Switching an active site helix in dihydrofolate reductase reveals limits to sub-domain modularity

Victor Y. Zhao[1,*], Joao V. Rodrigues[1,*], Elena R. Lozovsky[2], Daniel L. Hartl[2], and Eugene I. Shakhnovich[1]

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138

[2]Department Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

[*]These authors contributed equally to this work.

Corresponding author:

Eugene I. Shakhnovich
Department of Chemistry and Chemical Biology
Harvard University
12 Oxford Street
Cambridge, MA 02138
Tel: 617-495-4130
Email: shakhnovich@chemistry.harvard.edu

Abstract

To what degree are individual structural elements within proteins modular such that similar structures from unrelated proteins can be interchanged? We study sub-domain modularity by creating 20 chimeras of an enzyme, *E. coli* dihydrofolate reductase (DHFR), in which a catalytically important, 10-residue α-helical sequence is replaced by α-helical sequences from a diverse set of proteins. The chimeras stably fold but have a range of diminished thermal stabilities and catalytic activities. Evolutionary coupling analysis indicates that the residues of this α-helix are under selection pressure to maintain catalytic activity in DHFR. We performed molecular dynamics simulations using replica exchange with solute-tempering. Chimeras with low catalytic activity exhibit non-helical conformations that block the binding site and disrupt the positioning of the catalytically essential residue D27. Simulation observables and *in vitro* measurements of thermal stability and substrate binding affinity are strongly correlated. Several *E. coli* strains with chromosomally integrated chimeric DHFRs can grow, with growth rates that follow predictions from a kinetic flux model that depends on the intracellular abundance and catalytic activity of DHFR. Our findings show that although α-helices are not universally substitutable, the molecular and fitness effects of modular segments can be predicted by the biophysical compatibility of the replacement segment.

Keywords: protein modularity, DHFR, direct coupling analysis, MD simulation, cellular fitness

Abbreviations: DHFR: dihydrofolate reductase; MD: molecular dynamics; REST2: replica exchange with solute tempering; RMSD: root mean square deviation; SASA: solvent-accessible surface area.

Introduction

Protein evolution proceeds through a process of genetic change, selection, and fixation. Changes to protein sequences involve not just point mutations but also larger-scale variations such as insertions, deletions, and fusions. These rearrangements often occur at the level of protein domains [1–3]. By their ability to fold independently and carry out specific functions either alone or as part of a larger polypeptide, protein domains demonstrate properties of modularity. Modularity in biology is characterized by a portion of a larger biological unit evolving and functioning as a distinct unit [4]. The modularity of protein domains facilitates evolution by enabling their combination and shuffling to create multi-domain proteins with novel functions [5–7].

While the evidence that protein domains can behave as independent units is clear, less is known about modularity within domains [8]. Although short peptide sequences are unlikely to have defined structures in isolation, insertions, deletions, and shuffling of short segments commonly occur in protein evolution [9–13]. One suggestion of sub-domain modularity is the occurrence of similar substructures in proteins from across the protein universe, from contiguous segments to recurring tertiary interactions [14–20]. This reuse of structure reflects both natural thermodynamic basins of stability as well as inheritance of ancestral, peptide building blocks in evolution [18,19,21–23]. One method to create designed proteins, inspired by structural repetition in natural proteins [24–26], is the composition of a larger protein by repeating a small, modular subunit [27].

At the level of secondary structure, there is evidence that α-helices within proteins can act as distinct units with localized folding and unfolding transitions [28–31], and some α-helices exist as standalone structures [32–34]. Compared to other structures in proteins, α-helices are also more tolerant of mutations [35,36]. A few past studies examined the replacement of whole α-helices in enzymes [37–39]. In *E. coli* alkaline phosphatase, α-helices were replaced with non-homologous helical sequences while maintaining enzyme activity in most cases [37]. More recently, we replaced the catalytically important α_B-helix in *E. coli* dihydrofolate reductase (DHFR) with orthologous sequences from DHFRs of six other species with little effect on catalytic activity and organism growth rate but some changes to thermal stability and antibiotic resistance [39]. For these reasons, we hypothesize that α-helices represent modular segments within protein domains.

In this work, we explore the modularity of α-helices using *E. coli* DHFR as a model protein. We performed non-homologous replacement by switching the α_B-helix sequence in *E. coli* DHFR with sequences taken from helices in unrelated, non-DHFR proteins while leaving the key catalytic residue, D27, unchanged. The chimeras vary in catalytic activity, although all of them have reduced activity compared with wildtype DHFR. The chimeras with weaker catalytic activity have residues that appear infrequently among homologous DHFR sequences and exhibit non-helical conformations. The ability of replacement helix sequences to maintain a wildtype-like helical conformation determines cellular fitness through the dual effect of helix conformation on catalytic activity and *in vivo* protein abundance, which is determined by the balance between rates of protein production and degradation [40]. These findings support qualified modularity in this specific α-helix in which tertiary interactions govern the functional compatibility of replacement sequences. More generally, our work shows, in mechanistic detail, how structural variation in the vicinity of the active site broadly mediates the activity of an enzyme. On a cellular level, we show how large-scale sequence variation pleiotropically affects the fitness landscape through its dual effect on enzymatic activity and in-cell turnover, which cause pronounced variation in total DHFR activity in the cytoplasm.

Results

Chimera generation process

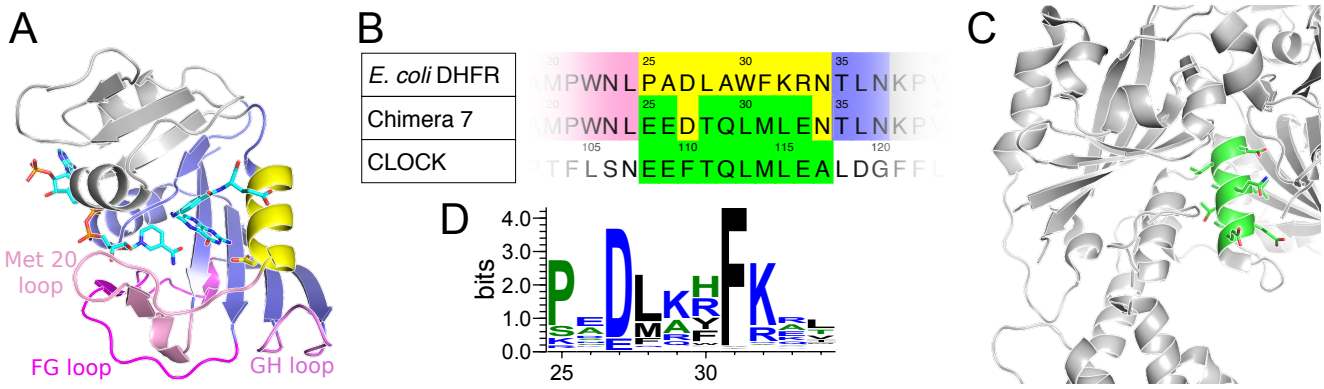

Figure 1: **Overview of *E. coli* DHFR and chimera generation.** (A) Structure of *E. coli* DHFR (PDB ID 4NX6), with NADP+ and folic acid in teal, adenosine-binding subdomain in gray, and major subdomain in purple and various other colors (Met 20, FG, and GH loops in magenta shades and the α_B-helix in yellow). The D27 side chain is illustrated using stick rendering. (B) Chimera 7 as an example: alignment of *E. coli* DHFR wildtype, Chimera 7, and helix donor (*M. musculus* CLOCK) sequences. In Chimera 7, the donor helix sequence from CLOCK (green) takes the place of residues 25-34 (yellow) in *E. coli* DHFR, with residues 27 and 34 maintaining their *E. coli* wildtype identities. (C) Structure of CLOCK (PDB ID 4F3L), donor protein for Chimera 7. Residues 108-117, highlighted in green, are substituted into E. coli DHFR to create Chimera 7. (D) Sequence logo for the region of *E. coli* DHFR undergoing replacement, generated by WebLogo [41]. Input consists of 19,912 sequences homologous to DHFR, found and aligned by EVcouplings [42].

We designed 20 chimeric variants of *E. coli* DHFR by replacing all but one residue (T35) of the α_B-helix (residues 25-34) with the sequences of α-helices taken from non-DHFR proteins in various species (see Materials and Methods). The full list of chimeric DHFR variants is presented in Table S1. DHFR is a globular protein consisting of a β-sheet surrounded by loops and helices (Figure 1A). The protein structure can be divided into two subdomains: the adenosine-binding subdomain (residues 38-88, colored gray) and the major subdomain (residues 1-37 and 89-159, various colors). The major subdomain consists of a central β-sheet along with two of the three α-helices in DHFR as well as three loops: the Met 20 loop (residues 9-24), the FG loop (residues 116-132), and the GH loop (residues 142-150) [43]. The 10 residues of the α_B-helix targeted for replacement are highlighted in yellow. D27 is catalytically essential [44] and is shown in stick representation.

The chimera generation process is illustrated in Figure 1B and Figure 1C using Chimera 7 as an example. The donor sequence, from a helix in *M. musculus* protein CLOCK, replaces nearly all of the *E. coli* DHFR α_B-helix sequence. The catalytically essential residue, D27, was left unmutated for all chimeras. Additionally, in the case of Chimera 7, N34 is also unchanged. For 4 of 20 chimeras, including Chimera 7, some C-terminal positions were left unmutated, due to shorter donor helix sequences. This serves to vary the length of the replacement region. Figure 1D shows a DHFR sequence logo [41] for this 10-residue segment, based on multiple sequence alignment of 19,912 sequences homologous to DHFR found using the EVcouplings web server [42]. Several positions, which are P25, D27, L28, F31, and K32 in *E. coli* DHFR, show a substantial preference for specific residues. The sequence identities of the replacement helical sequences with the original wildtype sequence range from 10% to 50%. In our previous study, we studied six chimeras in which the same residues 25-34 of *E. coli* DHFR were replaced with sequences from six DHFR orthologs [39]. We refer to these original six chimeras as the

orthologous chimeras and the new chimeras generated in this study as the non-orthologous chimeras. Table 1 lists the sequences and properties of key DHFR chimeras discussed in this work.

Table 1: Sequences and properties of select DHFR variants in this work. Residues that match wildtype *E. coli* DHFR sequence are bolded. EVcouplings Δ*E* is calculated according to the EVcouplings non-epistatic model (see below). k_{cat}/K_M values are ± standard error of the mean. Mean α_B-helix RMSD (root mean square deviation) values are ± standard deviations.

	Sequence, residues 25-34	Sequence identity (%)	EVcouplings Δ*E*	k_{cat}/K_M ($s^{-1}\mu M^{-1}$)	Mean α_B-helix RMSD (Å)
E. coli DHFR	**PADLAWFKRN**	100	0	$1.1\times10^{1} \pm 9.8\times10^{-1}$	1.00 ± 0.42
Chimera 1	VQ**D**DTEELKS	10	-46.8	$1.1\times10^{-2} \pm 7.1\times10^{-4}$	5.47 ± 1.25
Chimera 2	K**AD**TLTTFDA	20	-29.8	$3.2\times10^{-3} \pm 1.4\times10^{-3}$	1.55 ± 0.40
Chimera 6	EQ**DL**RFVM**RN**	40	-16.4	$2.9\times10^{-3} \pm 4.5\times10^{-4}$	3.35 ± 0.78
Chimera 7	EE**D**TQLMLE**N**	20	-22.7	$3.0\times10^{-5} \pm 2.6\times10^{-5}$	3.34 ± 1.48
Chimera 8	SN**D**SDLHAH**N**	20	-31.1	$2.1\times10^{-4} \pm 8.1\times10^{-5}$	2.08 ± 1.63
Chimera 13	GR**D**YEQ**FKRN**	50	-16.9	$3.6\times10^{-3} \pm 7.3\times10^{-4}$	1.31 ± 0.46
Chimera 17	DS**D**MDIICEK	10	-30.0	$2.1\times10^{-5} \pm 8.3\times10^{-6}$	5.31 ± 1.00
Chimera 18	QQ**D**VSNRL**RE**	20	-33.0	$4.1\times10^{-5} \pm 4.8\times10^{-6}$	1.52 ± 0.69
Chimera 19	TD**D**VL**W**LRSV	20	-19.3	$1.6\times10^{-2} \pm 3.0\times10^{-3}$	1.09 ± 0.32
Chimera 20	KC**DA**LLQSF	20	-25.4	$1.4\times10^{-3} \pm 3.6\times10^{-4}$	4.29 ± 1.37

Non-orthologous DHFR chimeras have reduced catalytic activity

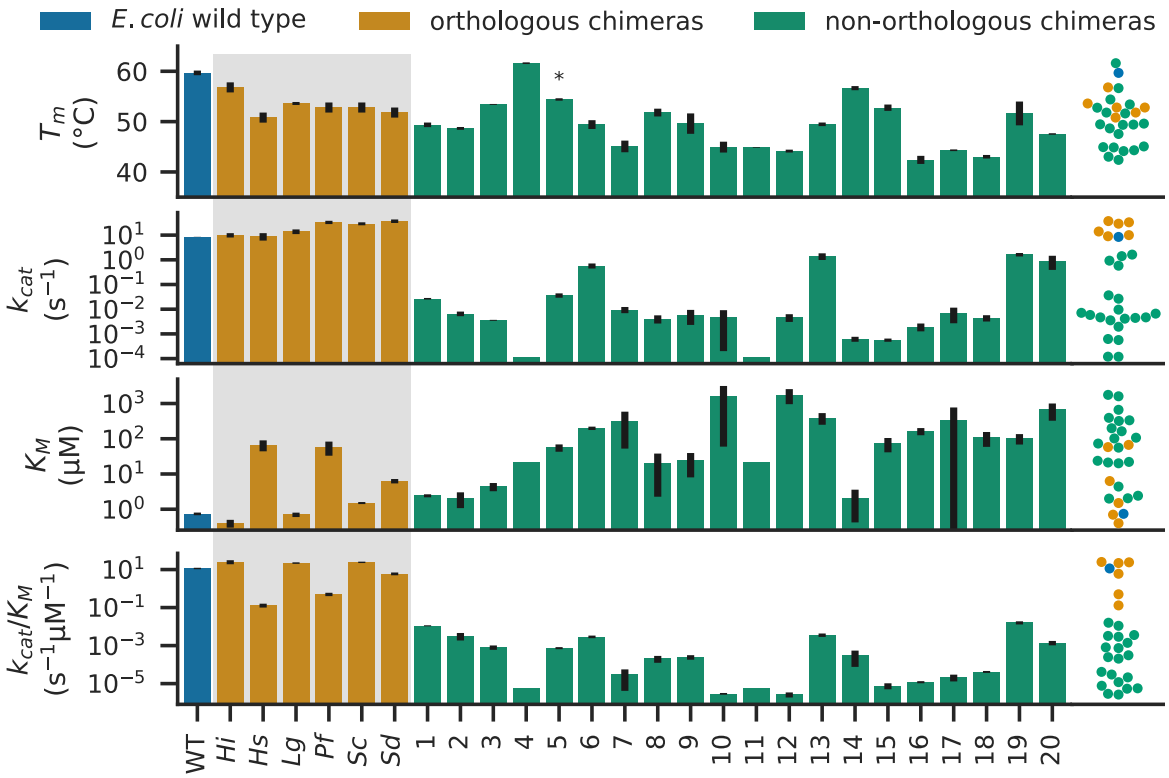

Figure 2: **Stabilities and catalytic activities of non-orthologous chimeras of DHFR variants are generally decreased.** Error bars indicate the standard error of the mean. Right-hand side swarm plots show distributions of corresponding properties (values are plotted against the vertical axis with horizontal offsets to avoid overlap). Orthologous chimeras (orange bars, gray background) were characterized in previous work [39]. *: Size exclusion chromatography found that Chimera 5 is partly dimeric (Figure S1).

Twenty chimeric proteins were overexpressed in *E. coli*, purified to homogeneity, and characterized biophysically and biochemically. The melting temperature (T_m) of each chimera was measured by following the changes in intrinsic tryptophan fluorescence upon thermal denaturation (see Materials and Methods). Except for Chimera 4, the thermal stabilities of the non-orthologous chimeras are reduced (Figure 2). For 14 of 20 non-orthologous chimeras, T_m is at least 10 ˚C lower than that of the wild type (WT), and the majority have T_m values lower than those of the orthologous chimeras. Size exclusion chromatography revealed that all tested chimeras elute as monomers, except for Chimera 5, which elutes as a mixture of monomer and dimer (Supplementary Figure S1). The relatively high T_m (54.4 ˚C) that was measured for Chimera 5 is likely associated with additional stabilization conferred by dimerization.

Substitution with non-orthologous helical sequences compromises the catalytic function of the protein. The catalytic rates (k_{cat}) and catalytic efficiencies (k_{cat}/K_M) of non-orthologous chimeras are all lower than those of wildtype DHFR and the orthologous chimeras (Figure 2). k_{cat}/K_M values are at least three orders of magnitude lower than that of wildtype DHFR. Substrate affinity (K_M) is also reduced (higher values) for non-orthologous chimeras. One possible reason for the differences among the chimeras could be that the replacement segments differ in their helical propensities. We found, however, no correlations between measured *in vitro* properties and the Agadir score (Supplementary Figure S2), a computational prediction for the helical propensity of the

sequences in isolation [45–49]. This suggests that a mechanism that accounts for interactions with the rest of the protein is more likely to explain the observed differences.

Helix residues are under evolutionary selection for catalytic activity

The striking difference in catalytic activity (several orders of magnitude) between orthologous and non-orthologous chimeras suggests that all or some residues in the α_B-helix were evolutionarily selected to maintain the catalytic activity of the enzyme. To this end, we assessed how divergent our replacement sequences are from DHFR. The probability that a given sequence, $\{\sigma\}$, belongs to an ensemble of aligned, homologous sequences with selected biophysical or functional properties can be modeled using a Boltzmann distribution [50,51]. For instance, direct coupling analysis uses the following model [52,53]:

$$P(\{\sigma\}) = \frac{e^{H_{\text{eff}}(\{\sigma\})}}{Z}$$

$$H_{\text{eff}}(\{\sigma\}) = \sum_{i}^{N} \boldsymbol{h}_i(\sigma_i) + \sum_{i<j}^{N} \boldsymbol{J}_{ij}(\sigma_i, \sigma_j)$$

$$Z = \sum_{\{\sigma\}} e^{H_{\text{eff}}(\{\sigma\})}$$

where N is the alignment length, $\boldsymbol{h}_i$ and $\boldsymbol{J}_{ij}$ are site-specific bias and site-site coupling terms, respectively, and Z is the normalizing factor. $\boldsymbol{J}_{ij}$ accounts for epistatic interactions between sites by weighing occurrences of specific residue pairs among the aligned sequences. This method has been used to predict the fitness effect of mutations as the difference in the statistical energy (H_{eff}) between two sequences [54] and is available as the EVcouplings web server and Python package [42]. Mutational effects, or ΔE, are calculated as $\Delta E = H_{\text{eff}}(\{\sigma^{\text{ch}}\}) - H_{\text{eff}}(\{\sigma^{\text{wt}}\})$. We calculated mutational effects using the epistatic model described above. We also used a non-epistatic model which has a similar form but does not include $\boldsymbol{J}_{ij}$ terms: $H_{\text{eff}}(\{\sigma\}) = \sum_i^N \boldsymbol{h}(\sigma_i)$. This simpler model reflects sequence conservation at each position [54]. In both cases, a more negative ΔE indicates a sequence with residues that occur infrequently among DHFR homologs. ΔE, therefore, reports on the evolutionary compatibility of replacement fragments with DHFR.

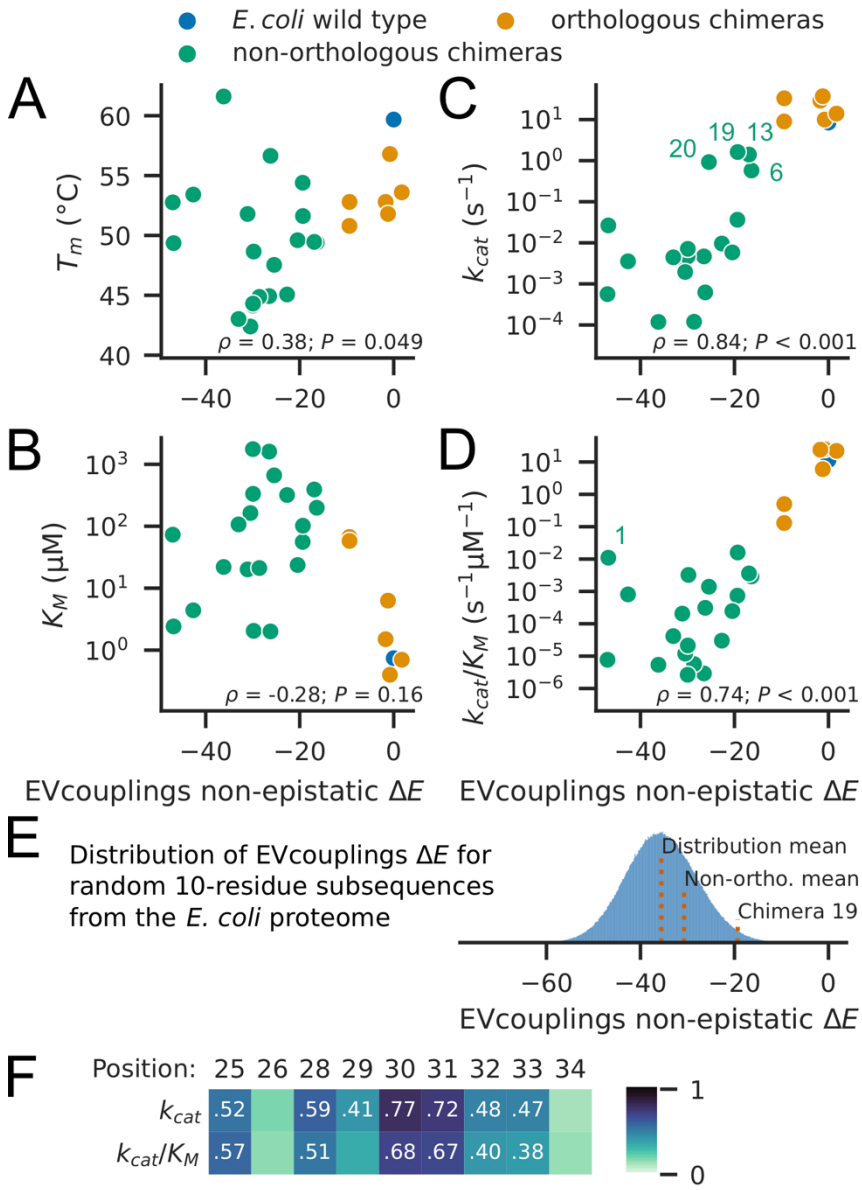

Figure 3: **EVcouplings non-epistatic model Δ*E* correlates with catalytic turnover and efficiency of DHFR chimeras.** Non-epistatic Δ*E* is compared with (A) T_m, (B) K_M, (C) k_{cat}, with Chimeras 6, 13, 19, and 20 labeled, and (D) k_{cat}/K_M, with Chimera 1 labeled. Spearman rank correlation coefficients and *P* values are given in the lower portion of each plot. (E) Distribution of non-epistatic Δ*E* from substituting α_B-helix residues 25-34 in *E. coli* DHFR with all 10-residue subsequences from all *E. coli* coding sequences while retaining D27 (n = 1282235). The mean of the distribution, the mean Δ*E* from the 16 non-orthologous chimeras (Non-ortho.) in which residues 25-34 were substituted (excludes Chimeras 6, 7, 8, and 13), and Δ*E* for Chimera 19 are indicated. (F) Heatmap of Spearman rank correlation coefficients between catalytic parameters, k_{cat} and k_{cat}/K_M, and non-epistatic Δ*E* values of point mutations at each position. Correlation coefficients are printed for statistically significant ($P < 0.05$) combinations.

EVcouplings non-epistatic model Δ*E* has strong correlations with catalytic parameters k_{cat} and k_{cat}/K_M. Scatterplots between the non-epistatic Δ*E* and experimental properties are shown in Figure 3. While Spearman correlations between Δ*E* and T_m ($\rho = 0.38$, $P = 0.049$, Figure 3A) and between Δ*E* and K_M are weak ($\rho = -0.28$, *P*

= 0.16, Figure 3B), there are strong correlations between Δ*E* and k_{cat} (ρ = 0.84, P < 0.001, Figure 3C) and between Δ*E* and k_{cat}/K_M (ρ = 0.74, P < 0.001, Figure 3D). Similar but slightly weaker correlations are obtained using Δ*E* according to the epistatic model (Supplementary Figure S3). These correlations relate evolutionary data to experimental protein properties to indicate that the residues of the α_B-helix have evolved under selection pressure to maintain a high enzyme catalytic rate and efficiency. The correlations between catalytic activity and Δ*E* are stronger than the correlations calculated using sequence identity (k_{cat} and sequence identity: ρ = 0.63, P < 0.001) or sequence similarity (k_{cat} and sequence similarity: ρ = 0.63, P < 0.001) (SI Table S2, Supplementary Figure S4). Sequence similarity was calculated according to Grantham's distance between amino acids, which is a model based on residue properties [55].

Examining individual chimeras, we find that four chimeras have at least one order of magnitude higher k_{cat} than those of all other non-orthologous chimeras: Chimeras 6, 13, 19, and 20 (Figure 3C). Of these four, Chimeras 6, 13, and 19 also have the highest Δ*E* values among non-orthologous chimeras. Chimeras 6 and 13 have higher wildtype sequence identities (40% and 50%, respectively), as a result of shorter helix replacement lengths (25-32 and 25-30, respectively), which likely explain their higher k_{cat} values and less negative Δ*E* values. By contrast, Chimeras 19 and 20 both have only 20% identity (Table 1). Chimera 19 notably has the highest k_{cat} and k_{cat}/K_M of all chimeras. The Chimera 19 sequence matches the wildtype DHFR sequence at position 30 (W30); this, alongside the overall compatibility of other residues, likely explains its high activity. To gauge the rarity of sequences like Chimera 19, we calculated non-epistatic Δ*E* values for substitution of α_B-helix residues 25-34 with all valid 10-residue subsequences (n = 1282235) from protein-coding sequences in *E. coli*, while keeping position 27 as D (see Materials and Methods). We find that Chimera 19-like sequences are rare: 1.80 % of the sequences have a higher Δ*E* than that of Chimera 19 (Figure 3E). We also compared these protein subsequences to the 16 non-orthologous helical sequences in which residues 25-34 were substituted (excluding Chimeras 6, 7, 8, and 13). The non-orthologous chimeras have a slightly less negative mean Δ*E* than that of random protein subsequences (-30.7 versus -35.7). The difference between the distributions was statistically significant at the P = 0.05 level by Mann-Whitney U test (P = 0.0058) and Kolmogorov-Smirnov test (P = 0.0027). The data suggest that α-helical sequences have slightly better sequence compatibility with the replaced DHFR fragment than random protein segments.

In considering k_{cat}/K_M (Figure 3D), one outlier stands out: Chimera 1, which has the second-lowest non-epistatic model Δ*E* and the second-highest k_{cat}/K_M among non-orthologous chimeras. The α_B-helix sequence for chimera 1, VQDDTEELKS$^{25\text{-}34}$ (10% identity), switches residue hydrophobicity at positions 28, 30, 31, and 32, all sites with substantial sequence conservation (Figure 1D). The relatively high activity of this variant is therefore surprising and is further explored in the next section.

Finally, we examined the contributions to Δ*E* from each position in the α_B-helix. We calculated the correlation of Δ*E* for point mutations to the chimeric DHFR residues at each α_B-helix position with k_{cat} and k_{cat}/K_M of the chimeras. For all sites except positions 26 and 34, there are statistically significant correlations between point mutation effects and at least one of k_{cat} and k_{cat}/K_M (Figure 3F). Mutational effects at positions 30 and 31 have the strongest correlations with k_{cat} and k_{cat}/K_M. The residue identities of positions 30 and 31 appear most important in determining the catalytic activity of helix-swapped DHFR chimeras. These two sites are inward-facing and hydrophobic in the wildtype protein (W30 and F31, respectively), with F31 forming part of the substrate-binding pocket. Notably, Chimera 19 retains W30, and all orthologous chimeras retain F31. Overall, EVcouplings analysis shows that the residues of the non-orthologous replacement sequences are less likely to appear among sequences homologous to DHFR, and our *in vitro* measurements of catalytic activity indicate that such residues were selected against in evolution due to deleterious effects on catalysis. We next try to understand the low catalytic activity of these DHFR variants on a structural level.

Simulations find chimeras have destabilized α_B-helices

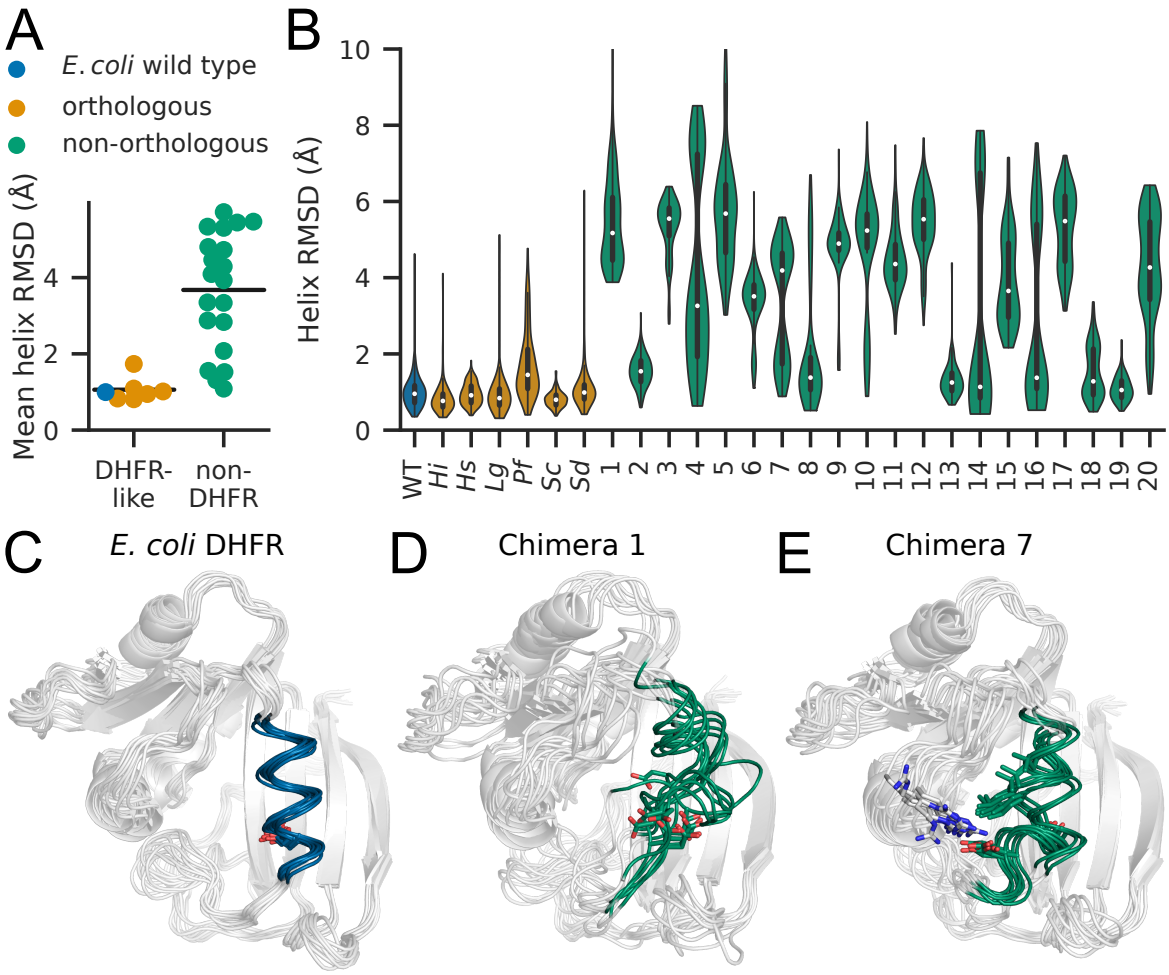

Figure 4: **Replica exchange MD simulations of DHFR variants show non-orthologous chimeras deviate from *E. coli* wildtype conformations.** (A) Mean helix RMSDs of *E. coli* wildtype DHFR and orthologous chimeras (DHFR-like) are generally lower than those of non-orthologous chimeras (non-DHFR). Horizontal lines indicate the mean value for each group. (B) Violin plots of helix RMSD show distributions of helix RMSD for each chimera. Inner white marker denotes median value, and ends of thicker vertical bars denote 25th and 75th percentiles. (C, D, and E) Superimposed structures of *E. coli* DHFR, Chimera 1, and Chimera 7 taken from the last 100 ns of each simulation, 10 samples per image, evenly separated in time. Helix backbones (25-35) are colored and depicted in loop representation. The D27 sidechains of all three variants as well as the R52 and L32 sidechains of Chimera 7 are illustrated using stick representation. Samples have been aligned using PyMOL [56].

To investigate the structures of the DHFR chimeras and understand the physical reason behind their decreased catalytic activity, we turned to molecular dynamics (MD) simulation. We built homology models of the chimeras [57] and used NAMD to simulate all 27 DHFR variants [58], using replica exchange with solute tempering (REST2) to enhance sampling of helix conformations [59]. Whole-protein REST2 simulations have been previously used to compare DHFR from mesophilic and thermophilic species [60]. To avoid inconsistency in folate binding between chimeras, all proteins were simulated in the apo state. All chimeras were simulated for at least 300 ns, with some simulations extended to 375 ns to achieve convergence; our analyses use simulation data from the

last 100 ns of simulation. Supplementary Figure S5 demonstrates that the root mean square deviation (RMSD) of the Cα atoms of the α_B-helix reaches steady state over at least the last 100 ns of simulations. The RMSD of the α_B-helix is measured after alignment of the major subdomain core residues (the β-sheet), using the DHFR crystal structure as the reference. This enables detection of conformational differences in which the entire α_B-helix is displaced relative to the core of the protein.

The non-orthologous chimeras generally show greater deviation from the *E. coli* DHFR crystal structure reference. Figure 4A compares the mean α_B-helix RMSDs of wildtype DHFR and orthologous chimeras and non-orthologous chimeras. All six orthologous chimeras maintain the α_B-helix in the wildtype-like conformation despite three of the six orthologous chimeras having only 20% α_B-helix sequence identity with *E. coli* DHFR, demonstrating evolutionary structural conservation. Only the *P. falciparum* (*Pf*) orthologous chimera shows a slightly higher helix RMSD (1.73 Å). By contrast, just 4 of the 20 non-orthologous chimeras (Chimeras 2, 13, 18, and 19) have mean helix RMSDs (1.55 Å, 1.31 Å, 1.52 Å, and 1.09 Å, respectively) approaching those of wildtype DHFR and the orthologous chimeras (mean RMSD 1.06 Å).

The distributions of helix RMSDs of the DHFR variants (Figure 4B) suggest that the non-orthologous chimeras can be divided into two groups, those with predominantly helical, wildtype-like α_B-helix conformations and those exhibiting non-wildtype conformations. The median helix RMSD (Figure 4B) is one differentiator; seven chimeras (2, 8, 13, 14, 16, 18, and 19) have values below 2 Å. The α_B-helices in these chimeras are fully helical in the majority of simulation frames: the median percent helicities of these chimeras are all 100% (Supplementary Figure S6), as calculated by DSSP [61,62]. Compared to the orthologous chimeras, however, the α_B-helices in these wildtype-like non-orthologous chimeras can be subtly different from the wildtype structure or exhibit more disordered conformations, which both result in higher mean helix RMSDs. As previously discussed, Chimeras 6, 13, and 19 are the variants with relatively higher ΔE and k_{cat}. Higher helix RMSD in Chimera 6 suggests a partial discrepancy between helical structure in simulations and evolutionary sequence compatibility. Nonetheless, Chimera 19 (highest k_{cat} and k_{cat}/K_M of all non-orthologous chimeras) is particularly wildtype-like (1.09 Å mean helix RMSD), demonstrating one case in which structure and function can be maintained after a helix swap with an evolutionarily unrelated sequence.

The second category of non-orthologous chimeras shows greater disorder in the α_B-helix. Two examples from this second group of non-orthologous chimeras, Chimeras 1 and 7, are compared to wildtype DHFR in Figure 4C, Figure 4D, and Figure 4E. Each image shows 10 samples from the last 100 ns of simulation (evenly separated in time). While wildtype DHFR maintains the crystallographic helix conformation, with D27 in its ligand-binding orientation (Figure 4C), Chimeras 1 and 7 both show ensembles of more variable helix conformations, and the three loops of the major subdomain are also visibly more disordered (Figure 4D and Figure 4E, respectively). Chimera 1 maintains partial helicity in just the C-terminal half of the α_B-helix, and the D27 sidechain faces outward (Figure 4D). Since Chimera 1 has relatively high catalytic activity (second highest k_{cat}/K_M among non-orthologous chimeras), it is remarkable that its α_B-helix conformations are among the least wildtype-like (second highest mean helix RMSD). These observations are unlikely to be due to simulation inaccuracy given that the Chimera 1 α_B-helix sequence contains several polar residues (Table 1). Secondly, Chimera 7 shows both wildtype-like and non-wildtype α_B-helix conformations (Figure 4E). The median helix RMSD of 4.2 Å for Chimera 7 (Figure 4B) indicates, however, that the non-wildtype conformations are more populated than the wildtype-like conformations. The non-wildtype conformation is characterized by occlusion of the active site as a result of L32 occupying the active site pocket and D27 interacting with R52 (Figure 4E). Since loss of the wildtype α_B-helix conformation can block the binding site and disrupts the positioning of the catalytically essential D27 (Supplementary Figure S6), the non-wildtype helix conformations in non-orthologous chimeras likely explains a large part of the loss of catalytic activity. We next examine the correlation of these simulation measurements with experimental properties.

Simulation measurements correlate most strongly with thermal stability and K_M

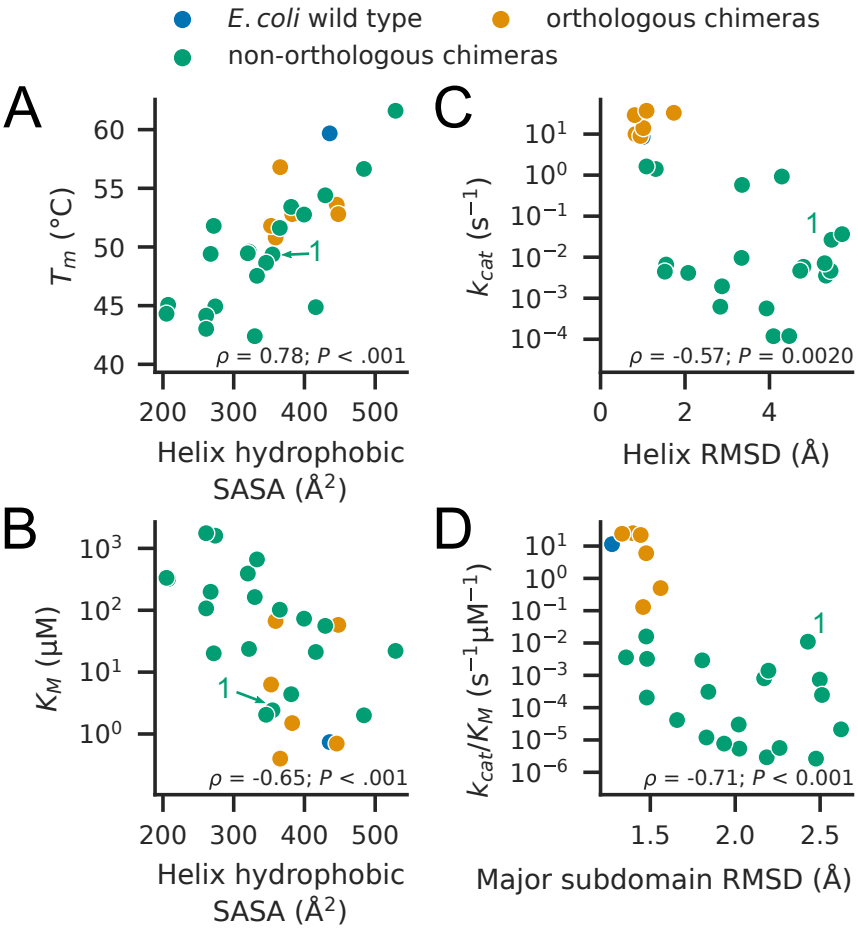

Figure 5: **Measures of hydrophobic solvent accessible surface area (SASA) correlate with T_m and K_M, and RMSD measurements correlate with k_{cat} and k_{cat}/K_M.** Scatterplots between (A) helix hydrophobic SASA and T_m, (B) helix hydrophobic SASA and K_M, (C) α_B-helix RMSD and k_{cat}, and (D) major subdomain RMSD and k_{cat}/K_M. Simulation quantities are mean values. Chimera 1 is labeled in each plot. Spearman rank correlation coefficients and P values are given in the lower portion of each plot. When only non-orthologous chimeras are considered, the correlations with helix hydrophobic SASA remain moderately high and statistically significant: (A) $\rho = 0.67$, $P = 0.0013$, (B) $\rho = -0.63$, $P = 0.0027$, (C) $\rho = -0.08$, $P = 0.75$, (D) $\rho = -0.36$, $P = 0.12$.

The different conformations of the DHFR variants prompted an investigation of the correlation between simulation measurements and *in vitro* properties of our variants. In simulations, non-helical conformations in the replacement segment are often characterized by the burial of hydrophobic sidechains into the substrate binding pocket or the hydrophobic surface of the major subdomain β-sheet, as was seen for Chimera 7 (Fig 4E). This prompted investigation of the differences in hydrophobic solvent accessible surface area (SASA) among the DHFR variants. This quantity was measured by summing up the SASA of nonpolar atoms belonging to the α_B-helix. This measurement is done in the context of the full protein structure; buried nonpolar helix atoms do not have any contribution to hydrophobic SASA (see Materials and Methods). Helix hydrophobic SASA has moderately high correlations with both T_m ($\rho = 0.78$, $P < 0.001$, Figure 5A) and K_M ($\rho = -0.65$, $P < 0.001$, Figure 5B). That is, DHFR variants with greater solvent-exposed hydrophobic surface areas on helix residues have

better thermal stabilities and substrate binding affinities. This is a trend that both orthologous and non-orthologous chimeras follow, and correlations remain moderately high and statistically significant for both *in vitro* quantities when only non-orthologous chimeras are considered (helix hydrophobic SASA and T_m: $\rho = 0.67$, $P = 0.0013$; helix hydrophobic SASA and K_M: $\rho = -0.63$, $P = 0.0027$). T_m and K_M are correlated properties ($\rho = -0.68$, $P < 0.001$), with more destabilized DHFR variants also having higher K_M values (Supplementary Figure S7). Decreased thermal stability likely reduces substrate affinity by making the optimal substrate-binding conformation less probable. But it is unintuitive that DHFR variants with more solvent-exposed nonpolar surface area would be associated with greater thermal stability. We note that helix hydrophobic SASA is not an indicator of wildtype-like α_B-helix conformations, since the correlation between helix hydrophobic SASA and helix RMSD is weak ($\rho = -0.21$, $P = 0.28$, Supplementary Figure S8).

We also examined simulation order parameters for different parts of the protein. Differences in disorder among DHFR variants occur in the major subdomain, while the adenosine-binding subdomain remains stable (Supplementary Figure S6). The adenosine-binding subdomain is more distant from the α_B-helix and exhibits greater stability in native state proteolysis experiments [63]. We thus focused our analysis on the various regions of the major subdomain. We find statistically significant correlations between mean RMSD measurements of different regions of the DHFR major subdomain and the experimentally determined *in vitro* properties of T_m, k_{cat}, K_M, and k_{cat}/K_M (Supplementary Figure S9). The RMSD measurements of different regions are also substantially correlated with each other (Supplementary Figure S8), reflecting the cooperative nature of protein stability. Figure 5C and Figure 5D show the strongest correlations seen in Supplementary Figure S9 between the RMSD measurements and k_{cat} and between the RMSD measurements and k_{cat}/K_M, respectively. The correlation between helix RMSD and k_{cat} is moderate ($\rho = -0.57$, $P = 0.002$, Figure 5C), while the correlation between RMSD of the entire major subdomain and k_{cat}/K_M is moderately high ($\rho = -0.71$, $P < 0.001$, Figure 5D). These trends are driven by the difference between orthologous and non-orthologous chimeras, however. Among non-orthologous chimeras, there is no correlation between helix RMSD and k_{cat} ($\rho = -0.08$, $P = 0.78$), and the correlation between major subdomain RMSD and k_{cat}/K_M is weak and no longer statistically significant ($\rho = -0.36$, $P = 0.12$). These trends show that while maintenance of the wildtype DHFR conformation is associated with high k_{cat} and k_{cat}/K_M, the degree of deviation from the wildtype conformation weakly predicts these catalytic parameters.

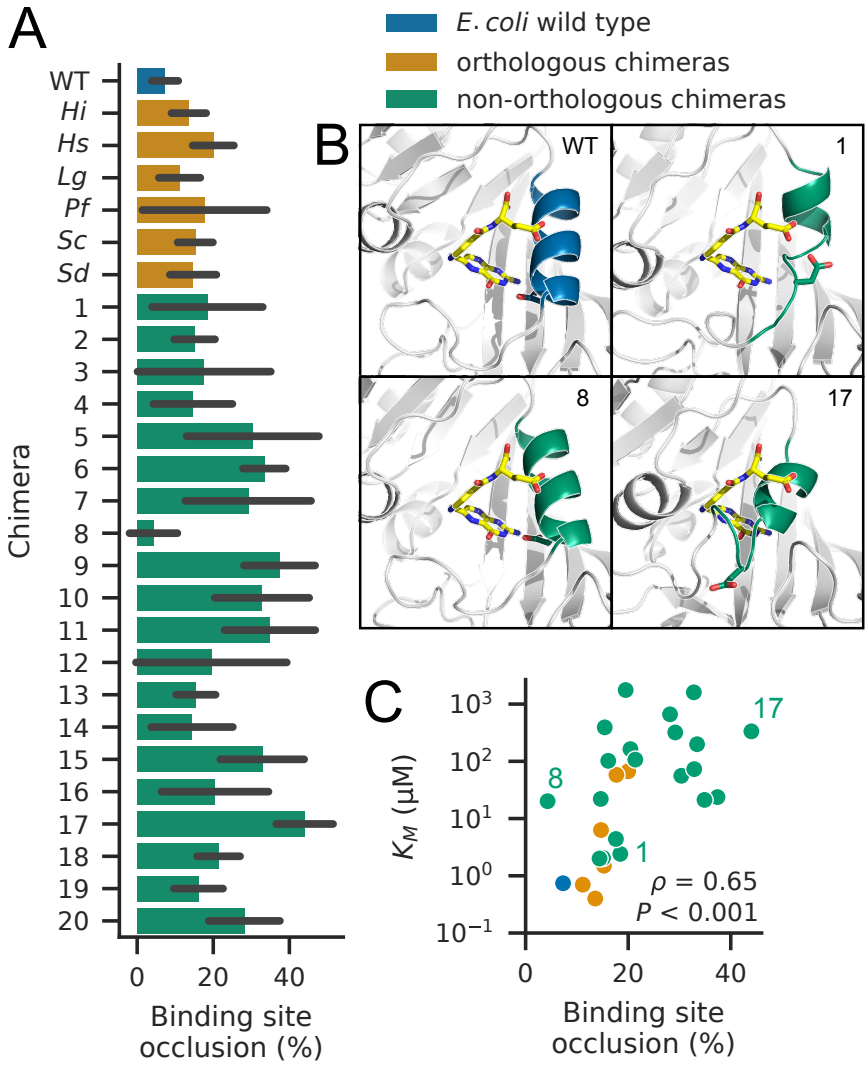

Figure 6: **DHFR variants with greater DHF binding site occlusion have higher K_M values.** Binding site occlusion was measured by calculating the amount of volume available to the DHF substrate. (A) Average binding site occlusion for each DHFR variant. Error bars represent standard deviations. (B) Representative structures from simulations for *E. coli* wildtype DHFR and Chimeras 1, 8, and 17 are illustrated. Conformations adopted by Chimeras 1 and 8 do not block the substrate binding site, whereas Chimera 17 conformations occupy the substrate binding site. The yellow folate ligand is overlaid from alignment to the DHFR crystal structure using the core residues of the major subdomain, and the α_B-helix (residues 25-35) is colored. (C) Scatterplot of binding site occlusion versus K_M. Chimeras 1, 8, and 17 are labeled. Spearman rank correlation coefficient and P value are given in the lower right of the plot. When only non-orthologous chimeras are considered, correlation is weaker (ρ = 0.42, P = 0.069).

The previously discussed Chimera 1 is an outlier with respect to RMSD and k_{cat}/K_M (Figure 5D). Chimera 1 also has an unusually low K_M (3rd lowest among non-orthologous chimeras), given its overall deviation from the wildtype conformation (Supplementary Figure S9), although Chimera 1 is not an outlier when considering helix hydrophobic SASA and K_M (Figure 5B). The conformations of Chimera 1 (Figure 4D) show that the α_B-helix conformations in Chimera 1 are not wildtype-like but do leave the active site relatively open. By contrast, many of the other disordered-helix non-orthologous chimeras have conformations that block the DHF binding site.

To quantify this, we measured DHF binding site occlusion by overlaying the folate ligand from the *E. coli* DHFR crystal structure into simulation frames after structural alignment using the major subdomain core residues. Occlusion was calculated as the proportion of ligand volume that is occupied by protein atoms (see Materials and Methods). Chimera 1 exhibits only moderate occlusion, comparable to chimeras with high helicity and wildtype-like α_B-helix conformations (Figure 6A). The wildtype DHFR and Chimera 8 are also notable for low binding site occlusion.

Figure 6B highlights representative structures from simulations in which the folate ligand from the DHFR crystal structure is superimposed on the simulation frames for wildtype DHFR and Chimeras 1, 8, and 17. Apo state wildtype DHFR maintains the crystallographic conformation, with D27 well-aligned with the superimposed folate ligand. For Chimera 1, the α_B-helix residues adopt a partly helical, open conformation, leaving the binding site free. Chimera 8 is also illustrated because it has exceptionally low binding site occlusion, which can be explained by an α_B-helix conformation that is 100% helical but is tilted away compared to wildtype DHFR; D27 is displaced outward relative to the superimposed folate ligand. By contrast, Chimera 17 exhibits the highest binding site occlusion and typifies how most non-helical conformations block the DHF binding site. The residues of Chimera 17 overlap with the atoms of the superimposed folate ligand. Since blocking the binding site would reduce enzyme-substrate affinity, we tested whether binding site occlusion correlates with K_M, and indeed, we find our measure of binding site occlusion has a moderately high correlation with K_M ($\rho = 0.65$, $P < 0.001$, Figure 6C), although the correlation is weaker when only non-orthologous chimeras are considered ($\rho = 0.42$, $P = 0.069$). Thus, Chimera 1 differs from other non-orthologous chimeras exhibiting disordered α_B-helices by maintaining an open DHF binding site. And overall, MD simulation measurements show moderately high correlations with thermal stability and substrate binding affinity and can differentiate between high and low catalytic rate and efficiency.

Low catalytic efficiency and decreased intracellular abundance strongly impact the fitness of *E. coli* strains expressing non-orthologous chimeras.

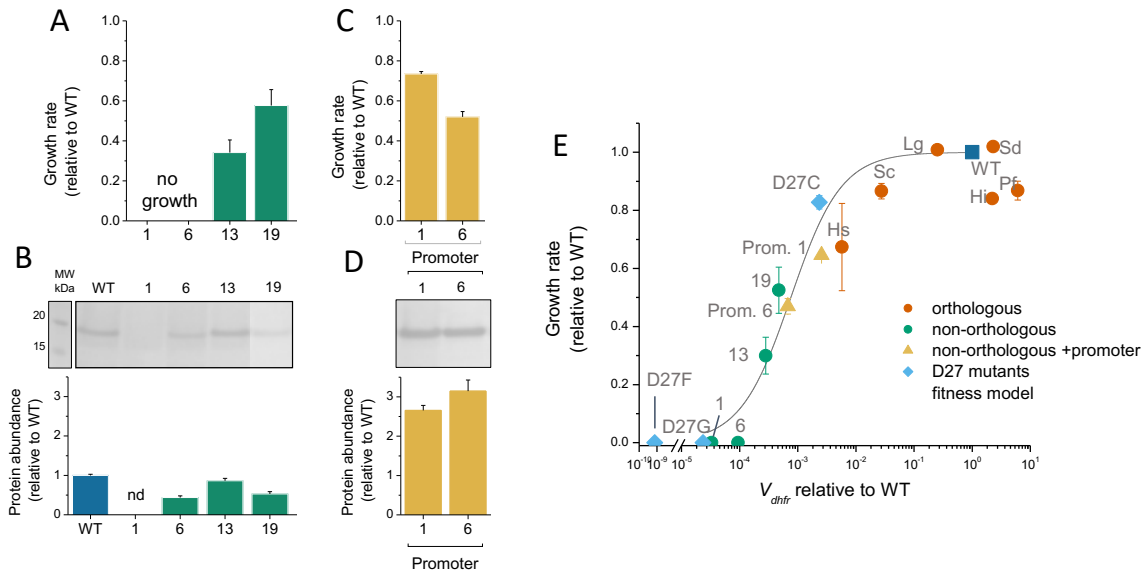

Figure 7: **The fitness of non-orthologous chimeric DHFR strains is markedly reduced due to a combination of decreased catalytic activity and low intracellular protein abundance.** (A) Chromosomal replacement of the gene *folA* with non-orthologous chimeric sequences produces strains that are non-viable or that have decreased fitness. No detectable growth was observed for strains with

Chimeras 1 and 6 after 48 h. The growth rates were normalized with respect to that of the strain with wildtype DHFR, and the represented values are the mean ± standard deviation of three replicates. (B) The intracellular abundance of DHFR is decreased in chimeric strains. The amount of DHFR protein present in the soluble fraction of lysates of chimeric strains was determined by Western blot using anti-*E. coli* DHFR polyclonal antibodies (top panel and Supplementary Figure S10). The represented values are the mean ± standard deviation of three replicates and were normalized with respect to the levels of wild type DHFR. (C), (D) A mutation (C-35T) in the promoter region of the *folA* gene that increases the levels of DHFR expression rescues the growth of strains with Chimeras 1 and 6. (E) The fitness determined for chimeric strains agrees with the DHFR fitness model based on the protein biophysical properties [64]. V_{dhfr}, the normalized flux through DHFR enzymatic reaction was estimated for each chimeric strain using the equation $V_{\mathrm{dhfr}} = k_{cat}/K_M \cdot [\mathrm{DHFR}]$, where [DHFR] is the intracellular concentration DHFR determined for each chimeric strain. The solid line represents the best fit of the data with the equation $y = V_{\mathrm{norm}} / (B + V_{\mathrm{norm}})$ where $V_{\mathrm{norm}} = V_{\mathrm{dhfr}}/V_{\mathrm{dhfr,WT}}$ and $B = 0.001$ [64] (see Materials and Methods).

To determine the extent of modularity in α-helices and its relevance to evolution, it is crucial to evaluate how changes in protein biophysical properties caused by the swapping of structural segments can affect the fitness of an organism. To investigate this, we carried out the chromosomal replacement of the wildtype DHFR-coding gene in *E. coli* (*folA*) with the sequences of the five non-orthologous chimeras with the highest k_{cat}/K_M (Chimeras 1, 2, 6, 13, and 19). Anticipating that the low catalytic function of these proteins would make the strains non-viable, all gene editing steps were made in the presence of folAmix, a metabolic supplement that rescues cells from DHFR deficiency [65]. Upon successful *folA* gene replacement with Chimeras 1, 6, 13, and 19 (but not Chimera 2), we tested for the ability of these strains to grow in the absence of folAmix supplement. Under such conditions, only Chimeras 13 and 19 are viable, though their growth rates are lower than that of the wild type (Figure 7A).

The observation that Chimera 1 cannot grow without supplementation, despite having higher k_{cat}/K_M than that of Chimera 13 (0.011 $s^{-1}\mu M^{-1}$ vs 0.0036 $s^{-1}\mu M^{-1}$), suggested that its intracellular abundance is considerably decreased due to increased degradation by cellular proteases [64,66]. We confirmed this was the case by Western blots with anti-DHFR polyclonal antibodies (Figure 7B and Supplementary Figure S10). No DHFR-specific band was detected in the soluble fraction of lysates from the strain carrying the Chimera 1 gene, whereas such bands were visible for the remaining chimeric strains, albeit less intense than that of the wild type. As in previous studies [40], the intracellular abundance determined for the tested chimeras follows a general inverse correlation with the fluorescence intensity upon *bis*-ANS binding to the purified proteins (Supplementary Figure S11). The *bis*-ANS dye binds preferentially to hydrophobic regions exposed in molten-globule conformations. These results are compatible with the non-native α_B-helix conformations observed in simulations of Chimeras 1 and 6, which may increase the likelihood of molten globule states. The Chimera 19 strain, however, also has a lower intracellular abundance of DHFR despite native-like simulation conformations, which points to aspects of DHFR conformation and *in vivo* interactions either not captured or not accurately rendered by simulations.

We reasoned that enhancing the intracellular protein abundance of strains with Chimeras 1 and 6 could rescue their growth. We engineered versions of these chimeric strains with a promoter mutation (C-35T) that is known to increase the expression of wildtype DHFR several-fold [67]. As expected, the intracellular abundances of Chimeras 1 and 6 increased considerably on the background of the *folA* promoter mutation (Figure 7C), and both strains were able to grow without metabolic supplementation (Figure 7D). The fitness of the chimeric strains is in good agreement with the predictions made by a DHFR fitness model we described previously [64] that is based on intracellular abundance and biophysical properties of the variants (Figure 7E). This model

predicts that none of the other non-orthologous chimeras studied in this work would result in a strain that can grow without folAmix supplementation. These results show that the replacement of the α_B-helix in DHFR with non-orthologous sequences has a strongly deleterious effect on fitness that is due to decreases in catalytic function and stability against intracellular degradation. Nevertheless, the viability of strains with Chimeras 13 and 19 shows that certain, sequence-compatible α-helix swaps can be tolerated *in vivo*, at least to some extent, leading to solutions that are functional in the biological context.

Discussion

The decomposition of protein structure into discrete segments has been a productive method in protein structural analysis due to structural degeneracy in sequence space [68,69]. Sub-domain modularity may be a fundamental property of proteins that facilitates their evolution by promoting robustness and evolvability [10,19,70]. We define modules as independent and transferable protein segments that maintain consistent structure despite changes in structural context. We examine the extent to which this kind of modularity is present in α-helices using the DHFR α_B-helix. The α_B-helix of DHFR has several typical characteristics seen in helix-sheet structures, including alignment of the helix axis with β-strands and the presence of hydrophobic aromatic residues on the helix and hydrophobic aliphatic residues on the β-sheet [71]. Studying this helix provides a highly sensitive way to probe the effects of helix replacements since conformational changes that affect the positioning of the D27 sidechain are expected to impact catalytic activity considerably.

Substitution of an α-helix for structurally similar sequences does not preserve the wildtype-like helix conformation in most cases, with pleiotropic and deleterious effects on thermal stability, catalytic activity, and cellular abundance, which together determine cellular fitness [64,66]. We find principles that govern the outcome of whole-segment substitutions, namely biophysical compatibility, as characterized by EVcouplings and MD simulations. Evolutionary coupling analysis has been shown to predict the experimentally-measured fitness effects of mutations [54]. Here, we show that EVcouplings can also predict the effects of larger scale sequence changes on enzyme catalysis, which partly determines fitness. In our study, the EVcouplings non-epistatic model ΔE correlates slightly better with *in vitro* properties than ΔE calculated by the epistatic model. This may be because simultaneously mutating several residues in the α_B-helix puts chimeric DHFRs in less populated sequence space where epistatic coupling parameters are less accurate.

With regard to MD simulations, we find the strongest correlations between simulation observables and the experimental properties of thermal stability and K_M. This appears to be in line with past results in which computational techniques predict thermal stability more easily than catalytic activity [72,73]. The finding that increased hydrophobic surface area on helix residues strongly correlates with increased thermal stability is curious. The general expectation is that introduction of hydrophobic surface residues is destabilizing [74], but there are examples in which the introduction of hydrophobic surface residues stabilizes a protein [75–77]. Here, changes in helix conformation accompany changes in hydrophobic surface area, and it is difficult to identify a mechanism relating changes hydrophobic surface area to thermal stability. A past study in our group characterized destabilizing mutations in DHFR and found that such mutants formed soluble oligomers at 42 °C that prevented aggregation [78]. Although this previous study found that increased oligomerization at 42 °C was associated with thermally destabilized mutants, oligomerization at higher temperatures was not measured. Variation in hydrophobic surface area may relate to oligomerization of DHFR at increasing temperatures, which would be expected to protect against thermal denaturation.

Of the non-orthologous chimeras, Chimera 19 is particularly notable in that this variant has low sequence identity with corresponding helical region of *E. coli* DHFR and is viable when engineered into *E. coli*, demonstrating that in some cases, the DHFR α_B-helix can be switched for an evolutionarily unrelated sequence.

On the other hand, the low catalytic activity and non-wildtype conformations in most non-orthologous chimeras demonstrate incompatibility. One key issue with low activity chimeras appears to be conformations that occlude the binding site, resulting from non-native hydrophobic interactions with the DHF binding site pocket. This suggests that one constraint on the α_B-helix sequence is avoidance of errant interactions with the hydrophobic DHF binding pocket next to the helix. Chimera 1 is a major exception to the trends. The numerous charged sidechains in the Chimera 1 α_B-helix lead to flexible non-helical conformations that do not block the DHF binding site. The binding of DHF may be sufficient to induce transient folding of the α_B-helix to catalyze the enzymatic reaction, or it may be possible that one of the three other carboxylate side chain residues beside D27 in the Chimera 1 replacement sequence promotes DHF reduction. This structural instability comes at the cost of *in vivo* intracellular abundance, however, and increased expression levels are necessary for *in vivo* growth of the Chimera 1 *E. coli* strain. Thus, although the disordered Chimera 1 α_B-helix is not a sequence that that retains helicity in a new context, Chimera 1 demonstrates an alternative solution for catalytic activity.

The results of α_B-helix substitution in this work can be compared with previously characterized point mutations in the DHFR α_B-helix. Such point mutations have much smaller effects on stability and catalysis compared to helix substitution. For instance, L28R and A26T are trimethoprim resistance mutations that result in a 50% decrease in k_{cat}/K_M, with L28R improving thermal stability [64,79,80]. Single mutants L28F and L28Y are also within an order of magnitude of wildtype k_{cat}/K_M [81]. In this work, we identify W30 and F31 as key residues for maintaining enzyme function and wildtype-like helix conformations, since the chimeras with the highest activity retain at least one of W30 and F31. Single mutants W30G and W30R are comparable in catalytic activity to wildtype DHFR [79]. Mutations at position 31 are more deleterious however. F31Y, F31V, and L28A/F31A have been characterized, with the latter two resulting in two orders of magnitude loss in k_{cat}/K_M [82–84]. All but one of the non-orthologous chimeras are mutated at position 31, and the combined effect of multiple mutations over the entire helix produces a severe loss of catalytic activity.

One framework with which to interpret our results is fold polarity [85,86]. This hypothesis states that a protein fold with higher polarity—a fold characterized by both a well-ordered scaffold and an active site with flexible loops—is simultaneously robust to mutations and also more capable of evolving new functions. *E. coli* DHFR was cited as an example of a low polarity fold since the DHF binding site in DHFR is composed primarily of residues in β-sheets and α-helices. In support of this hypothesis, evolution has found only one function for the DHFR fold [85]. This framework suggests that the DHF binding site in DHFR cannot easily adjust to mutations compared to proteins with more polar folds. It is possible that whole-helix substitutions in locations further from the active site or in proteins with greater fold polarity would be less deleterious to protein function, as was seen with *E. coli* alkaline phosphatase [37]. Our results show that catalysis is more sensitive to helix substitution than thermal stability: although the k_{cat}/K_M values of non-orthologous chimeras are uniformly worse than those of orthologous chimeras, the T_m values of several non-orthologous chimeras are comparable to some of the orthologous chimeras. Furthermore, preserving helical secondary structure upon switching an α-helix is more easily achieved than preserving catalytic activity. MD simulations of non-orthologous chimeras find seven variants that maintain high helicity, and EVcouplings analysis shows that the helical sequences in this study are slightly more compatible replacements for the α_B-helix than are random subsequences taken from the entire *E. coli* proteome.

By testing the fungibility of α-helices, we investigated the principles governing α-helical conformations in proteins: maintenance of interactions by key residues and avoidance of errant interactions. Our results highlight how local destabilization caused by the non-compatibility of the replacement sequences, which leads to low catalytic function and higher intracellular protein degradation, places limits on helix modularity. More broadly, our work reveals the relationship between broad sequence variation and molecular and fitness effects for an essential enzyme that is central to key metabolic pathways.

Materials and Methods

In vitro studies

Selection of non-orthologous α-helices

The α-helical module associated with enzyme activity of *E. coli* DHFR (residues 25-34, Figure 1) was substituted with α-helical modules from non-orthologous proteins taken from a wide variety of species. Only the key catalytic residue, D27, was left unchanged. We deliberately chose proteins with completely different functions unrelated to the function of DHFR. The chimeric proteins that we studied included non-orthologous helices from other bacteria, *Drosophila*, plants, mammals including humans, as well as transposable elements and mitochondria (Table S1).

DHFR chimeric constructs

To be able to carry out in vitro characterization of a considerable number of chimeras, we first developed an efficient two-step PCR strategy to generate various DNA constructs and clone them into an expression plasmid. DNA constructs of chimeric DHFR fused to a C-terminal (6x) His-tag were generated and cloned into the plasmid pFLAG-CTC by a series of PCR reactions using as template the plasmid pFLAG-CTC-wtDHFR, in which wild type E. coli DHFR had been cloned using NdeI/HindIII restriction sites [64]. First, separate fragments of ~550 bp size that flank the helix segment to be modified, designated pFLAG_A and pFLAG_B, were amplified using the set of primers Ampl_pFLAG_H1/ H1_pKD_rev and H1_pKD_for/Ampl_pFLAG_H1_rev, respectively. The backbone of pFLAG-CTC-wtDHFR was amplified using H1_downstream_for/ H1_upstream_rev to produce a linear product (TpFB) where the sequence of the wild type helix segment is absent. Second, for each helix sequence, we designed pairs of complementary primers containing the helix-specific sequence and fixed overhangs with homology either end flanking the helix segment. These pairs of primers were combined with fragments pFLAG_A and pFLAG_B, and a ~1kb linear fragment was amplified by PCR using Ampl_pFLAG_H1/ Ampl_pFLAG_H1_rev. Finally, the amplified product, containing the desired chimeric DHFR sequence, was used to extend the fragment TpB by circular polymerase extension using Phusion polymerase. The non-ligated circular product was used to transform DH5alpha cells, and the sequence of the transformants was verified by Sanger sequencing. For a single colony being sequenced per each chimeric construct, an approximately 75 % success rate was achieved with this methodology. See Supplementary Table S3 for a list of primers used in this work.

Overexpression and purification of chimeric DHFR

DHFR chimeras fused to C-terminal (6x) His-tag were overexpressed using pFLAG expression vector under isopropyl b-D-1-thiogalactopyranoside (IPTG) inducible T7 promoter described previously [64]. Transformed BL21-Gold (DE3) cells were grown in Terrific Broth media (BD) until the cultures reached an optical density of ~0.5, and protein expression was induced overnight at 18˚C following the addition of 100µM IPTG. The recombinant proteins were purified from lysates on Ni-NTA columns (Qiagen).

Steady-state kinetic measurements

DHFR kinetic parameters were measured by progress-curve kinetics as previously described [64]. The measurements were carried at 25˚C in a 96-well plate format, and the decrease in absorbance at 340 nm due to NADPH oxidation was monitored spectrophotometrically using a Biotek Powerwave HT plate reader. Each assay contained 50 mM MES (2-(N-morpholino)ethanesulfonic acid) pH 7.0, 1mM dithiothreitol, and 150 µM NADPH and purified DHFR chimeras. The reactions were started by the addition of 10/60 µM dihydrofolate (100 µL final volume in the assay).

Thermal denaturation

Melting temperatures were determined by differential scanning fluorimetry using the fluorescent dye Sypro Orange (Invitrogen). DHFR solutions (5 µM) were prepared in 50 mM phosphate buffer, 1 mM DTT at pH 7.0 in the presence of 100 µM NADPH, and the dye sypro orange was added at 0.5× final concentration). Thermal denaturation was performed using a Biorad CFX96 Touch with a temperature ramp of 1˚C/min set between 25 and 90˚C and fluorescence intensity was measured every 0.2˚C using the instrument's FRET channel. The denaturation curves were analyzed using the online software Calfitter v1.3 [87].

Bis-ANS fluorescence

Purified chimeric proteins (2 µM) were incubated for 5 minutes at 37˚C in the presence of 12 µM *bis*-ANS, in 20 mM potassium phosphate buffer pH7.2. The fluorescence emission spectra were then recorded between 400 and 600 nm (slit 5 nm) upon excitation at 395 nm (slit 2.5 nm). The emission band was integrated, and the fluorescence of *bis*-ANS alone was subtracted. Results are normalized with respect to the wild type protein and represent the average of 3 measurements.

Size exclusion chromatography

A Superdex 10/300 GL analytical column (Cytiva) was equilibrated with 20mM tris, 150 mM NaCl, pH 8.0 at a flow rate of 0.8 mL/min. 75 µL of a solution of each protein (20 µM) and NADPH (100 µM) were sequentially injected using an Akta 905 Autosampler. As molecular standards, we used a mixture containing bovine serum albumin (66 kDa), carbonic anhydrase (29 kDa), cytochrome c (12.3 kDa), and aprotinin (6.5 kDa).

Computational methods

Agadir score

To predict the helical content of the sequences that were used to replace the α-helix in DHFR, we used the Agadir score [45–49], which was computed based on the 10-residue (positions 25-34) subsequence corresponding to each replacement helix. The parameters used in the calculations were: temperature = 37˚C, ionic strength = 0.1 M, pH = 7.0.

EVcouplings

We queried the EVcouplings web server [42] with the *E. coli* DHFR sequence, which yielded 19,912 aligned sequences, and we then used plmc from the Marks Lab [54] to obtain the parameters $\boldsymbol{h}_i$ and $\boldsymbol{J}_{ij}$, calculated over 12,682 effective (non-redundant) sequences. Mutational effects, or Δ*E*, were calculated using the EVcouplings Python package [42] as $\Delta E = H_{\text{eff}}(\{\sigma^{\text{mut}}\}) - H_{\text{eff}}(\{\sigma^{\text{wt}}\})$ where H_{eff} is defined according to the epistatic or non-epistatic model.

To calculate EVcouplings Δ*E* for random protein sequences, *E. coli* K-12 MG1655 coding sequences (assembly ASM584v2) were obtained from the RefSeq database [88]. For all 10-residue subsequences in all non-pseudo gene coding sequences, EVcouplings Δ*E* values were calculated for the substitution of each subsequence for residues 25-34 in *E. coli* DHFR, while preserving position 27 as D.

Molecular dynamics simulation

Replica exchange with solute tempering (REST2) is an enhanced sampling simulation method in which different replicas use scaled versions of the potential energy [59]. For a replica m, the potential energy function in a part of the system is scaled by $\lambda_m \equiv \beta_m/\beta_0$:

$$E_m(X) = \frac{\beta_m}{\beta_0} E_{pp}(X) + \sqrt{\frac{\beta_m}{\beta_0}} E_{pw}(X) + E_{ww}(X)$$

where w corresponds to the remainder of the system outside the scaled region of interest, p. All replicas are simulated at (inverse) temperature $\beta_0 = 1/k_B T_0$. β_m represents an effective temperature for replica m within the scaled region. Replica exchange solute tempering simulations used an effective temperature ladder of 300-550 K and an exchange interval of 500 steps (1 ps). For Chimeras 15-20, an effective temperature ladder of 300-600 K was used instead. All replica exchange simulations used 8 replicas, and the acceptance rate between the two lowest rungs was at least 0.15 (acceptance rates varied by DHFR variant).

Structure preparation began with an X-ray crystal structure of E. coli DHFR with NADP+ and folic acid bound (PDB ID 4NX6). MOE [57] was used to perform structure preparation, with the PFROSST forcefield loaded. The crystallographic waters, manganese ion, and beta-mercaptoethanol molecule were deleted. The Structure Preparation tool was used to optimize side chain protonation states. His 45 and 114 were protonated at the ε nitrogen, His 124 and 149 were protonated at the δ nitrogen, and His 141 was doubly protonated. From this prepared structure, all DHFR chimeras were prepared using the Homology Model tool. FASTA sequences of chimeric DHFRs were loaded, and the prepared DHFR complex was used as the template for homology modeling, with ligand atoms set as part of the molecular environment. The NADP+ and folate ligands were then removed to obtain the apo structures of each DHFR variant.

Proteins structures were solvated with water in cubic boxes with at least a 10 Å distance between box edges and protein atoms. Na+ and Cl- ions were added to the system at a concentration of 150 mM such that the system charge was neutralized. The AMBER ff14SB forcefield was used to model protein atoms, and the TIP3P forcefield was used for the water [89,90]. All MD simulations were performed using NAMD version 2.14 [58], and the included REST2 script was used for replica exchange simulations. All atoms in protein residues 24 through 36 were tempered. The nonbonded interaction cutoff distance was set to 9 Å; the simulation timestep was 2 fs, and all bonds to hydrogens were constrained. Long-range electrostatics were calculated using particle-mesh Ewald with a grid spacing of 1.0 Å. Simulations were run in the NPT ensemble at 300 K and 1 atm. The temperature was maintained by a Langevin thermostat with a damping value of 1 ps^{-1}, and the pressure was maintained by a Langevin piston with a period of 100 fs and decay of 50 fs. Simulation coordinates were saved every 0.5 ps.

Initial equilibration was performed by minimizing the system for 5000 steps and then running 5 ns of simulation. Simulation coordinates from the end of the equilibration trajectory were used to initiate replica exchange simulations. Replica exchange simulations were run for 300 ns, with some simulations extended to 375 ns to achieve converged helix RMSD values in the last 100 ns of simulations (Supplementary Figure S5). The last 100 ns of simulations of the non-scaled replicas ($\beta_m = \beta_0$) were used for measurements.

Simulation measurements and analysis

Simulation trajectories were analyzed using MDTraj version 1.9.4 [62]. The *E. coli* wildtype DHFR crystal structure (PDB ID 4NX6) was used as the reference for RMSD measurements, and the following are the residue selections: adenosine-binding subdomain: 38-88; adenosine-binding subdomain core: 39-43, 58-62, 73-85; major subdomain: 1-37, 89-159; major subdomain core: 2-8, 91-93, 109-115, 133-141, 151-158; α_B-helix: 25-35; Met 20 loop: 9-24; FG loop: 116-132; GH loop: 142-150.

For all RMSD measurements except for RMSD of the entire adenosine-binding and major subdomains, measurements were performed by first aligning simulation structures to the crystal structure reference using the Cα-atoms of core residues of the major subdomain. Then, RMSD measurements were performed over the Cα-atoms of the selected residues of interest.

The Shrake-Rupley method [91] within MDTraj was used to measure the solvent accessible surface area (SASA) with default parameters (probe radius 1.4 Å and 960 sphere points). This measures the SASA of each atom in a

protein structure. A simple heuristic, following Shrake and Rupley, was used to classify protein atoms as either non-polar or polar: A hydrogen atom is nonpolar only if it is bonded to carbon. A carbon atom is nonpolar except when it is bonded to two or more heteroatoms. Sulfur atoms are nonpolar. All other atoms are polar. The SASA values of non-polar atoms of helix atoms (residues 25-35) were summed to determine the hydrophobic SASA of the helix.

Binding site occlusion was calculated as follows. For each simulation frame, the protein coordinates were aligned to the *E. coli* DHFR crystal structure (PDB ID 4NX6). The coordinates of the folate ligand from the crystal structure were then added to the simulation frame. A 3D grid with a spacing of 0.1 Å was overlaid in the vicinity of the ligand. The volume available to the folate ligand was then measured by the number of grid points that are within the van der Waals radius of any ligand atom but not within the van der Waals radius of any protein atom. Occlusion was calculated as the fraction of ligand volume blocked by protein atoms.

In vivo studies

Chromosomal replacement of non-orthologous chimeras

Non-orthologous chimeric strains were created by lambda-red recombination [92] according to a previously described procedure [65]. We used an approach similar to the one described above to introduce the chimeric sequences in a pKD13-derived plasmid containing the entire regulatory and coding sequence of the folA gene, flanked by kanR and cmR gene markers and approximately 1kb homologous region of both upstream and downstream chromosomal genes flanking folA gene. The entire cassette was amplified using primers PCRseq_KefC_for2 and PCRseq_apaH_rev and transformed into BW25113 cells with induced Red helper plasmid pKD46, and cells were recovered in SOC medium containing 0.25x folAmix (adenine 5 ug/mL, inosine 20 ug/mL, thymine 50 ug/mL, methionine 5 ug/mL and glycine 5 ug/mL). Transformants were selected in agar plates containing kanamycin, chloramphenicol, and folAmix. To confirm the correct integration of the desired mutations, folA locus was amplified by PCR (primers Ampl_RRfolA_for, Ampl_RRfolA_rev) and Sanger-sequenced using primer PCRseq_RRfolA_rev. Plasmid pKD46 was removed by plating cells at 37 °C twice in the absence of antibiotic selection. The “wild type strain” referred to in this study represents a BW25113 strain in which the original folA locus was likewise replaced by a replacement cassette containing the sequence of wild type DHFR flanked by kanR and cmR antibiotic resistance markers. Three colonies of each chimeric or wild type E coli strain were picked from agar plates and grown in M9 minimal media supplemented with 0.25x folAmix until saturation. Glycerol was added (10% final volume) to the grown cultures and various aliquots were stored at -80˚C for subsequent in *vivo studies.*

Growth measurements

Overnight cultures were prepared by inoculating 200 µL of M9 minimal media. supplemented with 0.8 g/L glucose, 1mM $MgSO_4$, 0.1mM $CaCl_2$, 50 µg/mL kanamycin, 34µg/mL chloramphenicol and 0.25x folAmix, with 5 uL of thawed -80 ˚C glycerol stocks. The next morning, the cultures were diluted into fresh M9 minimal media (composition as above) to an optical density (OD) of approximately 0.01. After a period of 6-8h growth, the cultures were washed with M9 minimal medial without folAmix by 3 cycles of centrifugation/resuspending the cell pellets 3 times. After the final resuspension, the OD of the cultures was normalized to 0.0015. A 96-well microplate containing 100 uL of M9 minimal media, supplemented with 0.8g/L glucose, 1mM $MgSO_4$, 0.1mM $CaCl_2$, 50 µg/mL kanamycin, 34µg/mL chloramphenicol, and different concentrations of folAmix, was inoculated with 50 uL/well of the normalized cultures. The growth measurements were performed at 37˚C with a Tecan Infinite M200 Pro microplate reader.

Intracellular protein abundance

The intracellular abundance of chimeric DHFR was determined by Western blotting. Chimeric strains were grown at 37˚C in M9 minimal media supplemented with 2 g/L glucose, 50 µg/mL kanamycin, 34µg/mL chloramphenicol, and 0.25x folAmix. When the optical density (OD) was approximately 0.3 the cultures (5mL) were centrifuged and the supernatant discarded. Based on the OD of the cultures before centrifugation, the pellets were resuspended to an OD of 11 in lysis buffer consisting of 100mM HEPES pH 7.3, 1x Popculture reagent (Millipore), 10x Complete protease inhibitor cocktail (Sigma), 1 mg/mL lysozyme, 1U/mL Benzonase (Sigma) and 1mM DTT. The tubes were sonicated in a water bath for 20-30 min until complete lysis and the insoluble material was removed by centrifugation at 20.000xg for 10 min. 30 µL of the cleared supernatants were mixed with 15 µL SDS-PAGE loading buffer (3X) containing β-mercaptoethanol. The samples were incubated for 20 min at 37˚C and loaded on an SDS-PAGE gel together with colored MW markers. After the separation, the gel was cut in half and the low-MW part (below 37kDa) was transferred to a nitrocellulose membrane. The DHFR-specific band was identified using rabbit polyclonal anti-DHFR antibodies (1/20000 dilution) followed by alkaline phosphatase-linked anti-rabbit secondary antibodies (1/10000 dilution). The band was revealed with Novex™ AP Chromogenic Substrate (BCIP⁄ NBT) (Thermofisher Scientific). The other half of the SDS-PAGE gel with the high-MW bands was stained with Bio-Safe™ Coomassie Stain (Biorad). The intensity of Western-blot bands was quantified by the ImageJ software [93], and was normalized by the total protein content, which was estimated by the total intensity of each corresponding lane in the Coomassie-stained high-MW half of the SDS-PAGE gel.

DHFR fitness model

A fitness model for DHFR was described previously [64]. The flux through DHFR enzymatic reaction was estimated for each chimeric strain using the equation $V_{\mathrm{dhfr}} = k_{cat}/K_M \cdot [\mathrm{DHFR}]$, where [DHFR] is the intracellular concentration determined for the chimeric protein. Flux values were normalized with respect to that of the wild type strain. To estimate the flux for the wild type strain at different concentrations of trimethoprim, the calculated flux was corrected by a factor, $1 / (1 + \alpha \cdot [\mathrm{TMP}]_{\mathrm{medium}} / K_{\mathrm{i}})$, where $[\mathrm{TMP}]_{\mathrm{medium}}$ is the concentration of trimethoprim in the growth medium, α is the ratio between the intracellular trimethoprim concentration and that of the growth medium, and K_{i} is the trimethoprim inhibition constant (0.94 nM for wild type DHFR). The parameter α must be determined independently for different strains since cell permeability and overall sensitivity to trimethoprim are expected to vary. For the strain BW25113 used in this work, α = 0.4. The growth data determined at different trimethoprim concentrations were fitted to the equation $y = V_{\mathrm{norm}} / (B + V_{\mathrm{norm}})$ where $V_{\mathrm{norm}} = V_{\mathrm{dhfr}}/V_{\mathrm{dhfr,WT}}$ and $B = 0.001$.

CRediT authorship contribution statement

Victor Y. Zhao: Investigation, Methodology, Software, Visualization, Writing – original draft. **Joao V. Rodrigues**: Investigation, Methodology, Visualization, Writing – original draft. **Elena R. Lozovsky**: Investigation, Methodology. **Daniel L. Hartl**: Conceptualization, Supervision, Writing - Review & Editing. **Eugene I. Shakhnovich**: Conceptualization, Funding acquisition, Supervision, Writing - Review & Editing.

Acknowledgments

This work was supported by the National Institutes of Health (NIGMS 5R01GM068670-16). The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562 [94]. Specifically, this work used the Bridges cluster at the Pittsburg Supercomputing Center and the Comet cluster

at the San Diego Supercomputer Center through allocation BIO200028. Plots were created using the Seaborn and Matplotlib Python packages [95,96]. We are grateful to Rostam M. Razban for a critical reading of the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Bornberg-Bauer, E., and Albà, M.M. (2013). Dynamics and adaptive benefits of modular protein evolution. *Curr. Opin. Struct. Biol.* 23, 459–466.
2. Levitt, M. (2009). Nature of the protein universe. *Proc. Natl. Acad. Sci.* 106, 11079–11084.
3. Dohmen, E., Klasberg, S., Bornberg-Bauer, E., Perrey, S., and Kemena, C. (2020). The modular nature of protein evolution: domain rearrangement rates across eukaryotic life. *BMC Evol. Biol.* 20, 30.
4. Wagner, G.P., Pavlicev, M., and Cheverud, J.M. (2007). The road to modularity. *Nat. Rev. Genet.* 8, 921–931.
5. Lees, J.G., Dawson, N.L., Sillitoe, I., and Orengo, C.A. (2016). Functional innovation from changes in protein domains and their combinations. *Curr. Opin. Struct. Biol.* 38, 44–52.
6. Bashton, M., and Chothia, C. (2007). The Generation of New Protein Functions by the Combination of Domains. *Structure* 15, 85–99.
7. Peisajovich, S.G., Garbarino, J.E., Wei, P., and Lim, W.A. (2010). Rapid Diversification of Cell Signaling Phenotypes by Modular Domain Recombination. *Science* 328, 368–372.
8. Hleap, J.S., Susko, E., and Blouin, C. (2013). Defining structural and evolutionary modules in proteins: a community detection approach to explore sub-domain architecture. *BMC Struct. Biol.* 13, 20.
9. Grishin, N.V. (2001). Fold Change in Evolution of Protein Structures. *J. Struct. Biol.* 134, 167–185.
10. Peisajovich, S.G., Rockah, L., and Tawfik, D.S. (2006). Evolution of new protein topologies through multistep gene rearrangements. *Nat. Genet.* 38, 168–174.
11. Jiang, H., and Blouin, C. (2007). Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. *BMC Bioinformatics* 8, 444.
12. Zhang, Z., Wang, Y., Wang, L., and Gao, P. (2010). The Combined Effects of Amino Acid Substitutions and Indels on the Evolution of Structure within Protein Families. *PLOS ONE* 5, e14316.
13. Guo, B., Zou, M., and Wagner, A. (2012). Pervasive Indels and Their Evolutionary Dynamics after the Fish-Specific Genome Duplication. *Mol. Biol. Evol.* 29, 3005–3022.
14. Han, K.F., and Baker, D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl. Acad. Sci.* 93, 5814–5818.
15. Tendulkar, A.V., Krallinger, M., Torre, V. de la, López, G., Wangikar, P.P., and Valencia, A. (2010). FragKB: Structural and Literature Annotation Resource of Conserved Peptide Fragments and Residues. *PLOS ONE* 5, e9679.
16. Vanhee, P., Verschueren, E., Baeten, L., Stricher, F., Serrano, L., Rousseau, F., and Schymkowitz, J. (2011). BriX: a database of protein building blocks for structural analysis, modeling and design. *Nucleic Acids Res.* 39, D435–D442.
17. Vallat, B., Madrid-Aliste, C., and Fiser, A. (2015). Modularity of Protein Folds as a Tool for Template-Free Modeling of Structures. *PLOS Comput. Biol.* 11, e1004419.
18. Mackenzie, C.O., Zhou, J., and Grigoryan, G. (2016). Tertiary alphabet for the observable protein structural universe. *Proc. Natl. Acad. Sci.* 113, E7438–E7447.
19. Nepomnyachiy, S., Ben-Tal, N., and Kolodny, R. (2017). Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proc. Natl. Acad. Sci.* 114, 11703–11708.

20. Konagurthu, A.S., Subramanian, R., Allison, L., Abramson, D., Stuckey, P.J., Garcia de la Banda, M., and Lesk, A.M. (2021). Universal Architectural Concepts Underlying Protein Folding Patterns. *Front. Mol. Biosci.* 7.
21. Söding, J., and Lupas, A.N. (2003). More than the sum of their parts: On the evolution of proteins from peptides. *BioEssays* 25, 837–846.
22. Alva, V., Söding, J., and Lupas, A.N. (2015). A vocabulary of ancient peptides at the origin of folded proteins. *eLife* 4, e09410.
23. Sobolevsky, Y., Frenkel, Z.M., and Trifonov, E.N. (2007). Combinations of Ancestral Modules in Proteins. *J. Mol. Evol.* 65, 640–650.
24. Yadid, I., and Tawfik, D.S. (2007). Reconstruction of Functional β-Propeller Lectins via Homo-oligomeric Assembly of Shorter Fragments. *J. Mol. Biol.* 365, 10–17.
25. Romero Romero, M.L., Rabin, A., and Tawfik, D.S. (2016). Functional Proteins from Short Peptides: Dayhoff's Hypothesis Turns 50. *Angew. Chem. Int. Ed.* 55, 15966–15971.
26. Smock, R.G., Yadid, I., Dym, O., Clarke, J., and Tawfik, D.S. (2016). De Novo Evolutionary Emergence of a Symmetrical Protein Is Shaped by Folding Constraints. *Cell* 164, 476–486.
27. Broom, A., Trainor, K., MacKenzie, D.W., and Meiering, E.M. (2016). Using natural sequences and modularity to design common and novel protein topologies. *Curr. Opin. Struct. Biol.* 38, 26–36.
28. Bai, Y., Sosnick, T.R., Mayne, L., and Englander, S.W. (1995). Protein folding intermediates: native-state hydrogen exchange. *Science* 269, 192–197.
29. Petersen, J.M., Skalicky, J.J., Donaldson, L.W., McIntosh, L.P., Alber, T., and Graves, B.J. (1995). Modulation of transcription factor Ets-1 DNA binding: DNA-induced unfolding of an alpha helix. *Science* 269, 1866–1869.
30. Cheng, P., Liu, D., Chee, P.X., Yang, D., and Long, D. (2019). Atomistic Insights into the Functional Instability of the Second Helix of Fatty Acid Binding Protein. *Biophys. J.* 117, 239–246.
31. Wang, H., Logan, D.T., Danielsson, J., and Oliveberg, M. (2020). Exposing the distinctive modular behavior of β-strands and α-helices in folded proteins. *Proc. Natl. Acad. Sci.* 117, 28775–28783.
32. Sivaramakrishnan, S., Spink, B.J., Sim, A.Y.L., Doniach, S., and Spudich, J.A. (2008). Dynamic charge interactions create surprising rigidity in the ER/K α-helical protein motif. *Proc. Natl. Acad. Sci.* 105, 13356–13361.
33. Swanson, C.J., and Sivaramakrishnan, S. (2014). Harnessing the Unique Structural Properties of Isolated α-Helices *. *J. Biol. Chem.* 289, 25460–25467.
34. Wolny, M., Batchelor, M., Knight, P.J., Paci, E., Dougan, L., and Peckham, M. (2014). Stable Single α-Helices Are Constant Force Springs in Proteins*. *J. Biol. Chem.* 289, 27825–27835.
35. Abrusán, G., and Marsh, J.A. (2016). Alpha Helices Are More Robust to Mutations than Beta Strands. *PLOS Comput. Biol.* 12, e1005242.
36. Goldman, N., Thorne, J.L., and Jones, D.T. (1998). Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution. *Genetics* 149, 445–458.
37. DuBose, R.F., and Hartl, D.L. (1989). An experimental approach to testing modular evolution: directed replacement of alpha-helices in a bacterial protein. *Proc. Natl. Acad. Sci. U. S. A.* 86, 9966–9970.
38. Teterina, N.L., Gorbalenya, A.E., Egger, D., Bienz, K., Rinaudo, M.S., and Ehrenfeld, E. (2006). Testing the modularity of the N-terminal amphipathic helix conserved in picornavirus 2C proteins and hepatitis C NS5A protein. *Virology* 344, 453–467.
39. Rodrigues, J.V., Ogbunugafor, C.B., Hartl, D.L., and Shakhnovich, E.I. (2019). Chimeric dihydrofolate reductases display properties of modularity and biophysical diversity. *Protein Sci.* 28, 1359–1367.
40. Bershtein, S., Mu, W., Serohijos, A.W.R., Zhou, J., and Shakhnovich, E.I. (2013). Protein Quality Control Acts on Folding Intermediates to Shape the Effects of Mutations on Organismal Fitness. *Mol. Cell* 49, 133–144.
41. Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190.
42. Hopf, T.A., Green, A.G., Schubert, B., Mersmann, S., Schärfe, C.P.I., Ingraham, J.B., Toth-Petroczy, A., Brock, K., Riesselman, A.J., Palmedo, P., Kang, C., Sheridan, R., Draizen, E.J., Dallago, C., Sander, C., and Marks, D.S.

(2019). The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* 35, 1582–1584.
43. Schnell, J.R., Dyson, H.J., and Wright, P.E. (2004). Structure, Dynamics, and Catalytic Function of Dihydrofolate Reductase. *Annu. Rev. Biophys. Biomol. Struct.* 33, 119–140.
44. Villafranca, J.E., Howell, E.E., Voet, D.H., Strobel, M.S., Ogden, R.C., Abelson, J.N., and Kraut, J. (1983). Directed mutagenesis of dihydrofolate reductase. *Science* 222, 782–788.
45. Muñoz, V., and Serrano, L. (1994). Elucidating the folding problem of helical peptides using empirical parameters. *Nat. Struct. Biol.* 1, 399–409.
46. Muñoz, V., and Serrano, L. (1995). Elucidating the Folding Problem of Helical Peptides using Empirical Parameters. II. Helix Macrodipole Effects and Rational Modification of the Helical Content of Natural Peptides. *J. Mol. Biol.* 245, 275–296.
47. Muñoz, V., and Serrano, L. (1995). Elucidating the Folding Problem of Helical Peptides using Empirical Parameters. III. Temperature and pH Dependence. *J. Mol. Biol.* 245, 297–308.
48. Muñoz, V., and Serrano, L. (1997). Development of the multiple sequence approximation within the AGADIR model of α-helix formation: Comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers* 41, 495–509.
49. Lacroix, E., Viguera, A.R., and Serrano, L. (1998). Elucidating the folding problem of α-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters11Edited by A. R. Fersht. *J. Mol. Biol.* 284, 173–191.
50. Shakhnovich, E.I., and Gutin, A.M. (1993). A new approach to the design of stable proteins. *Protein Eng. Des. Sel.* 6, 793–800.
51. Dokholyan, N.V., and Shakhnovich, E.I. (2001). Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* 312, 289–307.
52. Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE* 6, e28766.
53. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.* 108, E1293–E1301.
54. Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P.I., Springer, M., Sander, C., and Marks, D.S. (2017). Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35, 128–135.
55. Grantham, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* 185, 862–864.
56. Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8. .
57. (2021). *Molecular Operating Environment (MOE)*, Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.
58. Phillips, J.C., Hardy, D.J., Maia, J.D.C., Stone, J.E., Ribeiro, J.V., Bernardi, R.C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., McGreevy, R., Melo, M.C.R., Radak, B.K., Skeel, R.D., Singharoy, A., Wang, Y., Roux, B., Aksimentiev, A., Luthey-Schulten, Z., Kalé, L.V., Schulten, K., Chipot, C., and Tajkhorshid, E. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* 153, 044130.
59. Wang, L., Friesner, R.A., and Berne, B.J. (2011). Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J. Phys. Chem. B* 115, 9431–9438.
60. Maffucci, I., Laage, D., Stirnemann, G., and Sterpone, F. (2020). Differences in thermal structural changes and melting between mesophilic and thermophilic dihydrofolate reductase enzymes. *Phys. Chem. Chem. Phys.* 22, 18361–18373.
61. Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637.

62. McGibbon, R.T., Beauchamp, K.A., Harrigan, M.P., Klein, C., Swails, J.M., Hernández, C.X., Schwantes, C.R., Wang, L.-P., Lane, T.J., and Pande, V.S. (2015). MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* 109, 1528–1532.
63. Kasper, J.R., Liu, P.-F., and Park, C. (2014). Structure of a partially unfolded form of Escherichia coli dihydrofolate reductase provides insight into its folding pathway. *Protein Sci.* 23, 1728–1737.
64. Rodrigues, J.V., Bershtein, S., Li, A., Lozovsky, E.R., Hartl, D.L., and Shakhnovich, E.I. (2016). Biophysical principles predict fitness landscapes of drug resistance. *Proc. Natl. Acad. Sci.* 113, E1470–E1478.
65. Rodrigues, J.V., and Shakhnovich, E.I. (2019). Adaptation to mutational inactivation of an essential gene converges to an accessible suboptimal fitness peak. *eLife* 8.
66. Bershtein, S., Serohijos, A.W.R., Bhattacharyya, S., Manhart, M., Choi, J.-M., Mu, W., Zhou, J., and Shakhnovich, E.I. (2015). Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria. *PLOS Genet.* 11, e1005612.
67. Toprak, E., Veres, A., Michel, J.-B., Chait, R., Hartl, D.L., and Kishony, R. (2011). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat. Genet.* 44, 101–105.
68. Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* 20, 681–697.
69. Lapenta, F., and Jerala, R. (2020). Design of novel protein building modules and modular architectures. *Curr. Opin. Struct. Biol.* 63, 90–96.
70. Rorick, M.M., and Wagner, G.P. (2011). Protein Structural Modularity and Robustness Are Associated with Evolvability. *Genome Biol. Evol.* 3, 456–475.
71. Hu, C., and Koehl, P. (2010). Helix-sheet packing in proteins. *Proteins* 78, 1736–1747.
72. Yang, J., Naik, N., Patel, J.S., Wylie, C.S., Gu, W., Huang, J., Ytreberg, F.M., Naik, M.T., Weinreich, D.M., and Rubenstein, B.M. (2020). Predicting the viability of beta-lactamase: How folding and binding free energies correlate with beta-lactamase fitness. *PLOS ONE* 15, e0233509.
73. Broom, A., Trainor, K., Jacobi, Z., and Meiering, E.M. (2020). Computational Modeling of Protein Stability: Quantitative Analysis Reveals Solutions to Pervasive Problems. *Structure* .
74. Schwehm, J.M., Kristyanne, E.S., Biggers, C.C., and Stites, W.E. (1998). Stability Effects of Increasing the Hydrophobicity of Solvent-Exposed Side Chains in Staphylococcal Nuclease. *Biochemistry* 37, 6939–6948.
75. van den Burg, B., Dijkstra, B.W., Vriend, G., van der Vinne, B., Venema, G., and Eijsink, V.G.H. (1994). Protein stabilization by hydrophobic interactions at the surface. *Eur. J. Biochem.* 220, 981–985.
76. Machius, M., Declerck, N., Huber, R., and Wiegand, G. (2003). Kinetic Stabilization of Bacillus licheniformis α-Amylase through Introduction of Hydrophobic Residues at the Surface*. *J. Biol. Chem.* 278, 11546–11553.
77. Islam, M.M., Kobayashi, K., Kidokoro, S.-I., and Kuroda, Y. (2019). Hydrophobic surface residues can stabilize a protein through improved water–protein interactions. *FEBS J.* 286, 4122–4134.
78. Bershtein, S., Mu, W., and Shakhnovich, E.I. (2012). Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. *Proc. Natl. Acad. Sci.* 109, 4857–4862.
79. Tamer, Y.T., Gaszek, I.K., Abdizadeh, H., Batur, T.A., Reynolds, K.A., Atilgan, A.R., Atilgan, C., and Toprak, E. (2019). High-Order Epistasis in Catalytic Power of Dihydrofolate Reductase Gives Rise to a Rugged Fitness Landscape in the Presence of Trimethoprim Selection. *Mol. Biol. Evol.* 36, 1533–1550.
80. Abdizadeh, H., Tamer, Y.T., Acar, O., Toprak, E., Atilgan, A.R., and Atilgan, C. (2017). Increased substrate affinity in the Escherichia coli L28R dihydrofolate reductase mutant causes trimethoprim resistance. *Phys. Chem. Chem. Phys.* 19, 11416–11428.
81. Thompson, S., Zhang, Y., Ingle, C., Reynolds, K.A., and Kortemme, T. (2020). Altered expression of a quality control protease in E. coli reshapes the in vivo mutational landscape of a model enzyme. *eLife* 9, e53476.
82. Chen, J.-T., Taira, K., Tu, C.-P.D., and Benkovic, S.J. (1987). Probing the functional role of phenylalanine-31 of Escherichia coli dihydrofolate reductase by site-directed mutagenesis. *Biochemistry* 26, 4093–4100.
83. Huang, Z., Wagner, C.R., and Benkovic, S.J. (1994). Nonadditivity of mutational effects at the folate binding site of Escherichia coli dihydrofolate reductase. *Biochemistry* 33, 11576–11585.

84. Reynolds, K.A., McLaughlin, R.N., and Ranganathan, R. (2011). Hot spots for allosteric regulation on protein surfaces. *Cell* 147, 1564–1575.
85. Dellus-Gur, E., Toth-Petroczy, A., Elias, M., and Tawfik, D.S. (2013). What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-offs. *J. Mol. Biol.* 425, 2609–2621.
86. Tóth-Petróczy, Á., and Tawfik, D.S. (2014). The robustness and innovability of protein folds. *Curr. Opin. Struct. Biol.* 26, 131–138.
87. Mazurenko, S., Stourac, J., Kunka, A., Nedeljkovic, S., Bednar, D., Prokop, Z., and Damborsky, J. (2018). CalFitter: a web server for analysis of protein thermal denaturation data. *Nucleic Acids Res.* 46, W344–W349.
88. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., and Pruitt, K.D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733-745.
89. Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* 11, 3696–3713.
90. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79, 926–935.
91. Shrake, A., and Rupley, J.A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* 79, 351–371.
92. Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci.* 97, 6640–6645.
93. Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* 9, 671–675.
94. Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G.D., Roskies, R., Scott, J.R., and Wilkins-Diehr, N. (2014). XSEDE: Accelerating Scientific Discovery. *Comput. Sci. Eng.* 16, 62–74.
95. Waskom, M.L. (2021). seaborn: statistical data visualization. *J. Open Source Softw.* 6, 3021.
96. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 9, 90–95.