli* as refined to an atomic resolution. *J. Mol. Biol.* **223**, 1029–1052 (1992).
66. E. F. Pettersen *et al*., UCSF Chimera-a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
67. G. Dantas *et al*., High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. *J. Mol. Biol.* **366**, 1209–1221 (2007).
68. J. Lee *et al*., Computer-based engineering of thermostabilized antibody fragments. *AIChE J.* **66**, e16864 (2020).
69. J. Z. Vilseck, K. A. Armacost, R. L. Hayes, G. B. Goh, C. L. Brooks III, Predicting binding free energies in a large combinatorial chemical space using multisite λ dynamics. *J. Phys. Chem. Lett.* **9**, 3328–3332 (2018).
70. N. C. Wu, L. Dai, C. A. Olson, J. O. Lloyd-Smith, R. Sun, Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife* **5**, e16965 (2016).
71. P. Tian, R. B. Best, How many protein sequences fold to a given structure? A coevolutionary analysis *Biophys. J.* **113**, 1719–1730 (2017).
72. M. Ekeberg, T. Hartonen, E. Aurell, Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.* **276**, 341–356 (2014).
73. D. U. Ferreiro, E. A. Komives, P. G. Wolynes, Frustration, function and folding. *Curr. Opin. Struct. Biol.* **48**, 68–73 (2018).
74. R. L. Hayes, J. Buckner, C. L. Brooks III, Blade: A basic lambda dynamics engine for GPU accelerated molecular dynamics free energy calculations. *J. Chem. Theory Comput.* **17**, 6799–6807 (2021).

75. C. N. Bedbrook *et al.*, Structure-guided SCHEMA recombination generates diverse chimeric channelrhodopsins. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E2624–E2633 (2017).
76. U. T. Bornscheuer *et al.*, Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).
77. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
78. V. Opuu *et al.*, A physics-based energy function allows the computational redesign of a PDZ domain. *Sci. Rep.* **10**, 11150 (2020).
79. T. M. Raschke, J. Kho, S. Marqusee, Confirmation of the hierarchical folding of RNase H: A protein engineering study. *Nat. Struct. Biol.* **6**, 825–831 (1999).
80. T. O. Street, N. Courtemanche, D. Barrick, Protein folding and stability using denaturants. *Methods Cell Biol.* **84**, 295–325 (2008).
81. L. Wang *et al.*, Accurate modeling of scaffold hopping transformations in drug discovery. *J. Chem. Theory Comput.* **13**, 42–54 (2017).
82. H. S. Yu *et al.*, Accurate and reliable prediction of the binding affinities of macrocycles to their protein targets. *J. Chem. Theory Comput.* **13**, 6290–6300 (2017).
83. S. Liu, L. Wang, D. L. Mobley, Is ring breaking feasible in relative binding free energy calculations? *J. Chem. Inf. Model.* **55**, 727–735 (2015).
84. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
85. U. Essmann *et al.*, A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).
86. G. J. Rocklin, D. L. Mobley, K. A. Dill, P. H. Hünenberger, Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *J. Chem. Phys.* **139**, 184103 (2013).
87. F. Morcos *et al.*, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).
88. M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E* **87**, 012707 (2013).
89. B. R. Brooks *et al.*, CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217 (1983).
90. B. R. Brooks *et al.*, CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).
91. A. P. Hynninen, M. F. Crowley, New faster CHARMM molecular dynamics engine. *J. Comput. Chem.* **35**, 406–413 (2014).
92. R. B. Best *et al.*, Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).
93. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
94. M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, J. H. Jensen, PROPKA3: Consistent treatment of internal and surface residues in empirical pK$_a$ predictions. *J. Chem. Theory Comput.* **7**, 525–537 (2011).
95. P. J. Steinbach, B. R. Brooks, New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comput. Chem.* **15**, 667–683 (1994).
96. B. Frenz *et al.*, Prediction of protein mutational free energy: Benchmark and sampling improvements increase classification accuracy. *Front. Bioeng. Biotechnol.* **8**, 558247 (2020).
97. J. Mistry *et al.*, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
98. R. L. Hayes, C. L. Brooks III, Scripts for Selection pressures on evolution of ribonuclease H explored with rigorous free-energy-based design. GitHub. https://github.com/RyanLeeHayes/PublicationScripts/blob/main/2023RNaseH.tgz. Deposited 13 December 2023.
99. R. L. Hayes, ALF, GitHub. https://github.com/RyanLeeHayes/ALF. Accessed 20 July 2023.