peds
protein engineering design & selection

Protein Engineering, Design and Selection, 2019, vol. 32 no. 11, pp. 489–501
doi: 10.1093/protein/gzaa012
Advance Access Publication Date: 6 July 2020
Original Article

OXFORD

Original Article

Protein tolerance to random circular permutation correlates with thermostability and local energetics of residue-residue contacts

Joshua T. Atkinson[1,2], Alicia M. Jones[3], Vikas Nanda[4], and Jonathan J. Silberg[2,5,6,*]

[1]Systems, Synthetic, and Physical Biology Graduate Program, Rice University, 6100 Main Street, MS-180, Houston, TX 77005, USA, [2]Department of BioSciences, Rice University, 6100 Main Street, MS-140, Houston, TX 77005, USA, [3]Biochemistry and Cell Biology Graduate Program, Rice University, 6100 Main Street, MS-140, Houston, TX 77005, USA, [4]Center for Advanced Biotechnology and Medicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA, [5]Department of Bioengineering, Rice University, 6100 Main Street, MS-142, Houston, TX 77005, USA, and [6]Department of Chemical and Biomolecular Engineering, Rice University, 6100 Main Street, MS-362, Houston, TX 77005, USA

*To whom correspondence should be addressed. E-mail: joff@rice.edu

Paper edited by: Daniel Otzen

Received 18 February 2020; Revised 13 April 2020; Accepted 15 April 2020

Abstract

Adenylate kinase (AK) orthologs with a range of thermostabilities were subjected to random circular permutation, and deep mutational scanning was used to evaluate where new protein termini were nondisruptive to activity. The fraction of circularly permuted variants that retained function in each library correlated with AK thermostability. In addition, analysis of the positional tolerance to new termini, which increase local conformational flexibility, showed that bonds were either functionally sensitive to cleavage across all homologs, differentially sensitive, or uniformly tolerant. The mobile AMP-binding domain, which displays the highest calculated contact energies, presented the greatest tolerance to new termini across all AKs. In contrast, retention of function in the lid and core domains was more dependent upon AK melting temperature. These results show that family permutation profiling identifies primary structure that has been selected by evolution for dynamics that are critical to activity within an enzyme family. These findings also illustrate how deep mutational scanning can be applied to protein homologs in parallel to differentiate how topology, stability, and local energetics govern mutational tolerance.

Key words: combinatorial library, deep mutational scanning, mutational tolerance, topological mutation, transposon mutagenesis

Introduction

Protein folding occurs through a funneled energy landscape where the top of the funnel represents the ensemble of unfolded polypeptide conformations and the bottom represents the folded native-state ensemble with the lowest free energy (Zhuravlev and Papoian, 2010; Wolynes, 2015). To encode the native-state ensemble, the residues at each native position have been selected through evolution to minimize the average free energy of their contacts. However, some residue-residue contacts can present free energies that are high relative to all other possible contacts that could exist at the same location (Ferreiro *et al.*, 2007). In cases where a given contact has a high energy relative to other possible residues at that location, it has been described as 'highly frustrated' (Bryngelson and Wolynes, 1987; Onuchic *et al.*, 1995). Regions of high frustration have been implicated in facilitating protein folding, dynamics, and binding to other molecules (Ferreiro *et al*., 2011; Zheng *et al*., 2013; Bandyopadhyay *et al*., 2017; Li *et al*., 2017; Lindström and Dogan, 2018). These studies suggest that nature selects for high frustration in regions where high conformational flexibility is critical to protein function and low frustration in regions that are more critical to supporting the stability of the native-state ensemble. While there is evidence that some regions of primary structure are under selection for specific levels of energetic frustration (Whitford *et al*., 2007; Li *et al*., 2011), it remains unclear whether selection for contacts with low versus high energetic frustration affects tolerance to any specific classes of mutational lesions.

The interplay between stability and dynamics has been intensively studied in the adenylate kinase (AK) family (Bae and Phillips, 2006; Whitford *et al*., 2007; Rundqvist *et al*., 2009; Schrank *et al*., 2009; Li *et al*., 2011; Kerns *et al*., 2015). During the AK catalytic cycle, a reaction that involves reversible phosphoryl transfer (ATP + AMP ↔ 2 ADP), the lid and AMP-binding domains undergo coordinated conformational changes that involve local unfolding (Miyashita *et al*., 2003; Whitford *et al*., 2007; Schrank *et al*., 2009; Olsson and Wolf-Watz, 2010; Saavedra *et al*., 2018), while the core domain remains more rigid and is thought to determine overall thermostability (Bae and Phillips, 2006; Henzler-Wildman *et al*., 2007). Rational design studies have shown that modulation of AK dynamics affect the temperatures where maximal activities are observed (Bae and Phillips, 2006; Schrank *et al*., 2009; Saavedra *et al*., 2018). Mutational studies have also provided evidence that the lid domain may be less tolerant to changes in conformational flexibility because the dynamics of this domain have been fine-tuned for substrate binding (Saavedra *et al*., 2018). The AMP-binding domain, in contrast, appears more tolerant to enhanced conformational dynamics as local unfolding is thought to control catalytic turnover (Saavedra *et al*., 2018). While these studies have provided insight into the effects of a handful of mutations on AK structure and function in a small number of model systems, it remains unclear how these trends relate to other AK family members with distinct primary structures and thermostabilities.

One challenge with studying the effects of mutations on structure, function, and dynamics across a protein family is the limited throughput of *in vitro* measurements. In contrast, deep mutational scanning represents a simple approach for rapidly assessing the functional tolerance of many mutations in parallel (Araya and Fowler, 2011; Firnberg *et al*., 2014; Higgins and Savage, 2018), although this approach yields more limited biophysical insight. With deep mutational scanning, profiles of functional tolerance can be rapidly generated by creating a library of mutants, using a high-throughput assay to enrich for mutants that are active, sequencing the library of mutants before and after functional analysis, and calculating the relative abundance of each sequence following the functional enrichment (Araya and Fowler, 2011; Fowler and Fields, 2014; Higgins and Savage, 2018). Changes in the abundances of each sequence can be used to estimate the relative activities of each mutant by calculating a fitness score (Fowler *et al*., 2011; Bloom, 2015; Starita and Fields, 2015). To date, this approach has been applied to a wide range of proteins (Araya and Fowler, 2011; Fowler *et al*., 2011; Firnberg *et al*., 2014; Fowler and Fields, 2014; Bloom, 2015; Starita and Fields, 2015; Higgins and Savage, 2018). However, the effects of primary structure and protein stability within a protein family have not yet been examined, and it is unclear whether there are conserved patterns of mutational tolerance across protein orthologs that differ in primary structure and thermostability.

Recently, a method for creating combinatorial libraries was described that alters conformational flexibility at different locations in a protein. This approach, which is called circular permutation profiling with DNA sequencing (CPP-seq), randomly samples all sequence permutations in a protein by covalently connecting the original termini and creating new termini elsewhere in the primary structure (Atkinson *et al*., 2018). When proteins are created with this type of topological mutation, the local chain entropy is increased at the location where the new termini are created by cleaving the peptide backbone (Reitinger *et al*., 2010; Guntas *et al*., 2012; Daugherty *et al*., 2013). The application of CPP-seq to an AK with extreme thermostability, $T_m > 99°C$ (Vieille *et al*., 2003), revealed that more than half of all possible circularly permuted variants retain biological activity at 42°C (Atkinson *et al*., 2018). While this study identified diverse regions in all three AK domains that are functionally tolerant to changes in conformational flexibility that can arise from new protein termini, it remains unclear whether the observed pattern of tolerance depends upon AK sequence, local energetic frustration, or thermostability. Because proteins with enhanced thermostability are buffered from other classes of mutations (Bloom *et al*., 2005; Bershtein *et al*., 2006; Radestock and Gohlke, 2011; Elias *et al*., 2014), it seems likely that AKs with lower thermostability will be more sensitive to this class of topological mutation.

To better understand how thermostability influences protein tolerance to circular permutation, we subjected three AK homologs to CPP-seq and used a cellular assay to quantify protein fitness at a fixed temperature. We targeted AKs with a range of thermostabilities ($T_m = 43–74°C$) and reported crystal structures, including *Bacillus globisporus* AK (*Bg*-AK), *Bacillus subtilis* AK (*Bs*-AK), and *Geobacillus stearothermophilus* AK (*Gs*-AK) (Glaser *et al*., 1992; Bae and Phillips, 2004). The results from these experiments revealed that the fraction of functional variants in each library correlates with parental protein thermostability. At the residue level, we found that a large fraction of the permutated proteins were either nonfunctional across all three AKs or displayed fitness that correlates with AK thermostability. However, a subset of circularly permuted AK retained fitness across all three variants. A comparison of the positional tolerance to permutation revealed that this latter group of variants arose from the creation of new protein termini at native sites within the AMP-binding domain, which has contacts with consistently high energetic frustration across all three AKs. This correlation suggests that energetic frustration may be useful for predicting regions of homologous proteins that are functionally tolerant to increases in conformational flexibility arising from circular permutation.

Materials and methods

Vector construction

The genes encoding *Bacillus globisporus*, *Bacillus subtilis*, and *Geobacillus stearothermophilus* AK were PCR amplified from pNIC28-BgAK, pNIC28-BsAK, and pNIC28-GsAK to create amplicons flanked by NotI sites. In these amplicons, a spacer base pair was added before the start codon to maintain in-frame translation and the stop codon was removed. The amplicon was cloned into pET26b vectors using Golden Gate assembly to yield pET26b-BgAK, pET26b-BsAK, and pET26b-GsAK. All vectors were sequence verified.

Library construction

Each AK vector (~4 μg each) was digested with NotI overnight at 37°C, agarose gel electrophoresis was used to separate the AK genes from other reaction products, and each AK gene was purified using a Zymoclean^{TM} Gel DNA Recovery Kit (Zymo Research) and eluted with DNA-grade water. Purified AK genes (~400 ng) were circularized using T4 DNA ligase in a 20 μL reaction incubated at 16°C for 16 hours. Following ligation, circularized AK genes were further purified using a DNA Clean & Concentrator Kit (Zymo Research) and eluted with 12 μL of DNA-grade water. The yield of each circularized AK was assessed using a NanoDrop spectrophotometer (220–350 ng). To create each library, the circular genes were mixed with BglII-linearized pMT-P1 (50–150 ng), an artificial transposon with all of the attributes of a protein expression vector (Addgene #120863), and one unit of MuA transposase (Thermo Fisher Scientific) in a 20 μL reaction and incubated at 37°C for 16 hours. Following incubation at 75°C for 10 minutes, total DNA from each reaction was purified using a DNA Clean and Concentrator Kit and transformed (~200–400 ng) into library grade MegaX DH10B Ultracompetent cells (Thermo Fisher Scientific) using electroporation. Following electroporation, cells were suspended in recovery media (Thermo Fisher Scientific) for 45 minutes at 37°C while shaking before plating 200 μL onto five LB-agar plates (150 × 15 mm) containing 25 μg/mL of kanamycin. After incubation at 37°C for 24 hours, colony forming units (CFU) were quantified visually. Estimates of unselected library sampling based on colony counts on plates indicated that sampling exceeded the number of possible variants (1320) by >20-fold. To harvest the final libraries, 3 mL of LB was added to each plate, a sterile spreader was used to homogenize colonies, the cell slurries from each library were pooled, and a QIAprep Spin Miniprep Kit (Qiagen) was used to isolate each unselected library. The library construction reaction yielded three types of DNA, circularized permuteposon (pMT2), permuteposon-AK gene hybrids in the parallel orientation, and permuteposon-AK gene hybrids in the antiparallel orientation (Fig. 1A). The band corresponding to the permuteposon-AK hybrids was cut out from the agarose gel, and this DNA was purified using a Zymoclean^{TM} Gel DNA Recovery Kit and DNA-grade water elution. Restriction digest analysis of these purified size-selected libraries reveals that they were homogeneous for permuteposon-AK gene hybrids (Fig. S1, Supplementary data are available at *PEDS* online).

Library selections

Libraries were selected using *Escherichia coli* CV2 [Hfr(PO2A), fhuA22, ΔphoA8, adk-2(ts), ompF627(T2R), fadL701(T2R), relA1, glpR2(glp^c), glpD3, pitA10, spoT1, rrnB-2, mcrB1, creC510] (Cronan *et al.*, 1970). The chromosomally-encoded AK in this strain has a Pro87Ser mutation that results in temperature-sensitive protein (T_m = 38.5°C) that is inactive at 42°C (Haase *et al.*, 1989). To quantify the fraction of functional AKs in each library, the purified unselected libraries (300 ng) were transformed into *E. coli* CV2 using electroporation, cells were spread on LB-agar plates containing 25 μg/mL kanamycin, and five plates containing each library were incubated at 30 and 42°C for 48 hours. Visual inspection of these plates was used to estimate total CFU, and the fraction functional was calculated as the number of CFU at 42°C divided by the number at 30°C. The number of unique selected CFU in each library was also estimated by dividing the observed CFU by two to account for the doubling of cells that occurs during the growth recovery following transformation. Among the three libraries, this analysis revealed that we sampled 60,080 *Bg*-AK, 104,720 *Bs*-AK, and 48,200 *Gs*-AK vectors. The colonies from each library were harvested and pooled, and a QIAprep Spin Miniprep Kit was used to isolate each selected library. The incubation times were determined by performing controls with *E. coli* CV2 transformed with a circularized transposon lacking an AK gene. When these cells were grown on LB-agar plates containing 25 μg/mL kanamycin at 42°C for 48 hours, no colonies were observed. In a prior study, revertants to wild-type *Ec*-AK were isolated at a frequency $\leq 10^{-9}$ when plated at 42°C (Haase *et al.*, 1989). Based on the CFU plated in our library selections (1–3 × 10^8), each selected library should contain <1 revertant colony. To minimize revertants, nonselective strain manipulations were performed at 30°C prior to selections because this temperature is below the *Ec*-AK(P87S) denaturation temperature (Haase *et al.*, 1989).

Library sequencing

Each unselected and selected library (≥700 ng) was digested with restriction endonucleases that cut ~750 base pairs upstream of the permuted genes (ClaI) and ~150 base pairs downstream of the permuted genes (PciI). Agarose electrophoresis was used to separate the permuted AK genes from the vector backbone, and the bands encoding each ensemble of permuted AK genes (~1.6 kb) were excised and purified. All downstream processing and sequencing of the unselected and selected libraries was performed by the Baylor College of Medicine Genomic and RNA Profiling Core. A Nextera XT kit was used to fragment the DNA ensemble in each library and attach unique sequencing adapters to the ends. An Illumina MiSeq System was then used to collect single-end sequencing reads (150 base pairs) on all of the libraries in parallel.

Sequence analysis

Sequencing data was analyzed using a previously described custom Python pipeline, which is available at github.com/SilbergLab/CPP-seq. In brief, this pipeline first identifies reads containing a permuteposon using nine base pair sequence motifs at the end of each permuteposon. The pipeline then determines if each read is at the beginning or end of the permuteposon by analyzing for the presence of a larger 54 base pair sequence. Reads including the start codon are designated 'start motif' reads, while those including the stop codon are designated the 'stop motif' reads. Each of these reads is then further divided into sense and antisense reads. We next evaluated the first 11 base pairs of AK-derived sequence adjacent the start/stop motifs and determined how they related to all possible 11 base pair sequences within each of the AK genes. This comparison allowed us to then designate AK genes as parallel (P) and in an orientation that can be transcribed and translated or antiparallel (AP) and in an orientation that does not allow for expression.

Statistical analysis

To analyze whether vectors were significantly enriched by the selection, we took a similar approach to the DESeq method (Anders and Huber, 2010). DEseq was developed to analyze if there are significant changes in gene expression with RNA-seq and ChIP-seq datasets, using no change as the frame of reference and the null hypothesis. With our dataset, we used the ratio of selected to unselected AP variant counts as a frame of reference for the expected effect of selective pressure as the null hypothesis instead of the no-change null hypothesis that DESeq assumes. We modeled the AP variant counts (i = 1–220 for each unique variant) using a negative binomial model where we defined X_i and Y_i as the unselected and selected counts, respectively. In our model, the AP counts in the unselected library are defined as m_i, and the AP counts in the selected library

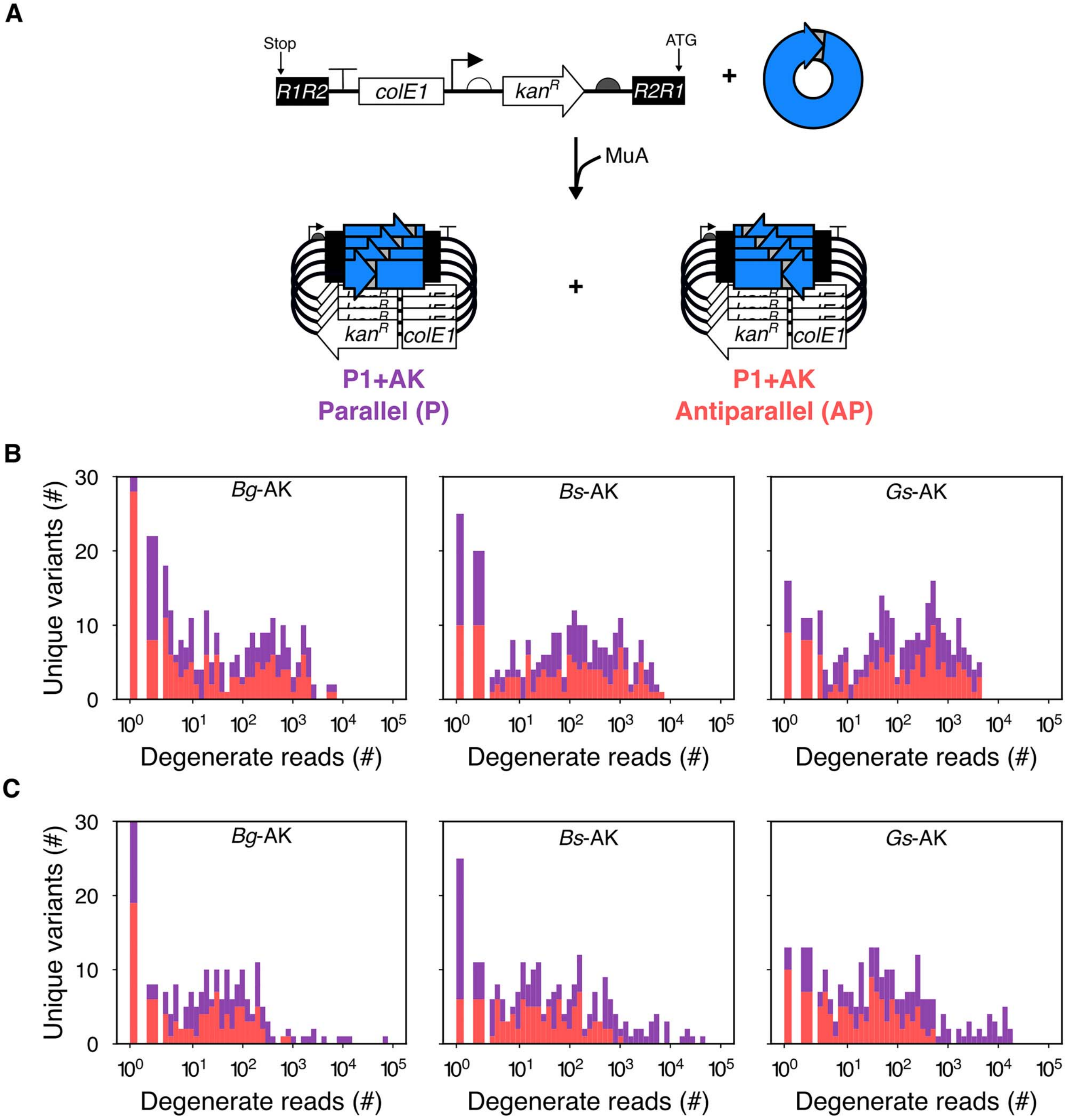

Fig. 1 Permuted gene abundances versus gene orientation. (**A**) Generation of libraries using PERMUTE requires mixing a circular gene, a permuteposon, and MuA transposase. The permuteposon inserts in two orientations with near equal frequencies (Atkinson *et al*., 2018). The orientation where the regulatory elements (i.e. promoter, RBS, and terminator, and the permuted genes) are in the same direction is designated as *parallel*. Conversely, when the permuteposon inserts in the opposite direction such that the regulatory elements and the permuted genes are oriented in opposite directions, this is designated as *antiparallel*. Only when the permuteposon inserts in the *parallel* orientation does it generate a vector capable of expressing a circularly permuted protein with an 18-residue peptide amended to the new N-terminus. The DNA encoding this peptide maintains the genetic context of the RBS so that translation initiation is constant across different protein variants (Jones *et al*., 2016). For each (**B**) unselected library and (**C**) selected library, we evaluated the number of degenerate, in-frame sequences at each position in both the P (purple) and AP (red) orientations.

are m_i multiplied by a scalar dilution factor (β), such that $X_i = m_i$, $Y_i = \beta m_i$, Var $(X_i) = \gamma m_i$, and Var $(Y_i) = \gamma \beta m_i$. There are two global parameters shared by all the loci, β and γ. The former reflects the average selection effect on all the AP variants, and the latter reflects the overdispersion compared with Poisson distribution. β was estimated using sample median, and γ was estimated by assuming a negative binomial distributions for X_i and Y_i and then maximizing the likelihood. Using the estimated negative binomial distribution, the *P*-value of each *P* variant pair (X_i, Y_i) was computed by conditioning on the sum of X_i and Y_i as outlined in Equation 11 of DESeq (Anders and Huber, 2010). Finally, P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure.

Fitness calculations

To relate the relative activity of all three AK orthologs, we converted the $\log_2$(fold change) into a measure of fitness relative to the native AK. This was done by normalizing each variant's fold change by the fold change of the native AK present in the same library according to:

$$\text{Fitness} = \frac{\text{Fold change}_{\text{variant}}}{\text{Fold change}_{\text{native}}}$$

Data availability

Deep sequencing data are available from the NCBI Sequence Read Archive (SRA) under accession numbers SAMN13192615 (*Bg*-AK, unselected), SAMN13192616 (*Bg*-AK, selected), SAMN13192617 (*Bs*-AK, unselected), SAMN13192618 (*Bs*-AK, selected), SAMN13192619 (*Gs*-AK, unselected), and SAMN13192620 (*Gs*-AK, selected). Python scripts used to analyze the data are available at github.com/SilbergLab. Analyzed data for figures is available upon request.

Results

Library construction and characterization

A major challenge with using cellular assays to compare the mutational tolerance of proteins is achieving uniform expression in cells, such that the retention of function only depends upon protein stability and total activity. This is challenging with topological mutations like circular permutation where the mutation alters the genetic context of the ribosomal binding site (RBS) that controls translation initiation (Jones *et al.*, 2016). In a previous study, we showed that transposon mutagenesis can be used to build vector libraries that express different circularly permuted variants of any protein with the same translation initiation rate (Jones *et al.*, 2016). With this approach, which is called PERMUTE, libraries of permuted genes are created by randomly inserting a linear vector, called a permuteposon (Mehta *et al.*, 2012), into a gene that has been circularized by connecting the first and last codons using a nondisruptive linker sequence.

Three different AKs were subjected to random circular permutation using PERMUTE (Atkinson *et al.*, 2018), including *Bg*-AK (T_m = 43.3°C), *Bs*-AK (T_m = 47.6°C), and *Gs*-AK (T_m = 74.5°C). These AK orthologs were chosen because they exhibit melting temperatures that span over 30°C (Glaser *et al.*, 1992; Bae and Phillips, 2004). A sequence alignment reveals that these AKs are identical in length and all contain the Cys-X_2-Cys-X_{16}-Cys-X_2-Cys/Asp zinc-binding motif (residues 130–153) in their lid domain (Fig. S2, Supplementary data are available at *PEDS* online), which is typical of AKs from gram-positive bacteria (Glaser *et al.*, 1992). Additionally, these AKs exhibit high pairwise sequence identities (66–74%), and structures with low RMSD (0.73–1.76 Å) (Bae and Phillips, 2004). Because their native termini are proximal, all of the circularly permuted AKs were expressed with their native termini linked using a tripeptide (Ala-Ala-Ala). This peptide was previously found to be compatible with creating functional permuted *Thermotoga neapolitana* AK (*Tn*-AK) (Mehta *et al.*, 2012).

In total, each AK library contains up to 1320 different vectors. This diversity is produced because PERMUTE randomly inserts the permuteposon at every location in the AK genes (651 base pairs) that were circularized using a nine-base pair linker. Half of the insertions (n = 660) into the circular genes occur such that the regulatory elements that drive expression in the permuteposon are parallel (P) with the open reading frames (ORFs) encoding each circularly permuted gene (Fig. 1A). In contrast, the other half are in an antiparallel (AP) orientation and cannot express circularly permuted genes. Among the P variants, only one third of the permuted genes are in frame (n = 220), and only 216 of these express permuted proteins. Four of the variants occur within the linker and express full-length AKs, which serve as a frame of reference for parent protein fitness in all experiments.

We transformed each library into *E. coli* DH10B and quantified the number of unique colony forming units (CFU) following an overnight incubation under nonselective conditions. This analysis revealed that the libraries sampled more than 174 000 (*Bg*-AK), 164 000 (*Bs*-AK), and 28 000 (*Gs*-AK) vectors containing AK genes. To evaluate the sequence diversity created, each library was purified, barcoded, and subjected to deep sequencing using MiSeq (Atkinson *et al.*, 2018). This analysis yielded more than 3 million total sequence reads for each library. Table I lists the abundances for the P and AP reads in each library. The numbers of desired expression vectors sampled, i.e. those that are in frame and parallel, were 52 332 (*Bg*-AK), 67 193 (*Bs*-AK), and 70 258 (*Gs*-AK).

When creating libraries using transposon mutagenesis, the abundances of each permuted gene can vary widely (Atkinson *et al.*, 2018), although the relative abundance of P and AP variants of each permuted gene is typically similar. Figure 1B shows that the relative abundance of each permuted gene varied by up to three orders of magnitude within each library. In contrast, the abundances of P and AP variants arising from insertion at the same location presented abundances that differed by <7% (Fig. S3, Supplementary data are available at *PEDS* online). The variants with the highest abundances in each library were different. In the *Bg*-AK library, the permuted gene that started with the Pro-165 codon was most abundant (n = 7307), while the AK genes that started with codons encoding Val-115 (n = 5039) and Thr-31 (n = 4456) were most prevalent in the *Bs*-AK and *Gs*-AK libraries, respectively.

Thermostability correlates with mutational tolerance

To identify permuted AKs in each library that retain catalytic activity at 42°C, each library was enriched for vectors that express an active AK using bacterial complementation with *E. coli* CV2 (Haase *et al.*, 1989), a strain with a temperature-sensitive AK that cannot grow above 40°C unless a functional AK is expressed in trans (Segal-l-Shapiro *et al.*, 2011). The vector ensembles were purified following the selection and deep sequenced, and retention of function was quantified by analyzing the ratio of P and AP sequence abundances before and after selection as previously described (Atkinson *et al.*, 2018). By quantifying the ratio of P to AP variants for each permuted AK, we identified variants with a range of enrichments, including P variants that were not enriched compared with cognate AP variants, P variants that were enriched to a similar extent as the P parental AKs, and variants with intermediate enrichment.

Following the selection of each library for functional AKs using *E. coli* CV2, we quantified the number of unique vectors sampled. All three libraries yielded >48 000 unique vectors. Table I shows that MiSeq analysis of the selected reads yielded similar total counts as the unselected libraries. Figure 1C shows that genes with the highest abundances in the selected libraries were in the P orientation. Across the selected libraries, the fraction of in frame

Table I. Parallel and antiparallel sequence read counts in each library

	Bg-AK			*Bs*-AK			*Gs*-AK		
	Unselected	Selected	S/US ratio	Unselected	Selected	S/US ratio	Unselected	Selected	S/US ratio
Total reads	3 345 761	2 943 111	0.88	3 541 540	3 583 510	1.01	3 486 607	2 872 462	0.82
+11 bp AK	224 235	245 313	1.09	230 619	248 489	1.08	232 916	224 023	0.96
P in frame	52 332	224 028	4.28	67 193	226 735	3.37	70 258	208 547	2.97
AP in frame	48 893	5699	0.12	67 747	9361	0.14	67 197	7003	0.10
P +1 frame	49 708	6422	0.13	34 633	4179	0.12	32 411	2544	0.08
AP +1 frame	48 340	5380	0.11	34 362	4675	0.14	33 684	3259	0.10
P −1 frame	12 583	2344	0.19	13 191	1820	0.14	14 634	1264	0.09
AP −1 frame	12 379	1440	0.12	13 493	1719	0.13	14 732	1406	0.10

The data generated by MiSeq analysis, including the total reads having the barcode corresponding to each AK ortholog in both the unselected (US) and selected (S) libraries, the number of reads with a permuteposon barcode and 11 base pairs (bp) of AK gene used to identify the location of the new termini. For reads with permuteposon barcode and 11 bp of AK gene, we list the number of parallel (P) and antiparallel (AP) that are in frame, +1 frame, and −1 frame. The ratio of selected to unselected abundances (S/US ratio) is also shown.

P vectors (0.968 ± 0.006) was 30-fold greater than the AP fraction (0.032 ± 0.006). This result can be contrasted with the unselected libraries, where the in frame P and AP fractions were similar (0.509 ± 0.008 and 0.491 ± 0.008, respectively). These results illustrate how the selective pressure applied by the cellular assay dilutes the AP variants that cannot express a permuted AK relative to the P variants that can express active permuted AKs.

To establish which permuted AKs are active in cells, we calculated the fold change in abundances of each unique permuted AK variant in the selected and unselected libraries in both the P and AP orientations. In PERMUTE libraries, cognate P and AP variants are synthesized at similar proportions (Fig. S3, Supplementary data are available at *PEDS* online) and, thus, represent paired data whose ratios can be used to assess the significance of sequence enrichment across the wide range of initial abundances observed (Atkinson *et al*., 2018). While the P variants express the different permuted proteins, the AP variants are unable to express proteins and cannot be enriched by the selection. In this way, the AP variants serve as an internal frame of reference for evaluating whether their cognate P variants are biologically active (enriched more than AP variants) or inactive (similarly diluted as AP variants).

We used a negative binomial distribution to model the mean and the variance of the fold change of AP variants in the selected and unselected conditions, i.e. $\log_2\left(\frac{\text{Selected}}{\text{Unselected}}\right)$, relative to each variant's initial abundance in the unselected library (Atkinson *et al*., 2018). Using these values, we investigated which P variants are biologically active by identifying those that were significantly enriched by the selection (two-tailed t-test, $P < 0.01$). With this analysis (Fig. S4, Supplementary data are available at *PEDS* online), 25.8% (n = 38/147) of the sampled P variants were significantly enriched in the *Bg*-AK library, while 43.9% (n = 65/148) and 59.7% (n = 86/144) were significantly enriched in *Bs*-AK and *Gs*-AK libraries, respectively. Comparing the fraction of functional variants with T_m reveals a positive correlation (Fig. 2A, Pearson's correlation = 0.91). This trend shows that thermostability buffers AKs from the destabilizing effects of circular permutation, similar to the buffering effects observed with other classes of mutations (Bloom *et al*., 2005; Bershtein *et al*., 2006; Radestock and Gohlke, 2011; Elias *et al*., 2014). This buffering is thought to occur because circular permutation has similar destabilizing effects on structurally-related AK homologs (Fig. 2B), yielding a similar ensemble of $\Delta\Delta G_f$ in each topological mutant library.

Family-level fitness of topological mutants

To determine if functional tolerance to new termini is dependent upon domain location, we next compared the enrichment of each in frame P variant with AK structure (Fig. 3). This analysis reveals that new termini are differentially tolerated at diverse locations within the different AK orthologs. Circularly permuted AKs were uniformly inactive when new termini were generated within the glycine rich p-loop (residues 7–15), the region of the core domain that constitutes the active site in AKs and other related kinases (Bae and Phillips, 2004; Romero *et al*., 2018). In contrast, the lid (residues 128–159) and core (residues 1–30, 60–127, and 160–217) domains both varied in their tolerance to new termini across the different AK orthologs, with some positions being uniformly intolerant and other positions exhibiting tolerance that varied across the orthologs. Surprisingly, a large fraction of the permuted AK having new termini within the mobile AMP-binding domain (residues 31–59) presented biological function in all three AK homologs. In *E. coli* AK *(Ec*-AK), this domain uses local unfolding to control the rate-limiting step of the catalytic cycle, product release (Whitford *et al*., 2007; Saavedra *et al*., 2018).

In our libraries, there are two frames of reference for biological activity, including (i) AP variants that cannot express AKs, which are diluted by the selection (Fig. S4, Supplementary data are available at *PEDS* online), and (ii) native AK, whose enrichment is dependent upon total activity. Because each AK library was selected in a different experiment, we evaluated the relative enrichment of each parental AK, which was encoded and observed in all three libraries following the selections. This analysis revealed that the fold change of each native AK decreases linearly (r^2 = 0.99) as the fraction of functional variants in each library increases (Fig. S5, Supplementary data are available at *PEDS* online). This finding illustrates the additional selective pressure of increased competition that arises as the fraction of functional variants competing with native AKs changes in each experiment. To enable us to compare the sequence enrichment trends across the three libraries, we converted the fold enrichment value into a measure of parental fitness by normalizing the enrichment value for each permuted variant to that observed with the native AKs. Due to partial sampling in the individual libraries, only 38.7% (n = 84/217) of the circularly permuted AKs were observed in all three libraries and thus available to create a family permutation profile. Those circularly permuted AKs sampled in all three libraries were distributed across the AMP-binding, lid, and core domains.

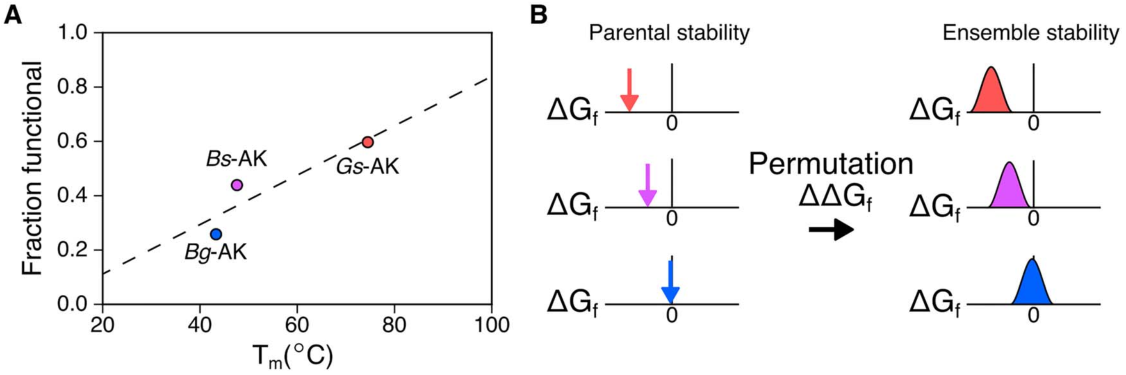

Fig. 2 Thermodynamic stability correlates with tolerance to permutation. (**A**) The fraction of P variants sampled in the libraries that were biologically active following selection is plotted relative to the melting temperature of each AK parent. A linear fit ($y = 0.01x - 0.07$, $R^2 = 0.822$) is shown as a dashed line. (**B**) This trend suggests that circular permutation has similar destabilizing effects on structurally-related AK homologs yielding a similar ensemble of $\Delta\Delta G_f$.

To first evaluate how the fitness of each of the permuted AK homologs relates to the location of their protein termini, we compared the fitness of each variant to the number of AK orthologs in which that position was observed to be significantly enriched. With this family-level analysis, we observed different trends when looking at homologous circular-permuted variants derived from the three parental AKs (Fig. 4A). Some permuted AKs presented low fitness values across all three AK homologs that could not be distinguished from cells lacking an AK. This trend indicates that these positions are uniformly intolerant to permutation across all three AK homologs. Other permuted AKs presented fitness values that varied across the different parents. At some locations, we observed one out of the three structurally-related variants as enriched, while two out of three of the variants were enriched at other locations. Surprisingly, some AK variants presented parent-like fitness values across all three AK homologs.

To visualize all of the family-level data, we analyzed the *average* fitness of each circularly permuted variant observed in the *Bg*-AK, *Bs*-AK, and *Gs*-AK libraries (Fig. 4B). This analysis revealed that circularly permuted proteins that were inactive in all three libraries primarily arise from the creation of new protein termini in the core domain, although inactive variants having protein termini in all three domains were observed. In contrast, those circularly permuted variants that were active in all three libraries arose from new termini in either the AMP-binding or core domains. The variants arising from protein termini in the AMP-binding domain exhibited the highest average fitness among these variants. Permuted AKs that were significantly enriched in either one or two orthologs contained protein termini in all three domains.

Relationship between fitness and energetics

Because circular permutation creates increased conformational flexibility at the location of the new termini, we hypothesized that new termini might be tolerated to a greater extent at locations that have been selected to support local flexibility. One way that such locations have been identified is by identifying the residue-residue contacts in a protein that exhibit higher pairwise energies (frustration) than other possible amino acids at that same position. To evaluate how frustration changes across each domain within the two conformations, we used the frustratometer server to generate profiles of inhibitor-bound and substrate-free *Ec*-AK (Fig. 5A) (Parra *et al.*, 2016). At native locations where the amino acids make contacts that are lower energy compared to all other amino acid possibilities at that contact, the profile yields a low frustration value. In contrast, native locations having amino acids that present higher contact energies compared with all other theoretical amino acid combinations at that contact yield high frustration values. In the inhibitor-bound closed conformational state, there is a shift in this energetic profile with an increase in the density of highly frustrated contacts within the AMP-binding domain and a subtle decrease in the lid domain. These state-specific frustrated contacts are thought to facilitate the local unfolding events in the AMP-binding domain that are coupled to the lid opening at the end of the AK catalytic cycle (Li *et al.*, 2011).

To determine if the AKs used to build our libraries exhibit high frustration in the AMP-binding domain, we calculated profiles of their energetic frustration using inhibitor-bound AK structures (Fig. 5B). These calculations revealed that all three proteins exhibit similar patterns of energetic frustration even though their thermostability varies widely. This result suggests that this pattern of energetic frustration has been selected during evolution to support catalysis across different ranges of temperatures. Comparing these profiles with AK structure reveals that regions of high energetic frustration are localized in the AMP-binding and lid domains as well as surface-exposed sites in the core domain (Fig. 5C). In contrast, regions of minimal energetic frustration are localized primarily within the core domain. These residue-residue contacts are thought to play an important role in controlling protein folding and maintaining a pre-organized active site that is poised for catalysis (Kerns *et al.*, 2015). Together, these results suggest that AKs have selected residue-residue contacts with specific levels of frustration at individual native sites to support protein folding, substrate binding, and conformational dynamics critical to catalysis (Fig. 5D).

We next investigated how the locations of the termini in variants having different levels of fitness relate to primary structure, tertiary structure, and energetic frustration (Fig. 6A and B). This comparison revealed that the region of the AMP-binding domain where new termini are functionally tolerated to the greatest extent overlaps with the region containing the highest density of frustrated contact energies. When looking at the positions within the AMP-binding domain that are enriched across either zero, one, two, or three AK orthologs (Fig. S6, Supplementary data are available at *PEDS* online), the average fitness increases as the energy landscape becomes more frustrated. Like the AMP-binding domain, regions proximal to the native protein termini exhibit high functional tolerance to new termini across all three AKs, even though this region does not

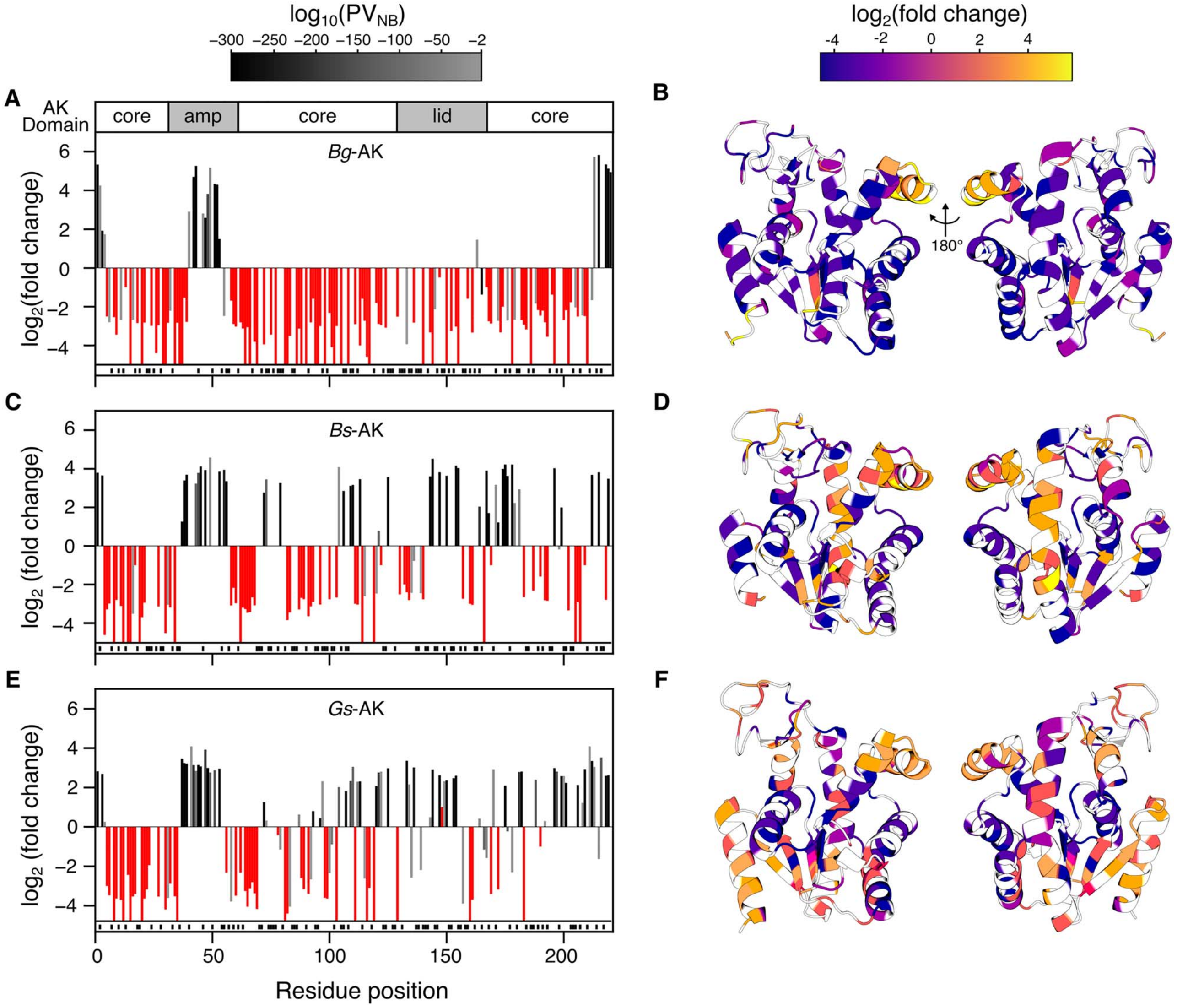

Fig. 3 Comparison of circular permutation profiles of each AK. For each P variant, the $\log_2$(fold change) is shown as a function of the AK residue found at the N-terminus of the circularly permuted protein derived from (**A**) *Bg*-AK, (**C**) *Bs*-AK, and (**E**) *Gs*-AK. *P*-values obtained from the negative binomial model (PV_{NB}) are color coded with values $>10^{-2}$ in red and variants having values $\leq 10^{-300}$ in black and those variants displaying intermediate values shaded as indicated by the bar. The AK domain structure is shown at the top as a frame of reference. Variants no longer observed following selection are shown as bars that reach the line at the bottom of the graph. The cognate P and AP variant pairs absent from both the unselected and selected libraries (*n* = 73, 72, and 76 for *Bg*-AK, *Bs*-AK, and *Gs*-AK, respectively) are indicated as black lines shown below the x-axis. The $\log_2$(fold change) is compared with AK structure for (**B**) *Bg*-AK (PDB: 1S3G), (**D**) *Bs*-AK (PDB: 1P3J), and (**F**) *Gs*-AK (PDB: 1ZIO) with the residues colored according to their $\log_2$(fold change) as indicated by the bar. Unsampled positions are in white.

exhibit high frustration like the AMP-binding domain. Similarly, a patch within the middle of the $\alpha 7$ helix that is surface exposed on the back of the core domain is uniformly tolerant to new termini, even though this region exhibits moderate frustration across all three AK homologs. We next compared the average fitness of each variant with the average density of highly frustrated contacts found in the ‘closed’ state structures of each AK ortholog (Fig. 6C). This analysis revealed a significant correlation (Spearman’s rank correlation R_{SR} = 0.512, two-tailed *P*-value = 6.13×10^{-7}) between the density of highly frustrated contacts and fitness across the different permuted variants.

To further explore how protein dynamics influence tolerance to permutation, we compared average fitness to several other experimental and computational metrics. Comparison of the average fitness with normalized B-factors from the inhibitor-bound crystal structures of the three orthologs (Bae and Phillips, 2004) revealed a significant correlation (Fig. S7A, Supplementary data are available at *PEDS* online, Spearman’s rank correlation R_{SR} = 0.439, two-tailed *P*-value = 2.95×10^{-5}), although the *P*-value for this correlation was lower than that observed with frustration. Comparison of the average fitness with nuclear Overhauser effect (NOE) data from nuclear magnetic resonance (NMR) studies of *Ec*-AK (Tugarinov *et al.*, 2002) revealed an inverse correlation (Fig. S7B, Supplementary data are available at *PEDS* online, Spearman’s rank correlation R_{SR} = −0.324, two-tailed *P*-value = 6.00×10^{-3}). To take an entropic perspective, we compared the average fitness with predictions of residue-level conformational entropy in the native-state ensemble using

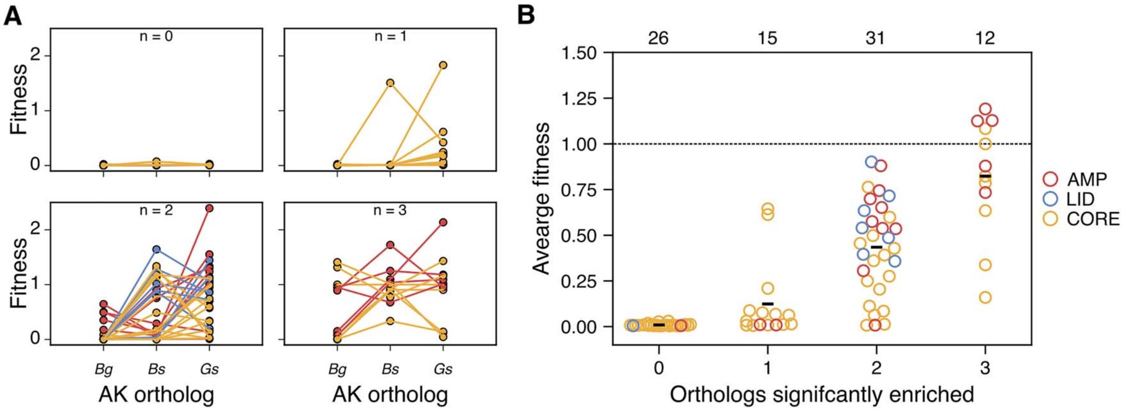

Fig. 4 Comparison of family enrichment and fitness. (**A**) For variants that were sampled in all three AK libraries ($n = 84$), the fitness score (fold change relative to native AK) is shown. In each box, the number of AK homologs across the three libraries that were found to be significantly enriched using PV_{NB} is noted as $n = 0$, $n = 1$, $n = 2$, or $n = 3$. Data for each topological variant is connected by a line to illustrate the trend across homologous permuted proteins derived from parent AKs having distinct thermostabilities. Those topological mutants that were uniformly biologically active across all three libraries ($n = 3$) presented a wide range of fitness values (ranging from 0.004 to 2.13). (**B**) For the same set of circularly permuted protein homologs, the average fitness of all three orthologs is compared to the number of orthologs found to be significantly enriched using PV_{NB}. The number of data points in each bin is noted at the top. The average fitness of all the orthologs in a column is noted with short horizontal black lines, while the average fitness of the native AKs is indicated by the dashed line. The domain location of the new protein termini in each permuted protein trio is indicated by the color of the ring according to the legend.

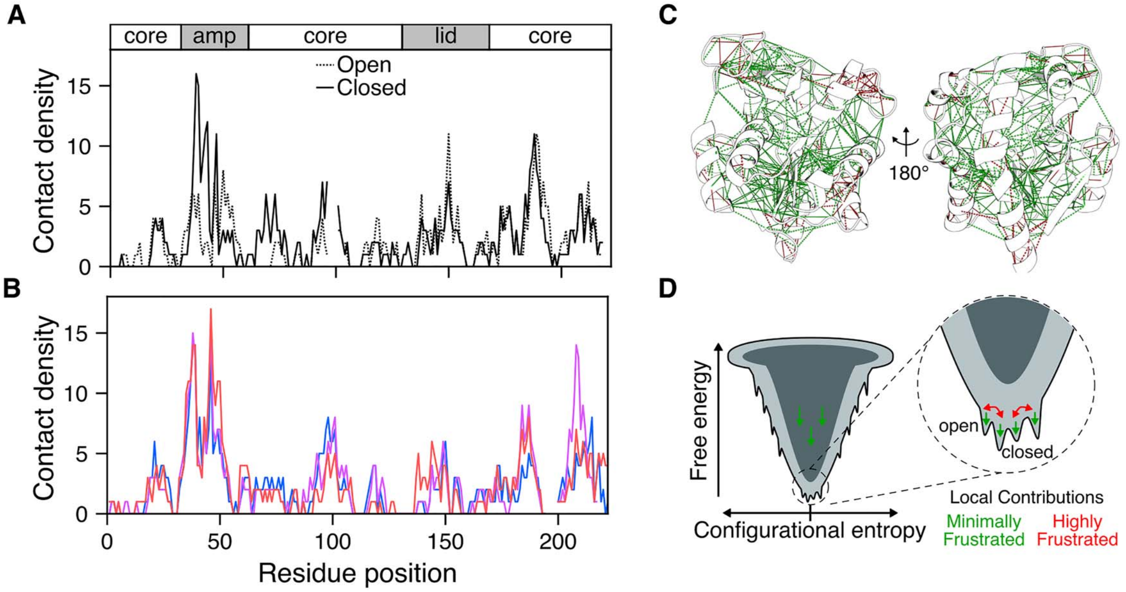

Fig. 5 Energetic frustration of intramolecular contacts in adenylate kinase. (**A**) Frustratograms showing the calculated configurational frustration of the open (dashed) and closed (solid) conformations of *Ec*-AK (calculated from PDB: 4AKE and 1AKE, respectively). The density of contacts in a 5 Å sphere with highly frustrated contact energies is plotted at every AK residue position. (**B**) A comparison of the frustratograms of the closed conformations of *Bg*-AK (blue), *Bs*-AK (purple), and *Gs*-AK (red). (**C**) The contact energy frustration is mapped onto the tertiary structure of a closed structure (PDB: 1AKE). Highly frustrated contacts are displayed in red, while minimally frustrated are shown in green. Direct contacts are solid lines, while water mediated contacts are dashed lines. (**D**) Influence of contact frustration on the global protein folding and functional energy landscape. Minimally frustrated contacts promote protein folding into the low-energy native ensemble, while within this ensemble, highly frustrated contacts facilitate the sampling of conformations including those important for catalysis.

POPCOEN, a neural-network trained on molecular dynamics simulations of ~1000 proteins (Goethe *et al*., 2018). This analysis also revealed a significant correlation (Fig. S7C, Supplementary data are available at *PEDS* online, Spearman's rank correlation R_{SR} = 0.380, two-tailed *P*-value = 3.71 × 10^{-4}). However, when we compared average fitness to the residue-level predictions of conformational entropy loss upon folding from the denatured state ensemble to the native-state ensemble using the calculator PLOPS (Baxa *et al*., 2014), no significant correlation was observed (Fig. S7D, Supplementary data are available at *PEDS* online).

To establish how the different protein dynamics metrics relate to one another, we evaluated the pairwise relationship of each metric.

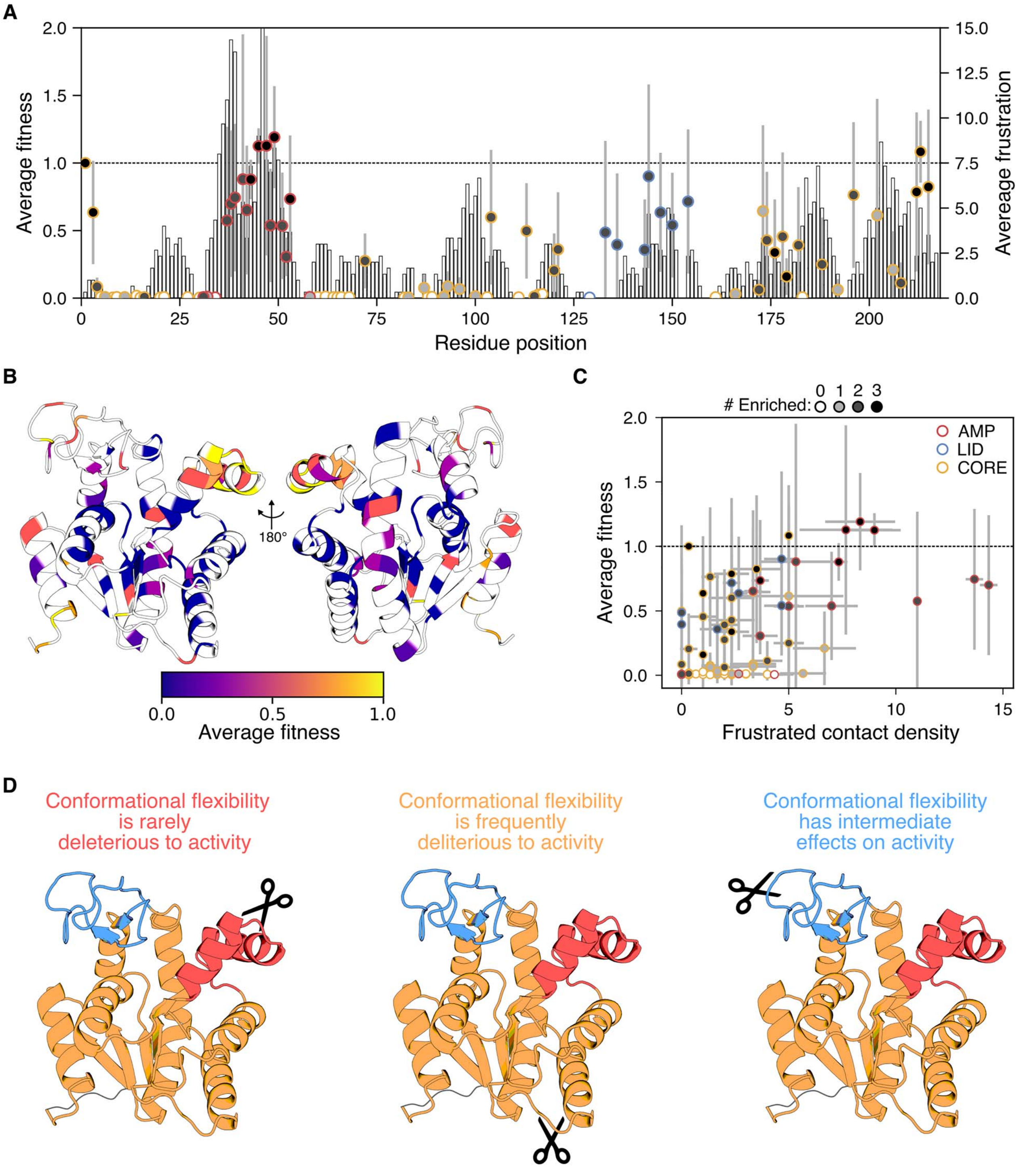

Fig. 6 Comparison of family permutation profiles and energetic frustration. (**A**) For topological mutants that were sampled in all three libraries ($n = 84$), the average fitness and average frustration are compared to the position in the primary structure. The circles represent the average fitness of each variant with the standard deviation shown in gray. The shading of each circular data symbol represents the number of orthologs in which the variant was found to be enriched. The bars represent the average density of highly frustrated contacts at each position observed across all three structures. The dotted line represents the native AK fitness. (**B**) The average fitness of all three orthologs is compared to the tertiary structure (PDB: 1S3G). The color of the residue corresponds to the average fitness of a permutation at each site according to the color bar. Positions that were not observed in all three libraries are shaded white. (**C**) The average fitness of each trio of homologous permuted AK is compared to the average density of highly frustrated contacts at that position in the native AK structures. Each circle represents the average fitness and the average frustration observed across all three structures. The standard deviations are shown in gray. The shading of each circular data symbol represents the number of orthologs in which the variant was found to be enriched. The dotted line represents the native AK fitness observed in each library. Spearman's rank correlation coefficient between the fitness and frustrated contact density was 0.512 with a two-tailed *P*-value = 6.13×10^{-7}. (**D**) Higher conformational flexibility caused by permutation in the AMP-binding domain is not deleterious to enzyme fitness, while permutation in the core and lid domains negatively impacts enzyme fitness

These comparisons revealed significant positive correlations between B-factors and the density of highly frustrated contacts calculated using the frustratometer as well as B-factors and the residue-level conformational entropy calculated by POPCOEN (Fig. S8A, Supplementary data are available at *PEDS* online); B-factors did not correlate with the entropy loss calculated by PLOPS. In addition, inverse correlations were revealed between NOEs and both frustration and POPCOEN calculated entropies (Fig. S8B, Supplementary data are available at *PEDS* online). However, no significant correlation was observed between NOEs and PLOPS. Furthermore, no significant correlations were observed when comparing the three computational metrics with NOE and B-factors.

Discussion

Deep mutational scanning is increasingly used to analyze the contributions that individual residues make to protein function (Fowler and Fields, 2014; Wrenbeck *et al.*, 2017b; Higgins and Savage, 2018). This approach has previously been used to identify residues that underlie protein solubility (Klesmith *et al.*, 2017), protein-protein interactions (Richard *et al.*, 2012; Doolan and Colby, 2015), membrane protein insertion (Elazar *et al.*, 2016), thermostability (Araya *et al.*, 2012; Romero *et al.*, 2015; Nisthal *et al.*, 2019), substrate specificity (Wrenbeck *et al.*, 2017a), enzymatic function (Romero *et al.*, 2015), and mutational epistasis (Olson *et al.*, 2014). Such efforts have provided fundamental insight into sequence-function relationships within individual proteins. By analyzing the effects of mutations on protein functions across different conditions, deep mutational scanning has shown that environmental conditions can affect how mutations contribute to fitness (Melnikov *et al.*, 2014; Romero *et al.*, 2015; Stiffler *et al.*, 2015; Wrenbeck *et al.*, 2017a). However, the extent to which the results from any individual study are generalizable to related protein homologs is not known. The family permutation profile generated through our study identified sequence-function relationships that are consistent across multiple protein homologs. These results suggest that family mutational profiles can be used to glean permutation rules for protein homologs that exhibit similar biological activities but differ in sequence and stability.

In our experiments, we chose to vary AK thermostability, while assessing protein function at a single temperature. Previous studies have shown that thermostability buffers proteins from the disruptive effects of random amino acid substitutions (Bloom *et al.*, 2005; Bershtein *et al.*, 2006) and backbone fission (Segall-Shapiro *et al.*, 2011), suggesting that it should be generalizable to topological mutations like circular permutation that alter local conformational flexibility. Our measurements extend this trend to circular permutation by showing that the fraction of functional AK increases with the T_m of each AK homolog, from 25.8% (*Bg*-AK) to 43.9% (*Bs*-AK) to 59.7% (*Gs*-AK). This observation is consistent with a previous study examining tolerance to permutation in *Tn*-AK, an AK with even higher thermostability (T_m = 99.5°C) than the homologs studied herein. Using the same cellular assay, this study found that new termini were tolerated in 65.5% of the circular-permuted *Tn*-AK sampled (Atkinson *et al.*, 2018).

Using AK homologs of varying thermostability for mutational studies at a single temperature is akin to altering the temperature of a reaction being studied with a single enzyme (Hobbs *et al.*, 2013; Arcus *et al.*, 2016). Those proteins that evolved to function closest to the assay temperature typically exhibit enhanced flexibility at that temperature compared to the more thermostable variants (Radestock and Gohlke, 2011; Elias *et al.*, 2014). This additional mobility is thought to allow these proteins to sample more conformations in the native basin of the energy landscape resulting in lower occupancy of the functionally folded state, while more rigid thermostable variants occupy a narrower region of the energy landscape increasing the fraction of functionally folded enzyme (Závodszky *et al.*, 1998; Chan *et al.*, 2004; Radestock and Gohlke, 2011; Elias *et al.*, 2014). Mutations that locally perturb this flexibility/rigidity trade-off can have different fitness effects depending on their domain locations (Tokuriki and Tawfik, 2009; Dong *et al.*, 2018). This trade-off has been observed in chimeras created by recombining *Bs*-AK (T_m = 47.6°C) and *Gs*-AK (T_m = 74.5°C) (Bae and Phillips, 2006). A chimera made up of the *Gs*-AK core domain and the *Bs*-AK AMP-binding and lid domains exhibited stability like *Gs*-AK but activity that is higher than both parental proteins (Bae and Phillips, 2006). In contrast, a chimera having the *Bs*-AK core domain and the *Gs*-AK mobile domains displayed stability like *Bs*-AK and activity that is lower than both parent proteins (Bae and Phillips, 2006).

Our results suggest that circular permutation profiling across enzymes exhibiting a range of stabilities represents a simple way to systematically perturb flexibility and identify regions that have been selected to promote folding stabilization versus conformational flexibility. Permuted AKs having protein termini within the core domain presented fitness that is largely dependent upon thermostability. In contrast, permuted AKs having protein termini within the mobile AMP-binding domain frequently exhibited parent-like fitness across all three AK homologs, and thus, was independent of protein thermostability. The lid domain presented tolerance to protein termini that was intermediate to that observed with the core and AMP-binding domains. These findings suggest that there is a benefit to using thermostability as a variable when performing family permutation profiling, since it allows for measurements that decouple the mutational fitness effects on folding and dynamics.

Comparing protein fitness following permutation with the contact energies at each position where protein termini were introduced revealed a correlation. Sites having high energetic frustration were more functionally tolerant to new termini created by permutation compared with low frustration sites. Permutations in the AMP-binding domain, which has been selected for the highest contact energies, yielded variants that exhibit near WT levels of total activity in cells. Since the AMP-binding domain is thought to control product release (Whitford *et al.*, 2007; Saavedra *et al.*, 2018), the rate-limiting step of catalysis in AK, our findings suggest that the extra conformational flexibility created by permutation does not affect this mechanistic role. Additionally, our results suggest that the extra conformational flexibility arising from new termini in the AMP-binding domain does not disrupt folding, substrate binding, or catalytic activity (Fig. 6D). The variants with the lowest fitness across all three homologs arose from the creation of new protein termini at different locations within the core domain. Since these new termini were generated at locations that have been selected for low contact energies and previous biophysical studies have implicated the core domain as critical to overall protein thermostability (Bae and Phillips, 2006), these findings suggest that increases in conformational flexibility arising from permutation in this domain are deleterious to folding.

Our discovery of permuted variants with cellular fitness that exceeds the parental proteins suggests that these permuted AKs may differ in activity and/or folding from the parental AKs. In future studies, it will be interesting to characterize the biochemical and biophysical properties of these and other topological mutants discovered herein. Specifically, it will be interesting to investigate how permuted

AK activities vary with temperature and whether the mechanism of product release is altered by permutation in the AMP-binding and/or lid domain. Increased conformational flexibility in the AMP-binding domain caused by backbone cleavage may facilitate product release and improved enzyme turnover rates (Whitford *et al.*, 2007; Saavedra *et al.*, 2018) as has been observed in other enzymes that undergo substantial conformational changes as part of their enzymatic cycle (Guntas *et al.*, 2012; Daugherty *et al.*, 2013). Additionally, it will be interesting to explore which of the topological mutants that are inactive retain structures that are similar to one of the conformational states sampled by the native proteins but are unable to navigate the full catalytic cycle because of changes in substrate binding or dynamic conformational changes.

Our results indicate that CPP-seq is a powerful tool in protein engineering for scanning and identifying positions where increasing local flexibility enhances function. This is counter to the prevailing approach of targeting increased native-state stability through substitutions that improve packing (Korkegian *et al.*, 2005) or decrease conformational entropy (Anil *et al.*, 2006). Circular permutation scanning of AKs supports the use of such strategies for their core domains, but not for positions proximal to the active site. A broad survey of high-resolution enzyme structures finds that active sites consistently exhibit higher energetic frustration than the rest of the protein (Freiberger *et al.*, 2019), suggesting that increasing active site conformational entropy may be a general engineering strategy for enzymes beyond AK. Given that introducing chain termini in the middle of a protein sequence can cause structural and energetic perturbations other than increased local flexibility, i.e. the larger size of N- and C-termini relative to the peptide bond, and the unfavorable desolvation energy of burying formally charged termini, it is quite possible that the fitness increases observed from near-active-site permutations may underestimate the potential for functional optimization at such positions through point mutations. In the future, it will be interesting to investigate whether CPP-seq can be used to identify positions where conformational entropy may be enhanced through substitutions with glycine or other small amino acids that reduce local packing. To obtain family permutation profiles that more comprehensively sample each native position, oligo-synthesis technology can be used in the future to build libraries with more uniform coverage as was recently done for domain-insertion profiling with ion channel genes (Coyote-Maestas *et al.*, 2020).

Supplementary data

Supplementary data are available at *PEDS* online.

Acknowledgments

We are grateful to Dr. George Phillips and Jose L. Olmos for their gift of the plasmids pNIC28-BgAK, pNIC28-BsAK, and pNIC28-GsAK. We thank Drs. Kevin R. MacKenzie, Peter Wolynes, and George N. Phillips for helpful discussions.

Funding

This work was supported by the National Science Foundation, including Molecular and Cellular Biosciences grant number [1150138 to J.J.S.] and Graduate Research Fellowships to A.M.J. and J.T.A. Additionally, this work was supported by the Office of Naval Research grant [N00014-17-1-2639 to J.J.S.]. A.M.J. was partially funded by a training fellowship from the Keck Center of the Gulf Coast Consortia, on the Houston Area Molecular Biophysics Program, National Institute of General Medical Sciences (NIGMS) [T32GM008280]. J.T.A. was partially supported by a Lodieska Stockbridge Vaughn Fellowship.

Author contributions

A.M.J., J.T.A., and J.J.S. designed the study. A.M.J. and J.T.A. performed experiments. J.T.A. analyzed the data. J.T.A., V.N., and J.J.S. wrote the manuscript.

Competing interests

The authors declare no competing interests.

References

Anders, S. and Huber, W. (2010) *Genome Biol.*, **11**, R106.

Anil, B., Craig-Schapiro, R., Raleigh, D.P. (2006) *J. Am. Chem. Soc.*, **128**, 3144–3145.

Araya, C.L. and Fowler, D.M. (2011) *Trends Biotechnol.*, **29**, 435–442.

Araya, C.L., Fowler, D.M., Chen, W., Muniez, I., Kelly, J.W., Fields, S. (2012) *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 16858–16863.

Arcus, V.L., Prentice, E.J., Hobbs, J.K., Mulholland, A.J., Van der Kamp, M.W., Pudney, C.R., Parker, E.J., Schipper, L.A. (2016) *Biochemistry*, **55**, 1681–1688.

Atkinson, J.T., Jones, A.M., Zhou, Q., Silberg, J.J. (2018) *Nucleic Acids Res.*, **46**, e76.

Bae, E. and Phillips, G.N. (2004) *J. Biol. Chem.*, **279**, 28202–28208.

Bae, E. and Phillips, G.N. (2006) *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 2132–2137.

Bandyopadhyay, B., Goldenzweig, A., Unger, T., Adato, O., Fleishman, S.J., Unger, R., Horovitz, A. (2017) *J. Biol. Chem.*, **292**, 20583–20591.

Baxa, M.C., Haddadian, E.J., Jumper, J.M., Freed, K.F., Sosnick, T.R. (2014) *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 15396–15401.

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., Tawfik, D.S. (2006) *Nature*, **444**, 929–932.

Bloom, J.D. (2015) *BMC Bioinform.*, **16**, 168.

Bloom, J.D., Silberg, J.J., Wilke, C.O., Drummond, D., Adami, C., Arnold, F.H. (2005) *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 606–611.

Bryngelson, J. and Wolynes, P. (1987) *Proc. Natl. Acad. Sci. U. S. A.*, **84**, 7524–7528.

Chan, C.H.H., Liang, H.K.K., Hsiao, N.W.W., Ko, M.T.T., Lyu, P.C.C., Hwang, J.K.K. (2004) *Proteins*, **57**, 684–691.

Coyote-Maestas, W., Nedrud, D., Okorafor, S., He, Y., Schmidt, D. (2020) *Nucleic Acids Res.*, **48**, 1010.

Cronan, J.E., Ray, T.K., Vagelos, P.R. (1970) *Proc. Natl. Acad. Sci. U. S. A.*, **65**, 737–744.

Daugherty, A.B., Govindarajan, S., Lutz, S. (2013) *J. Am. Chem. Soc.*, **135**, 14425–14432.

Dong, Y.W.W., Liao, M.L.L., Meng, X.L.L., Somero, G.N. (2018) *Proc. Natl. Acad. Sci. U. S. A.*, **115**, 1274–1279.

Doolan, K.M. and Colby, D.W. (2015) *J. Mol. Biol.*, **427**, 328–340.

Elazar, A., Weinstein, J., Biran, I., Fridman, Y., Bibi, E., Fleishman, S.J. (2016) *Elife*, **5**, e12125.

Elias, M., Wieczorek, G., Rosenne, S., Tawfik, D.S. (2014) *Trends Biochem. Sci.*, **39**, 1–7.

Ferreiro, D.U., Hegler, J.A., Komives, E.A., Wolynes, P.G. (2007) *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 19819–19824.

Ferreiro, D.U., Hegler, J.A., Komives, E.A., Wolynes, P.G. (2011) *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 3499–3503.

Firnberg, E., Labonte, J.W., Gray, J.J., Ostermeier, M. (2014) *Mol. Biol. Evol.*, **31**, 1581–1592.

Fowler, D.M., Araya, C.L., Gerard, W., Fields, S. (2011) *Bioinformatics*, **27**, 3430–3431.

Fowler, D.M. and Fields, S. (2014) *Nat. Methods*, **11**, 801–807.

Freiberger, M.I., Guzovsky, A.B., Wolynes, P.G., Parra, R.G., Ferreiro, D.U. (2019) *Proc. Natl. Acad. Sci. U. S. A.*, **116**, 4037–4043.

Glaser, P., Presecan, E., Delepierre, M., Surewicz, W., Mantsch, H., Bârzu, O., Gilles, A. (1992) *Biochemistry*, **31**, 3038–3043.

Goethe, M., Gleixner, J., Fita, I., Rubi, J.M. (2018) *J Chem Theory Comput*, **14**, 1811–1819.

Guntas, G., Kanwar, M., Ostermeier, M. (2012) *PLoS One*, **7**, e35998.

Haase, G., Brune, M., Reinstein, J., Pai, E., Pingoud, A., Wittinghofer, A. (1989) *J. Mol. Biol.*, **207**, 151–162.

Henzler-Wildman, K.A., Lei, M., Thai, V., Kerns, S., Karplus, M., Kern, D. (2007) *Nature*, **450**, 913–916.

Higgins, S.A. and Savage, D.F. (2018) *Biochemistry*, **57**, 38–46.

Hobbs, J.K., Jiao, W., Easter, A.D., Parker, E.J., Schipper, L.A., Arcus, V.L. (2013) *ACS Chem. Biol.*, **8**, 2388–2393.

Jones, A.M., Mehta, M.M., Thomas, E.E., Atkinson, J.T., Segall-Shapiro, T.H., Liu, S., Silberg, J.J. (2016) *ACS Synth. Biol.*, **5**, 415–425.

Kerns, S., Agafonov, R.V., Cho, Y.J.J. *et al.* (2015) *Nat. Struct. Mol. Biol.*, **22**, 124–131.

Klesmith, J.R., Bacik, J.P.P., Wrenbeck, E.E., Michalczyk, R., Whitehead, T.A. (2017) *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 2265–2270.

Korkegian, A., Black, M.E., Baker, D., Stoddard, B.L. (2005) *Science*, **308**, 857–860.

Li, J., White, J.T., Saavedra, H. *et al.* (2017) *Elife*, **6**, e30688.

Li, W., Wolynes, P.G., Takada, S. (2011) *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 3504–3509.

Lindström, I. and Dogan, J. (2018) *ACS Chem. Biol.*, **13**, 1218–1227.

Mehta, M.M., Liu, S., Silberg, J.J. (2012) *Nucleic Acids Res.*, **40**, e71.

Melnikov, A., Rogov, P., Wang, L., Gnirke, A., Mikkelsen, T.S. (2014) *Nucleic Acids Res.*, **42**, e112.

Miyashita, O., Onuchic, J., Wolynes, P. (2003) *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 12570–12575.

Nisthal, A., Wang, C.Y., Ary, M.L., Mayo, S.L. (2019) *Proc. Natl. Acad. Sci.*, **116**, 16367–16377.

Olson, C., Wu, N.C., Sun, R. (2014) *Curr. Biol.*, **24**, 2643–2651.

Olsson, U. and Wolf-Watz, M. (2010) *Nat. Commun.*, **1**, 111.

Onuchic, J., Wolynes, P., Luthey-Schulten, Z., Socci, N. (1995) *Proc. Natl. Acad. Sci. U. S. A.*, **92**, 3626–3630.

Parra, R., Schafer, N.P., Radusky, L.G., Tsai, M.Y.Y., Guzovsky, A., Wolynes, P.G., Ferreiro, D.U. (2016) *Nucleic Acids Res.*, **44**, W356–W360.

Radestock, S. and Gohlke, H. (2011) *Proteins*, **79**, 1089–1108.

Reitinger, S., Yu, Y., Wicki, J. *et al.* (2010) *Biochemistry*, **49**, 2464–2474.

Richard, N.M., Poelwijk, F.J., Raman, A., Gosal, W.S., Ranganathan, R. (2012) *Nature*, **491**, 138–142.

Romero, M.L., Yang, F., Lin, Y.R.R. *et al.* (2018) *Proc. Natl. Acad. Sci. U. S. A.*, **115**, E11943–E11950.

Romero, P.A., Tran, T.M., Abate, A.R. (2015) *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 7159–7164.

Rundqvist, L., Adén, J., Sparrman, T., Wallgren, M., Olsson, U., Magnus, W.W. (2009) *Biochemistry*, **48**, 1911–1927.

Saavedra, H.G., Wrabl, J.O., Anderson, J.A., Li, J., Hilser, V.J. (2018) *Nature*, **558**, 324–328.

Schrank, T.P., Bolen, D., Hilser, V.J. (2009) *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 16984–16989.

Segall-Shapiro, T.H., Nguyen, P.Q., Dos Santos, E.D., Subedi, S., Judd, J., Suh, J., Silberg, J.J. (2011) *J. Mol. Biol.*, **406**, 135–148.

Starita, L.M. and Fields, S. (2015) *Cold Spring Harb. Protoc.*, **2015**, 781–783.

Stiffler, M.A., Hekstra, D.R., Ranganathan, R. (2015) *Cell*, **160**, 882–892.

Tokuriki, N. and Tawfik, D.S. (2009) *Science*, **324**, 203–207.

Tugarinov, V., Shapiro, Y.E., Liang, Z., Freed, J.H., Meirovitch, E. (2002) *J. Mol. Biol.*, **315**, 155–170.

Vieille, C., Krishnamurthy, H., Hyun, H.H.H., Savchenko, A., Yan, H., Zeikus, J. (2003) *Biochem. J.*, **372**, 577–585.

Whitford, P.C., Miyashita, O., Levy, Y., Onuchic, J.N.N. (2007) *J. Mol. Biol.*, **366**, 1661–1671.

Wolynes, P.G. (2015) *Biochimie*, **119**, 218–230.

Wrenbeck, E.E., Azouz, L.R., Whitehead, T.A. (2017a) *Nat. Commun.*, **8**, 15695.

Wrenbeck, E.E., Faber, M.S., Whitehead, T.A. (2017b) *Curr. Opin. Struct. Biol.*, **45**, 36–44.

Závodszky, P., Kardos, J., Svingor, Petsko, G. (1998) *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 7406–7411.

Zheng, W., Schafer, N.P., Wolynes, P.G. (2013) *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 1680–1685.

Zhuravlev, P.I. and Papoian, G.A. (2010) *Q. Rev. Biophys.*, **43**, 295–332.