PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

An amino-domino model described by a cross-peptide-bond Ramachandran plot defines amino acid pairs as local structural units

Aviv A. Rosenberg[a], Nitsan Yehishalom[b], Ailie Marx[a,1], and Alex M. Bronstein[a,1]

Edited by Rama Ranganathan, The University of Chicago, Chicago, IL; received January 20, 2023; accepted August 24, 2023 by Editorial Board Member Lila M. Gierasch

Protein structure, both at the global and local level, dictates function. Proteins fold from chains of amino acids, forming secondary structures, α-helices and β-strands, that, at least for globular proteins, subsequently fold into a three-dimensional structure. Here, we show that a Ramachandran-type plot focusing on the two dihedral angles separated by the peptide bond, and entirely contained within an amino acid pair, defines a local structural unit. We further demonstrate the usefulness of this cross-peptide-bond Ramachandran plot by showing that it captures β-turn conformations in coil regions, that traditional Ramachandran plot outliers fall into occupied regions of our plot, and that thermophilic proteins prefer specific amino acid pair conformations. Further, we demonstrate experimentally that the effect of a point mutation on backbone conformation and protein stability depends on the amino acid pair context, i.e., the identity of the adjacent amino acid, in a manner predictable by our method.

Ramachandran plot | dihedral angle | secondary structure | protein structure | Ramachandran outlier

Significance

A large-scale analysis of high-resolution protein structures suggests that amino acid pairs constitute another layer of ordered structure, more local than the conventionally defined secondary structures. We develop a cross-peptide-bond Ramachandran plot that captures the conformational preferences of the amino acid pairs and demonstrate that it conveys biologically meaningful information, not apparent in the traditional Ramachandran plot.

Protein function is enabled by structure, which ranges from being folded and compact to intrinsically disordered and unstable at the tertiary level. Proteins must adopt conformation(s) supporting their biological function through defined and reproducible interactions in the cellular network. This suggests that the apparently disordered regions of the amino acid chain may not be entirely random coil (1) and points toward the existence of order at a level more local than the conventionally defined secondary structures like helices, strands, and turns. Evidence for such local structure includes a work by Eker et al. showing that alanine-based tripeptides, AXA, have intrinsic structural propensities dependent on the identity of amino acid X (2). More recently, it was shown that scrambling the amino acid sequence (without altering overall amino acid composition) of the PUMA BH3 domain led to protein chains with FRET, SAXS, and SEC characteristics consistent with residual structure, even when the circular dichroism signal for the helicity of the wild-type protein was lost (3).

The first attempts to capture protein backbone structure more finely than the well-recognized secondary structures were made more than 30 y ago. Unger et al. suggested that consecutive pairs of ψ and φ dihedral angles are dependent and that these doublet pairs of dihedral angles maintain properties that are lost when a structure is fragmented to individual dihedral angle pairs (4). The same group showed that six-amino-acid-long sections of the backbone cluster into about 100 specific conformations and that entire protein structures can be well reconstructed from these building blocks (5). The idea of creating such structural alphabets from the conformations formed by consecutive residues has since been explored, refined, and used extensively, commonly as "Protein Blocks" of 16 prototypes defined by five consecutive amino acids (6–8). Another related concept is TERMs, where the protein structure is broken down into smaller structural entities which are then used to assess model quality, sequence/structure compatibility, and conformational transitions (9). Alternative to this fragment-based approach, attempts to capture local backbone structure have included describing rotation, twist, and rise-per-residue parameters in helical structures (10), using tripeptides and a measurement of their rigidity (11), and defining a $(\varphi, \psi)_2$ motif (12, 13). The $(\varphi, \psi)_2$ motif defines amino acid pairs as structural elements and is attractive in being very local and restricted to a discrete set of 400 elements. However, the apparent necessity to define four dihedral angles in describing the structure of an amino acid pair is a major disadvantage of the approach.

We suggest a different angle (pair) on defining amino acid pairs as structural elements and analyzing their preferred backbone conformations. We show that amino acid pair conformations are well captured in an "amino-domino" model where each pair is defined

Author affiliations: [a]Department of Computer Science, Technion–Israel Institute of Technology, Haifa 32000, Israel; and [b]Faculty of Biology, Technion–Israel Institute of Technology, Haifa 32000, Israel

Author contributions: A.A.R., A.M., and A.M.B. designed research; A.A.R., N.Y., A.M., and A.M.B. performed research; A.A.R., N.Y., A.M., and A.M.B. analyzed data; and A.A.R., A.M., and A.M.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. R.R. is a guest editor invited by the Editorial Board.

Copyright © 2023 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: ailie@technion.ac.il or alexbronst@gmail.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2301064120/-/DCSupplemental.

Published October 25, 2023.

Downloaded from https://www.pnas.org by Institute of Science and Technology Austria on September 22, 2026 from IP address 78.104.209.6.

by only two dihedral angles, those separated by the peptide bond, specifically ψ of the first amino acid and φ of the second. We denote these cross-bond angle pairs as (ψ_k, φ_{k+1}) where k is an index of the first amino acid. Two important advantages of the amino-domino model are that i) all atoms defining (ψ_k, φ_{k+1}) are part of the amino acid pair, making it the smallest self-contained structural unit, and ii) the description of the amino acid pair remains two-dimensional and has a clear geometric interpretation that can be captured using the familiar Ramachandran plot, but using these two angles (ψ_k, φ_{k+1}) instead of the standard (ψ_k, φ_k). To demonstrate the usefulness of this method for defining local structure, we show that i) the cross-bond Ramachandran plot captures β-turns, structural elements that are not readily identifiable in the conventional Ramachandran plot, ii) outliers in the traditional Ramachandran plot fall into occupied locations of the corresponding cross-bond Ramachandran plots, iii) thermophilic proteins have distinct preferences for amino acid pair conformations which may suggest that these structural elements have different intrinsic thermostabilities, and iv) single point mutations are sensitive to the identity of the adjacent residues in the amino acid chain in a manner which is predictable by comparison of the (ψ_k, φ_{k+1}) distributions of the relevant amino acid pairs.

Results and Discussion

The Two Central Dihedral Angles of an Amino Acid Pair Define a Wider Range of Structural Motifs than the Pair of Dihedral Angles Associated with a Single Amino Acid. The protein backbone is commonly described as a series of dihedral angle pairs, (φ_k, ψ_k). The joint distributions of these angle pairs fall into specific (allowed) regions, compatible with the stereochemical constraints of the molecule, and corresponding to the backbone conformations which are characteristic of the protein secondary structures as shown by Ramachandran almost 60 y ago (13). It is of note that although (φ_k, ψ_k) are conventionally assigned to an individual amino acid, they are in fact not defined solely by that amino acid since each of the dihedrals relies on the coordinates of an atom from each of the adjacent amino acids (Fig. 1*A*). Seeking to identify and readily describe the smallest, self-contained, structural element possible, we describe the protein backbone as a series of overlapping amino acid pairs, analogous to dominos, and captured by the dihedral angle pairs (ψ_k, φ_{k+1}) (Fig. 1*B*).

As a first step, we extracted the coordinates of amino acid pairs from high-resolution protein structures (better than 1.5 Å) and clustered them according to their backbone conformation. This was achieved by first bringing the pairs into a canonical orientation, as follows: i) the location of the C atom of the first amino acid was defined as the origin of the coordinate system; ii) the peptide bond was defined as the x axis; iii) the plane formed by the C^{α} and C atoms of the first amino acid and the N atom of the second amino acid was defined as the xy plane; and iv) the normal to this plane was defined as the z axis (Fig. 1*C*). The spatial coordinates of the canonically oriented C^{α}, C, and N atoms of both amino acids were then stacked into an 18-dimensional vector representing the pair, and the set of all such vectors was divided into twenty clusters. Each cluster was further clustered into 10 subclusters using the same procedure to obtain the 10 representative backbones depicted in Fig. 1*D* and *SI Appendix*, Fig. S2. Next, we calculated the (ψ_k, φ_{k+1}) dihedral angle distribution (Fig. 1*F*) and compared it to the traditional, single amino acid dihedral angle distributions of (φ_k, ψ_k) and $(\psi_{k+1}, \varphi_{k+1})$, coloring the points by cluster (Fig. 1 *E* and *G*). We emphasize that the choice of $k = 20$ clusters was arbitrary as well as their numbering. For comparison, we show $k = 6$, $k = 10$, and $k = 15$ in *SI Appendix*, Fig. S1 and defer exploration into which number of clusters best reflects the underlying physical phenomena for future work.

A clearly observable difference is that the cross-peptide-bond Ramachandran plot of (ψ_k, φ_{k+1}) covers more of the conformational space than the traditional (φ_k, ψ_k) plot. Moreover, the clustered amino acid pair backbone conformations map into more distinctly separate regions in the cross-bond Ramachandran plot compared to the traditional Ramachandran plot.

The increased occupancy of the two-dimensional conformational space might be expected due to the reduced overlap between the sets of atoms defining the angles in the cross-bond pair, separated by the peptide bond. Specifically, the two adjacent dihedrals, (φ_k, ψ_k), share three of the four atoms which define each (Fig. 1*A*), while the cross-bond pair (ψ_k, φ_{k+1}) share only two atoms (Fig. 1*B*). Importantly, this greater independence does not necessarily imply that the trajectories of cross-bond angles under thermodynamic fluctuations would be less dependent than the trajectories of (φ_k, ψ_k), especially at a specific location within a formed and folded protein chain. To test this, we performed molecular dynamics simulations of high-resolution, small proteins. The results indicate that the constituents of the cross-bond pair (ψ_k, φ_{k+1}) are more correlated as those of the regular pair (φ_k,ψ_k) in about 74% of the times (Fig. 2). This observation seems reasonable given the rigidity of the peptide bond that ensures these two angles produce a highly correlated distribution. To understand how secondary structure context affects the correlation coefficients, we plotted the distribution of correlation coefficients for the regular and cross-bond Ramachandran dihedral pairs for ten secondary structure contexts; entrance to, exit from, and the middle of a helix, strand and turn, as well as random coil not belonging to any of the former groups. It is evident from *SI Appendix*, Fig. S3 that in all but one case (entrance to helix), the cross-bond angles exhibit stronger negative correlations compared to their traditional counterparts, a phenomenon particularly prominent in strands.

In sum, we observe that in comparison to the traditional Ramachandran plot, the distribution of angle pairs in the cross-peptide version covers more conformational space while maintaining higher correlation in thermodynamic fluctuations at the local level. Another indication that these observations are physically relevant comes from analysis of outliers in the traditional, (φ_k, ψ_k), Ramachandran plot. We analyzed amino acids with dihedral angles in the disallowed region of the usual Ramachandran plot, defined indisputably in very high (sub 1 Å) resolution crystal structures (14). We found that when plotted in the cross-bond Ramachandran plot, almost all these conformations now fall into occupied regions (Fig. 3). This observation suggests that the cross-bond Ramachandran plot might be more useful than the traditional Ramachandran plot for structure verification. Structures having all dihedrals within the Ramachandran allowed regions are commonly equated as being high quality despite recognition that explainable Ramchandran outliers are not uncommon (15).

The Amino-Domino Model Readily Captures Structural Motifs and Their Features Not Evident in the Conventional Ramachandran Plot. An important feature obtained by shifting the conventional reference frame, from the immediately adjacent angle pair to the cross-bond dihedral angle pair, is that both angles are now entirely contained within the amino acid pair they are being used to describe. This allows for sequence context-free assignment of amino acid pair propensities to different structural motifs. To that end, we calculated, per cluster, the propensity of each amino acid to be in each position of an

Downloaded from https://www.pnas.org by Institute of Science and Technology Austria on September 22, 2026 from IP address 78.104.209.6.

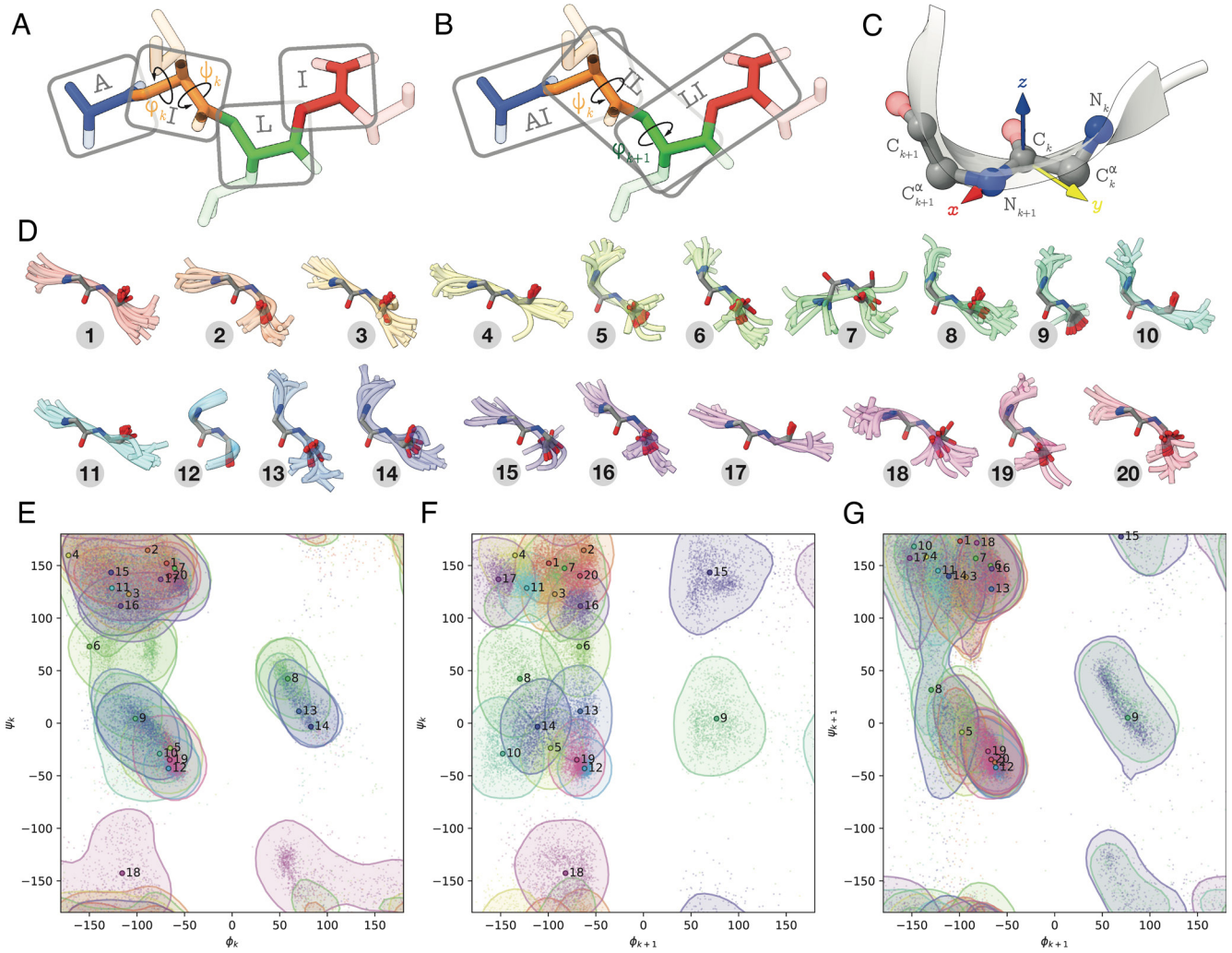


Fig. 1. Describing the protein backbone through a different angle pair better explains conformations adopted by amino acid pairs. (*Top*) rather than treating each amino acid as a single structural unit described by (φ_k, ψ_k) (*A*), we propose overlapping pairs of amino acids described by (ψ_k, φ_{k+1}) as a form of structural "domino blocks" (*B*). Traditionally, dihedral angle pairs are defined by the two atoms on either side of the bond around which they are centered, making the dihedral angles (φ_k, ψ_k) describing a single amino acid dependent on atoms from the two neighboring amino acids: φ_k depends on the preceding nitrogen atom N_{k-1}, while ψ_k depends on the following carbon atom C_{k+1}. Contrastingly, the pair of dihedral angles (ψ_k, φ_{k+1}) bridged by the peptide bond connecting two consecutive amino acids k and $k + 1$ depends only on the atoms contained within the pair. We brought pairs of amino acids from 4,291 nonredundant proteins into a canonical reference frame in which C_k defines the origin, the peptide bond to the consecutive amino acid $k + 1$ is aligned with the x axis, and the normal to the C^{α}_k, C_k, N_{k+1} plane is aligned with the positive direction of the z axis (*C*). The canonically oriented atoms were then clustered by Euclidean (rigid body) similarity into 20 clusters using k-means (*D*). Gray stick plots visualize cluster centers in the context of two surrounding amino acids depicted as cartoons. Standard secondary structures (such as left and right α-helices, β-strands, and turns) are clearly recognizable. (*Bottom*) clusters are visualized using the different two-dimensional marginals of the four-dimensional joint distribution (φ_k, ψ_k, φ_{k+1}, ψ_{k+1}), namely, (*E*): (φ_k, ψ_k), (*F*): (ψ_k, φ_{k+1}), and (*G*): (φ_{k+1}, ψ_{k+1}). The variety of conformations assumed by the amino acid pairs is clearly discernible in the cross-bond Ramachandran plot (*F*), which has less cluster overlap compared to the traditional plots (*E* and *G*) in which distinct conformations are undiscernible. Angles are indicated in degrees and have a periodic boundary at ±180°. Contours indicate 90% probability regions. Colors indicate cluster assignment and numbered points indicate cluster centers.

amino acid pair. Fig. 4 shows these propensities for a selection of exemplary clusters, together with a cartoon depiction of their cross-bond and two regular Ramachandran plots (see also *SI Appendix*, Fig. S2 for all clusters). Cluster 12 exemplifies a clear agreement between the cluster propensities and the known α-helix propensities. This cluster corresponds to the structure adopted in the middle of an alpha helix, as demonstrated by all three marginals, (φ_k, ψ_k), (ψ_k, φ_{k+1}), and (φ_{k+1}, ψ_{k+1}) of the amino acid pair, having the same, coinciding, characteristic α-helix mode. In accordance, the amino acid propensities found in this cluster correlate well with the known amino acid propensities in α-helices, for both positions of the pair: higher for alanine, arginine, glutamic acid, leucine, methionine, and glutamine and lower for proline, glycine, asparagine, threonine and serine (16). More generally, it appears that many of the clusters defined in Fig. 1 represent structural motifs in the transition into and out of secondary structures. Cluster 9 and cluster 10, shown in Fig. 4, represent, different transition into an α-helix structure, as apparent by inspection of the marginal dihedral angle distributions; the individual conventional Ramachandran plots obtained from the first and second position in the amino-domino pair. What is immediately apparent here is that these different transition motifs have distinct amino acid preferences. The detail and ease with which local structural features are obtained will surely find use in protein design and in a better understanding of how mutations affect the protein backbone stability as we experimentally demonstrate below.

Another advantage of the cross-bond Ramachandran plot is that it captures structural features not discernable in the conventional Ramachandran plot. For example, cluster 15 occupies a

Downloaded from https://www.pnas.org by Institute of Science and Technology Austria on September 22, 2026 from IP address 78.104.209.6.

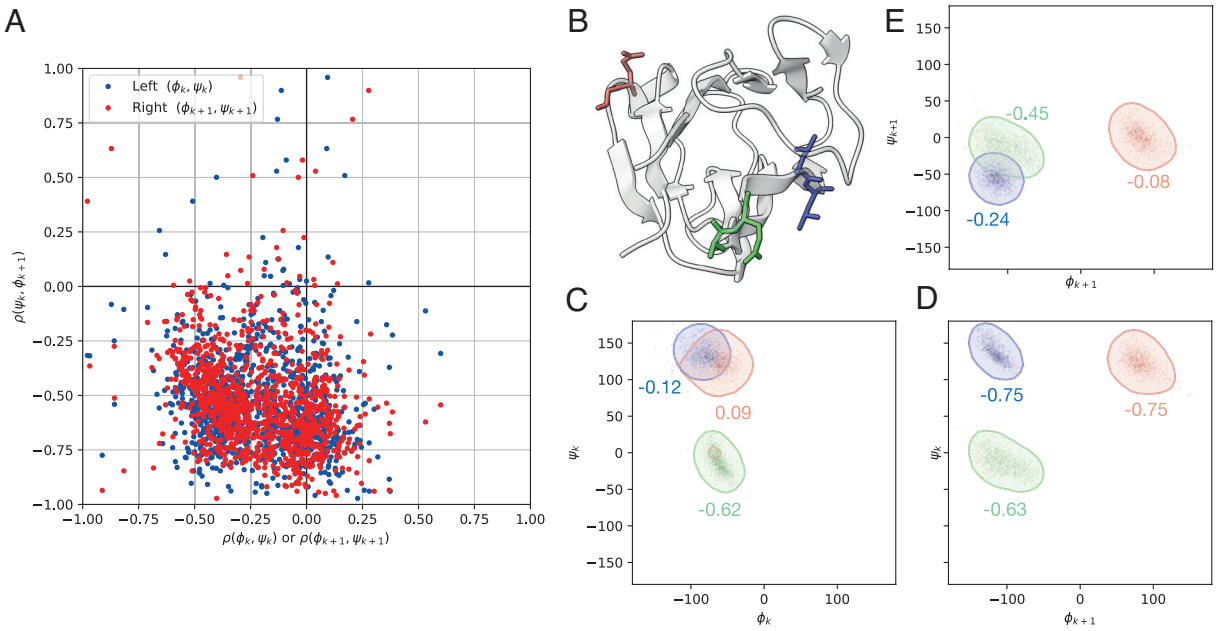


Fig. 2. Molecular dynamics simulations demonstrate that given a specific location, the cross-bond dihedral angle pair is more correlated than the traditional dihedral angle pair. (*A*) Correlation coefficients ρ between simulated dihedral angle pairs (φ_k, ψ_k)(ψ_k, φ_{k+1}) and (φ_{k+1}, ψ_{k+1}) in 1,095 locations in 13 protein structures. The cross-bond pairs (ψ_k, φ_{k+1}) exhibit higher correlation (in absolute value) than the maximum correlation of their left context (φ_k, ψ_k) and right context (φ_{k+1}, ψ_{k+1}) about 74% of the locations. Note the strong negative correlation of the cross-bond pairs. Correlations were calculated on a torus using 5,000 molecular dynamics temporal samples. (*B*) Three locations in an exemplary simulated structure, each location contributing a blue and a red dot to the plot in panel (*A*). (*C*–*E*) joint distributions of the dihedral angle pairs (φ_k, ψ_k), (ψ_k, φ_{k+1}), (φ_{k+1}, ψ_{k+1}), respectively. Contours indicate 90% probability regions. Numbers indicate correlation values. Negative correlations are visible in the form of diagonally elongated regions.

distinct position in the upper right-hand side of the cross-bond Ramachandran plot which is not populated in the traditional Ramachandran plot. This position captures the recognizable and characteristic (ψ_{k+1}, φ_{k+2}) conformation of the apex of a type II β-turn which is formed when the $k + 1$ and $k + 2$ amino acids adopt $\psi \cong 120°$ and $\varphi \cong 80°$ angles, respectively (17, 18). The interconversion between a type I and type II β-turn involves a peptide flip between this pair of amino acids, a phenomenon thought to play a role in the early stages of protein folding (19). In our recent work, aimed at identifying alternate protein folding pathways, we

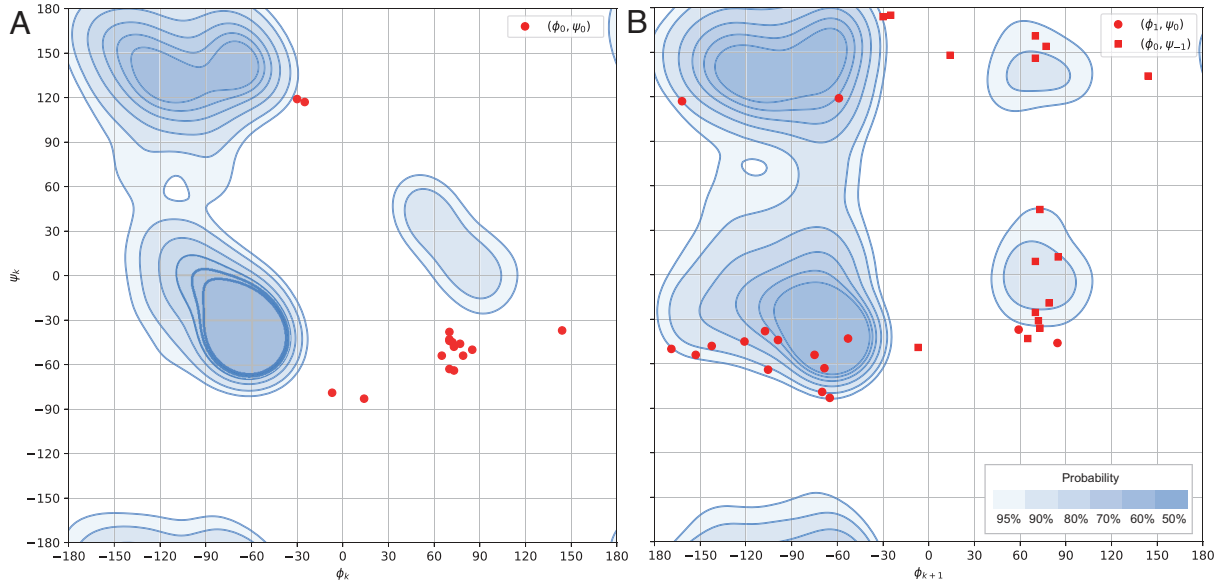


Fig. 3. Outliers in the traditional Ramachandran plot appear as inliers in the cross-bond plot, indicating that it represents the conformation space better than the traditional plot. Sixteen locations in different sub-1 Å structures that appear as outliers in the traditional (φ_k, ψ_k) Ramachandran plot (*A*) result to be inliers when considered in the context of a pair in the cross-bond Ramachandran plot (*B*, 316 consecutive amino acid pairs in 896 matching mesophile and thermophile bacterial proteins were assigned the nearest of the 20 cluster labels from Fig. 1 using AlphaFold-predicted structures. The frequencies of an amino acid pair in cluster *i* in a mesophile being assigned to cluster *j* in the corresponding location in the corresponding extremophile were recorded and normalized into a transition probability matrix that is visualized here as a directed graph. Edges represent the mesophile to extremophile transition probability, with higher probabilities indicated in darker black. Each cluster *k* is depicted as a graph node *k* colored according to the difference between in-degree (the probability of any mesophile cluster $i \neq k$ becoming cluster *k* in the extremophile) and the out-degree (the probability of cluster *k* in a mesophile to become any cluster $j \neq k$ in the extremophile). Node locations are set to the (φ_{n+1}, ψ_n) coordinates of the corresponding cluster centers. The *Inset* shows the two most common mesophile-to-extremophile label transitions identified by this analysis.

Downloaded from https://www.pnas.org by Institute of Science and Technology Austria on September 22, 2026 from IP address 78.104.209.6.

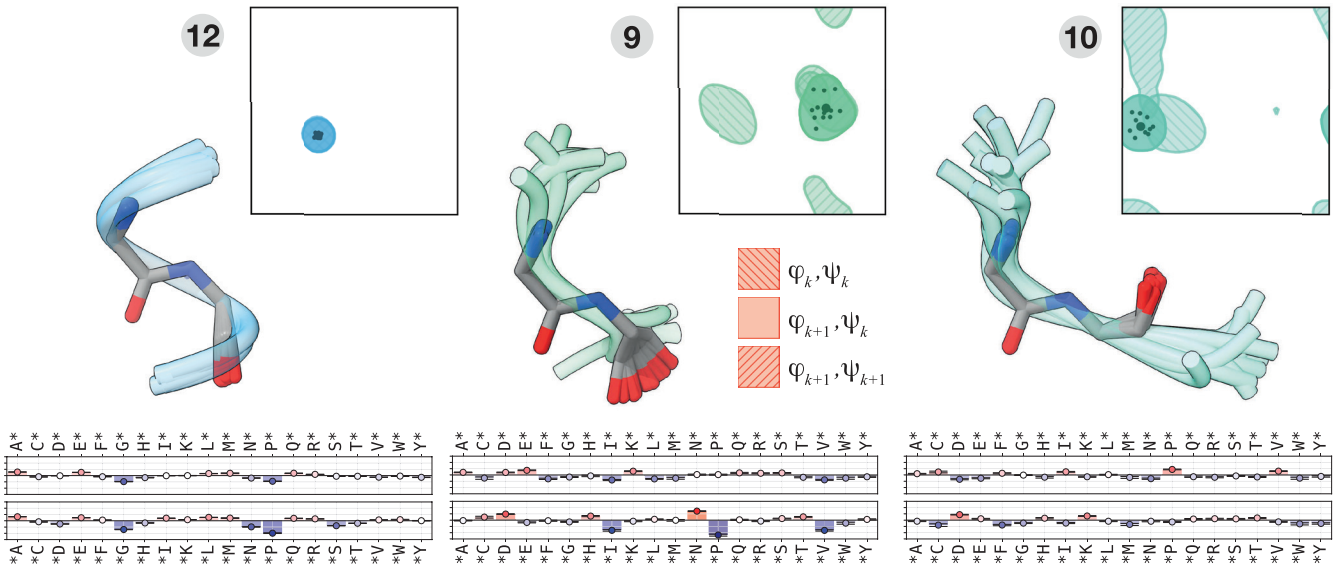


Fig. 4. Structural clusters have specific amino acid pair propensities. Three of the twenty canonically aligned clusters from Fig. 1 are displayed. The *Insets* show the regions occupied by each cluster in the three marginal plots (solid fill indicates the proposed cross-bond Ramachandran plot, while hatched fills depict the "conventional" plots). Note that for cluster 12, the occupied region in all three plots coincides. The colored bars below each cluster indicate the log propensity of different amino acid pairs of the form X* and *X to belong to the cluster. Cluster 12 corresponds to the center of a α-helix and cluster 9 and 10 different transitions into a helix. The calculated amino acid propensities of cluster 12 correspond to the known propensities as explained in the main text. We suggest that the amino-domino model unveils unrecognized amino acid propensities for other structural contexts including transitions between secondary structures.

have already found use for this cross-peptide bond angle pair in identifying peptide flips in equivalent environments of homologous proteins. Analysis of the representative amino acid pair backbones from cluster 15 shows that this type II β-turn apex conformation appears not only to facilitate a sharp change in direction of the protein backbone to connect two secondary structure elements but is also common in other secondary structure motifs, especially entering turns or exiting strands (Fig. 5 and *SI Appendix*, Fig. S4). This suggests that this conformation might have intrinsic stability as a structural unit.

Evidence That Amino Acid Pair Conformations Have Different Thermostabilities. Intrinsic to the protein structure–function relationship is the association between structure and stability. To facilitate different functions, proteins adopt global and local structure having varied degree of conformational flexibility. To probe for inherent differences in conformational flexibility at the level of the amino acid pair, we analyzed the clusters of cross-bond Ramachandran plots obtained from AlphaFold-predicted structures belonging to thermophilic and mesophilic bacteria species. It is well recognized that the proteome of thermophiles, species which thrive in higher-than-normal temperatures, has adapted to prevent denaturation of proteins at these temperatures. Compared to their mesophilic counterparts, thermophiles often have higher core hydrophobicity, greater numbers of ionic interactions, increased packing density, additional networks of hydrogen bonds, decreased lengths of surface loops, etc. (20). To probe for differences in the propensities of thermophiles and mesophiles toward specific amino acid pair conformations, which would be indicative of different intrinsic conformational stabilities, we analyzed 160,316 amino acid pairs from 896 different proteins belonging to the proteomes of six mesophilic bacteria and aligned them to homologous proteins in thermophilic bacteria. The (ψ_k, φ_{k+1}) dihedral angles were then extracted from AlphaFold-predicted structures and used to assign amino acid pairs to the 20 cluster labels shown in Fig. 1. Next, for each amino acid pair from aligned locations in homologous proteins from mesophilic and thermophilic species, we recorded differences in cluster assignments as transitions and calculated a normalized transition probability matrix that is visualized as a directed graph in Fig. 6. Our analysis shows that certain clusters are preferred in thermophilic proteins. It is of note that the preferred clusters in thermophiles, presumably the more thermostable ones, are those which are more ordered, having a lower RMS of constituent structures (refer to Fig. 6, *Inset* and *SI Appendix*, Fig. S5). Since this analysis relies on predicted structures, we probed for error biases toward either meso- or thermophiles in our AlphaFold-predicted data. We calculated confusion matrices for cluster label assignment for X-ray crystal structures vs. the AlphaFold-predicted structure. The matrices for mesophilic proteins and thermophilic proteins were highly similar indicating no bias (*SI Appendix*, Fig. S6).

Evidence That the Effect of a Single-Point Mutation Predictably Depends on Its Context. An implication of the amino-domino model is that immediately adjacent amino acids in the sequence can influence the effect of a single point mutation. While it is blatantly clear that mutations altering the side chain at a particular position can affect the interaction network and chemistry around that position, there is little recognition or understanding as to how that mutation might affect the conformation and stability of the protein backbone. Indeed, most approaches and tools consider mutations independent of the immediate sequence context, taking mainly the side chain identity into account. First, at the level of a single amino acid, we find that the distance between the (φ_k, ψ_k) dihedral angle distributions of two amino acids is predictive of their BLOSUM62 score (*SI Appendix*, Fig. S7), possibly capturing how amino acids having similar side chains can be accommodated in similar structural locations. Second, we suggest that beyond the effect of the side chain, an additional effect of mutation may result from altered backbone conformational preferences.

Probing within our data for how mutations might affect backbone conformational preferences, we calculated the distances between the cross-bond (ψ_k, φ_{k+1}) distributions of different amino acid pairs, and observed that an amino acid substitution within

Downloaded from https://www.pnas.org by Institute of Science and Technology Austria on September 22, 2026 from IP address 78.104.209.6.

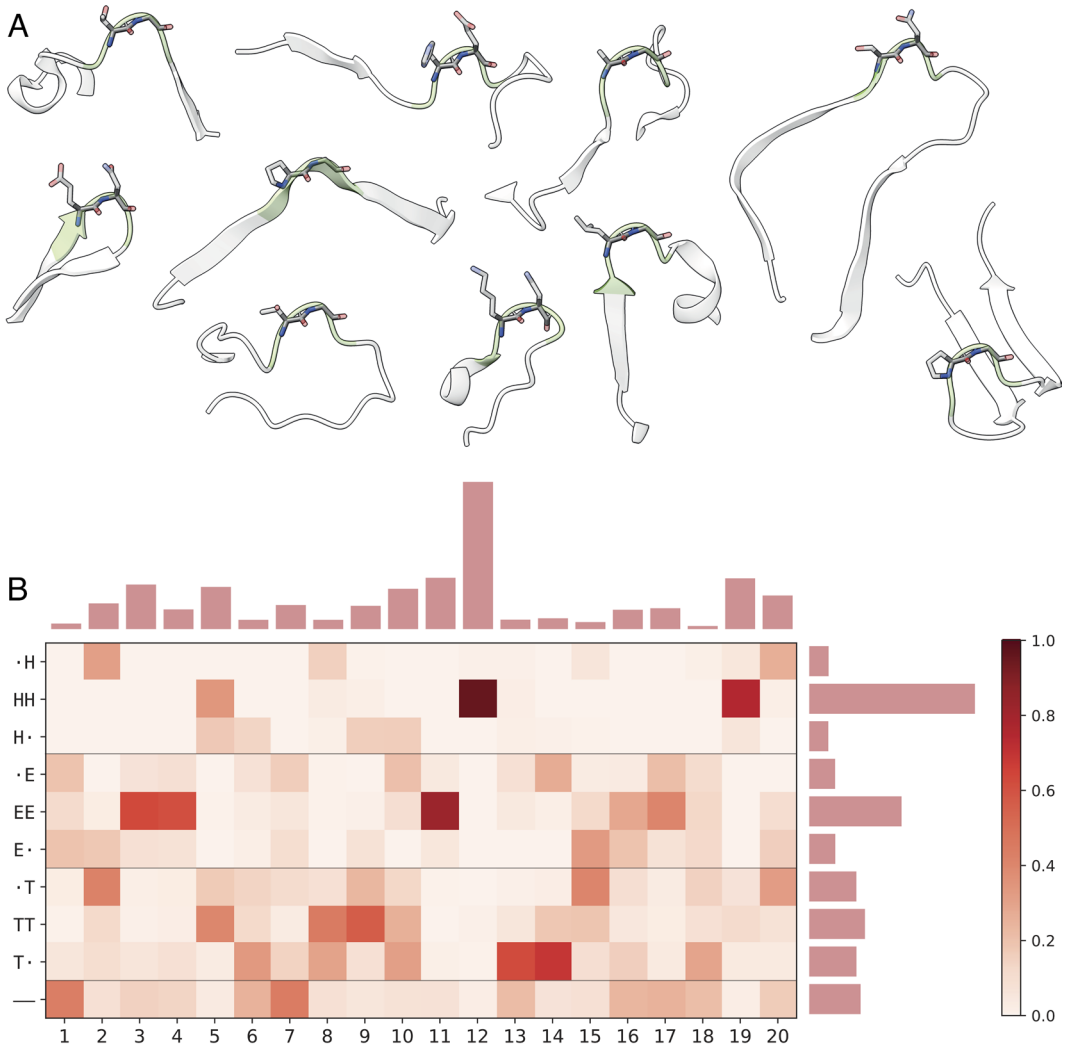


Fig. 5. Clusters of amino acid pair conformations encompass a range of larger structural contexts. (*A*) Amino acid pairs (displayed as sticks) from ten representative structures of cluster 15 are shown in their greater structural context (displayed as cartoons). (*B*) Affinity of each cluster to groups of secondary structures. Each column shows the probability of the corresponding cluster to be classified as different secondary structures. H indicates helices (DSSP labels H, G, and I), E indicates strands (DSSP E and B), T indicates turns (DSSP T and S), while – denotes lack of secondary structure (loops). ·H represents entrance to a helix with the dot · indicating every type of structure except H (same for E and T, mutatis mutandis); analogously, H· represents exit from a helix. Horizontal and vertical bars indicate marginal probabilities of secondary structures and clusters, respectively.

the pair may have vastly different impact on the (ψ_k, φ_{k+1}) distribution depending on the partner amino acid in the left or right position in the pair. For example, as visualized in *SI Appendix*, Fig. S8 and quantified in *SI Appendix*, Fig. S9, the pairs FI and YI have indistinguishable cross-bond distributions, while FH and YH exhibit very distinct distributions. We hypothesized that this distance may be predictive of how structurally disruptive a mutation may be, even in cases where the side chain substitution has little impact on the environment, such as our F > Y example. To explore this idea further, we analyzed published studies having melting temperature (T_m) data accompanied by crystal structures for lysozyme (21, 22). In both studies, isomorphous structures were obtained by the same experimental protocol, in the same crystallization conditions. In the first study on T4 lysozyme (21), helical position R96 (*SI Appendix*, Fig. S10*A*) was mutated to each of the other 19 amino acids and the difference in melting temperature between the mutant and wild type, ΔT_m, was measured. Plotting the difference in T_m against the difference in calculated pairwise Ramachandran distributions shows a correlation coefficient of 0.41 (*SI Appendix*, Fig. S10*B*). To associate differences in T_m more directly to changes in backbone conformation, we ran MD simulations using the crystal structures from Mooers et al. (21). A correlation coefficient of 0.15 between T_m change and the Wasserstein distance between the distributions of the backbone during simulation upon mutation was observed (*SI Appendix*, Fig. S10*C*). This example demonstrates, by providing an equivalent environment for analysis, that different amino acids have different backbone preferences; however, it does not show that different contexts can affect these preferences. To explore this question, we used the data from a study on hen egg white lysozyme (22) that included five G to A mutations within loop regions (*SI Appendix*, Fig. S10*D*). Plotting again the difference in T_m against the difference in calculated pairwise Ramachandran distributions gave a correlation coefficient of 0.18 (*SI Appendix*, Fig. S10*E*). However, the G102A mutation (sequence context DGN), that resulted in a peptide flip, a drastic structural change which as noted by the authors of ref. (22) may restore local stability. When this outlier was removed from our analysis, the correlation coefficient rose to 0.89. Plotting, as above, differences in T_m against differences in MD dihedral angle distributions between

Downloaded from https://www.pnas.org by Institute of Science and Technology Austria on September 22, 2026 from IP address 78.104.209.6.

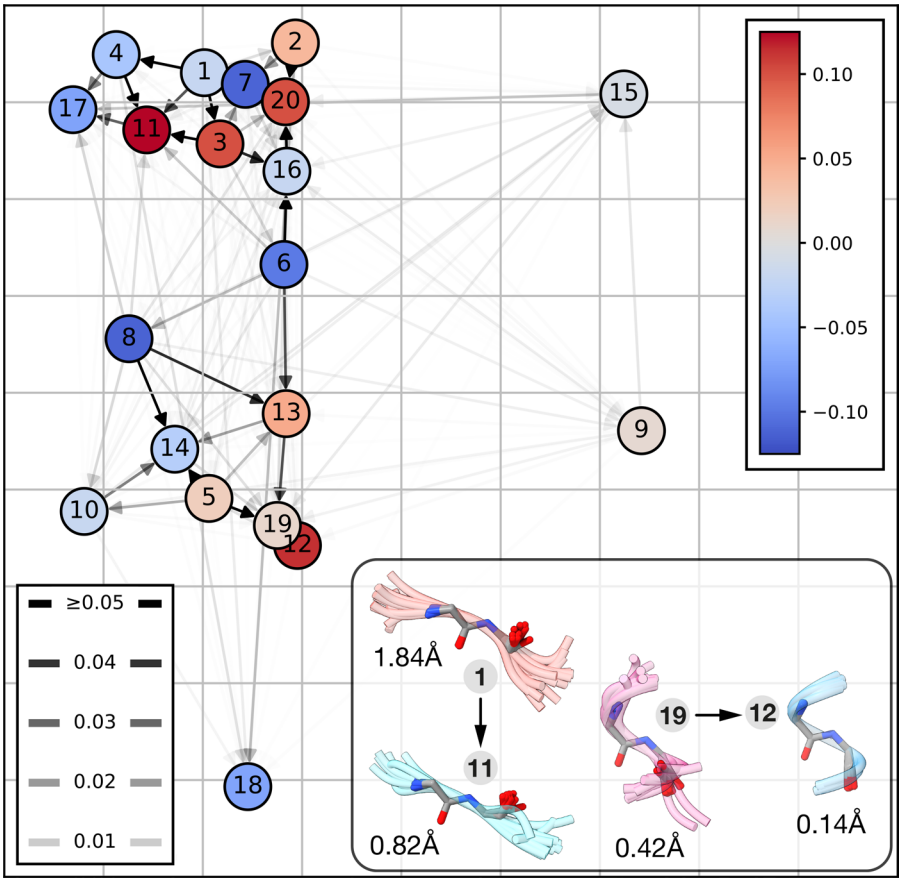


Fig. 6. Cross-bond Ramachandran plots show local structure preferences between corresponding proteins in mesophiles and extremophiles. A total of 160,316 consecutive amino acid pairs in 896 matching mesophile and thermophile bacterial proteins were assigned the nearest of the 20 cluster labels from Fig. 1 using AlphaFold-predicted structures. The frequencies of an amino acid pair in cluster *i* in a mesophile being assigned to cluster *j* in the corresponding location in the corresponding extremophile were recorded and normalized into a transition probability matrix that is visualized here as a directed graph. Edges represent the mesophile to extremophile transition probability, with higher probabilities indicated in darker black. Each cluster k is depicted as a graph node *k* colored according to the difference between in-degree (the probability of any mesophile cluster $i \neq k$ becoming cluster *k* in the extremophile) and the out-degree (the probability of cluster *k* in a mesophile to become any cluster $j \neq k$ in the extremophile). Node locations are set to the (φ_{n+1}, ψ_n) coordinates of the corresponding cluster centers. The insert shows the two most common mesophile-to-extremophile label transitions identified by this analysis.

the wild type and each mutation gave a correlation coefficient of 0.74 (*SI Appendix*, Fig. S10*F*).

To further corroborate the hypothesis that sequence context affects different the backbone conformational change resulting from a mutation, we performed mutagenesis studies in green fluorescent protein (GFP). GFP was chosen as having many surface-exposed, noninteracting amino acids in equivalent, β-sheet, secondary structure (Fig. 7*C*). Pairs of the same mutation, X>Y, were made in different adjacent amino acid contexts. The mutations were chosen, based on our analysis of the cross-bond angle distributions, to give maximally different outcomes in each sequence context. Specifically, for an amino acid, X, present in the GFP structure as noninteracting and surface exposed at least twice, we chose mutation to Y (A_1XB_1>A_1YB_1 and A_2XB_2>A_2YB_2) such that the mutation in the first context, A_1 and B_1 would have a relatively small calculated difference between the cross-bond Ramachandran plots of (A_1X, A_1Y) and (XB_1, YB_1), while the same mutation in the second context, A_2 and B_2, would have a relatively large distance between (A_2X, A_2Y) and (XB_2, YB_2). Melting temperature relative to the wild type, ΔT_m, was again used to quantify the change in protein stability. Fig. 7 shows that the distance between cross-bond Ramachandran plots is predictive of ΔT_m (correlation coefficient 0.58; removing HKF > HEF outlier, the coefficient rises to 0.77). The K26E outlier may be explained by the fact that the K26 is the only mutated site located at the end of a β-strand (single DSSP E annotation to the left), unlike the other sites that a located in the middle (surrounded by at least two Es). Additionally, the K26E mutation is located in a β-strand region hydrogen-bonded only on one side, unlike the other mutations located in β-sheets. This might allow the region to better recover from the local backbone strain introduced by the mutation. For comparison, the Q95R mutation has approximately the same predicted distribution distance but since it located in the middle of a β-sheet, the actual effect on the ΔT_m is bigger (refer to Fig. 7*C* for a visualization).

Our work demonstrates that the cross-bond Ramachandran plot is useful in describing the structure of amino acid pairs and for identifying features characteristic of these amino-dominos. A natural application is for understanding the effects of mutations on protein stability, and for rational design of proteins, where engineering or increasing stability is often a prime concern. We further suggest that the amino-domino model may open avenues for defining structure i disordered proteins and regions or as an additional aspect of homology modeling. In conclusion, we show here that amino acid pairs can be thought of as structural units which are effectively described by the unconventional, cross-peptide-bond pair of dihedral angles. We believe that this approach is an accessible tool which will enable previously unexplored methods to understand and predict local protein structure.

Downloaded from https://www.pnas.org by Institute of Science and Technology Austria on September 22, 2026 from IP address 78.104.209.6.

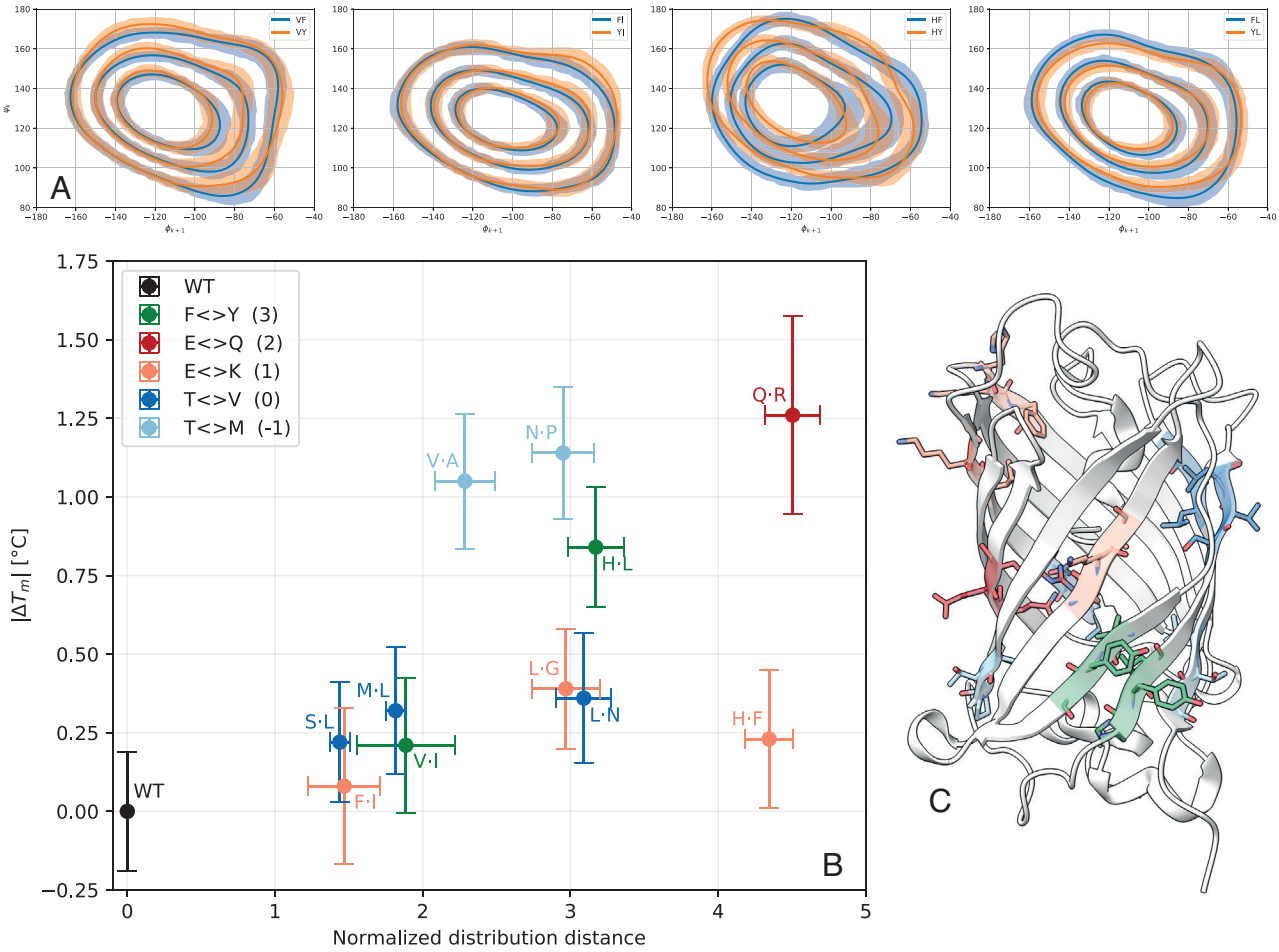


Fig. 7. Similarity of cross-bond Ramachandran plot is predictive of the impact of a conservative mutation on protein stability. Four conservative mutations were made in the GFP (PDB:2B3P) in contexts either minimizing or maximizing distances between (ψ_k, φ_{k+1}) dihedral angle distributions of two consecutive amino acid pairs containing the mutation. For example, in VYI>VFIVY>VF and YI>FI have similar cross-bond Ramachandran plots, while in HYL>HFLHY>HF and YL>FL have dissimilar plots (*A*). Protein stability, quantified as ΔT_m relative to the wild type, is correlated to the distance between the cross-bond angle distributions (*B*). Distance is defined as the maximum over the normalized distance between the left and the right pair distributions, including the mutation (e.g., for VYI>VFI, it is the maximum over the distances between VYVF and YIFI). Mutations are color-coded; amino acid contexts are indicated in black next to each data point. CIs of 1σ are shown. Legend reports BLOSUM62 scores of different mutations in parentheses. Mutations in noninteracting locations in β-strands are depicted in panel (*C*).

Materials and Methods

Most of the analysis procedure is based on the data collection and mathematical tools described in the methods section of Rosenberg et al. (23). Here, we highlight mainly the additional details and modifications.

Protein Structures Collection. Nonredundant protein chains were collected using the method described in ref. 19, with a slightly different initial PDB query. The query parameters were as follows: i) Method: X-Ray Diffraction; ii) X-Ray Resolution: less than or equal to 1.5 Å; iii) R_{free}: less than or equal to 0.24; and iv) Expression system contains the phrase "*Escherichia Coli*". After performing the query, the structures were obtained from PDB-redo, which provides optimized and refined structures for most PDB entries (24). The PDB structures were processed using the biopython package in order to obtain the backbone-atom coordinates at each residue. In total, we collected 870,266 consecutive amino acid pairs from 4,291 nonredundant proteins, discarding all pairs that were closer than 5 positions to the chain termini.

Clustering of Amino Acid Pairs. In order to obtain the 20 clusters depicted in Fig. 1 and *SI Appendix*, Fig. S1, each amino acid pair was first brought into a canonical orientation (Fig. 1*C*). This was achieved by setting the location of the C atom of the first amino acid at the origin, aligning the peptide bond in the direction of the *x* axis, and the plane formed by the C^{α} and C atoms of the first amino acid and the N atom of the second amino acid with the *xy* plane, with the normal pointing in the *z* axis direction. The coordinates of the canonically oriented C^{α}, C, and N atoms of both amino acids were then stacked into an 18-dimensional vector representing the pair. The set of all such vectors from all the pairs was then clustered into $k = 20$ clusters using Lloyd's *k*-means algorithm with the Euclidean metric implemented in the sklearn python package. Ten runs with different centroid seed initializations were used, with the maximum of 300 iterations in each run. Amino acid pairs closest (in the Euclidean sense) to the calculated cluster centers were selected to represent the cluster. Each cluster was further clustered into 10 subclusters using the same procedure to obtain the 10 representative backbones depicted in *SI Appendix*, Fig. S2. Although the labeling of the clusters with numbers is arbitrary, we used a multidimensional scaling (MDS) procedure to ensure that closely labeled cluster centers are closer to each other in the sense of the above-defined Euclidean distance.

Calculation of Cluster Propensities. The propensity $\pi(\mathrm{A}*\,|\,c)$ of an amino acid pair A* (with the star indicating that any amino acid can occupy the second position) to cluster *c* (reported in *SI Appendix*, Fig. S2) was calculated as the log of the ratio of the probabilities which were, in turn, estimated from the empirical frequencies in the dataset:

$$\pi(\mathrm{A}*\,|\,c) = \log_{10}\frac{P(\mathrm{A}*\,|\,c)}{P(\mathrm{A}*)} = \log_{10}\frac{N(\mathrm{A}*\,|\,c)}{N(c)}\frac{N}{N(\mathrm{A}*)}.$$

Here, N denotes the total number of pairs, $N(c)$ the number of pairs in cluster *c*, $N(\mathrm{A}*)$ the total number of pairs with A in the first position, and $N(\mathrm{A}*\,|c)$ the number of pairs with A in the first position in cluster *c*. The propensities of amino acid pairs of the form *A were calculated in a corresponding manner.

Downloaded from https://www.pnas.org by Institute of Science and Technology Austria on September 22, 2026 from IP address 78.104.209.6.

Dihedral Angle Density Estimation. Dihedral angle densities depicted in the Ramachandran plots (Figs. 1 *E–G*, 2 *C–E*, 4, and 7 and *SI Appendix*, Figs. S1, S2, and S8) were calculated using the kernel density estimation (KDE) procedure described in detail in ref. 23. We used square bins of width 2° and a Gaussian kernel with $\sigma = 12°$. The regions containing 90% probability and contours containing various levels of the probability were also calculated using the procedures detailed in ref. 23.

Calculation of Distances between Ramachandran Plots. Distances between Ramachandran plots were estimated as L_1 distances between KDEs estimated from their points. To gauge the effect of finite sample sizes, average distances and their SDs were calculated on 100 random samples of 1,000 angle pairs bootstrapped independently with replacements.

To calculate the normalized distances between pairs of amino acids AX and AY reported in Fig. 7 and *SI Appendix*, Figs. S8 and S10, we calculated the ratio of the L_1 distances between the corresponding KDEs and the KDEs of *X and *Y,

$$\frac{d(\mathrm{AX}, \mathrm{AY})}{d(*\mathrm{X}, *\mathrm{Y})},$$

where the KDEs of *X and *Y are estimated from points of any pairs with X and Y at the second position, respectively. As with the unnormalized distances, 100 random samples of 1,000 angle pairs bootstrapped independently with replacements were used to estimate the CIs. The distance between two triplets AXB and AYB was defined as the maximum of the two distances between AX, AY and XB, YB.

Calculation of Wasserstein Distance in *SI Appendix*, Fig. S10. To measure the distances between quadruplets of dihedral angles (φ_k, ψ_k, φ_{k+1}, ψ_{k+1}), we used torus statistics in order to calculate the mean μ and covariance C parameters. We then used the closed form expression for the Wasserstein distance between normal distributions,

$$d_W^2 = \|\mu_1 - \mu_2\|_2^2 + \mathrm{trace}\left(C_1 + C_2 - 2\left(C_2^{1/2} C_1 C_2^{1/2}\right)^{1/2}\right).$$

The squared root of d_W^2 was expressed in degrees.

Calculation of Correlation Coefficients between Dihedral angles. Correlation coefficients were calculated on the MD simulation data using the flat torus statistics described in detail in ref. 23.

MDS Plots. The similarity plots in *SI Appendix*, Fig. S8 were calculated using the probabilistic MDS procedure proposed and described in ref. 23.

Calculation of Mesophile/Thermophile Local Structure Transitions. We used full proteomes of six similar mesophile/thermophile bacterial pairs from ref. 25. For each bacterial pairs, proteins expressed from matching genes with over 70% sequence identity were selected resulting in a total of 896 protein pairs. Each protein pair was aligned using global sequence alignment, and locations of consecutive pairs of matching amino acids were recorded. For each amino acid pair, corresponding dihedral angles (ψ_k, φ_{k+1}) were extracted from AlphaFold-predicted structures and assigned the nearest (in the sense of the flat torus distance) of the 20 cluster labels from Fig. 1. This resulted in a total of 160,316 matching mesophile/thermophile cluster label pairs. The frequencies of an amino acid pair in cluster *i* in a mesophile being assigned to cluster *j* in the corresponding location in the corresponding extremophile were recorded and normalized into a transition probability matrix that was visualized as a directed graph in Fig. 6.

Molecular Dynamics Simulations. The PDB was queried for protein structures, solved by X-ray crystallography to a resolution of equal to or better than 0.85 Å, coming from a sequence no longer than 125 amino acids and having no disulfide bonds. After removing redundant structures (belonging to the same Uniprot entry), the following PDB codes remained: 1NWZ, 1R6J, 1YK4, 2O7A, 2O9S, 2PVE, 3UI4, 4EIC, 5DHV, 5NFM, 7A5M, 7AVK, and 7BNH. Single chains without water molecules were prepared for molecular dynamic simulations using the PDB2PQE server (https://server.poissonboltzmann.org/, 26). Molecular dynamics was preformed using GROMACS (27, 28). Prior to measuring dynamics an equilibrated and energy-minimized system was prepared by standard procedures including solvation and ion addition to achieve a neutral state. NVT (constant Number of particles, Volume, and Temperature) was carried out for 100-ps and 300 K. NPT (constant Number of particles, Pressure, and Temperature) was carried out for 100-ps. We ran 50-ns MD simulations with trajectory images being recorded every 10-ps. Relevant backbone φ and ψ angles were extracted and analyzed. The molecular dynamics input files are provided as a supplementary file.

Wild-Type and Mutant GFP Melting Temperature Measurements. sfGFP his-tagged at both the N and C terminals in expression plasmid pET28a was obtained as a gift from Avi Schroeder. Point mutations were introduced using complementary primers designed in the NEBaseChanger program and following the standard Q5 site-directed mutagenesis protocol. The introduction of mutations was verified by sequencing. Wild-type and mutant (V120T, V206T, V219T, K26E, K126E, K166E, T186M, T225M, Y151F, Y200F, and E95Q) sfGFP protein was overexpressed in BL21 cells. A 10-mL overnight starter culture was added to 500 mL LB supplemented with kanamycin and grown to OD600 ~ 0.6. Protein expression was induced by the addition of 1 mM IPTG, and cells were grown for a further 4 h before being collected by centrifugation and stored at −20 °C until purification. Frozen cells were resuspended in lysis buffer (150 mM NaCl and 50 mM Tris pH 8) before disruption in a microfluidizer. Following centrifugation to remove cellular debris, the supernatant was passed through a nickel column which was then first washed with 10 CV lysis buffer and then another 5 CV lysis buffer supplemented with 10 mM imidazole. The protein was eluted in lysis buffer containing 250 mM imidazole which was subsequently removed by overnight dialysis against lysis buffer. The purity of the protein was confirmed by SDS-PAGE, and the protein samples were concentrated to 5 mg/mL. Melting temperature was determined by differential scanning fluorimetry, measuring the fluorescence every 0.5 °C upon heating from 20 °C to 95 °C using 80% power in a Prometheus NT.48 nano-DSF instrument (Nanotemper Technologies). The results shown in Fig. 7 represent 5 independent measurements from each of two biological repeats.

Data, Materials, and Software Availability. Data used in this work can be found at 10.7910/DVN/NADTTJ (29), and the code is available at 10.5281/zenodo.8161469 (30).

ACKNOWLEDGMENTS. This research was partially supported by the Council For Higher Education–Planning & Budgeting Committee, and the Schmidt Career Advancement Chair in AI research grant. We thank Shlomi Reuveni and Noam Adir for their helpful discussions, as well as Yael Pazy Benhar and Dikla Hiya at the Technion Center for Structural Biology for their guidance and technical support with the T_m experiments.

1. N. C. Fitzkee, G. D. Rose, Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 12497–12502 (2004).
2. F. Eker *et al.*, Preferred peptide backbone conformations in the unfolded state revealed by the structure analysis of alanine-based (AXA) tripeptides in aqueous solution. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 10054–10059 (2004).
3. D. Moses *et al.*, Structural biases in disordered proteins are prevalent in the cell. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.11.24.469609 (Accessed 12 October 2023).
4. R. Unger, D. Harel, S. Wherland, J. Sussman, Analysis of dihedral angles distribution: The doublets distribution determines polypeptides conformations. *Biopolymers* **30**, 499–508 (1990).
5. R. Unger, D. Harel, S. Wherland, J. L. Sussman, A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* **5**, 355–373 (1989).
6. A. G. de Brevern, C. Etchebest, S. Hazout, Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* **41**, 271–287 (2000).
7. C. Etchebest, C. Benros, S. Hazout, A. G. de Brevern, A structural alphabet for local protein structures: Improved prediction methods. *Proteins* **59**, 810–827 (2005), 10.1002/prot.20458.
8. M. M. Mirjana, S. M. Nenad, A. G. de Brevern, Prediction of structural alphabet protein blocks using data mining. *Biochimie* **197**, 74–85 (2022), 10.1016/j.biochi.2022.01.019.
9. F. Zheng, J. Zhang, G. Grigoryan, Tertiary structural propensities reveal fundamental sequence/structure relationships. *Structure* **23**, 961–971 (2015), 10.1016/j.str.2015.03.015.
10. J. Zacharias, E. W. Knapp, Geometry motivated alternative view on local protein backbone structures. *Protein Sci.* **22**, 1669–1674 (2013), 10.1002/pro.2364.
11. S. Anishetty, G. Pennathur, R. Anishetty, Tripeptide analysis of protein structures. *BMC Struct. Biol.* **2**, 9 (2002), 10.1186/1472-6807-2-9.
12. S. A. Hollingsworth, M. C. Lewis, D. S. Berkholz, W.-K. Wong, P. Andrew Karplus, (φ,ψ)2 Motifs: A purely conformation-based fine-grained enumeration of protein parts at the two-residue level. *J. Mol. Biol.* **416**, 78–93 (2012), 10.1016/j.jmb.2011.12.022.

Downloaded from https://www.pnas.org by Institute of Science and Technology Austria on September 22, 2026 from IP address 78.104.209.6.

13. G. N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963).
14. A. Ravikumar, C. Ramakrishnan, N. Srinivasan, Stereochemical assessment of (φ, ψ) outliers in protein structures using bond geometry-specific ramachandran steric-maps. *Structure* **27**, 1875–1884.e2 (2019).
15. O. Sobolev *et al.*, A global ramachandran score identifies protein structures with unlikely stereochemistry. *Structure* **28**, 1249–1258 (2020), 10.1016/j.str.2020.08.005.
16. S. Costantini, G. Colonna, A. M. Facchiano, Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochem. Biophys. Res. Commun.* **342**, 441–451 (2006), 10.1016/j.bbrc.2006.01.159.
17. A. de Brevern, Extension of the classical classification of β-turns. *Sci. Rep.* **6**, 33191 (2016), 10.1038/srep33191.
18. P. Kountouris, J. D. Hirst, Predicting β-turns and their types using predicted backbone dihedral angles and secondary structures. *BMC Bioinform.* **11**, 407 (2010).
19. A. Fuller *et al.*, Evaluating β-turn mimics as β-sheet folding nucleators. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 11067–11072 (2009).
20. R. Lieph, F. A. Veloso, D. S. Holmes, Thermophiles like hot T. *Trends Microbiol.* **14**, 423–426 (2006).
21. B. H. Mooers, W. A. Baase, J. W. Wray, B. W. Matthews, Contributions of all 20 amino acids at site 96 to the stability and structure of T4 lysozyme. *Protein Sci.* **18**, 871–880 (2009), 10.1002/pro.94.
22. K. Masumoto, T. Ueda, H. Motoshima, T. Imoto, Relationship between local structure and stability in hen egg white lysozyme mutant with alanine substituted for glycine. *Protein Eng.* **13**, 691–695 (2000), 10.1093/protein/13.10.691.
23. A. A. Rosenberg, A. Marx, A. M. Bronstein, Codon-specific Ramachandran plots show amino acid backbone conformation depends on identity of the translated codon. *Nat. Commun.* **13**, 1–11 (2022).
24. R. P. Joosten, F. Long, G. N. Murshudov, A. Perrakis, The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213–20 (2014).
25. J. H. McDonald, Temperature adaptation at homologous sites in proteins from nine thermophile-mesophile species pairs. *Genome Biol. Evol.* **2**, 267–76 (2010).
26. E. Jurrus *et al.*, Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **27**, 112–128 (2018), 10.1002/pro.3280.
27. H. J. C. Berendsen, D. van der Spoel, R. van Drunen, GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.* **91**, 43–56 (1995).
28. E. Lindahl, B. Hess, D. van der Spoel, GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.* **7**, 306–317 (2001).
29. A. Bronstein, A. Rosenberg, A. Marx, N. Yehishalom, "Amino Domino". Harvard Dataverse. https://doi.org/10.7910/DVN/NADTTJ. Deposited 18 July 2023.
30. A. A. Rosenberg, A. Marx, A. M. Bronstein, pp5: Estimation and comparison of protein backbone angle distributions. Zenodo. https://zenodo.org/record/8161469. Deposited 18 July 2023.

Downloaded from https://www.pnas.org by Institute of Science and Technology Austria on September 22, 2026 from IP address 78.104.209.6.