

Enzyme Stability-Activity Trade-Off: New Insights from Protein Stability Weaknesses and Evolutionary Conservation

Qingzhen Hou, Marianne Rooman,^{*,||} and Fabrizio Pucci^{*,||}



Cite This: *J. Chem. Theory Comput.* 2023, 19, 3664–3671



Read Online

ACCESS |



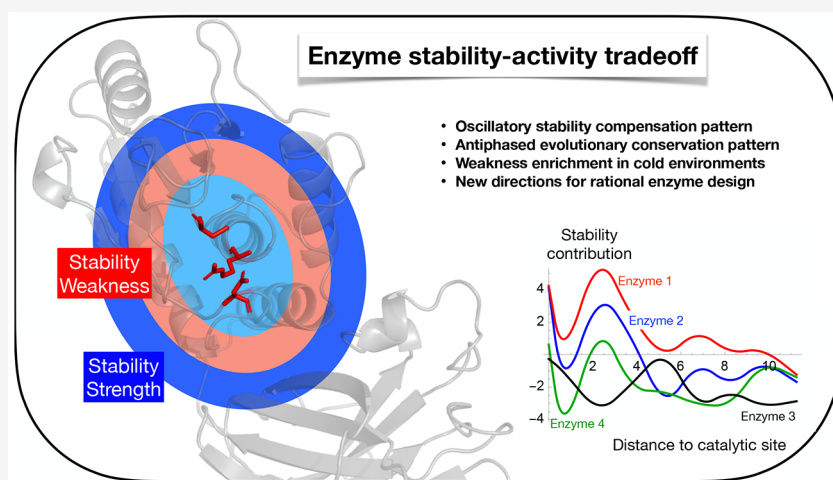
Metrics & More



Article Recommendations



Supporting Information



ABSTRACT: A general limitation of the use of enzymes in biotechnological processes under sometimes nonphysiological conditions is the complex interplay between two key quantities, enzyme activity and stability, where the increase of one is often associated with the decrease of the other. A precise stability-activity trade-off is necessary for the enzymes to be fully functional, but its weight in different protein regions and its dependence on environmental conditions is not yet elucidated. To advance this issue, we used the formalism that we have recently developed to effectively identify stability strength and weakness regions in protein structures and applied it to a large set of globular enzymes with known experimental structure and catalytic sites. Our analysis showed a striking oscillatory pattern of free energy compensation centered on the catalytic region. Indeed, catalytic residues are usually nonoptimal with respect to stability, but residues in the first shell around the catalytic site are, on the average, stability strengths and thus compensate for this lack of stability; residues in the second shell are weaker again, and so on. This trend is consistent across all enzyme families. It is accompanied by a similar, but less pronounced, pattern of residue conservation across evolution. In addition, we analyzed cold- and heat-adapted enzymes separately and highlighted different patterns of stability strengths and weaknesses, which provide insight into the longstanding problem of catalytic rate enhancement in cold environments. The successful comparison of our stability and conservation results with experimental fitness data, obtained by deep mutagenesis scanning, led us to propose criteria for improving catalytic activity while maintaining enzyme stability, a key goal in enzyme design.

INTRODUCTION

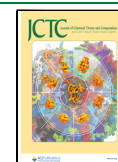
Enzymes are widely used as efficient biological catalysts in a large series of biotechnological and biopharmaceutical applications.¹ In the last decades, a lot of studies have been devoted to design new enzymes with improved stability and turnover by computational and/or experimental approaches.^{2–9} Despite these valuable contributions, it is still unclear how the biophysical principles have shaped the structural stability and dynamics of the catalytic regions of enzymes, which in turn determine their characteristic high efficiency and substrate specificity.

Enzyme catalytic sites are further constrained by the environment in which they must be fully functional since

their host organisms have adapted to sometimes extreme environmental conditions in terms of, e.g., pH, temperature, and ion concentration.^{10–15} For example, some enzymes have evolved to achieve high activity at low temperatures by tuning the flexibility of their catalytic regions,^{16,17} even though recent

Received: January 9, 2023

Published: June 5, 2023



computational approaches have instead suggested that mostly certain regions outside the active site boost entropic contributions to achieve enzymatic rate enhancements.^{18,19}

More generally, the stability-function trade-off of enzymes leads to catalytic regions that are clearly not optimized for stability. They correspond to stability weaknesses in the language we introduced earlier^{20,21} or, almost equivalently, to highly frustrated regions.²² While the key catalytic residues usually remain intact during natural evolution, mutations that modulate the affinity and the turnover rates for substrates are often localized in the periphery of the active site.²³ However, the complex interaction between the stability strengths and weaknesses of the catalytic residues, their periphery, and the rest of the enzymes remains elusive because no experimental method can have direct access to this information. Elucidating this could be of great help in designing new enzymes with increased activity and specificity.

To advance this issue, we used a computational approach based on the well-known formalism of statistical potentials²⁴ to determine strengths and weakness regions^{20,21} in enzymes. This method has proven to be accurate but also much faster than standard molecular dynamics simulations and can therefore be applied for proteome-scale analyses. By doing so, we have better understood not only the stability-function trade-off of enzymes but also how evolution and environmental constraints act to tune this interplay and improve catalytic properties.

METHODS

Data Set of Enzyme Structures. We set up a data set of globular enzymes with annotated catalytic residues and mechanisms. For this purpose, we started by downloading the Catalytic Site Atlas (CSA)²⁵ and identified the list of enzymes with known catalytic sites and properties of which the experimental 3-dimensional (3D) structure is available in the Protein DataBank (PDB).²⁶ From this initial set, we removed the structures that have mutations or post-translational modifications in the catalytic site. This resulted in a set of 883 enzymes.

We then used the software PISCES²⁷ to select the subset of structures that were determined by X-ray crystallography with a resolution of 2.5 Å at most and have a maximum sequence identity of 25% with any other entry of the data set. We thus obtained a set of 633 enzyme structures. Importantly, we considered the biologically active quaternary structure and therefore selected the structures of the biological units (biounits) assigned by the authors of the submission in the PDB; if this information was unavailable, we took the biounit predicted by the PISA program.²⁸

We applied a last filter to the enzyme data set by limiting our analysis to globular (nonmembrane) proteins and dropping hetero-oligomers as well as large homo-oligomeric structures containing more than eight chains. The final data set of enzyme structures, referred to as \mathcal{D}_{CSA} , consists of 551 entries among which 240 monomers and 311 homo-oligomers. The list of all enzymes, their PDB structures, and catalytic residues are available in our repository github.com/3BioCompBio/EnzymeStability.

A crucial parameter in our analyses is the spatial distance of a residue i in a given chain of an enzyme to the active site. We defined this distance as the closest distance between any heavy atom of residue i and any heavy atom of the catalytic residues in all the chains of the enzyme. We chose this distance

definition to avoid as much as possible the impact of the size and geometry of the catalytic site. Indeed, considering instead the distance to the geometric center of the catalytic site would mix up, e.g., residues that are close to a large catalytic site and residues that are distant from a small catalytic site.

Identification of Stability Strength and Weakness Regions. In order to identify stability strengths and weaknesses in a protein structure, we used the SWOTein program that we developed earlier.^{21,24} We will briefly review the main ideas behind SWOTein and refer the reader to Supplementary Section S1 and to the original papers for details.

Statistical potentials are one of the key ingredients of the SWOTein algorithm. They are knowledge-based mean force potentials derived from frequencies of sequence-structure associations computed from a well curated data set of experimentally resolved protein structures. More precisely, the free energy of the association between a given structure element c and a sequence element s is obtained from the frequency of observation of their association (c, s) in the structure data set using the inverse Boltzmann law.

SWOTein²⁴ uses three different kinds of statistical potentials, noted *acc*, *tor*, and *dis*. In these potentials, the c element is either the solvent accessibility of a given residue (*acc*), its backbone torsion angle domain (*tor*) or the inter-residue distance between two residues (*dis*), and the sequence element s is a pair of amino acid types. Using these potentials and partitioning the free energy into per-residue contributions, we defined three folding free energy values for each residue i , ΔG_i^{acc} , ΔG_i^{dis} , and ΔG_i^{tor} , which we linearly combined into a unique per-residue folding free energy contribution as

$$\Delta G_i = \frac{1}{3}(\Delta G_i^{acc} + \Delta G_i^{dis} + \Delta G_i^{tor}) \quad (1)$$

Note that we just summed the free energy contributions without adding weight factors. We made this choice because we prefer avoiding additional parameters and because there are no experimental measurements of the folding free energy per residue that can be used to identify these parameters.

With our conventions, negative ΔG_i values identify regions called stability strengths, which strongly contribute to the stability of the overall protein structure. Positive ΔG_i values, however, indicate stability weaknesses, which correspond to regions that are not stable by themselves but are often optimized for functional reasons rather than for their contribution to the stability of the native structure.^{20,29}

RESULTS

Investigating the biophysical and structural features of catalytic regions in enzymes is crucial to understand their role in the reactions they catalyze, and this understanding is necessary to be able to rationally modify enzyme specificity and turnover. Here we performed a systematic analysis of the patterns of stability and evolutionary conservation inside and outside catalytic regions in a large data set of enzyme structures. We also examined whether these patterns differ between cold- or heat-adapted enzymes and whether they correlate with experimentally measured fitness.

Strengths and Weaknesses in Catalytic Regions. We first analyzed the average pattern of stability strengths and weaknesses in catalytic sites and their periphery. For this purpose, we considered three types of statistical potentials labeled *dis*, *acc*, and *tor*, which are based on the preferences of

amino acids to be separated by a certain distance, to have a certain solvent accessibility, or to adopt a certain backbone torsion angle (see [Methods](#)). Using these three potentials, we computed the folding free energy contributions ΔG_i^{dis} , ΔG_i^{acc} , and ΔG_i^{tor} of every residue i of the enzyme data set \mathcal{D}_{CSA} as well as ΔG_i , the combination of these contributions defined in [eq 1](#). In parallel, we computed the distance d_i of each residue i to the closest catalytic residue (see [Methods](#)).

The per-residue folding free energy contributions as a function of the distance to the active site, averaged over all proteins of the \mathcal{D}_{CSA} set are shown in [Figure 1](#). Examples of

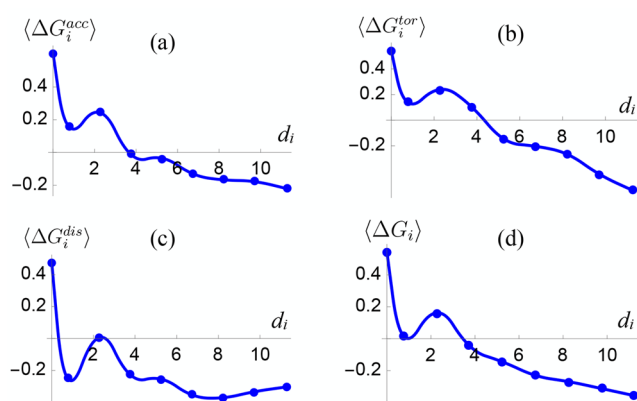


Figure 1. Overall folding free energy contribution $\langle \Delta G_i \rangle$ (in kcal/mol) of each residue i as a function of its distance d_i (in Å) from the closest catalytic residue. The folding free energy contributions were averaged over bins of 1.5 Å width and over all proteins from the \mathcal{D}_{CSA} set. The blue dots are these average values and the curves were obtained by interpolation using Wolfram Mathematica.³⁰ The folding free energy contributions were computed with (a) the solvent accessibility potential ($\langle \Delta G_i^{acc} \rangle$), (b) the backbone torsion angle potential ($\langle \Delta G_i^{tor} \rangle$), (c) the inter-residue distance potential ($\langle \Delta G_i^{dis} \rangle$), and (d) their combination ($\langle \Delta G_i \rangle$).

per-protein curves are given in [Figure 2](#) and [Supplementary Figure S1](#) for four enzymes; the plots for every enzyme from \mathcal{D}_{CSA} are available on github.com/3BioCompBio/EnzymeStability.

The first result we learn from [Figures 1](#) and [2](#) is that catalytic residues are, on the average, stability weaknesses for all the

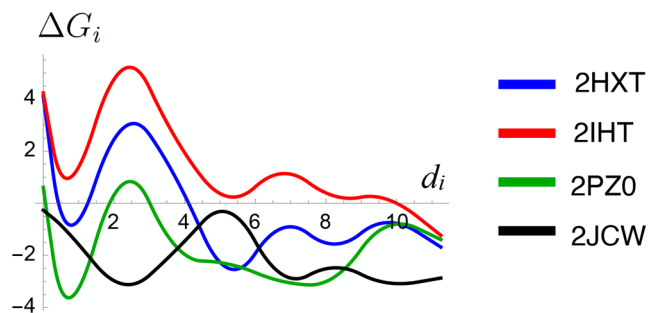


Figure 2. Per-enzyme folding free energy contribution ΔG_i (in kcal/mol) of each residue i as a function of its distance d_i (in Å) from the closest catalytic residue, averaged over bins of 1.5 Å width. The curves were obtained by interpolation³⁰ of the averaged binned values. The enzymes considered are superoxide dismutase (PDB code 2JCW), carboxethylarginine synthase (2IHT), L-fuconate dehydratase (2HXT), and aldo-keto reductase (2PZO).

statistical potentials considered. This is not surprising as they are optimized for their functional role rather than for stability, in agreement with previous reports.^{21,22} For example, the average ΔG_i value for all catalytic residues is equal to 0.54 kcal/mol, which is much higher than the average ΔG_i value of -0.37 kcal/mol obtained for all noncatalytic residues ([Supplementary Table S1](#)).

Interestingly, we observe a clear oscillatory pattern of the average per-residue folding free energy as a function of the distance to the catalytic site for most enzymes in \mathcal{D}_{CSA} taken separately ([Figure 2](#) and github.com/3BioCompBio/EnzymeStability). The residues in the first shell around the catalytic site are more stabilizing, in the next shell they are weaker again, and so on. This effect is damped at larger distances, as the energetic compensation is more needed in the vicinity the active site to maintain the overall structure. The amplitude, period, vertical shift, and damping of the oscillations differ according to the protein and, to a smaller extent, according to the potential ([Supplementary Figure S2](#)).

In spite of this variability, an average oscillatory pattern emerges when considering all enzymes in \mathcal{D}_{CSA} together, and this pattern is almost the same for all the potentials considered, as shown in [Figure 1](#). In the first shell, at about 1–2 Å from the catalytic center, the residues are less weak than the catalytic residues; they are sometimes even strengths. These residues still play a role in the catalytic reaction by energetically compensating for the important stability weaknesses of the catalytic residues. In the second shell, at distances of 2–3 Å, the residues become weak again, while at distances greater than 4 Å their contribution becomes increasingly stabilizing.

This stability compensation pattern appears as an emerging signature of enzymes in which residues close to the active site tend to counterbalance the stability weaknesses of the catalytic residues. Although the three considered potentials are defined from different conformational descriptors, they show a similar stability pattern, which gives further support to its generality. We would like to point out that the average oscillatory pattern of [Figure 1](#) appears despite the large variability of the per-enzyme folding free energy contributions visible in [Figure 2](#), with a standard deviation of about 0.5 kcal/mol. Note also that the damped oscillatory behavior is statistically more probable than a linear model, in spite of this high standard deviation, as we showed in [Supplementary Section S2](#).

One could argue that the observed pattern is the result of a nontrivial bias related to solvent accessibility, as residues in the core usually contribute more to stability than those at the surface. We thus repeated the same computation by averaging the ΔG_i values over all residues i that are situated either in the core (accessibility $\leq 15\%$) or at the surface (accessibility $\leq 15\%$). The results plotted in [Figure 3](#) clearly show the same stability compensation patterns regardless of solvent accessibility. The only difference is the magnitude of the compensation, in other words, the height of the bump, which is about 0.2 kcal/mol in the core and less than 0.1 kcal/mol at the surface. The stability compensation in the core is thus stronger than at the surface, on average. Note that this could be partly due to the averaging of the per-protein oscillatory patterns.

Finally, we investigated the possible differences in the observed strength/weakness patterns when focusing on different classes of enzymes. We considered for that purpose the enzyme commission (EC) nomenclature. We plotted in [Figure 4](#) the average $\langle \Delta G_i \rangle$ contributions as a function of the

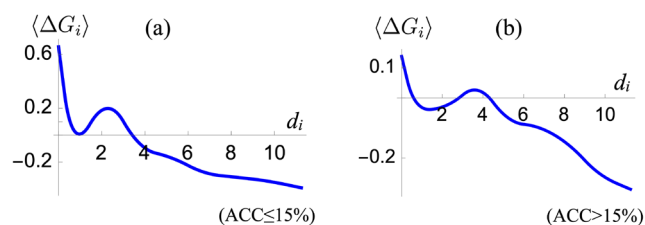


Figure 3. Folding free energy contribution $\langle \Delta G_i \rangle$ (in kcal/mol) for each residue i as a function of the distance d_i (in Å) from the closest catalytic residue. The folding free energy contributions were averaged over bins of 1.5 Å width and over all proteins from the \mathcal{D}_{CSA} set. The curves were obtained by interpolation³⁰ of the averaged binned values. The residues i are limited to (a) core residues with solvent accessibility smaller than or equal to 15%; (b) surface residues with solvent accessibility greater than 15%.

residue distance from the catalytic residues, separately for all enzymes belonging to the classes: EC1 (oxidoreductases), EC2 (transferases), EC3 (hydrolases), EC4 (lyases), EC5 (isomerases), and EC6 (ligases). The results of Figure 4 clearly confirm that the stability compensation patterns are a universal trend shared by all enzymes and are independent of the type of chemical reaction they catalyze. The strength of the compensation, however, slightly varies between enzyme classes.

Strengths and Weaknesses across Evolution. Strength and weakness regions play a pivotal role in stability and function of proteins. It is therefore interesting to analyze how natural evolution shapes their interplay. Previous investigations have speculated that evolution tends to minimize the stability weaknesses of macromolecules or, in other words, their level of frustration.^{31,32} However, residual weaknesses are necessary for the enzymes to be functional.²² To analyze the role of natural evolution, we estimated per-residue evolutionary rates using the ConSurf Web server,³³ which generates and analyses multiple sequence alignments of protein families and outputs an evolutionary score S_i for each residue i , with the convention that the lower S_i , the more conserved residue i in the protein family.

First of all, we plotted the average evolutionary score $\langle S_i \rangle$ as a function of residue distance d_i from the closest catalytic residue (Figure 5a). We also computed the Pearson correlation coefficient r between $\langle S_i \rangle$ and d_i and found a positive correlation with r equal to 0.39. This is expected, as catalytic

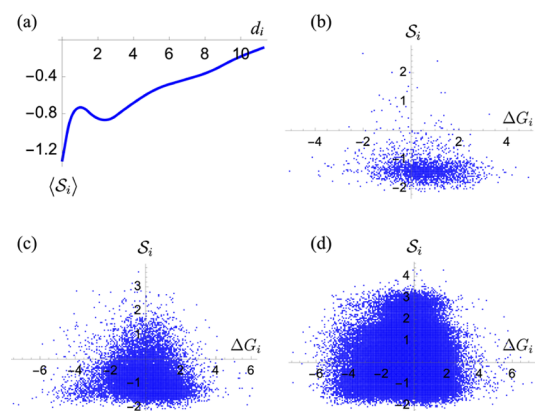


Figure 5. Relation between evolutionary conservation, folding free energy and distance to the active site. (a) Evolutionary score $\langle S_i \rangle$ of residue i as a function of its distance d_i (in Å) from the closest catalytic residue, averaged over bins of 1.5 Å and all proteins from the \mathcal{D}_{CSA} set. (b-d) Evolutionary score S_i as a function of the folding free energy ΔG_i (in kcal/mol) for (b) catalytic residues, (c) residues close to the catalytic site ($0 \text{ Å} < d_i \leq 5 \text{ Å}$) and (d) all other residues.

residues are generally very well conserved for obvious functional reasons. However, what is less expected is that there is a kind of nontrivial compensation similar to what we observed in the stability-distance plot of Figure 1: catalytic residues are highly conserved; the closest residues in the 1–2 Å shell are a bit less conserved; the residues of the second shell, at distances of 2–3 Å from the catalytic residues, are again more conserved; and for the residues distant of more than 4 Å, conservation gradually decreases.

To analyze this result in more detail, we examined the patterns that appear for individual enzymes. As shown in Supplementary Figure S2, an oscillatory pattern emerges of $\langle S_i \rangle$ as a function of the residue distance d_i , which is, however, much less pronounced than the ΔG_i oscillatory pattern. Interestingly, the oscillations are antiphased: shells of weaknesses correspond, on the average, to shells of well conserved residues, and shells of strengths, to shells of less conserved residues.

Although we found the distance from the catalytic site (d_i) to be correlated with the residue conservation (S_i) and weakly anticorrelated with the folding free energy (ΔG_i), we did not find any significant direct correlation or anticorrelation between S_i and ΔG_i . Indeed, the Pearson correlation

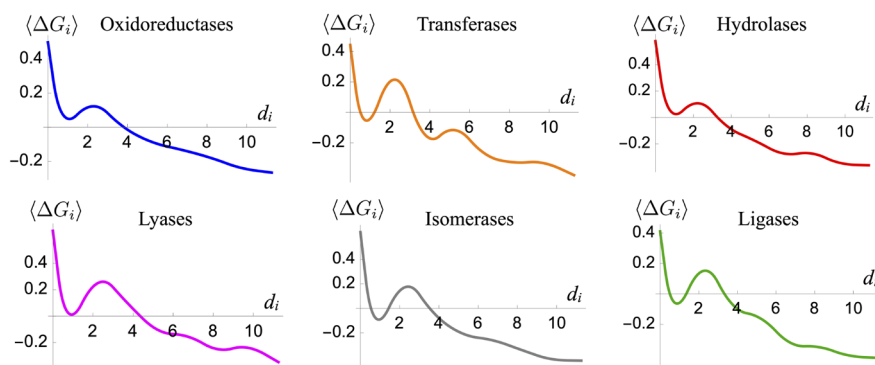


Figure 4. Folding free energy contribution $\langle \Delta G_i \rangle$ (in kcal/mol) of residue i as a function of its distance d_i (in Å) from the closest catalytic residue, averaged over bins of 1.5 Å width, for different EC classes of enzymes. The curves are obtained by interpolation³⁰ of the averaged binned values.

coefficient is equal to $r = -0.03$ and remains low when considering the three free energy contributions ΔG_i^{dis} , ΔG_i^{acc} , and ΔG_i^{or} separately. This correlation is not surprising: a good anticorrelation would mean that all the weak residues are well conserved, and a good correlation, that all the strong residues are well conserved. Obviously, neither is true.

To explore this issue further, we plotted S_i as a function of ΔG_i for the whole \mathcal{D}_{CSA} data set containing more than 200k residues; we divided these residues into three groups: catalytic residues (Figure 5b), residues in the periphery of the active site, i.e. at distances of 5 Å at most (Figure 5c), and residues far from all active sites, at distances of more than 5 Å (Figure 5d). We observe that catalytic residues are well conserved ($S_i < 0$) and weak ($\Delta G_i > 0$ kcal/mol). Note, however, that the S_i and ΔG_i values for these residues are only poorly anticorrelated with a Pearson correlation coefficient of -0.14 . This is because catalytic residues are weaker only on average and, moreover, the most conserved residues are not necessarily the weakest.

For the residues in the neighborhood of the catalytic sites, there are two categories: conserved residues ($S_i < 0$) and poorly conserved residues ($0 < S_i < 1.5$); the former span a large range of ΔG_i values and can thus be either strong or weak, while the latter usually have neutral stability contributions ($\Delta G_i \approx 0$ kcal/mol). Among the residues far from the active site, many are well conserved ($S_i < 0$) and many others not ($S_i > 0$), many are weak ($\Delta G_i > 0$ kcal/mol) and many others are strong ($\Delta G_i < 0$ kcal/mol). However, in general, there are more strong than weak residues in this category, and this is especially true for conserved residues ($S_i < 0$ and $\Delta G_i < 0$ kcal/mol).

We can thus conclude that, among the conserved residues, there is a substantial amount of weak residues and a substantial amount of strong residues. The former ensure the correct functioning of the enzyme and the latter, its overall structure and stability. This analysis clearly reflects the trade-off between stability and function and explains the insignificant correlation between folding free energy (ΔG_i) and evolutionary conservation and (S_i).

Strengths and Weaknesses as Environmental Adaptation Mechanisms. We investigated if and how the patterns of stability strengths and weaknesses are tuned to adapt to different environmental conditions. It has long been known that cold-adapted proteins have a different enthalpy–entropy balance than heat-adapted proteins, so that they can maintain catalytic activity at low temperatures.¹¹ Multiple studies suggested that an improved flexibility in the regions around the active site is necessary to allow such enthalpy–entropy rearrangement.^{34–37} Recent computational analyses, however, point out that the whole surface outside the active site, which is usually not well conserved, is adapted through mutations in order to boost entropy contributions and achieve enzymatic rate enhancement at low temperature.^{18,19}

To further investigate this issue, we first annotated all structures of the \mathcal{D}_{CSA} data set with the environmental temperature (T_{env}) of their host organisms using data collected earlier.³⁸ We then formed two groups of proteins according to whether their host organisms are cold-adapted ($T_{env} < 25$ °C) or heat-adapted ($T_{env} \geq 40$ °C). We computed the folding free energy value ΔG_i of all residues i in these two data sets as a

function of their distance d_i to the active site; the results are shown in Figure 6.

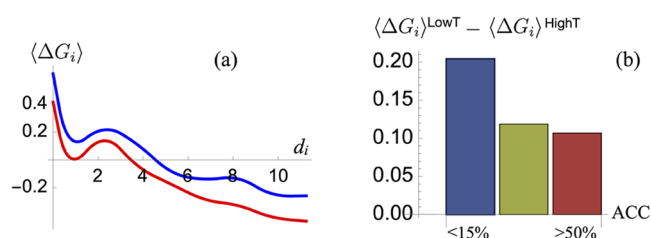


Figure 6. Difference in folding free energy between cold- and heat-adapted proteins from the \mathcal{D}_{CSA} set. (a) Average per-residue folding free energy value $\langle \Delta G_i \rangle$ (in kcal/mol) as a function of the residue distance (in Å) to the closest catalytic site for residues in cold-adapted (blue) and heat-adapted proteins (red). (b) Difference in per-residue folding free energy (in kcal/mol) averaged over all residues i between cold- and heat-adapted proteins in three protein regions: core (accessibility $\leq 15\%$), partially buried ($15\% < \text{accessibility} \leq 50\%$), and surface (accessibility $> 50\%$).

We found that residues occurring in proteins from cold-adapted organisms are, on the average, weaker than residues from heat-adapted organisms. The average ΔG_i difference between the two sets is about 0.2 kcal/mol for any distance to the catalytic site (Figure 6a). Residues close to the catalytic site in cold-adapted organisms are thus weaker, as they have to allow an enhanced flexibility that in turn allows them to maintain the catalytic activity at low temperature.

We also analyzed the ΔG_i distribution of surface residues for the sets of cold- and heat-adapted enzymes and found that the average ΔG_i of the former is higher by 0.1 kcal/mol than that of the latter. This trend is also observed in the partially buried region and in the core of the protein where it is even bigger, with a ΔG_i difference of 0.2 kcal/mol (Figure 6b).

Our computation suggests that both flexibility in the active site regions and a general weakening of the structure play an essential role in the entropy–enthalpy shift of cold-adapted enzymes that allow them to remain functional in such lower temperatures.

Comparison with Experimental Data. For comparing our computational results with experimental data, we used recently published fitness data obtained for the phosphatase and tensin homologue (PTEN).³⁹ This enzyme is an oncosuppressor, which plays a fundamental role in the negative regulation of the proliferative phosphatidylinositol-3 kinase (PI3K) signaling pathway, by dephosphorylating the signaling lipid phosphatidylinositol (3,4,5)-trisphosphate (PIP3).^{40–42} Due to its involvement in cancer, lots of efforts have been devoted to understand the impact of mutations on its catalytic function.

We exploited deep mutagenesis scanning data, in which the *in vivo* impact of 7,244 single amino acid variants on the lipid phosphatase activity of PTEN was measured.³⁹ We computed the average of the measured fitness values of all single amino acid substitutions at a given residue position i , noted \mathcal{F}_i , and compared it with the per-residue evolutionary conservation scores S_i computed using CONSURF³³ and with the structural stability values ΔG_i obtained using SWOTEin.²⁰ We also computed the distance d_i of every residue i in PTEN to the closest of the active site residues Cys124 and Arg130.

Evolution ($\langle S_i \rangle$) and fitness ($\langle \mathcal{F}_i \rangle$) values are, as expected, highly correlated with a Pearson correlation coefficient of 0.60 (P-value $< 10^{-31}$). Conserved residues are indeed those that most contribute to protein fitness. Instead, stability (ΔG_i) values do not show a statistically significant anticorrelation with fitness (P-value = 0.52). This is illustrated in Figure 7, where we

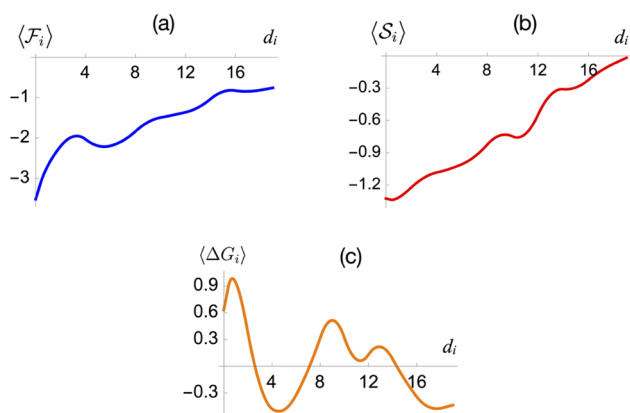


Figure 7. Fitness, stability, and conservation properties in PTEN. (a) Measured per-residue fitness values $\langle \mathcal{F}_i \rangle$ ³⁹ as a function of the distance d_i (in Å) from the catalytic site averaged over 1.5 Å bins; (b) Residue conservation $\langle S_i \rangle$ as a function of d_i (in Å) averaged over 1.5 Å bins; (c) Per-residue folding free energy contributions $\langle \Delta G_i \rangle$ (in kcal/mol) as a function of d_i (in Å) averaged over 1.5 Å bins. The curves were obtained by interpolation³⁰ of the average binned values.

observe their different behaviors when plotted as a function of d_i . Note that if we limit ourselves to residues far from the catalytic center ($d_i > 15$ Å), ΔG_i and \mathcal{F}_i show a good anticorrelation with a Pearson coefficient of -0.26 (P-value < 0.001). Protein stability is thus an important component of fitness outside the catalytic site region: the stronger the residues, the higher their contribution to protein fitness, which results in a good $\mathcal{F}_i - \Delta G_i$ anticorrelation. In the catalytic region, on the contrary, stability weaknesses and strengths both play a functional role and contribute to protein fitness; this explains the insignificant $\mathcal{F}_i - \Delta G_i$ anticorrelation.

We continued by analyzing the correlation between the residue distance d_i to the active site and the per-residue fitness score \mathcal{F}_i , conservation score S_i and stability value ΔG_i . As expected, both fitness \mathcal{F}_i and conservation S_i correlate positively with the distance d_i with a Pearson correlation coefficient of 0.50 (P-value $< 10^{-21}$) and 0.69 (P-value $< 10^{-44}$), respectively. The closer residues are to the active site, the more their mutation has a negative impact on fitness. Consequently, variants close to the active site are more conserved during evolution than variants far from it.

The stability score ΔG_i is anticorrelated with the distance d_i , as catalytic residues and residues in the vicinity are usually enriched in stability weaknesses. The correlation coefficient is weaker in absolute value than for fitness and evolution: -0.18 (P-value < 0.005). This is due to the nonlinear, oscillatory, behavior of this score as a function of the distance, as we pointed out in the previous subsection. We can clearly observe this oscillatory pattern in Figure 7: there are compensations between the stability weakness of the active site region, the stability strength of the first shell of residues at about 4–5 Å, and the stability weakness of the second shell of residues at

about 8–13 Å. Note that this stability weakness-strength compensation is even more complex, as specific compensation patterns also occur within the different residue shells, as shown in Supplementary Figure S3.

Mutations inserted close to the catalytic region have in general a negative impact on protein activity, but this is not the case for mutations further away from this region. It is therefore of interest to identify and analyze the positions in PTEN whose mutations lead to an increased catalytic activity. Here we focused on the ten positions of which the mutations most increase protein activity to find some shared characteristics of these mutations. We first noticed that they are all situated on the protein surface, generally far from the catalytic area, at an average distance of about $d_i \sim 13$ Å. Two main characteristics are shared by these positions: the wild-type amino acids are not conserved ($\langle S_i \rangle = 0.2$) and correspond to a stability weakness as their averaged ΔG_i is 0.6 kcal/mol. This result suggests that targeting nonconserved regions that are far from the catalytic area and that are stability weaknesses could be an interesting strategy to increase enzyme activity.

DISCUSSION

The precise understanding of the trade-off between the two key properties of enzymes, activity and stability, and of how this trade-off is shaped by natural evolution, is important for both theoretical and practical reasons. In particular, it would help design or redesign enzymes that are optimized to work in conditions that differ from the physiological conditions. In this paper we applied our previously developed algorithm, SWOTein,²⁰ to gain insights into this fundamental issue. We summarize our main results below.

First of all, we found a hallmark of the stability in the catalytic region, consisting of an energetic compensation between the catalytic residues, which are usually stability weaknesses, and their neighboring residues, which are rather stability strengths or at least much less weak. A second shell of weak residues is located at distances of 2–3 Å from the catalytic center; these shells are surrounded by damped oscillatory patterns of weaknesses and strengths. This hallmark is general for all types of enzymes, i.e. it does not depend on the catalytic reaction that they catalyze and does not depend on the solvent accessibility of the residues. Moreover, we showed that very similar stability patterns are observed for the three potentials ΔG_i^{dis} , ΔG_i^{acc} , and ΔG_i^{for} describing different types of residue interactions. This highlights that the stability compensation patterns involve not only tertiary interactions and hydrophobic forces but also local interactions along the polypeptide chain.

Our results are in general agreement with an earlier study on frustration in catalytic regions,²² where a slight compensatory behavior between frustrated and less frustrated residue-residue interactions was observed. Stability strengths/weaknesses and frustration are related concepts: the former estimates the favorable or unfavorable contribution of each residue to the stability of the overall fold, and the latter compares the strengths of wild-type and mutated residue-residue interactions.

We also found that residue conservation across evolution shows a nontrivial compensation pattern that is similar to, but much less marked than, the weakness and strength pattern: the catalytic residues are essentially conserved and weak; residues in the 1–2 Å shell around the catalytic center are less

conserved and less weak; residues in the 2–3 Å shell are again more conserved and more weak; at distances of more than 4 Å the residue conservation continues to decrease and the residues become more and more stability strengths. In contrast, a linear relationship between evolutionary conservation and distance to the nearest catalytic residue was found in a earlier study,⁴³ without any compensatory behavior. The disagreement between their results and ours is likely due to their use of wider shells of 5 Å, within which compensation takes place, whereas we considered thinner bins of 1.5 Å.

Interestingly, although the conservation and stability patterns have an antiphased behavior, they are only weakly correlated. This is due to the fact that evolutionary conservation is highly correlated with enzyme fitness, while stability is only one of its main ingredients, the other being function. Note that the weight of stability in fitness increases when moving away from the active site.

We also applied our approach to study the adaptation of enzymes to different temperature conditions. By comparing the stability patterns of cold-adapted and heat-adapted proteins, we found that the former are characterized by a global weakening of the protein structure. More precisely, we observed not only a weaker catalytic pocket but also weaker surface and core regions in cold-adapted proteins than in heat-adapted homologues.

Finally, our study has important perspectives in enzyme design. Indeed, we can take advantage of what we have learned about stability and evolution to identify positions to target in order to increase enzyme activity. These positions are often far from the catalytic site, poorly conserved, and, above all, stability weaknesses. The development of an enzyme activity improvement pipeline will be the subject of a forthcoming study. We would like to emphasize that, currently, the only way to identify functional positions to mutate in order to modulate function requires computationally expensive quantum mechanics/molecular mechanics methods.⁴⁴

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jctc.3c00036>.

Section S1. Statistical potentials and per-residue folding free energies. Section S2. Oscillatory versus linear stability patterns. Table S1. Per-residue folding free energy contributions averaged on all core residues, surface residues, catalytic residues and on all protein residues. Figure S1. Examples of per-enzyme folding free energy contributions ΔG_i^{tor} , ΔG_i^{dis} , and ΔG_i^{acc} as a function of the distance to the closest catalytic residue. Figure S2. Folding free energy contribution ΔG_i and evolutionary conservation S_i of each residue as a function of its distance to the closest catalytic residue. Figure S3. Folding free energy contributions ΔG_i for all residues of PTEN that belong to a given distance range to the closest catalytic residue. (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Marianne Rooman – Computational Biology and Bioinformatics, Université Libre de Bruxelles, 1050 Brussels, Belgium; Interuniversity Institute of Bioinformatics in

Brussels, 1050 Brussels, Belgium;
Email: Marianne.Rooman@ulb.be

Fabrizio Pucci – Computational Biology and Bioinformatics, Université Libre de Bruxelles, 1050 Brussels, Belgium; Interuniversity Institute of Bioinformatics in Brussels, 1050 Brussels, Belgium; orcid.org/0000-0003-2916-022X;
Email: Fabrizio.Pucci@ulb.be

Author

Qingzhen Hou – Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, Shandong 250012, China; National Institute of Health Data Science of China, Shandong University, Jinan, Shandong 250002, China

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.jctc.3c00036>

Author Contributions

^{||}M.R. and F.P. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge financial support from the FNRS - Fund for Scientific Research of Belgium through a PDR Research Project; M.R. is FNRS Research Director.

■ REFERENCES

- (1) Bell, E. L.; Finnigan, W.; France, S. P.; Green, A. P.; Hayes, M. A.; Hepworth, L. J.; Lovelock, S. L.; Niikura, H.; Osuna, S.; Romero, E.; Ryan, K. S.; Turner, N. J.; Flitsch, S. L. Biocatalysis. *Nat. Rev. Methods Primers* **2021**, *1*, 1–21.
- (2) Blomberg, R.; Kries, H.; Pinkas, D. M.; Mittl, P. R.; Grütter, M. G.; Privett, H. K.; Mayo, S. L.; Hilvert, D. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **2013**, *503*, 418–421.
- (3) Zeymer, C.; Hilvert, D. Directed evolution of protein catalysts. *Annual review of biochemistry* **2018**, *87*, 131–157.
- (4) Packer, M. S.; Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **2015**, *16*, 379–394.
- (5) Chica, R. A.; Doucet, N.; Pelletier, J. N. Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Curr. Opin. Biotechnol.* **2005**, *16*, 378–384.
- (6) Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Röthlisberger, D.; Baker, D. New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* **2006**, *15*, 2785–2794.
- (7) Privett, H. K.; Kiss, G.; Lee, T. M.; Blomberg, R.; Chica, R. A.; Thomas, L. M.; Hilvert, D.; Houk, K. N.; Mayo, S. L. Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 3790–3795.
- (8) Broom, A.; Rakotoharisoa, R. V.; Thompson, M. C.; Zarifi, N.; Nguyen, E.; Mukhametzhano, N.; Liu, L.; Fraser, J. S.; Chica, R. A. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **2020**, *11*, 1–10.
- (9) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. Computational enzyme design. *Angew. Chem., Int. Ed.* **2013**, *52*, 5700–5725.
- (10) Somero, G. N. Adaptation of enzymes to temperature: searching for basic “strategies. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **2004**, *139*, 321–333.
- (11) Åqvist, J.; Isaksen, G. V.; Brandsdal, B. O. Computation of enzyme cold adaptation. *Nat. Rev. Chem.* **2017**, *1*, 1–14.

- (12) Vieille, C.; Zeikus, G. J. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiology and molecular biology reviews* **2001**, *65*, 1–43.
- (13) DasSarma, S.; DasSarma, P. Halophiles and their enzymes: negativity put to good use. *Curr. Opin. Microbiol.* **2015**, *25*, 120–126.
- (14) Arcus, V. L.; Van Der Kamp, M. W.; Pudney, C. R.; Mulholland, A. J. Enzyme evolution and the temperature dependence of enzyme catalysis. *Curr. Opin. Struct. Biol.* **2020**, *65*, 96–101.
- (15) Xie, W. J.; Asadi, M.; Warshel, A. Enhancing computational enzyme design by a maximum entropy strategy. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119*, No. e2122355119.
- (16) Stark, C.; Bautista-Leung, T.; Siegfried, J.; Herschlag, D. Systematic investigation of the link between enzyme catalysis and cold adaptation. *Elife* **2022**, *11*, No. e72884.
- (17) Struvay, C.; Feller, G. Optimization to low temperature activity in psychrophilic enzymes. *International journal of molecular sciences* **2012**, *13*, 11643–11665.
- (18) Åqvist, J.; Kazemi, M.; Isaksen, G. V.; Brandsdal, B. O. Entropy and enzyme catalysis. *Accounts of chemical research* **2017**, *50*, 199–207.
- (19) Sočan, J.; Isaksen, G. V.; Brandsdal, B. O.; Åqvist, J. Towards rational computational engineering of psychrophilic enzymes. *Sci. Rep.* **2019**, *9*, 1–10.
- (20) Hou, Q.; Pucci, F.; Ancien, F.; Kwasigroch, J. M.; Bourgeois, R.; Rooman, M. SWOTEin: a structure-based approach to predict stability strengths and weaknesses of prOTEINs. *Bioinformatics* **2021**, *37*, 1963–1971.
- (21) De Laet, M.; Gilis, D.; Rooman, M. Stability strengths and weaknesses in protein structures detected by statistical potentials: Application to bovine seminal ribonuclease. *Proteins: Struct., Funct., Bioinf.* **2016**, *84*, 143–158.
- (22) Freiburger, M. I.; Guzovsky, A. B.; Wolynes, P. G.; Parra, R. G.; Ferreira, D. U. Local frustration around enzyme active sites. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 4037–4043.
- (23) Tokuriki, N.; Stricher, F.; Serrano, L.; Tawfik, D. S. How protein stability and new functions trade off. *PLoS computational biology* **2008**, *4*, No. e1000002.
- (24) Dehouck, Y.; Gilis, D.; Rooman, M. A new generation of statistical potentials for proteins. *Biophysical journal* **2006**, *90*, 4010–4017.
- (25) Ribeiro, A. J. M.; Holliday, G. L.; Furnham, N.; Tyzack, J. D.; Ferris, K.; Thornton, J. M. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic acids research* **2018**, *46*, D618–D623.
- (26) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Veerasamy, R.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. The protein data bank. *Acta Crystallogr. Section D: Biol. Crystallogr.* **2002**, *58*, 899–907.
- (27) Wang, G.; Dunbrack, R. L., Jr. PISCES: a protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589–1591.
- (28) Krissinel, E.; Henrick, K. Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology* **2007**, *372*, 774–797.
- (29) Basu, S.; Assaf, S. S.; Teheux, F.; Rooman, M.; Pucci, F. BRANEart: identify stability strength and weakness regions in membrane proteins. *Front. Bioinformatics* **2021**, *1*, 70.
- (30) Wolfram Inc. *Wolfram Mathematica*, Version 13.2.; Wolfram Inc.: Champaign, IL, 2022; <https://www.wolfram.com/mathematica>.
- (31) Röder, K.; Wales, D. J. Evolved minimal frustration in multifunctional biomolecules. *J. Phys. Chem. B* **2018**, *122*, 10989–10995.
- (32) Ferreira, D. U.; Hegler, J. A.; Komives, E. A.; Wolynes, P. G. Localizing frustration in native proteins and protein assemblies. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 19819–19824.
- (33) Ben Chorin, A.; Masrati, G.; Kessel, A.; Narunsky, A.; Sprinzak, J.; Lahav, S.; Ashkenazy, H.; Ben-Tal, N. ConSurf-DB: An accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. *Protein Sci.* **2020**, *29*, 258–267.
- (34) Feller, G.; Gerday, C. Psychrophilic enzymes: hot topics in cold adaptation. *Nature reviews microbiology* **2003**, *1*, 200–208.
- (35) Saunders, N. F.; Thomas, T.; Curmi, P. M.; Mattick, J. S.; Kuczek, E.; Slade, R.; Davis, J.; Franzmann, P. D.; Boone, D.; Rusterholtz, K.; Feldman, R.; Gates, C.; Bench, S.; Sowers, K.; Kadner, K.; Aerts, A.; Dehal, P.; Detter, C.; Glavina, T.; Lucas, S.; Richardson, P.; Larimer, F.; Hauser, L.; Land, M.; Cavicchioli, R. Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. *Genome research* **2003**, *13*, 1580–1588.
- (36) Siddiqui, K. S.; Cavicchioli, R. Cold-adapted enzymes. *Annu. Rev. Biochem.* **2006**, *75*, 403–433.
- (37) De Maayer, P.; Anderson, D.; Cary, C.; Cowan, D. A. Some like it cold: understanding the survival strategies of psychrophiles. *EMBO reports* **2014**, *15*, 508–517.
- (38) Pucci, F.; Kwasigroch, J. M.; Rooman, M. SCOP: an accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics* **2017**, *33*, 3415–3422.
- (39) Mighell, T. L.; Evans-Dutson, S.; O’Roak, B. J. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *American Journal of Human Genetics* **2018**, *102*, 943–955.
- (40) Chen, C.-Y.; Chen, J.; He, L.; Stiles, B. L. PTEN: tumor suppressor and metabolic regulator. *Front. Endocrinol.* **2018**, *9*, 338.
- (41) Lee, Y.-R.; Chen, M.; Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor: new modes and prospects. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 547–562.
- (42) Song, M. S.; Salmena, L.; Pandolfi, P. P. The functions and regulation of the PTEN tumour suppressor. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 283–296.
- (43) Jack, B. R.; Meyer, A. G.; Echave, J.; Wilke, C. O. Functional sites induce long-range evolutionary constraints in enzymes. *PLoS Biology* **2016**, *14*, No. e1002452.
- (44) Ishida, T. Effects of point mutation on enzymatic activity: correlation between protein electronic structure and motion in chorismate mutase reaction. *J. Am. Chem. Soc.* **2010**, *132*, 7104–7118.