AAontology: An Ontology of Amino Acid Scales for Interpretable Machine Learning

Stephan Breimann [1,2,3], **Frits Kamp** [2], **Harald Steiner** [2,3] **and Dmitrij Frishman** [1,*]

*1 - **Department of Bioinformatics,** School of Life Sciences, Technical University of Munich, Freising, Germany*

*2 - **Ludwig-Maximilians-University Munich,** Biomedical Center, Division of Metabolic Biochemistry, Munich, Germany*

*3 - **German Center for Neurodegenerative Diseases (DZNE),** Munich, Germany*

Correspondence to Dmitrij Frishman: *dimitri.frischmann@tum.de (D. Frishman)*

https://doi.org/10.1016/j.jmb.2024.168717

Edited by Rita Casadio

Abstract

Amino acid scales are crucial for protein prediction tasks, many of them being curated in the AAindex database. Despite various clustering attempts to organize them and to better understand their relationships, these approaches lack the fine-grained classification necessary for satisfactory interpretability in many protein prediction problems. To address this issue, we developed AAontology—a two-level classification for 586 amino acid scales (mainly from AAindex) together with an in-depth analysis of their relations—using bag-of-word-based classification, clustering, and manual refinement over multiple iterations. AAontology organizes physicochemical scales into 8 categories and 67 subcategories, enhancing the interpretability of scale-based machine learning methods in protein bioinformatics. Thereby it enables researchers to gain a deeper biological insight. We anticipate that AAontology will be a building block to link amino acid properties with protein function and dysfunctions as well as aid informed decision-making in mutation analysis or protein drug design.

Introduction

Amino acids are vital to numerous biological processes, and understanding their physicochemical properties is critical for protein bioinformatics research. The AAindex database[1–4] provides comprehensive quantifications of these properties (*e.g.*, volume, charge, or hydrophobicity) in form of 566 numerical indices (also called scales). Obtained by 149 studies from over six decades of research, these scales constitute valuable features for machine learning models.[5–7] However, the redundancy of AAindex—exemplified by the presence of over 30 scales for hydrophobicity alone—and its sometimes ambiguous annotations impede the development of highly-needed interpretable machine learning models.[8,9]

Efforts to cluster the AAindex database have aimed to organize amino acid properties and elucidate their relationships. In 1988, Kenta et al.[1] created the first version of AAindex by collating and hierarchically clustering 222 scales (using agglomerative clustering with single-linkage[10,11] into four groups: alpha and turn propensity, beta propensity, hydrophobicity, and other properties such as bulkiness. This work was extended by Tomii et al.[2] in 1996, introducing two new groups (amino acid composition and physicochemical properties) for the updated 402 indices, underscoring the importance of studying amino acid scale relationships. An update in 2008 expanded AAindex to 544 scales,[4] leading Saha et al.[12] to conduct a consensus fuzzy clustering analysis in 2012,[13] where scales were clustered by various algorithms and then assigned via majority voting, yielding eight clusters. They introduced three new groups: electric properties, residue propensity, and intrinsic properties—while eliminating 'other properties'. Moreover, they addressed the issue of scales not being clustered into main clusters by assigning them to the 'intrinsic property' group. In 2016, Simm et al.[14] clustered 98 hydrophobicity scales, mostly from AAindex, emphasizing their impact on secondary structures. Lastly, in 2018 Forghani and Khani[15] performed a multivariate clustering analysis on the latest version of AAindex (566 scales, 2017), revealing issues with clustering algorithms' performance and their dependence on settings, particularly for models with pre-defined cluster number such as *k*-means.[16] Using various clustering quality measures, including the silhouette coefficient[17] and the Calinski Harabasz score,[18] they determined optimal numbers of clusters: 2, 3, and 9. However, these clusters were not further biologically characterized.

While the existing studies have advanced our understanding of amino acid properties, two limitations hinder their direct use in protein prediction. The first limitation, redundancy, occurs when similar properties are grouped, potentially merging distinct but related properties such as charge and polarity. The second issue, limited interpretability, stems from clustering only based on statistical similarity, which may not reflect the biological meaning or functional relationships. Moreover, current efforts oversimplify the diversity of property scales by confining them to a few clusters without adequately mitigating their complexity. Therefore, a pressing need exists for a more fine-grained and biologically meaningful classification to meet diverse research requirements.

Here, we introduce AAontology, a two-level ontology[19,20] of amino acid scales, a systematic description of scales, and a comprehensive analysis of their relationships. AAontology organizes amino acid property scales into 8 categories and 67 subcategories based on their numerical similarity and physicochemical meaning. It enhances their interpretability, particularly for scale-based machine learning in protein prediction tasks.[21–24] This framework may serve as the foundation for systematically exploring the relationships between physicochemical properties and protein functions (*e.g.*, cellular signaling,[25,26] molecular recognition,[27] or membrane insertion[28]and dysfunctions (*e.g.*, oncogenicity[29–31] or aggregation linked to diseases such as Alzheimer's disease.[32–34] In addition, it may also propel informed decision-making in mutation analysis[35–40] or drug design of proteins such as antibodies.[41,42] Thus, AAontology promises to deepen our understanding of the multifaceted landscape of protein biology.

Results

Creation of AAontology

We developed AAontology by compiling a dataset of 586 amino acid scales (Supplementary Table 1), sourced from the AAindex database[4] and two additional studies.[43,44] Using clustering and knowledge-based criteria, these scales were first automatically assigned to categories and subcategories, and then manually refined to ensure that the classification accurately reflects biological meanings. This process, from initial classification to the refined two-level classification into 8 categories and 67 subcategories, is depicted in the graphical abstract.

Figure 1 further details our classification workflow through a Sankey diagram, beginning with a bag-of-words approach, where scales are assigned to categories based on key terms in their names or descriptions. Subsequently, scales within each category were clustered into subcategories using our AAclust clustering framework,[45] which serves as a wrapper for clustering models that require a pre-defined number of clusters, such as *k*-means.[16] AAclust optimizes the number of clusters (subcategories) to guarantee that each cluster meets a minimum similarity criterion, specifically ensuring that all pairwise Pearson correlations between members (scales) within a cluster are higher than a defined threshold (*e.g.*., 0.3 Pearson correlation coefficient, indicating at least week positive correlation). Finally, the assignments of scales to subcategories and categories were manually curated based on biological knowledge. These steps were iteratively repeated to refine and ensure consistency in the classification and the naming of subcategories. For comprehensive details, see Materials and methods and Supplementary Table 2.

To assess the coherence and validity of our two-level classification, we analyzed the Pearson correlations between scales within each category. Figure 2 shows a heatmap for every category, where each subcategory demonstrates a strong internal consistency with a minimum Pearson correlation of 0.3, visualizing the robustness of our semi-automatic classification approach. A Pearson correlation threshold of 0.3 was chosen for creating AAontology as a good tradeoff between numerical strictness (ensuring that scales within each subcategory exhibit at least a weak positive correlation) and the manageability of the resulting number of clusters, aiming for an outcome of 50–100 subcategories. While the average pairwise Pearson correlation within each subcategory is 0.68, a stricter threshold, such as 0.4, would result in over 100 subcategories, with many scales forming single-member clusters. While most subcategories (*e.g.*, volume or hydrophilicity) contain many highly correlated scales, a few subcategories (*e.g.*, charge or entropy) include only a few scales that reflect unique physicochemical properties. Scales that could not be classified based on these correlation criteria or lacked clear literature-based assignments were labeled as

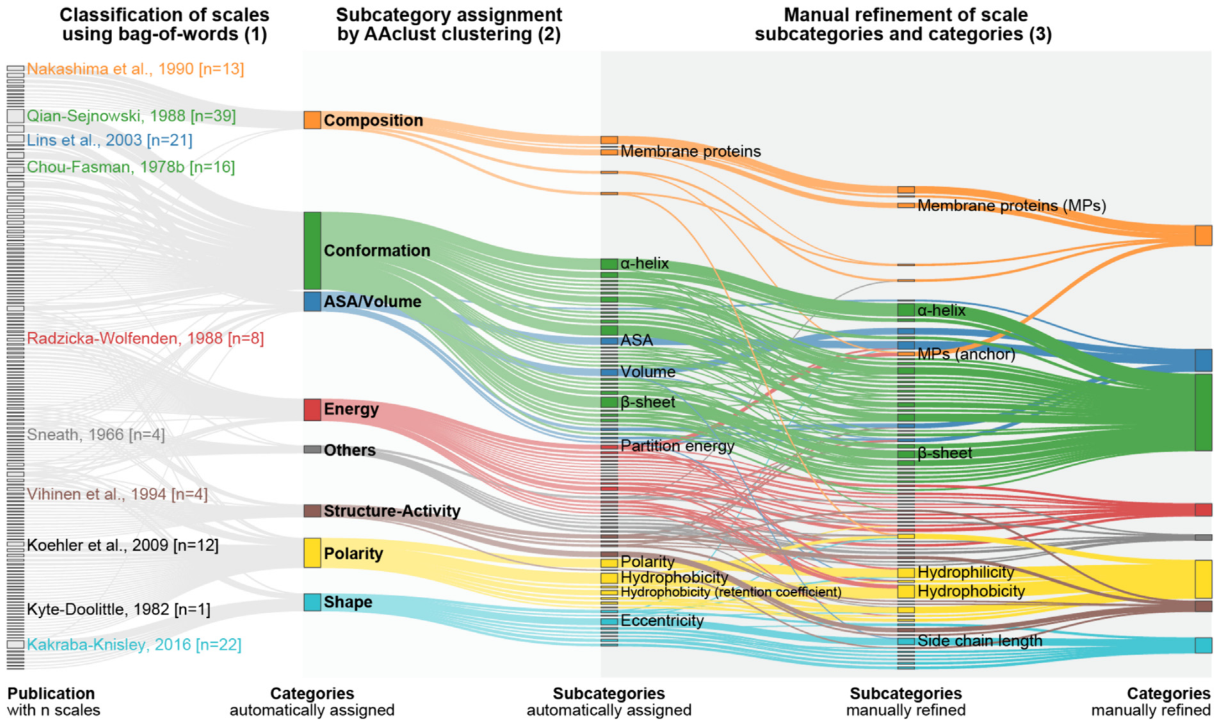
Classification of scales using bag-of-words (1)
Subcategory assignment by AAclust clustering (2)
Manual refinement of scale subcategories and categories (3)
Nakashima et al., 1990 [n=13]
Qian-Sejnowski, 1988 [n=39]
Lins et al., 2003 [n=21]
Chou-Fasman, 1978b [n=16]
Radzicka-Wolfenden, 1988 [n=8]
Sneath, 1966 [n=4]
Vihinen et al., 1994 [n=4]
Koehler et al., 2009 [n=12]
Kyte-Doolittle, 1982 [n=1]
Kakraba-Knisley, 2016 [n=22]
Composition
Conformation
ASA/Volume
Energy
Others
Structure-Activity
Polarity
Shape
Membrane proteins
α-helix
ASA
Volume
β-sheet
Partition energy
Polarity
Hydrophobicity
Hydrophobicity (retention coefficient)
Eccentricity
Membrane proteins (MPs)
α-helix
MPs (anchor)
β-sheet
Hydrophilicity
Hydrophobicity
Side chain length
Publication with n scales
Categories automatically assigned
Subcategories automatically assigned
Subcategories manually refined
Categories manually refined

Figure 1. Creation of AAontology. Sankey diagram showing the process of scale classification. 586 amino acid scales from 151 publications were first automatically assigned to 8 different categories using a bag-of-word approach based on their scale description. Key references with the number of containing scales are given. Next, an automatic subcategory assignment was performed using clustering with AAclust.[45] Finally, subcategory and category assignment were manually refined including the renaming of the automatically created subcategories. This procedure was repeated over multiple iterations to refine subcategory names and scale assignment.

'unclassified' and grouped separately within their respective category.

AAontology: Two-level classification of scale categories and subcategories

The two-level AAontology classification of 586 amino acids scales into 8 categories and 67 subcategories is depicted in Figure 3. This taxonomic hierarchy[19,20] provides meaningful relationships at the level of individual scales and scale subcategories.

Scale categories

Scale subcategories were manually named using our bag-of-word analysis as guidance and consolidated based on selected studies. Brief descriptions of each scale, scale subcategory, and scale category can be found in Supplementary Table 3. Each category (Figures 4–11) including their subordinated scale subcategories will be described in the following.

Accessible surface area (ASA)//Volume. The 'ASA/Volume' category comprises around 60 scales and 5 subcategories describing properties related to the general volume of amino acids and their preference of being either accessible to solvent, reflected by their ASA, or being not accessible to solvent, *i.e.*, being buried within a folded protein.

Accessible Surface Area (ASA) (n = 23) measures the residue surface area that is accessible/exposed to solvent (typically water), obtained from folded proteins. It indicates the ability of residues to interact with water, mainly at the protein surface. Residues with larger ASA often participate in protein–protein interactions, with higher ASA being more typical for polar residues with longer side chain (lysine > arginine > glutamic acid > glutamine).[44,46]

Hydrophobic ASA (n = 3) measures the residue surface area that is solvent-accessible and hydrophobic, obtained from folded proteins. This reflects the hydrophobic area exposed to water at the protein surface. This value is remarkably high for lysine—due to its long hydrophobic side chain ending in a polar ε-amino group—followed by proline, a hydrophobic residue often found in β-turns on the protein surface. Conversely, aromatic residues show moderately low values due to their less frequent occurrence on the protein surface.[44]

Buried (n = 12), as opposed to ASA, represents the propensity of amino acids to be buried within the protein core, shielded from the exterior

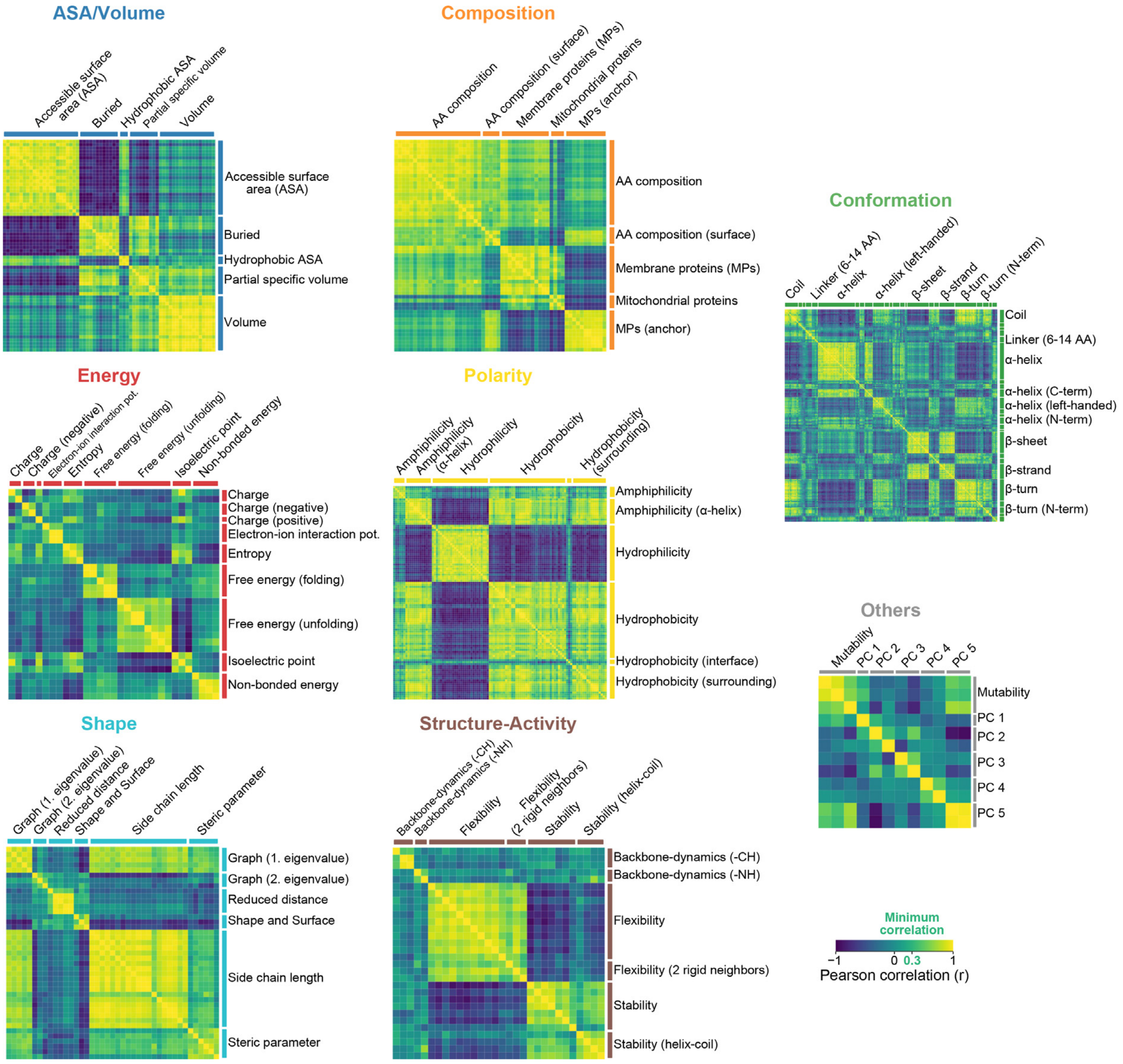

Figure 2. Correlations between scales in each category. Heatmaps for the eight scale categories (color-coded) comprising distinct subcategories with highly positive correlated scales. Each subcategory fulfills a minimum within Pearson correlation of 0.3, which was used as minimum quality criteria for the semi-automatic scale classification.

solvents. These concealed residues are crucial for the protein stability and folding, driven by hydrophobic interactions and disulfide bridges. Consequently, cysteine and large hydrophobic residues tend to be particularly buried.[46,47]

Volume ($n = 17$) is a direct measure of the amino acid size. Larger amino acids can enhance protein–protein interactions[48] and foster protein stability through long-range interactions.[49] However, their size requirements can impact chain packing.[50] This becomes crucial when mutations alter amino acid size in densely packed protein cores, which can lead to destabilization.[51] Aromatic amino acids (tryptophan > tyrosine > phenylalanine) and arginine are the largest, while glycine and alanine are the smallest.[52]

Partial specific volume ($n = 9$) reflects the effective amino acid volume in water, accounting for both physical volume and additional water displacement due to residue-solvent interactions.[53] Also included in this subcategory are hydrophobic interactivity potential[54] and bulkiness (*i.e.*, the side chain volume/length ratio).[55] These properties affect protein structure by promoting stability and introducing steric hindrances. Arginine, despite its high *Volume*, exhibits only a moderate *Partial specific volume* as it lacks additional hydrophobic water displacement. In contrast, large hydrophobic residues, particularly branched-chain amino acids (isoleucine, leucine, valine)[56] and aromatic residues (phenylalanine, tryptophan, tyrosine), score highest.[54,55,57]

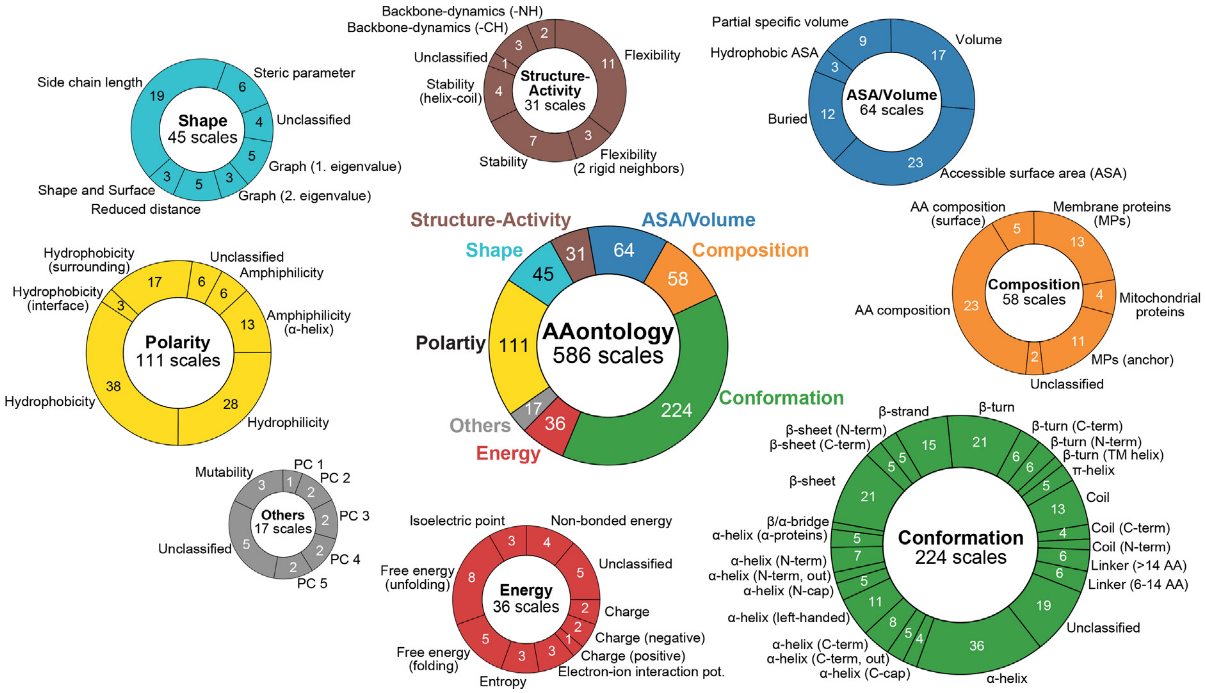

Figure 3. AAontology: Categories and subcategories. Two-level hierarchy of AAontology comprising 67 subcategories within the following 8 categories (excluding unclassified scales): ‘ASA/Volume’ (blue; ‘ASA’: Accessible Surface Area) with 5 subcategories, ‘Composition’ (orange) with 5 subcategories, ‘Conformation’ (green) with 24 subcategories, ‘Energy’ (red) with 9 subcategories, ‘Others’ (gray) with 6 subcategories, ‘Polarity’ (yellow) with 6 subcategories, ‘Shape’ (light blue) with 6 subcategories, and ‘Structure-Activity’ (brown) with 6 subcategories. Each category is visualized by a donut plot, with its size approximately reflecting the number of scales assigned to the respective category.

Composition. The ‘Composition’ category includes around 60 scales and 5 scale subcategories regarding the frequency of amino acid occurrence in different types of proteins, such as membrane proteins or the mitochondrial proteins.

AA Composition (n = 23) represents the overall frequency of amino acids (abbreviated here as ‘*AA*’) in proteins, largely independent of subcellular location or the residue position within a protein. This subcategory also comprises specific compositions, such as those of intra- or extracellular proteins, that correlate strongly with the general composition. Most abundant are alanine, leucine, and glycine, while amino acids containing a sulfur atom (methionine, cysteine) or a nitrogen atom within an aromatic ring (tryptophan, histidine) occur less frequently.[58–60]

AA Composition (Surface) (n = 5) reflects the propensity of amino acids to occure at the protein surface compared to the protein interior. Polar amino acids typically occur frequently on the protein surface, with aspartic acid and lysine being the most prevalent.[61]

Membrane Proteins (MPs) (n = 13) describes the amino acid frequency in transmembrane domains. These α-elical domains traverse cellular membranes, forming parts of either single-spanning or multi-spanning membrane proteins. Given their specialized roles within lipid-rich environments, these domains exhibit a distinct composition, predominantly non-polar amino acids (leucine > valine > alanine, phenylalanine), with methionine and tryptophan appearing least frequently.[62,63]

MPs (anchor) (n = 11) refers to the frequency of amino acids occurring in the N-/C-terminal regions flanking the transmembrane domains (TMD) of membrane proteins. The flanking regions fortify the anchoring of the hydrophobic TMDs[64] through strong partitioning between the hydrophobic and hydrophilic phase.[65,66] They terminate TMD helices by short motifs of hydrophilic amino acids, typically comprising charged helix-capping and/or helix-breaking residues (proline and glycine).[67] The N-terminal region is characterized by negatively charged/acidic residues, especially aspartic acid but also asparagine. In contrast, the C-terminal region follows the positive-inside rule,[68] with poly-basic motifs (arginine, lysine) enhancing protein-phospholipid interactions at the interface.[65,67,69]

Mitochondrial Proteins (n = 4) focuses on the frequency of occurrence of amino acids in mitochondrial proteins. Similar to membrane proteins, they are rich in hydrophobic residues, but typically with less valine.[62]

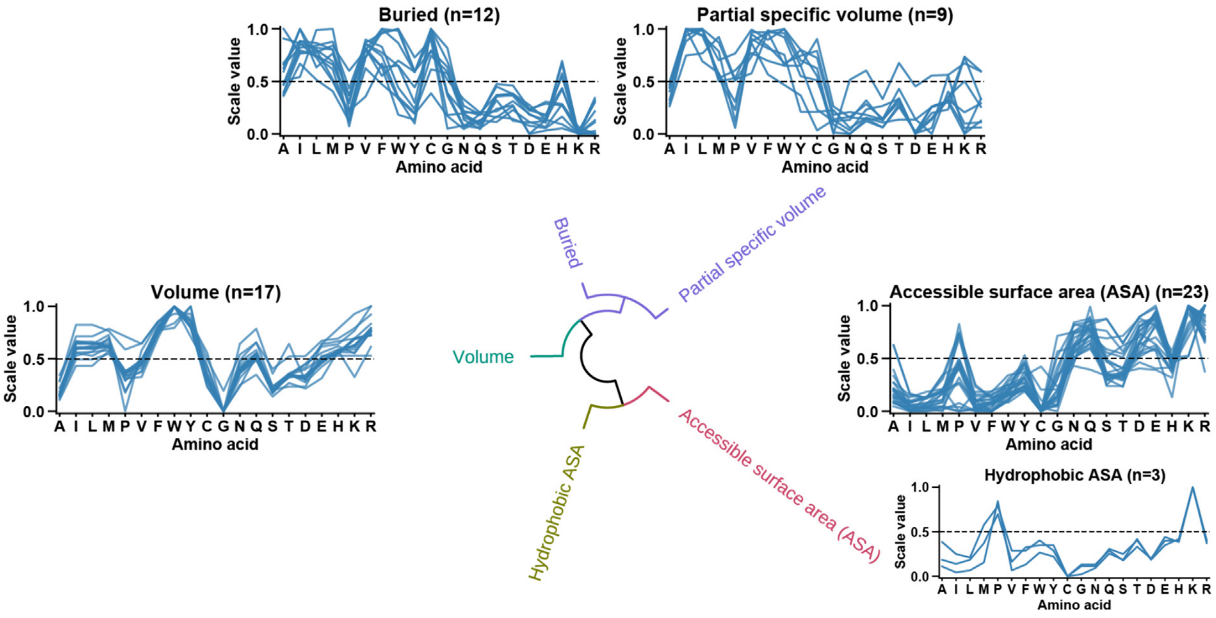

Figure 4. ASA/Volume category. Circular dendrogram showing the hierarchical clustering of all ‘ASA/Volume’ subcategories, based on the Euclidean distance between their average scales. Each subcategory is represented by a line plot, with scales depicted as lines and the number of scales assigned to a subcategory indicated in parentheses. The arrangement of line plots corresponds to the clustering results, and size modifications have been made if necessary, highlighting larger subcategories. Amino acids are given by their one-letter code and are grouped as follows: non-polar/hydrophobic amino acids comprising alanine (A), isoleucine (I), leucine (L), methionine (M), proline (P), and valine (V); aromatic amino acids of phenylalanine (F), tryptophan (W), and tyrosine (Y); polar/hydrophilic amino acids comprising cysteine (C), glycine (G), asparagine (N), glutamine (Q), serine (S), and threonine (T); acidic/negatively charged amino acids of aspartic acid (D) and glutamic acid (E); as well as basic/positively charged amino acids of histidine (H), lysine (K) and arginine (R).

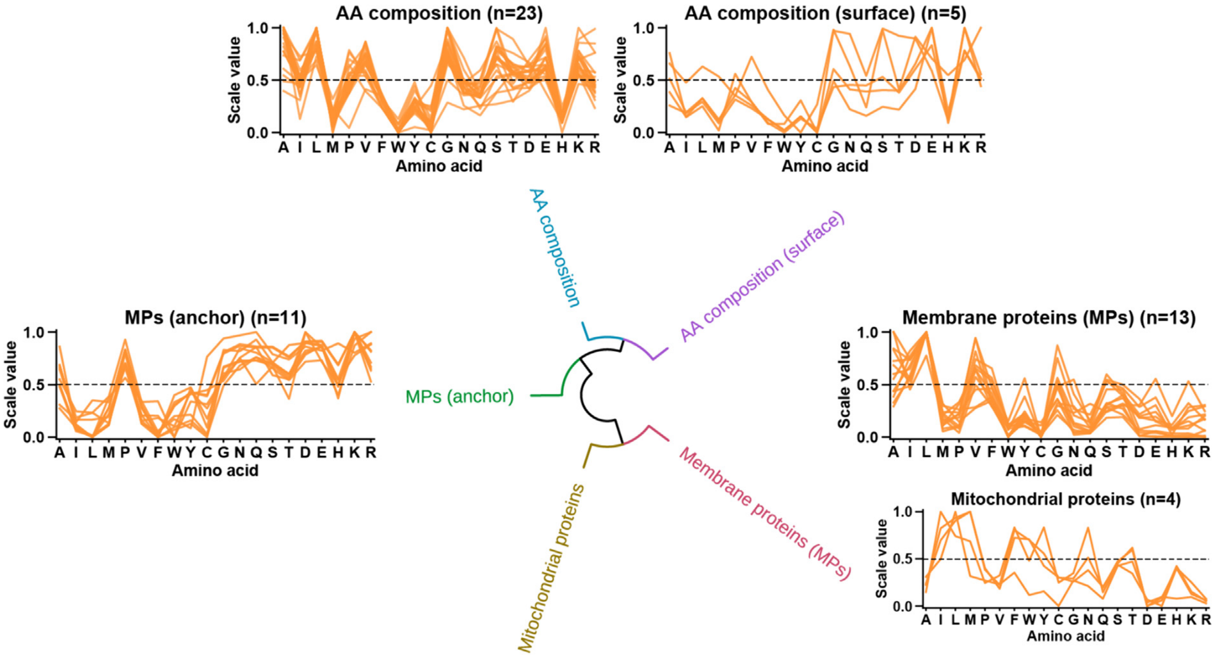

Figure 5. Composition category. Circular dendrogram showing hierarchical clustering of all ‘Composition’ subcategories. Further details are as described in Figure 4.

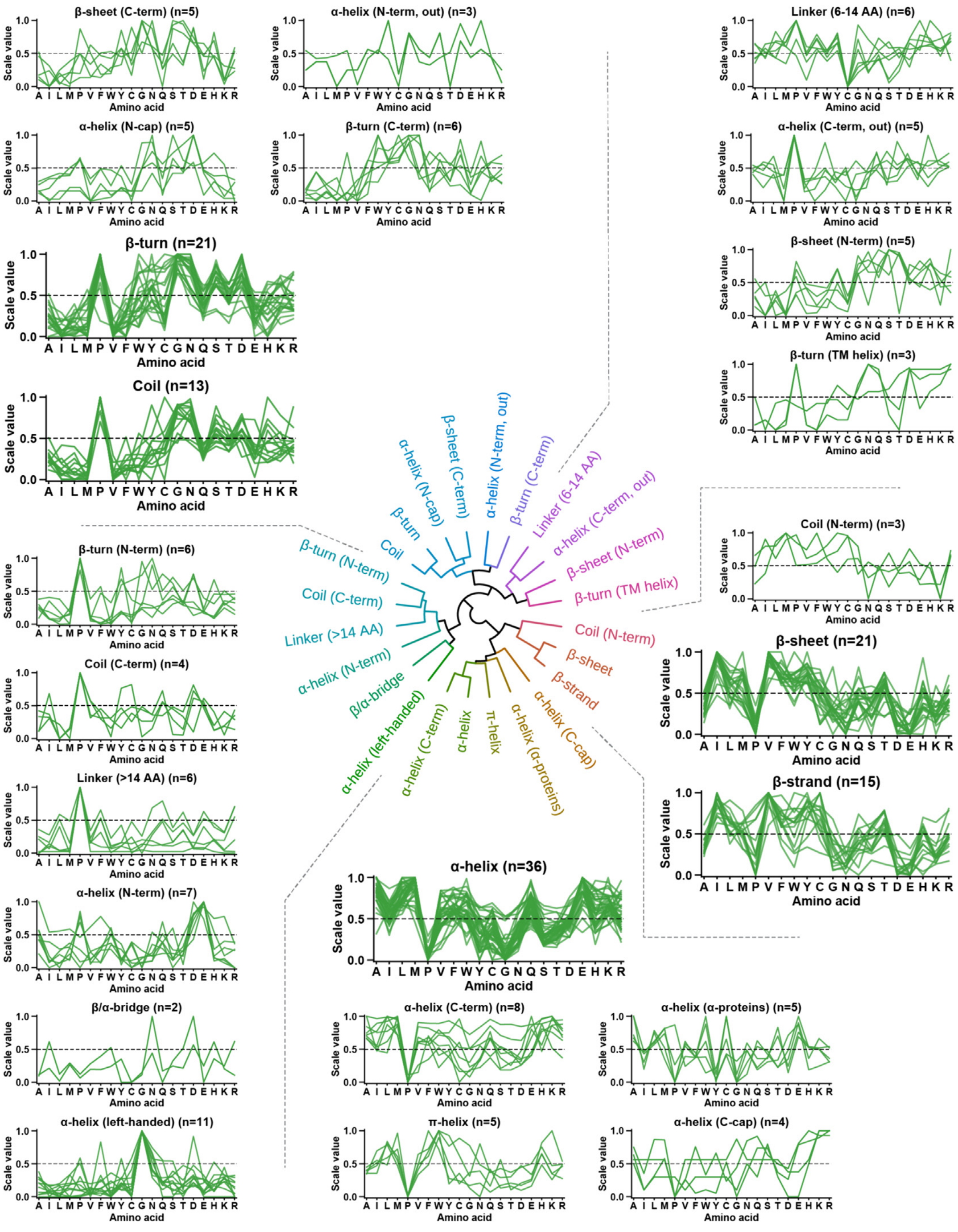

Figure 6. Conformation category. Circular dendrogram showing hierarchical clustering of all 'Conformation' subcategories. Grouping of line plots representing subcategories follows clustering results, as highlighted by dashed gray lines. An explanation of further details is given in Figure 4.

Conformation. The 'Conformation' category is the largest category, with over 200 scales across 24 subcategories. It covers four major conformations: helical, extended (β-sheet and β-strand) together with β-turn, and coil secondary structures. These account for almost all secondary structures—roughly 30–40% α-helix, 20–30% β-sheets, 20% β-turn, and 20% coils.[70] Pioneering work by Chou and Fasman,[71,72] Richardson and Richardson,[73,74] as well as Qian and Sejnowski[75] highlighted the distinct amino acid distribution across and within these conformations.

Helical and extended conformations are structured by hydrogen-bonding patterns, adopting specific dihedral angles within the polypeptide backbone.[76] Conversely, coils are defined by the

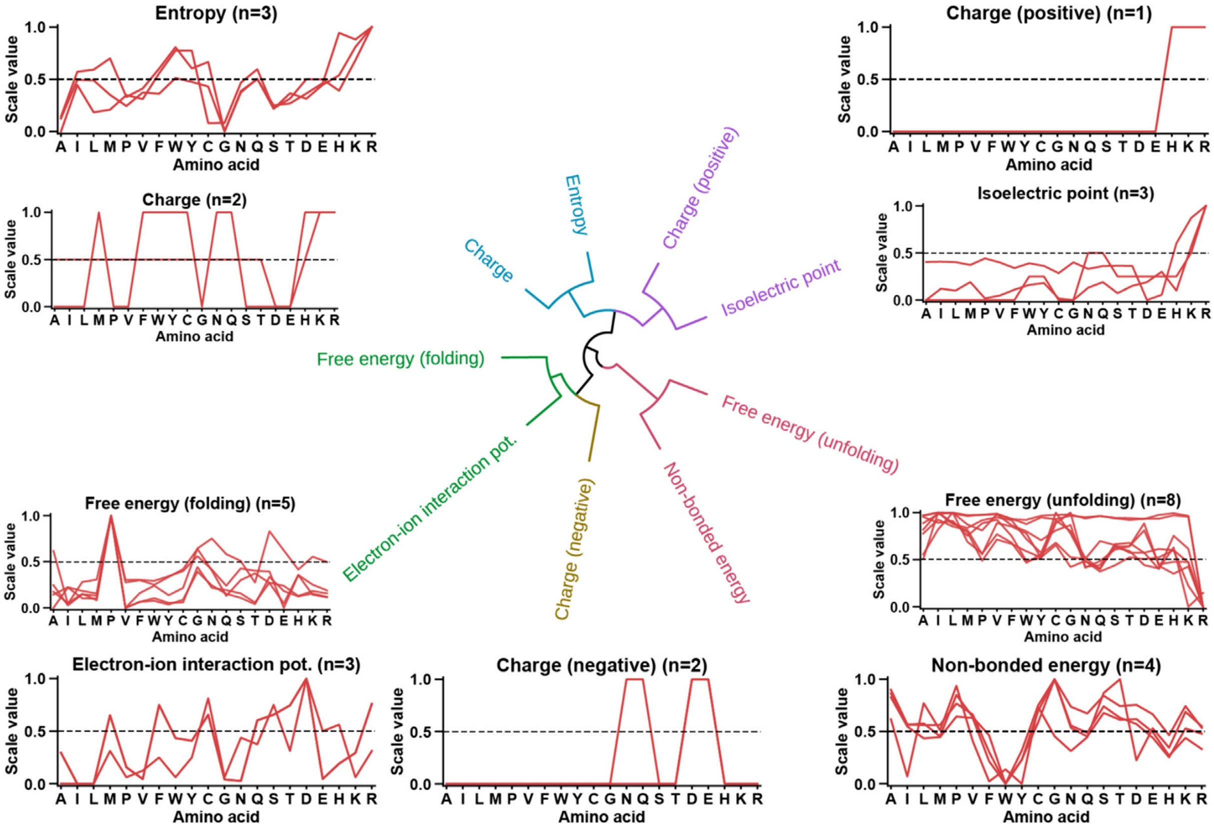

Figure 7. Energy category. Circular dendrogram showing hierarchical clustering of all ‘Energy’ subcategories. Further details are given in Figure 4. Note that we corrected the following transcription mistake in AAindex database for the ‘charge transfer capability’[115] scale (CHAM830107 scale id; *Charge (negative)* subcategory): glycine was erroneously scored 1 instead of glutamine.

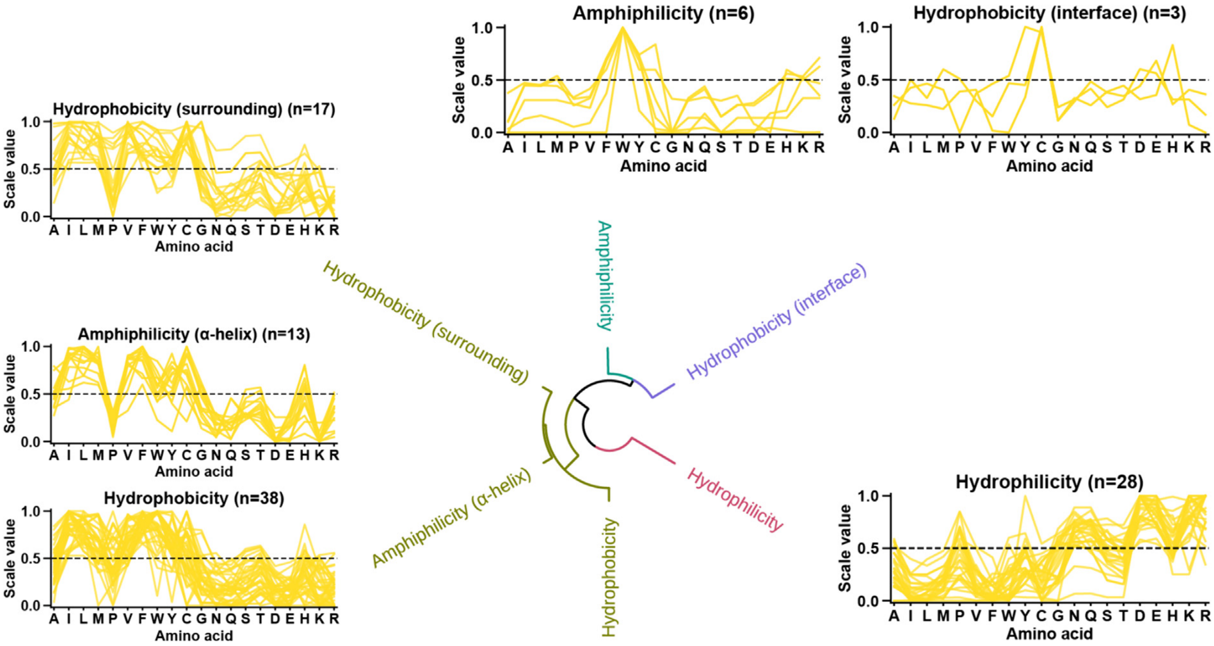

Figure 8. Polarity category. Circular dendrogram showing hierarchical clustering of all ‘Polarity’ subcategories. See Figure 4 for further details.

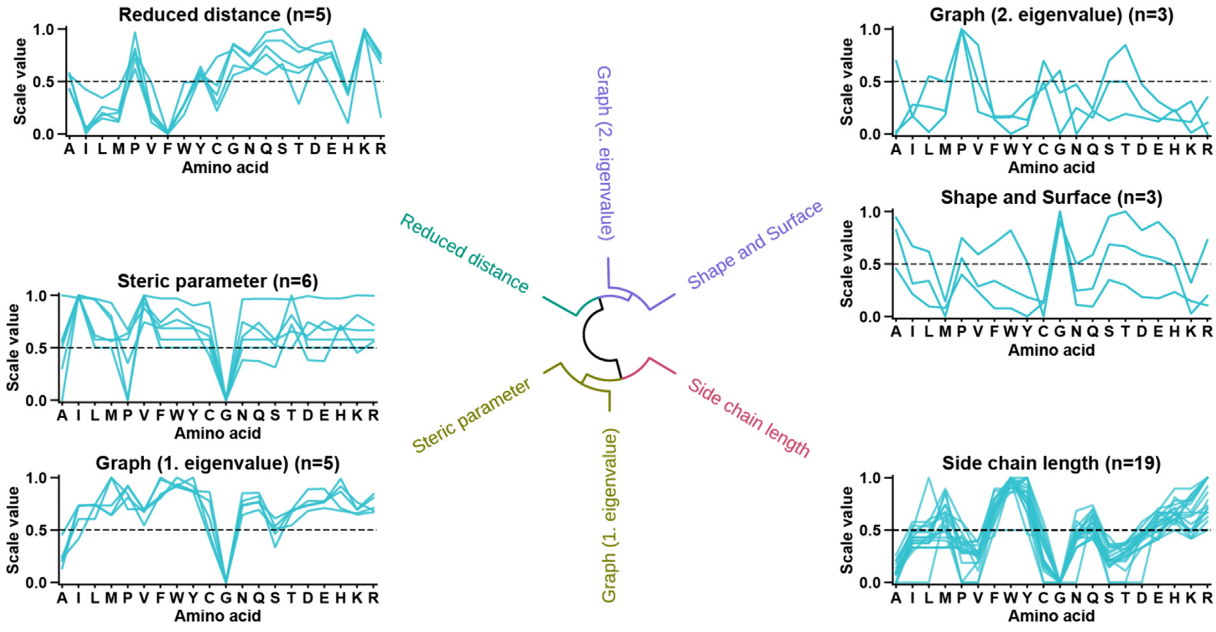

Figure 9. Shape category. Circular dendrogram showing hierarchical clustering of all 'Shape' subcategories. More details are explained in Figure 4.

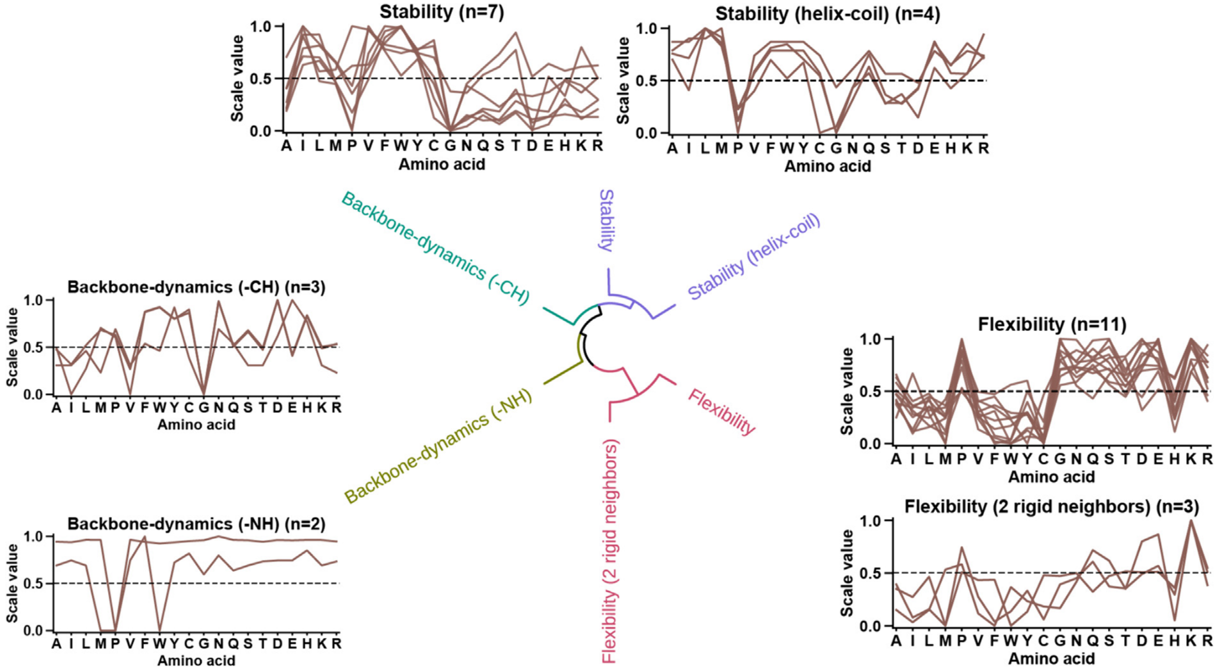

Figure 10. Structure-Activity category. Circular dendrogram showing hierarchical clustering of all 'Structure-Activity' subcategories. Details are described in Figure 4.

absence of these patterns according to the Define Secondary Structure of Proteins (DSSP) convention.[77] Coils can either adopt disordered conformations (called 'random' coils) or form structured loops (non-random coils[78] in folded proteins. Secondary structure formation is context-dependent, demonstrated by conformational transition studies for polyalanine[79] and polylysine.[80] With decreasing temperature, the conformational prevalence order is described by coil > β-sheet > helix, whereas an decrease in solvent hydrophobicity reorders this preference to coil > α-helix > β-sheet.[79,81] For

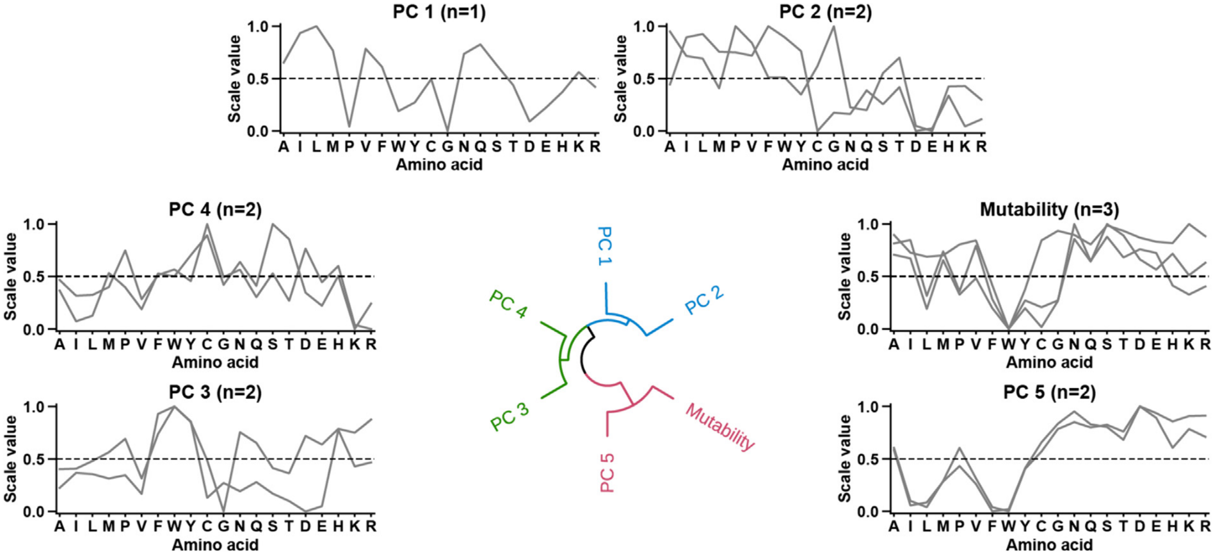

Figure 11. Others category. Circular dendrogram showing hierarchical clustering of all ‘Others’ subcategories. See Figure 4 for details.

example, in hydrophilic solvents, α-helices are less stable than β-sheets but vice versa in hydrophobic membranes. α-Helices can segregate hydrophobic and hydrophilic amino acids, while β-sheets shield hydrophobic residues from water,[44] albeit playing a multifaceted role in protein aggregation.[82] Intriguingly, proline and glycine (to a lesser extent), abundant in random coils, can interrupt α-helix formation and increase flexibility.[83] In β-sheets, these residues can destabilize the central structure but enhance peptide chain turns at the edges.

Each conformational subcategory describes the prevalence/tendency of residues to occur within the associated secondary structure. However, for the sake of clarity, only the secondary structure and their conformational implications will be described, while bearing in mind that these subcategories essentially represent residue frequencies.

Helical conformations. These subcategories describe α-helical and π-helical conformations, differing in the number of residues per 360° turn—3.6 and 4.4, respectively. The more compact 3_{10}-helix, comprising 3.0 residues per turn, is not described in AAindex, and thus not included here. Although 3_{10}-helices are rare (4%) due to their reduced stability,[84,85] short 3_{10}-helix segments often appear at the C-terminus of α-helices, tightening the final helical turn.[73] They can also switch entirely to an α-helical conformation.[84] The compactness order of these helix types is 3_{10}-helix > α-helix > π-helix, each characterized by distinct backbone hydrogen bonding patterns of *i* to *i* + 3, *i* + 4, *i* + 5, respectively.[86] They exhibit an altered stability order of α-helix > 3_{10}-helix > π-helix, explaining their respective prevalences of 30–40%, 4%, and 0.02%.[84] The π-helix prevalence might, however, be higher than suspected,[85] particularly for short (7–10 residue long) π-helical segments.[87]

α-helix (*n* = 36) refers to right-handed helical structures with 3.6 residues per turn. It is the most abundant protein helix type, accounting for 30–40% of all secondary structures.[85] α-Helices are formed by backbone hydrogen bonding (*i* + 4) and local side chain interaction between periodically neighbouring residues (*i* + 3/4), which increase helical stability via hydrophobic, polar, and aromatic stacking interactions.[88–90] Predominant residues include alanine, large hydrophobic residues (leucine, methionine), and glutamic acid alongside glutamine.[72]

α-helix (N-term) and *α-helix (C-term)* (*n* = 7, 8) denote the segments at the N-terminus and C-terminus inside the α-helix. These segments, especially the terminal residues, can notably influence the stability of the α-helix structure.[67,72,74]

α-helix (N-cap) and *α-helix (C-cap)* (*n* = 5, 4) refer to the positions at the exact termini of the α-helix where the helix is ‘capped’, *i.e.*, either begins (N-cap) or ends (C-cap), respectively. The N-terminus has a high prevalence for aspartic acid, but also asparagine and serine, while the C-terminus is characterized by positively charged/basic residues (arginine, lysine). These residues link α-helices to adjacent turns or unstructured regions.[67,74,91]

α-helix (N-term, out) and α-helix *(C-term, out)* (*n* = 3, 5) refer to the segments at the N-terminus and C-terminus outside the α-helix structure, critical for the termination of the helix. The N-terminus often harbors negatively charged/acidic residues (mainly aspartic acid) but also residues with large side chains (mainly tyrosine) and the helix-breaking glycine. In contrast, the C-terminus comprises positively charged/basic residues (histidine, arginine), residues with a large side chain (mainly phenylalanine), and, especially, the helix-breaking proline.[67,74]

α-helix (α-proteins) (*n* = 5) reflects the frequency of residues in α-helices within proteins that predominantly have α-helices as their secondary structural element.[92]

α-helix (left-handed) (*n* = 11) is a left-handed helical structure with 3.6 residues per turn, similar to right-handed helices, but winds in the opposite direction,[76] resulting in steric hindrance between side chain and main chain atoms. Although rare and typically 4-residue long, these helices contribute to protein stability and functions such as ligand binding or active site participation.[93] Glycine and aspartic acid are prevalent.[94]

π-helix (*n* = 5) is a rare and unstable[84] helix with 4.4 residues per turn, resulting in a 1 Å hole at its center too narrow to accommodate a water molecule. This hole diminishes backbone van der Waals interactions, destabilizing π-helices.[86] Mainly stabilized by side-chain interactions, including aromatic stacking and van der Waals forces, π-helices require a collinear alignment of every fourth side chain residue.[86] This unique structure, found in specific protein binding sites,[86] may enhance enzyme ligand coordination and dipole involvement in ion transport proteins.[85] Prevalent residues include aromatic and large aliphatic amino acids—particularly tryptophan, phenylalanine, and methionine.[85]

Extended conformations and β-turn. These subcategories comprise extended conformations (*i.e.*, β-strand and β-sheet) and β-turn, a sharp 4 residue turn often linking consecutive segments within anti-parallel β-sheets, which are characterized by adjacent β-strands running in opposite directions.

β-strand (*n* = 15) refers to a fully extended segment of a polypeptide chain, typically containing 3–10 residues. When multiple β-strands align, they form β-sheets, which are predominantly intra-molecular. Inter-molecular β-sheets (‘hybrid β-sheets’[95,96]can also mediate protein-protein interactions, facilitated by residue side chains fostering interaction specificity and affinity.[97] Prevalent residues are branched-chain amino acids (value, isoleucine > leucine) and cysteine.[72,98]

β-sheet (*n* = 21) is, after *α-helix*, the second most common secondary structure, characterized by its pleated structure, also called pleated sheet. It is composed of multiple β-strands forming distinct backbone hydrogen bonding patterns (either anti-parallel or parallel),[99] supported by hydrophobic side chain interactions and avoidance of steric side/main chain clashes.[100] β-Strands are connected by short loops of two to five residues or β-turns. Prevalent residues are branched-chain (valine, isoleucine > leucine) and aromatic (phenylalanine, tyrosine > tryptophan) amino acids.[72,75]

β-sheet (N-term) and *β-sheet (C-term)* (*n* = 5, 5) denote the N-terminal and C-terminal segment of a β-sheet, respectively. Generally, prevalent residues near or at the termini are serine and the destabilizing proline (inducing a 90° backbone turn) and glycine (enhancing backbone flexibility).[100] The C-terminus additionally contains asparagine and aspartic acid, while the N-terminus is also characterized by positively charged residues (histidine, arginine).[75]

β/α-bridge (*n* = 2), a new term introduced here, pertains to the frequency of residues in the ‘bridge region’ of the Ramachandran plot,[76] an area reflecting an energetically less favored conformation between β-sheet (top-left quadrant) and right-handed α-helix (bottom-left quadrant). Higher scores suggest that these residues can facilitate context-dependent folding, such as α-helix/β-sheet transition, by acting as a conformational switch by engaging in long-distance (intra)molecular hydrogen bonding.[101–103] Polar residues with longer side chains and moderate β-sheet or α-helix preferences are favored, *e.g.*, asparagine, aspartic acid, and histidine.[94,104,105]

β-turn (*n* = 21), also known as β-bend or hairpin loop, is the third most common secondary structure. Comprising 4 amino acids, β-turns sharply reverse the polypeptide chain by 180°. This reversal is often aided by a hydrogen bond between the 1st and the 4th residue, yielding multiple subclassifications.[106] Typically connecting consecutive segments within anti-parallel β-sheets, β-turns can overlap with other structural elements (*e.g.*, co-occurrence of 1st residue in α-helix or β-sheet[70] and frequently appear on protein surfaces, contributing to compact structures. Prevalent residues include proline, glycine, asparagine, and aspartic acid.[72,107]

β-turn (N-term) and *β-turn (C-term)* (*n* = 6, 6) represent the prevalence of residues for the 1st and 2nd positions, and the 3rd and 4th positions within the β-turn structure.[72,94]

β-turn (TM helix) (*n* = 3) refers to β-turns placed in the middle of a long single-spanning transmembrane helix. Unlike α-helical turns, completing 360° over 3 to 4 residues, β-turns accomplish a 180° turn over 4 residues. This chain reversal can transform a single long transmembrane helix into two closely spaced helices. Prevalent are charged residues (aspartic acid, arginine > glutamic acid, lysine) and the helix-breaking proline, while glycine, compared to its presence in *β-turn*, demonstrates only a moderate propensity.[108]

Coil conformation and linkers. These subcategories describe coil conformations and linker segments. Unlike helical and extended secondary structures, coils, devoid of a defined backbone hydrogen-bonding pattern, can adopt either unstructured coil ('random coils') or structured loop forms. Linkers, typically unstructured in isolation, may assume specific conformations upon interaction with other protein parts or under certain conditions. While random coils and unstructured linkers lack a defined structure, their roles in proteins differ. Coils are highly dynamic and can mediate various functions, whereas linkers specifically connect functional domains, providing flexibility.

Coil (n = 13), the fourth major secondary structure, is defined by the absence of well-defined backbone hydrogen-bonding patterns,[76] which are characteristic for α-helices or β-sheets. Coils can be 'random' coils, typically elusive in X-ray crystallography, or structured loops (non-random coils,[78] which are observable.[109] 'Random' coils are dynamic, flexible segments enabling critical conformational changes in proteins and playing a vital role in molecular interactions, especially in intrinsically disordered proteins.[110,111] In contrast, structured loops connect α-helix or β-sheet motifs in folded proteins,[112] contributing to molecular recognition at the protein surface.[109] Akin to β-turns,[113] prevalent residues are proline, glycine, asparagine, and serine.[75,107]

Coil (N-term) and *Coil (C-term)* (n = 3, 4) denotes the N-terminal and C-terminal sections of coil conformations, respectively. Notably, methionine is typically present in the N-terminus but absent at the N-terminus of a coil.[75]

Linker (>14 AA) and *Linker (6-14 AA)* (n = 6, 6) describes regions that connect functional domains, called linkers. The linker length significantly influences protein structure and function, enabling cooperative interactions between domains. Long linkers (>14 residues) predominantly adopt helical or coil structures, ensuring flexibility and domain separation. Medium-sized linkers (6–14 residues), however, can also form β-strand structures, striking a balance between flexibility and stability.[114]

Energy. The 'Energy' category comprises around 40 scales organized into 9 specific subcategories, each highlighting different energetic aspects of amino acids including free energy—determining conformational stability—and charge, playing an important role for protein structure and function such as enzymatic activity, protein interactions, or anchoring of transmembrane proteins.

Charge (n = 2) represents net charge and'charge transfer donor capability'.[115] The net charge reflects proton-related ionic charge, assigning scores 1, 0, 0.5 to positively charged (arginine, lysine), negatively charged (aspartic and glutamic acid), and all other residues, respectively..[116] The 'Charge transfer donor capability' marks the ability (1) or inability (0) to donate an electron with a certain ionization energy input. It is present in polar (cysteine, asparagine, glutamine), basic (histidine, arginine, lysine), and aromatic residues, as well as in methionine (due to its sulfur atom).[115]

Charge (negative) (n = 2) describes a residue´s negative charge (aspartic acid, glutamic acid)[117] and 'charge transfer capability'.[115] The latter involves polar side chain groups, specifically CONH2 in asparagine and glutamine and the COO– in deprotonated aspartic and glutamic acids, crucial for intermolecular interactions such as hydrogen bonding.[118] While the deprotonated aspartic and glutamic acid can accept protons,[117] asparagine and glutamine can accept electrons.[115,117] Presence is indicated as 1, absence as 0.[115,117]

Charge (positive) (n = 1) is dedicated to residues carrying a positive charge (arginine, lysine, histidine), indicating presence with 1 and absence with 0.[117]

Entropy (n = 3) refers to side chain conformational entropy, reflecting the number of potential conformations a residue can be part of. It is one opposing force to protein folding[119,120] and it also influences molecular interactions.[121] Alanine has a low *Entropy* due to its predominant α-helix propensity, while structure-breaking residues (glycine > proline) also exhibit low *Entropy*. In contrast, positively charged and aromatic residues have high *Entropy* due to their moderate propensities for various secondary structures such as α-helices or β-sheets.[122]

Free energy (unfolding) (n = 8) encompasses measures of conformational stability, including the 'Gibbs energy of unfolding' and 'activation Gibbs energy of unfolding'.[123] Higher values indicate enhanced stability and a larger energy barrier against unfolding. The energy required for residue transfer from hydrophobic to hydrophilic solvent, another component, reflects role of hydrophobicity role in stability.[124] Despite consistent high activation energy across residues, excluding arginine, other measures are more nuanced, with positively charged residues (arginine > lysine > histidine) scoring higher.[123,124]

Free energy (folding) (n = 5) comprises measures of the absolute free energy required for α-helix or β-strand formation, where higher values denote reduced stability, hence conformational instability.[125] While structure-disrupting residues (proline > glycine) show high values, asparagine and aspartic acid display a stronger tendency to destabilize β-strands.[126]

Isoelectric point (n = 3), often abbreviated as pI, designates the pH at which an amino acid is electrically neutral, indicating relative acidity or basicity. Basic residues (arginine > lysine > histidine) exhibit the highest values, while acidic residues show the lowest.[55] Additionally, residue basicity is also determined based on hydrogen bond donation, crucial for enzymatic reactions and protein–ligand interactions.[127] Quantified as the number of hydrogen bond donors within a side chain,[117] this parameter ranges from none (non-polar residues) to four (arginine), where each donated bond equates to 0.25.[55,117]

Electron-ion interaction potential ($n = 3$) denotes the capacity for electrostatic interactions, such as dipole–dipole, hydrogen bond, or ionic interactions. It is quantified by the average energy state of all valence electrons, the electrons within the outer shell of atoms. Essentially, it measures the potential of amino acids for engaging in electrostatic protein-protein and protein-DNA interactions.[128] Aspartic acid exhibits the highest score, followed by cysteine and serine, while glycine and the branched chain amino acids show the lowest scores.[129]

Non-bonded energy ($n = 4$) refers to the average energy per residue resulting from non-covalent interactions. Derived from X-ray crystallography using the Lennard-Jones potential, this energy encompasses electrostatic interactions and van der Waals forces, with the latter not being considered in the *Electron-ion interaction potential*. Therefore, smaller residues (glycine > proline > alanine) score highest due to their ability for close packing, fostering van der Waals interactions. Conversely, larger aromatic amino acids yield lower scores, as their size dilutes the energy contribution per atom.[130]

Polarity. The ‘Polarity’ category is the second largest category with over 100 scales organized into 6 subcategories. This category describes the foundational dichotomy of hydrophilicity and hydrophobicity, reflected by polar and non-polar residues which determine crucial biological phenomena such as protein folding and subcellular localization.

Hydrophilicity ($n = 28$) reflects the preference of an amino acid for a polar/hydrophilic environment. This is often gauged as the transfer free energy from water to a non-polar solvent. It is essential for protein-solvent interactions, influencing protein solubility, folding, and function. Charged residues exhibit the highest hydrophilicity.[124,131]

Hydrophobicity ($n = 38$) represents the amino acid preference for a non-polar/hydrophobic environment. It is often measured as the transfer free energy from a non-polar solvent to water or from the interior of a protein to the surface, playing a central role in protein folding and stability. Isoleucine and phenylalanine consistently score highest across all scales.[124,131,132]

Hydrophobicity (surrounding) ($n = 17$) describes the effective hydrophobicity of residues within globular proteins, accounting for both their own hydrophobicity and the hydrophobicity of the neighbouring residues within an 8-angstrom radius. Derived from protein crystal structures, it reflects *Hydrophobicity* in internal protein arrangements adjusted for steric hindrance, thereby gauging protein stability. Compared to *Hydrophobicity*, it is closer related to *Buried* measures, with particularly high scores for cysteine, slightly lower ones for aromatic residues, and similarly high scores for branched-chain amino acids (valine > isoleucine, leucine).[133]

Hydrophobicity (interface) ($n = 3$) focuses on the preference of residues for non-polar/hydrophobic environments at membrane interfaces, influencing the membrane-association behavior of proteins.[134] Cysteine exhibits the highest scores, followed by tyrosine.[43]

Amphiphilicity ($n = 6$) captures the preference of amino acids to occur at the interface of polar and non-polar solvents, such as the membrane-water boundary. Typically, high at the termini of transmembrane helices, this can substantially affect protein interactions and cellular processes. Amphiphilic residues consist generally of a polar and a non-polar group (alkyl side chain or aromatic ring). Tryptophan scores notably higher than other residues, followed by tyrosine and basic amino acids (arginine > lysine, histidine).[135]

Amphiphilicity (α-helix) ($n = 13$) denotes the amino acid propensity to form amphiphilic α-helices, characterized by segregated polar and non-polar faces.[136,137] These helices, when located at protein surfaces or membranes, often serve as signal sequences.[138] When interacting with membranes, they align parallel to the membrane, discerning curvature and aiding remodeling.[139] Prevalent are non-polar residues (methionine, tryptophan, and branched chain amino acids), along with certain polar (cysteine > tyrosine) and positively charged residues (histidine > arginine).[138,140]

Shape. The ‘Shape’ category, embracing 45 scales across 6 subcategories, delves into geometric and steric characteristics of amino acids. These subcategories describe side chain angles, symmetry, and unique parameters derived from amino acid representation based on graph theory. Here, amino acids are conceptualized as undirected, node-weighted graphs, with atoms as nodes and molecular bonds as edges.[141] From these graphs, different measures can be derived, such as the maximum eccentricity—the greatest number of bonds (edges) required to link the two furthest atoms within a graph. For instance, this results in 0 for glycine, 1 for alanine, or 3 for serine or cysteine. These metrics offer a mathematical fingerprint of each amino acid´s atomic structure.

Side chain length ($n = 19$) refers to the length of the amino acid side chain, quantified by the number of bonds in its longest chain. This subcategory includes also closely related graph–

based size measures such as the average eccentricity, offering an alternative perspective on the amino acid length.[115,141]

Graph (1. eigenvalue) and *Graph (2. eigenvalue)* ($n = 5$, 3) are graph-based measures, where amino acids are considered as undirected graphs, with atoms as nodes and bonds as edges. To capture the edge-bond relationships, the Laplacian matrix is computed from these graphs. Of interest are this matrix´s first (*i.e.*, smallest) and second smallest eigenvalues. These measures allow a nuanced differentiation between amino acids, even those with similar atomic configurations.[141]

Reduced distance (n = 5) reflects the relative distance of a residue from the protein's center of mass, as obtained for 14 native proteins.[142] To adjust for protein size and shape, this distance is divided by the root mean square of the 'radius of gyration', a metric depicting the average distance of all atoms from a protein´s center of mass.[143] When *Reduced distance* value exceeds 1 (0.8 for min–max normalized values), it suggests that a residue is located further from the center than the average. This may influence its function and interaction within the protein's overarching structure.[142]

Shape and Surface (n = 3) gauges relationships between the physical form of an amino acid and its solvent exposure in folded proteins. Two measures describe the rate at which the accessible surface increases relative to the protein core distance.[144] Also included is the typical torsion angle a side chain adopts within folded proteins.[145] While high for small (glycine > alanine > proline) and polar residues (*e.g.*, serine), these measures are low for cysteine and methionine, reflecting their tight packing within globular proteins.[144,145]

Steric parameter (n = 6) describes measures regarding the steric complexity of an amino acid side chain, such as branching, symmetry, or side chain angles. Low for glycine and high for isoleucine, these factors can influence how a residue fits into protein structures and its interactions with adjacent residues.[117]

Structure–Activity. The 'Structure–Activity' category, the second smallest category with 31 scales in 6 subcategories, encapsulates the spectrum of structural dynamics from flexibility to stability. Flexibility, often associated with surface residues and hydrophilicity, is key for interactions at sites such as catalytic centers, binding domains, or antigenic regions.[25,26] Conversely, structural stability is fundamental to protein folding and is generally high in buried, hydrophobic residues.

Flexibility (n = 11) refers to local residue movement within a protein (*e.g.*, positional changes or involvement in bending or twisting), comprising side chain and backbone flexibility.[146] Typically quantified by the B-factor in X-ray crystallography,[147] it reflects structural flexibility in form of atomic fluctuation and thermal vibrations. While side chain flexibility does not imply backbone flexibility, both types can contribute to conformational changes, allowing flexibility in disordered and ordered regions.[148] Side chain flexibility enhances molecular interactions such as protein–ligand binding,[149] whereas backbone flexibility is found in surface regions or epitopes.[150] Structure-breaking residues (proline, glycine), lysine, aspartic acid, and serine exhibit high flexibility.[150–153]

Flexibility (2 rigid neighbors) (n = 3) assesses the flexibility of a residue placed between two rigid neighboring residues, providing a context-specific view on structural flexibility. Under such conditions, lysine retains high flexibility, while the flexibility of aspartic acid, proline, serine moderately decreases, and that of glycine markedly drops.[150,152]

Stability (n = 7) denotes residues´ contribution to enhancing protein stability. In contrast to *Flexibility* reflecting local movement, stability is a global property of entire proteins governed by intermolecular forces, such as hydrophobic interactions. Typically, buried residues increase stability by forming β-sheets, measured, for example, by the β-coil equilibrium,[154] describing the preference of residues for extended over disordered coil structures. Prevalent residues include branched-chain (isoleucine, valine > leucine) and aromatic amino acids.[54,154,155]

Stability (helix-coil) (n = 4) quantifies the stability of residues using the helix-coil equilibrium. Higher scores indicate a preference for α-helices over disordered coil structures. Hence, residues common in α-helices, such as leucine and methionine, show higher *Stability (helix-coil)* scores. In contrast, residues more prevalent in β-sheets (*e.g.*, isoleucine and valine), tend to score lower in *Stability (helix-coil)* compared to *Stability*.[154,156]

Backbone-dynamics (—NH) (n = 2) reflects the mobility of a residue´s α-NH hydrogen atom within a polypeptide backbone, with higher values indicating backbone stability.[157] It is determined by α-NH chemical shifts using NMR, comparing hydrogen atom mobility in structured polypeptide to in random coil (highly flexible). High scores imply α-NH backbone stability, while low scores, typical in proline, indicate high mobility. Using an alternate method, spin–spin coupling, methionine and tryptophan score 0.[158]

Backbone-dynamics (—CH) (n = 3) gauges the mobility of a residue's α-CH hydrogen atom within a polypeptide, reflecting backbone stability.[157] These measures use α-CH chemical shifts to compare the hydrogen atom´s mobility between a structured backbone and a random coil. While tyrosine, asparagine, and lysine have high scores, low scores can be seen for valine, isoleucine, and especially glycine.[158,159]

Others. The 'Others' category, comprising 17 scales and 6 subcategories, contains scales which could not be reasonably assigned to other categories. It includes mutability and 5 groups of scales derived by principal component analysis (PCA). Essentially, PCA is a statistical technique that transforms a large set of correlated variables (*e.g.*, scales) into a smaller set of uncorrelated variables, called principal components (PC). These PCs encapsulate the maximum variance, thereby offering a succinct representation of the initial data and reducing the number of variables greatly. Moreover, it provides insight into the relative importance of these variables. This approach was employed by Sneath on 134 amino acid scales,[160] yielding four principal components (PC1–PC4), which we utilized for our subcategory naming.

Mutability (n = 3) refers to the likelihood of a residue to undergo mutations, calculated as the ratio of observed changes to its frequency of occurrence. This offers valuable insights into the evolutionary pressure exerted on specific residues within protein sequences.[59]

PC 1 ($n = 1$), termed 'aliphaticity' and 'typicalness' by Sneath, represents mainly a residue´s aliphatic properties (*i.e.*, signifying the presence of linear, non-aromatic carbon chains) and its general prevalence in proteins.[160]

PC 2 ($n = 2$), labeled 'hydrogenation' by Sneath, approximately corresponds to the inverse of the reactive group count in a residue, with lower values indicating higher reactivity. While proline and glycine (in Sneath´s PCA) have the highest values, acidic residues exhibit the lowest.[160]

PC 3 (n = 2), called 'aromaticity' by Sneath, denotes the aromatic properties of residues, elevated not only in aromatic residues but also residues with cyclic side chain such as histidine.[160]

PC 4 ($n = 2$), referred to as 'hydroxythiolation' and an ambiguous property by Sneath, potentially represent the ability of residues to form hydrogen bonds, prevalent in amino acids such as cysteine or serine.[160]

PC 5 ($n = 2$) is another principal component vector derived by Wold et al.,[161] but the precise properties it encapsulates are not explicitly detailed.

Relations of scale subcategories

Understanding the relationships between scale subcategories is crucial for interpretation of machine learning results. We assessed the relations between scale subcategories using hierarchical clustering, correlation analysis, and PCA. To make subcategories comparable, they were represented by their average scales (see '2.3 Representation of scale subcategories by average scales').

Hierarchical clustering of scale subcategories and amino acids. To explore the relations between all 67 subcategories, we hierarchically clustered their representative average scales based on their Euclidean distance using agglomerative clustering with complete distance[10,11] (Figure 12).

The 6 following clusters (C1–C6) were found: C1 reveals that hydrophobicity correlates with stability, frequently seen in buried residues, and driving the formation β-strands and β-sheets; C2 links α-helix propensities to conformational stability (*Free energy (unfolding)*) and steric complexity; C3 highlights the correlation between entropy, volume, side chain length, and charge, all characteristic in C-terminal residues of α-helices; C4 relates hydrophobic ASA, long linkers, and conformational instability (*Free energy (folding)*)—pronounced for proline; C5 demonstrates hydrophilicity relating to ASA, flexibility, and membrane anchor or surface residues; and C6 underlines a close conformational link between propensities to form coils, β-turns, and medium-sized linkers. Here, the connection between *Coil* and *α-helix (N-cap)* underscores that the α-helical N-terminus is less stable than the C-terminus,[162] while the association of *β-turn* with *β-sheet (C-term)* illustrates the process of anti-parallel β-sheet formation, characterized by β-strands that end at the C-terminus and sharp backbone reversal via a β-turn.[100,163]

In essence, C1–C2 represent core-forming properties, especially α-helix and β-sheet propensities. C3–C4 pertain to conformational transitions and molecular interactions, while C5 highlights surface-related properties. Finally, C6 underscores the contribution of coils and β-turns in terminating structured conformations.

Correlation analysis of subcategories. Scale-based machine learning can identify features not anticipated or seemingly inconsistent with our biological understanding. To enhance their interpretation, we investigated correlations and anti-correlations between average scales of subcategories (Figure 13).

Among strongly positively correlated subcategories (Pearson's r > 0.9–0.99, Figure 13a), three groups underline the connection between polarity and protein folding: (a) *Hydrophobicity*, *Partial specific volume*, and *Stability*, (b) *Hydrophilicity, Accessible surface area (ASA)*, and *MPs (anchor)*, which is additionally highly correlated with *Flexibility* and *Reduced distance*; and (c) *Amphiphilicity (α-helix)*, *Buried*, and *Hydrophobicity (surrounding)*. Other notable highly correlated pairs include *Volume*/*Side chain length*, *AA composition*/*MPs (single-spanning)*, and the conformation pairs *Coil*/*β-turn* and *β-sheet*/*β-strand*.

Strong negative correlations (Pearson's $r < -0.9$ to -0.99, Figure 13b) shed new light on three polarity groups: While *Hydrophobicity* anti-

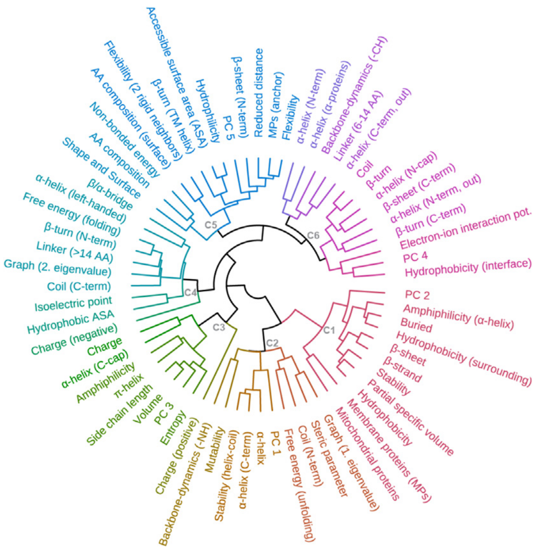

Figure 12. Hierarchical clustering of scale subcategories and amino acids. a, Circular dendrogram showing the following 6 clusters (C1–C6) of scale subcategories: C1 (*e.g.*, *Hydrophobicity* and *β-strand*), C2 (*e.g.*, *Steric parameter* and *α-helix*), C3 (*e.g.*, *Charge* and *α-helix (C-cap)*), C4 (*e.g.*, *Hydrophobic ASA* and *Linker (>14 AA)*), C5 (*e.g.*, *Hydrophilicity* and *Flexibility*), and C6 (*e.g.*, *Coil* and *β-turn*).

correlates with *MPs (anchor)*, *Hydrophilicity* anti-correlates with *Amphiphilicity (α-helix)*, *Hydrophobicity (surrounding)*, and *Buried*. In addition, the *Accessible surface area (ASA)*/*Buried* anti-correlation highlights the clear distinction between surface and buried residues. The anti-correlation of *Isoelectric point* and *Free energy (unfolding)* (conformational stability) shows that basic residues can potentially decrease stability due to the disruptive potential of their positive charge when buried, albeit context-dependent. Notably, *MPs (anchor)* inversely correlates with five other subcategories including *Buried* and *Amphiphilicity (α helix)*, underlaying its unique residue signature, dominated by proline, aspartic acid, and lysine.

Moderate positive correlations (Pearson's r = 0.75–0.9, Figure 13c) reveal a complex picture, especially for three major conformations: (a) *Coil* correlates with *Free energy (folding)* (conformational instability) and various β-turn subcategories; (b) *α-helix* correlates with *α-helix (C-term)* and *Stability (helix-coil)*, while *α-helix (C-cap)* correlates with *Isoelectric point* (basicity); (c) Extended conformations (*β-sheet*, *β-strand*) correlate with *Hydrophobicity*, *Stability* and *Partial specific volume*. Interestingly, structure termination (*β-sheet (C-term)*, *α-helix (N-cap)*) relates to *β-turn*, which highlights their structural overlapping. The correlation of *Entropy* with *Volume* and *Side chain length* indicates that larger residues contribute to structural variety (α-helix or β-sheet), unlike the helix-breaking proline or the helix-forming alanine. *Flexibility* shows broad correlations, including *Hydrophilicity* and *Accessible surface area (ASA)*.

We also identified diverse moderate negative correlations (Pearson's r = −0.75 to −0.9, Figure 13d). As expected, *Coil* anti-correlates with *α-helix*, as do extended conformations with *Hydrophilicity* and *Flexibility*. Notably, *Amphiphilicity* is negatively correlated with *Free energy unfolding* and *Mutability*, suggesting that amphiphilic residues can be less stable (*e.g.*, arginine) and less mutation-prone (*e.g.*, tryptophan). *Non-bonded energy* inversely correlates with *Volume* and *Side chain length*, showing that small residues have reduced van der Waals forces and electrostatic interactions. See Supplementary Table 4 for detailed results of our correlation analysis.

Principal component analysis of scale subcategories. PCA provides another way to visualize the relationships between scale subcategories. To obtain a two-dimensional 'snapshot' of these relationships, we divided the scale subcategories into four roughly equally sized sets from different categories, plotting the first two principal components for each (Figure 14a).

Our first PCA involved 'ASA/Volume' and 'Composition' subcategories. The first dimension, PC1 (56.2% variance explained), reveals a spectrum between buried and solvent accessible surface, relating *Buried* with *Mitochondrial Proteins*, and *Accessible surface area (ASA)* with *MPs (anchor)*. The second dimension, PC2 (23.3%), differentiates *Volume* from *Membrane proteins (MP)* and *AA composition*, implying that larger residues occur less frequently, especially in membrane proteins.

'Conformation' subcategories were analyzed together due to their large number. The first PC (38.8%) distinguishes unstructured (*Coil*, *Linker (>14 AA)*) and *β-turn* from extended (*β-strand*, *β-sheet*) and α-helical conformations, which are both separated along the second PC (15.9%). Interestingly, subcategories representing termination of β-sheets or α-helices are closely related to *Coil* and *β-turn*.

We grouped 'Energy' and 'Polarity' subcategories since hydrophobicity is often measured as transfer free energy from a non-polar solvent to water. Polarity (hydrophilicity vs. hydrophobicity) is reflected by the first PC (43.6%), relating *Free energy (folding)* to *Hydrophilicity* and *Free energy (unfolding)* to *Hydrophobicity*, reinforcing the polarity-conformational stability relationship. Interestingly, *Entropy* closely aligns with *Amphiphilicity*. The second PC (21.4%) differentiates positive from negative charge.

We also analyzed 'Shape', 'Structure-Activity', and 'Others' categories. The first PC (36.0%)

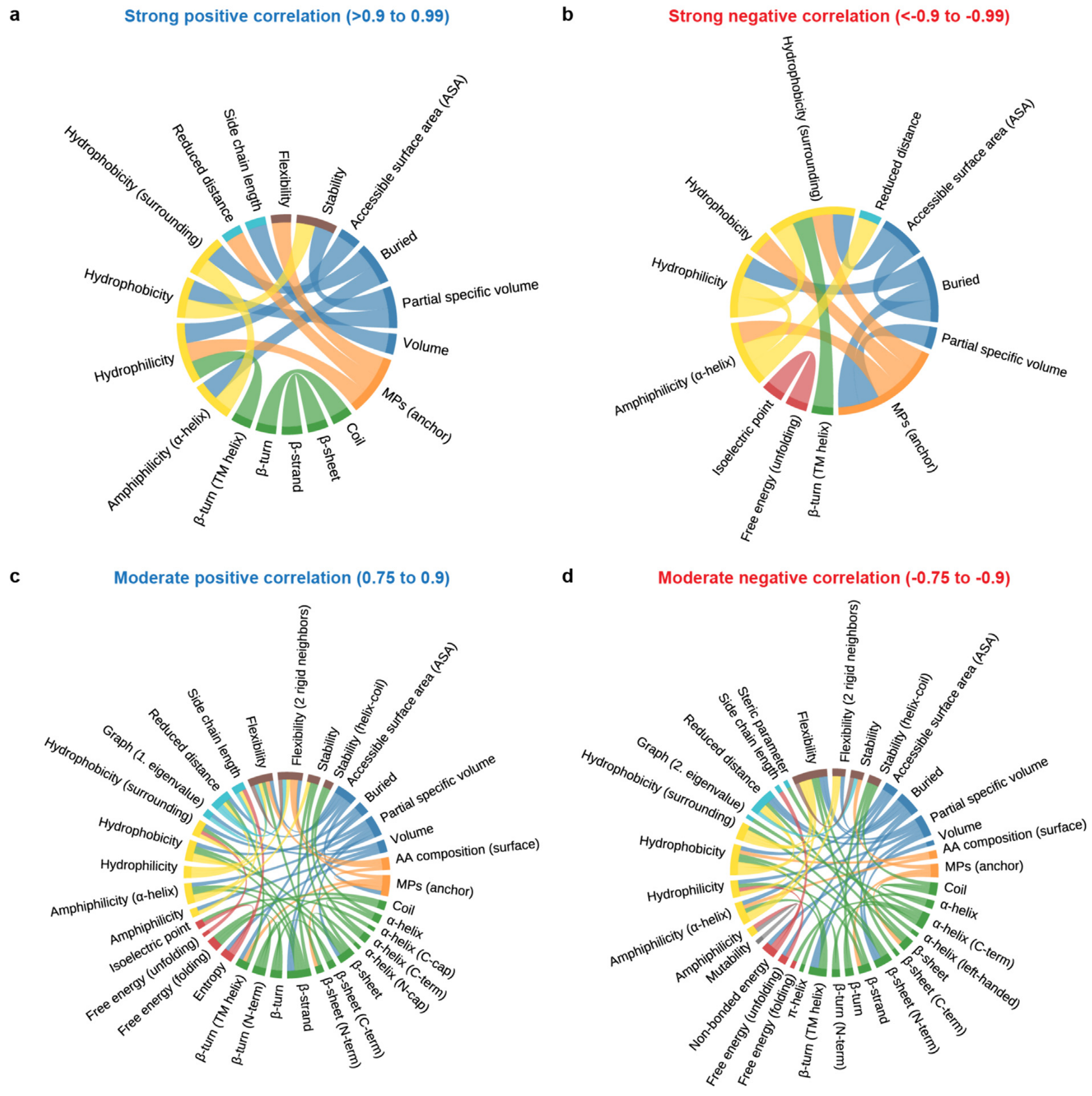

Figure 13. Correlations between average scales across scale categories. Chord diagrams showing the correlation and anti-correlation of subcategories filtered for ranges of Pearson correlation. Connections are formed between subcategories when their correlation lies within the indicated correlation range. Colors indicate the category to which the subcategories are assigned: ASA/Volume (blue), Composition (orange), Conformation (green), Energy (red), Polarity (yellow), Others (gray), Shape (light blue), and Structure-Activity (brown). Subcategories with fewer than 3 scales, such as those related to charge, were omitted for clarity. Self-referencing correlation of 1 or -1 were disregarded.

underscores the *Stability-Flexibility* dichotomy, relating *Stability* to steric complexity (*e.g.*, *Steric parameter* and *Side chain length*), and *Flexibility* to *Reduced distance*. The second PC (23.1%) separates orthogonal aspects of a residue's structural graph representation. embodied by *Graph (1. eigenvalue)* and *Graph (2. eigenvalue)*.

In a comprehensive PCA of all subcategories (Figure 14b), the polarity spectrum is represented in the PC1 (33.4%). *Hydrophilicity* relates to *Accessible surface area (ASA)*, *Flexibility*, *Coil*, and *β-turn*, while *Hydrophobicity* links with *Buried*, *Stability*, and structured conformations—*β-strand* and *β-sheet* more than *α-helix*. These ordered conformations are separated along the PC2 (19.1%), which also distinguishes the two complementary graph-theoretic measures of *Graph (1. eigenvalue)* and *Graph (2. eigenvalue)*.

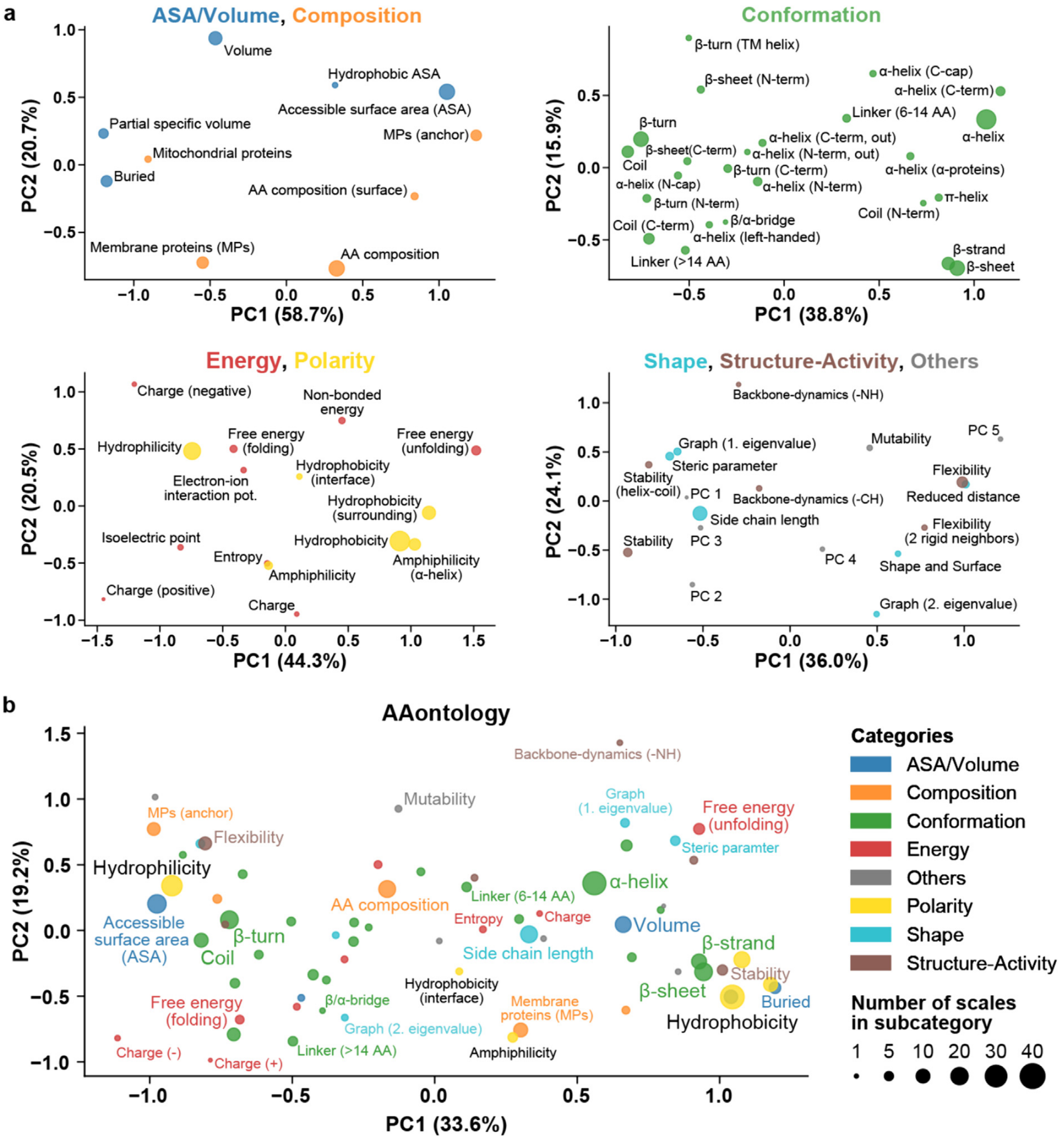

Figure 14. Relations between subcategories. Principal component (PC) analysis of average scales representing scale subcategories. **a**, Four separate PCA for subcategories of following sets of categories: ASA/Volume and Composition; Conformation; Energy and Polarity; Shape, Structure-Activity, and Others. **b**, PCA for all subcategories contained in AAontology with important ones being annotated. Each point represents a subcategory, color-coded by category; and their size is indicating the number of scales that are assigned to the subcategory.

Both PCs together discriminate *Free energy (unfolding)* from charge (negative and positive) and *Free energy (folding)*.

Overall, this analysis elucidates the interplay between higher-order physicochemical properties in protein folding, primarily driven by the hydrophilicity-hydrophobicity dichotomy. Protein folding is a two-phase process orchestrated by fundamental intermolecular forces (highlighted in *italic lowercase*). The first phase comprises the formation of secondary structures through a balance of attractive (mainly backbone *hydrogen bonding* and also *hydrophobic interactions*) and repulsive forces (*electrostatic repulsion* and *steric restrictions*),[164] reflected in the ‘ASA/Volume’, ‘Energy’, ‘Polarity’, and ‘Structure-Activity’ categories. The subsequent phase involves inter-amino acid forces, captured by the ‘Polarity’ and

'Energy' categories. Polar residues participate in diverse interactions (*dipole–dipole* or *ion–dipole interactions*, *hydrogen bonds*, or *ionic interactions*), while non-polar residues engage in weaker *van der Waals* and *hydrophobic interactions*. The latter drives hydrophobic clustering to minimize water exposure, known as the hydrophobic effect, which is crucial for protein stability.[165,166] Protein structures emerge from this interplay of intermolecular forces, influenced further by the molecular and cellular context.

Discussion

Advantages and challenges for AAontology

AAontology, with its two-level classification and in-depth subcategory analysis, enables a broader understanding of amino acid properties and provides a rigorous interpretation for scale-based machine learning. AAontology derives its strength from consensus-driven, robustly explained subcategories. Furthermore, it is a transparent, modular, and consistent framework to incorporate new amino acids properties. Its versatile nature permits applications across a wide range of computational biology problems, such as mutation analysis or protein design.

The abundance of property scales collated over six decades holds potential pitfalls. Specifically, issues may arise due to incomplete knowledge and the quality of the scales included. While AAontology helps to structure existing knowledge about amino acid properties, it is inherently limited by our current understanding of these characteristics. Any error or bias may inadvertently be carried over into computational approaches based on AAontology.

To tackle these pitfalls, scales can be revised or updated, as was done for CHAM830107 from the *Charge (negative)* subcategory and polarity scales, respectively. Furthermore, new subcategories could be defined, reflecting, for example, different types of β-strands[99] or participation in certain catalytic processes such as nucleophilic attacks (serine, cysteine, and histidine) or transition state stabilization by aromatic rings (*e.g.*, tyrosine) or charge (*e.g.*, aspartic acid or lysine).

A broader concern arises considering the general limitations of scale-based machine learning models. Their performance varies considerably depending on the prediction task, chosen model, and scale dataset used.[45] For instance, simple randomly created scales or one-hot encodings of residues have outperformed biologically meaningful property scales, however, only for a specific redundancy-reduced scale set.[167] Additionally, the array of scales and subcategories in AAontology could add complexity, challenging novice researchers.

Overcoming these limitations could involve knowledge-based pre-selection of categories from AAontology, coupled with the selection of redundancy-reduced scale sets via our AAclust framework.[45] Another solution is offered by large protein language models, overpassing simple scale-based approaches,[168–170] and thereby propelling the field of protein prediction.[171]

Do protein embeddings abolish scale-based machine learning?

Scale-based approaches (since the 1970s)[72,104] and Multiple Sequence Alignment (MSA)-based models (since the late 1980s)[172–174] leveraging evolutionary sequence information, have long been instrumental in protein structure and function prediction. However, the early 2010s saw the advent of deep learning models including Convolutional Neural Networks (CNNs)[175,176] and Long Short-Term Memory (LSTMs).[177] In the late 2010s, transformer models[178] emerged, combining CNN local structure capture and LSTM long-range dependency handling, which lead to breakthroughs such as AlphaFold[179] in protein structure prediction.

Recent state-of-the-art protein language models, such as ProtT5,[180,181] utilize transformer architectures to improve protein prediction tasks. These models, trained on billions of sequences, aim to decipher the 'language of life'[180,182] encoded by proteins. To this end, they produce a self-learned amino acid representation, called protein embeddings, that can be viewed as property scales enriched by both nearby and distant sequence context. Beside their superior prediction performance, these protein embeddings are advantageous in their simple application and unbiased nature because they are alignment-free[181] and do not rely on hand-crafted features.[170] They can also serve as input to other machine learning models via transfer learning,[183] enabling their application to moderately-sized datasets.

A significant drawback is, however, the lack of interpretability of protein embeddings,[8,9] as they are an opaque numerical representation of proteins. This challenge could potentially be addressed by AAontology. Mapping protein embeddings onto scale subcategories could merge the predictive power of embeddings with the interpretability provided by AAontology. Improved interpretability would enhance the biological insight derived from protein embeddings. The process of mapping might even expose yet unknown properties,[184] that one could call the 'dark matter' of protein biology, in an analogy previously used in this field.[185,186] Consequently, we foresee a synergistic relationship between AAontology and protein embeddings, fostering a novel understanding of proteins.

Aaontology as the fundament to understand protein biology

AAontology offers a framework to study how amino acids determine protein structure and function. While AlphaFold can produce accurate snapshots of protein structures, simulating complex and/or slow molecular processes remains still elusive.[187] AAontology can fill this gap by providing a deep understanding of amino acid properties to elucidate molecular mechanisms such as substrate or epitope recognition.

Additionally, AAontology facilitates a multidimensional understanding of amino acid changes, especially pertaining to disease-related mutations.[35–40] For instance, an alanine-to-lysine substitution not only introduces a positive charge and increases volume and hydrophilicity, but also increases the hydrophobic ASA, entropy, and flexibility within two rigid neighbors, while decreasing buriability, certain shape and surface properties and the propensity to form amphiphilic α-helices.

Moreover, we envision that AAontology serves as a cornerstone in linking physicochemical properties to protein functions and dysfunctions—as for example done for post-translational modifications.[188] Notable examples of connections to protein functions include (a) substrate cleavage facilitated by the formation of extended β-strands[189–192]; (b) epitope recognition influenced by hydrophobicity and hydrogen bonds[193,194]; and (c) protein–protein interactions[195,196] (*e.g.*, chaperone–client interactions,[197,198] protein–peptide interactions (involving peptides with extended structures, but also α-helix or β-turn conformation,[97] and protein interactions with other biomolecules such as DNA,[199] RNA,[199] lipids,[199] or small molecules.[200] Physiochemical properties are further important for subcellular localization,[21] protein stability,[201] or protein trafficking,[202] next to general cellular functions such as signaling[25,26] or cell division.[203]

Protein aggregation is a crucial example for protein dysfunction. Triggered by increased hydrophobicity and β-sheet tendency, protein aggregation is associated with over 40 human diseases including various neurodegenerative disease (*e.g.*, Alzheimer's, Parkinson, Huntington, and prion disease) and type II diabetes.[32–34] Numerous efforts have been made to predict aggregation–prone sequence regions,[24,204,205] yet they still need to be refined. Oncogenicity, another key example, is associated with altered protein folding due to changes in hydrophobicity, charge, or conformational characteristics.[29–31] Other examples encompass alterations in hydrophobicity leading to cystic fibrosis,[206] loss of charge in voltage-gated sodium channel associated with epilepsy,[207] or decreases α-helicity in myosin causing cardiomyopathy.[208] A systematic overview of the relationships between physicochemical properties and protein functions and dysfunction could lay the groundwork for a deeper understanding of protein biology. This would further foster functional annotations[209–211] of novel or so far poorly characterized proteins.

Overall, AAontology, in conjunction with tools like AlphaFold or ProtT5, is expected to unravel molecular mechanisms and unfold a systematic understanding of protein functions. This will augment numerous applications that rely on knowledge-based decisions. These include the development of biocompatible materials (*e.g.*, amino acid-based surfactants (*i.e.*, surface active agents)),[212,213] rational protein engineering[184] (*e.g.*, enzyme engineering),[214] and drug design of peptides[215,216] or proteins,[201,217] such as vaccines[218,219] or, especially, antibodies.[220–223]

Conclusion

Through a semi-automatic process involving clustering and manual refinement, we classified 586 amino acid scales into 8 categories and 67 subcategories, resulting in AAontology. This two-level classification also provides a detailed subcategory description and an extensive analysis of their relationships, bolstering interpretable machine learning models for protein bioinformatics. Beyond hierarchical clustering, PCA allowed to map these subcategories onto the major physicochemical spectra of polarity (hydrophobicity vs hydrophilicity) and conformational stability, thereby painting a holistic picture of the properties governing life.

AAontology, when combined with our AAclust framework, could overcome limitations of scale-based machine learning in protein prediction. Our plans include integrating AAontology with protein embeddings, such as ProtT5, enriching high predictive power with interpretability and possibly unveiling hitherto unknown amino acid properties. Complementing tools such as AlphaFold, AAontology could foster the deciphering of molecular mechanisms encapsulated in protein structures. Linking physicochemical properties and protein function/dysfunction, AAontology will be a useful decision support tool for mutation analysis and drug design.

Material and Methods

Dataset of amino acid scales

We obtained 566 property scales for amino acids from the AAindex database (version 9.2),[4] including their *scale id*, *scale name*, and their one-sentence description, referred to as *scale description*. A further 86 scales related to the solvent accessible surface area (ASA)[44] and hydrophobicity were manually collated from the literature (72 from Lins et al.[44] and 14 from Koehler et al.,[43] respectively) because of their general relevance for protein folding[224] and backbone dynamics.[225] After discarding scales due to missing values or complete redundancy, 586 scales remained in our dataset (553 from AAindex, 21 from Lins et al., and 12 from

Koehler et al.). For Lins et al. and Koehler at al., we created new *scale ids* by adopting the AAindex naming convention: first author´s last name, publication year, and the order of appearance in the publication (LINS030101, . . ., LINS030121 and KOEH090101, . . ., KOEH090112). A second publication in the same year is indicated by 02 such as in CHOP780211 (Chou-Fasman, 1978b). Moreover, a *scale name* and a *scale description* were created for each scale using the descriptions provided in the respective publications. Each of the 586 scales was min–max normalized to a [0,1] range (Supplementary Table 1).

Representation of amino acid scales

Property scales are represented as arrays $\mathbf{x}$ containing 20 numerical values, each corresponding to one of the 20 canonical amino acids: $\mathbf{x} = [\mathrm{a}_1, \mathrm{a}_2, \cdots \mathrm{a}_{20}]$. Multiple property scales are represented as an $m \times 20$ feature matrix X:

$$\boldsymbol{X} = \begin{bmatrix} \mathrm{a}_{1,1} & \cdots & \mathrm{a}_{1,20} \\ \vdots & \ddots & \vdots \\ \mathrm{a}_{\mathrm{m},1} & \cdots & a_{\mathrm{m},20} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_\mathbf{m} \end{bmatrix}$$

where m is the number of property scales.

Representation of scale subcategories by average scales

To compare different sets of scales, we created representative ‘average scales’. Given a predefined set of scales S = {$\mathbf{x}_1$, $\mathbf{x}_2$, $\cdots$ $\mathbf{x}_m$}, with m denoting the number of scales, an average scale $\bar{\mathbf{x}}$ was computed as the arithmetic average of scale values for each amino acid across all scales: $\bar{\mathbf{x}} = [\bar{\mathrm{a}}_1, \bar{\mathrm{a}}_2, \cdots, \bar{\mathrm{a}}_{20}]$, where $\bar{\mathrm{a}}_i$ is the arithmetic average value for the i-th amino acid over all m scales in S.

Such sets of scale can be derived using clustering algorithms. These algorithms generally partition sets of observations (in our case, the 566 property scales) into subsets or clusters by grouping numerically similar observations together. The central point of a cluster (called cluster center or *centroid*) is the arithmetic mean of the scales within that cluster and corresponds to our defined average scale $\bar{\mathbf{x}}$. The scale closest to the *centroid* is the *medoid*, which we refer to as ‘medoid scale’ in our study. Both these concepts will be instrumental in subsequent steps.

In this work, we define scale subcategories (indicated in *capital italic*) as sets of ‘similar’ scales. For instance, the *Volume* subcategory includes all volume-related scales, and its average scale is derived from the mean volume scores of the 20 amino acids across all scales. Scale similarity was determined both based on the literature and Pearson correlation analysis, ensuring a minimum correlation of 0.3 within each subcategory. For example, although hydrophobicity and hydrophilicity scales both reflect polarity concepts, they were kept separate to maintain numerical consistency. This approach lessens the cancellation of effects due to anti-correlated scales, allowing the average scales to provide a more robust consensus on certain properties drawn from various studies.

To this end, we performed the following three steps (Figure 1): scale assignment to categories using bag-of-words (a common technique in text classification,[226,227] automatic assignment of scales to subcategories through AAclust, and manual refinement of the scale assignment to subcategories and categories. To improve the clustering quality, we performed these steps multiple times with different clustering algorithms and AAclust settings, similar to an ensemble clustering approach.[228] The process is shown in Figure 2, with comprehensive details provided in Supplementary Table 2.

Scale classification and naming by automatic assignment and manual refinement

Leveraging our AAclust framework[45] as well as heuristic and knowledge-based criteria, we classified the 586 amino acid scales obtained from 151 studies curated in AAindex and two further studies (Lins et al.[44] and Koehler et al.[43]). We defined the following 8 categories (color-coded) based on previous results from the literature[1,2,12,14,15]: ‘ASA/Volume’ (blue), ‘Composition’ (orange), ‘Conformation’ (green), ‘Energy’ (red), ‘Others’ (gray), ‘Polarity’ (yellow), ‘Shape’ (light blue), and ‘Structure-Activity’ (brown). To increase the applicability and interpretability of our classification, we further aimed to subdivide each category into subcategories, each with an average size of 5–10 scales, ideally yielding a manageable number of between 50 and 100 subcategories in total.

Classification of scales using bag-of-words

In the field of natural language processing, bags-of-words are sets of terms associated with the same class and they are employed as a simple tool for text classification.[226,227] We defined a bag-of-words for each of the 8 categories: ASA/Volume ({‘accessible surface‘, ‘volume’, . . .}), Composition ({‘amino acid distribution’, ‘membrane-propensity’, . . .}), Conformation ({‘β-sheet’, ‘helix’, . . .}), Energy ({‘charge’, ‘entropy’, . . .}), Others ({‘mutability’, ‘principal component’, . . .}), Polarity ({‘hydrophobicity’, ‘polarity’, . . .}), Shape ({‘graph’, ‘steric parameter’, . . .}), and Structure-Activity ({‘flexibility’, ‘side chain interaction’, . . .}). The complete bag-of-words for each category are given in Table 1. For each scale and category, the occurrence of these words in the *scale name* and the *scale description* was counted, referred to as word count, and the scale was assigned to the category for which it had the highest word count.

525 out of 586 scales were classified by assigning a scale to a scale category when at least one word within the category´s bag-of-words occurred in the *scale name* or the *scale description*. If scales were assigned to multiple categories, the category with fewer scales was preferred to achieve a more balanced partition of scales. The remaining 61 scales were classified manually using additional information from the literature.

Subcategory assignment by AAclust clustering

Leveraging our AAclust framework,[45] we subdivided the categories hierarchically into subcategories, each named according to term counts from *scale names*, such as the *Entropy* subordinated to the 'Energy' category.

In the first step, scales within each category were clustered. We used AAclust in conjunction with *k*-means[16] and agglomerative clustering[10] (with average, complete, ward, single linkage methods.[11] These clustering models demonstrated excellent performance in our previous AAclust study, and were used as implemented in the scikit-learn library.[229] For AAclust, we set a minimum within-cluster Pearson correlation of 0.3, ensuring a positive correlation among all scales within each subcategory. Depending on the AAclust merging parameters, the number of clusters varied between 50 and 270 clusters. Cluster models were subsequently evaluated based on the silhouette coefficient,[230] Calinski Harabasz score,[17] and Bayesian information criterion (BIC),[230] as outlined in AAclust. To obtain a manageable number of 50–100 subcategories, we considered for further steps only the top-performing models that yielded a maximum of 100 clusters over all categories. This was achieved by employing cluster merging in the AAclust approach, a process that reduces cluster numbers by reassigning scales from smaller to larger clusters.

In the second step, we assigned names to each cluster by analyzing term counts in *scale names*, thereby naming subcategories. Term lists from each *scale name* were generated, considering both the *scale name* and whole-word substrings (without parentheses). For example, for the *scale names* 'α-helix' and 'α-helix (N-terminal)', the term lists would be ['α-helix'] and ['α-helix (N-terminal)', 'α-helix', 'N-terminal'], respectively. Term lists were then combined, and term counts within each subcategory were calculated. Additionally, we used AAclust to identify the medoid scale, having the highest correlation to the cluster center. The most frequently occurring term became the subcategory name, such as *Hydrophobicity* occurring in almost all hydrophobicity scales. Names were assigned in descending order of cluster size, and if two terms occurred equally frequently the medoid scale´s name or the shorter was preferred. Clusters with only one scale or a duplicate name were labeled 'unclassified' and merged into a single subcategory within their category, denoted as 'unclassified Category name'. We excluded five charge-related scales from the clustering process due to their binary nature and later assigned them to distinct subcategories.

This procedure was repeated multiple times with varying models and settings to improve naming consistency and clustering robustness.[228] This iterative process also allowed for the refinement of *scale names* based on subcategory classification, *scale description*, and literature review. For instance, scales such as the 'TOFT index' or 'PRIFT index' led to the subcategory of *Amphiphilicity (α-helix)*.

Overall, we obtained 7 'unclassified' and 73 meaningful subcategories (Supplementary Table 2). The subcategory names largely overlapped with the bag-of-words used for category classification, offering a convenient link between the scales and their respective subcategories.

Manual refinement of scale subcategories and categories

We reduced the 73 meaningful subcategories onto 67 subcategories by manual refinement using heuristic and knowledge–based criteria. In this step, for each scale one out of the following four different actions were performed:

- **No changing**: The scale classification is kept, as automatically derived using AAclust.
- **Changing category**: Change of the AAclust classification for the scale category and subcategory.
- **Changing subcategory**: Change of the AAclust classification for the scale subcategory.
- **Renaming subcategory**: Change of the semi-automatically derived subcategory name.

We manually renamed subcategories of 43 scales for clarity and brevity, applying heuristics including shortening (*e.g.*, 'Principal Component 1' to *PC1*), summarizing (*e.g.*, 'Charge donor' included into *Charge*), deleting unclear names (*e.g.*, 'Kerr-constant'), and aligning with naming conventions. Notably, we adhered to a general convention for the N- or C-terminal preference in conformational subcategories, as seen in the renaming of 'β-turn (3rd residue)' to *β-turn (C-term)*.

Scales were allotted to 'unclassified' subcategories when they had a Pearson correlation lower than 0.3 with any scale within their original subcategory and failed to attain a minimum within-cluster correlation of 0.3 with any other subcategory. Further, a scale was deemed 'unclassified' when a conclusive literature-based assignment failed. Consequently, 42 of the total 586 scales could not be classified.

Table 1 Scale categories with their bag-of-words and the number of respective subcategories before and after manual refinement (indicated in bold).

Category	Bag-of-words	# subcategories
ASA/Volume	‘accessibility’, ‘accessible surface’, ‘buriability’, ‘buried’, ‘bulkiness’, ‘exposed residue’, ‘size’, ‘van der waals’[#1], ‘volume’, ‘weight’	5 **5**
Composition	‘amino acid distribution’, ‘composition’, ‘frequency of occurrence’, ‘membrane preference’, ‘membrane proteins’, ‘mesophilic proteins’, ‘proteins of mesophiles’, ‘mt proteins’, ‘multi spanning proteins’, ‘nuclear proteins’, ‘proteins of thermophiles’, ‘proteins of mesophiles’, ‘sequence frequency’, ‘single spanning proteins’, ‘thermophilic proteins’, ‘transmembrane regions’	4 (+1 unclassified) **5 (+1 unclassified)**
Conformation	‘alpha-helix’, ‘aperiodic indices’, ‘average relative fractional occurrence’, ‘bend’, ‘beta-sheet’, ‘beta-strand’, ‘beta-structure’, ‘coil’, ‘conformational state’, ‘extended’, ‘helical’, ‘helix’, ‘linker’, ‘loop’, ‘normalized frequency’, ‘pi-helices’, ‘pleated-sheet’, ‘relative preference value’, ‘turn’, ‘chain reversal’	27 (+1 unclassified) **24 (+1 unclassified)**
Energy	‘charge’, ‘electrical effect’, ‘electron-ion interaction potential’, ‘energy transfer’, ‘entropy’, ‘free energies’, ‘free energy’, ‘gibbs energy’, ‘heat capacity’, ‘hydrogen bond donors’, ‘isoelectric point’[#2], ‘non-bonded energy’, ‘nonbonding orbitals’, ‘partition coefficient’, ‘partition energies’, ‘partition energy’, ‘polarizability’, ‘transfer energy’	15 (+1 unclassified) **9 (+1 unclassified)**
Polarity	‘amphiphilicity’, ‘cornette et al’[*], ‘hydration’, ‘hydrophobic moment’, ‘hydrophobic parameter’, ‘hydrophobicity’, ‘hydrophilicity’, ‘hydropathy’, ‘pK’, ‘polar requirement’, ‘polarity’, ‘refractivity’, ‘retention coefficient’, ‘surrounding residues’	5 (+1 unclassified) **6 (+1 unclassified)**
Shape	eccentricity’, ‘eigenvalue’, ‘graph’, ‘kakraba-knisley’[*], ‘longest chain’, ‘prabhakaran-ponnuswamy’[*], ‘reduced distance’, ‘side chain’, ‘steric parameter’, ‘value of theta’	6 (+1 unclassified) **6 (+1 unclassified)**
Structure-Activity	‘average interactions’, ‘b values’, ‘chemical shift’, ‘contact number’, ‘equilibrium constant’, ‘flexibility’, ‘interactivity’, ‘side chain interaction’, ‘signal sequence’, ‘site occupied by water’, ‘spin-spin coupling’, ‘stability’, ‘zimm-bragg parameter’	4 (+1 unclassified) **6 (+1 unclassified)**
Others	‘bitterness’, ‘kerr-constant’, ‘melting point’[#3], ‘mutability’, ‘optical rotation’, ‘principal component’, ‘principal property’, ‘rf rank’	7 (+1 unclassified) **6 (+1 unclassified)**

[#1] ‘van der waals’ refers not to van der Waals forces but rather to van der Waals volume or radius, properties related to ‘ASA/Volume’.

[#2] ‘isoelectric point’ is assigned to ‘Energy’ since it is related to the net charge of residues.

[#3] ‘melting point’ is attributed to ‘Others’ due to ambiguous assignments with other categories—it could be associated, for example, with ‘Energy’ (*i.e.*, as a measure of heat capacity) or to ‘Structure-Activity’ (*i.e.*, as a measure of stability).

[*] Names of authors contributing various scales to a specific category.

Scales were reassigned to different subcategories and categories to enhance the overall Pearson correlation within each subcategory and align better with existing literature. In some cases, scales were allocated to larger subcategories, resulting in a slightly lower minimum within-cluster Pearson correlation, but maintaining a high degree of specificity and consistency in smaller subcategories. For example, despite a higher minimum correlation with the *PC5* subcategory, the ’Relative mutability’ scale was placed into the broader *Mutability* subcategory.

Overall, we manually refined 70% of the automatically assigned scales using these heuristic and knowledge-based criteria, classifying 544 out of the 586 scales into 8 categories and 67 distinct subcategories. Scales within the same subcategory often displayed a high positive correlation (>0.9) or at least a minimum within Pearson correlation of 0.3 (Figure 3).

Code availability

AAontology is a foundational component of AAanalysis, a Python-based framework for interpretable protein prediction, freely accessible at https://github.com/breimanntools/aaanalysis and documented under https://aaanalysis.readthedocs.io/en/latest/.

Credit authorship contribution statement

Stephan Breimann: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Frits Kamp:** Writing – review & editing. **Harald Steiner:** Writing – review & editing, Funding acquisition. **Dmitrij Frishman:** Writing – review & editing, Supervision, Funding acquisition.

DATA AVAILABILITY

All data can be sourced from the Supplementary Tables, and the code is included in the freely accessible AAanalysis Python package.

DECLARATION OF COMPETING INTEREST

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

Acknowledgements

We thank Dieter Langosch for engaging in critical discussions regarding physicochemical properties, particularly on aspects related to polarity and flexibility.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.jmb.2024.168717.

Received 3 June 2024;
Accepted 19 July 2024;
Available online 24 July 2024

Keywords:
physicochemical properties;
protein evolution;
clustering;
interpretability;
protein prediction

Abbreviations:
ASA, Accessible Surface Area; AA, Amino Acids; MPs, Membrane Proteins; TMD, Transmembrane Domain; N-term, N-terminus; C-term, C-terminus; N-cap, N-terminal capped; C-cap, C-terminal capped; PCA, Principal Component Analysis; MSA, Multiple Sequence Alignment

References

1. Nakai, K., Kidera, A., Kanehisa, M., (1988). Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* **2**, 93–100.
2. Tomii, K., Kanehisa, M., (1996). Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* **9**, 27–36.
3. Kawashima, S., Kanehisa, M., (2000). AAindex: Amino acid index database. *Nucleic Acids Res.* **28**, 374.
4. Kawashima, S. et al, (2008). AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, 202–205.
5. Liu, B., Gao, X., Zhang, H., (2019). BioSeq-Analysis2.0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* **47**, 1–12.
6. Chen, Z. et al, (2021). ILearnPlus: A comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res.* **49**, 1–19.
7. Chen, Z. et al, (2018). IFeature: A python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **34**, 2499–2502.
8. Greener, J.G., Kandathil, S.M., Moffat, L., Jones, D.T., (2022). A guide to machine learning for biologists. *Nature Rev. Mol. Cell Biol.* **23**, 40–55.
9. Gosiewska, A., Kozak, A., Biecek, P., (2021). Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decis. Support Syst.* **150**, 1–10.
10. Ward, J.H., (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244.
11. Murtagh, F., Contreras, P., (2012). Algorithms for hierarchical clustering: An overview. *Data Min. Knowl. Discov.* **2**, 86–97.
12. Saha, I., Maulik, U., Bandyopadhyay, S., Plewczynski, D., (2012). Fuzzy clustering of physicochemical and biochemical properties of amino Acids. *Amino Acids* **43**, 583–594.
13. Bezdek, J.C., (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press.
14. Simm, S., Einloft, J., Mirus, O., Schleiff, E., (2016). 50 years of amino acid hydrophobicity scales: Revisiting the capacity for peptide classification. *Biol. Res.* **49**, 1–19.
15. Forghani, M., Khani, R., (2017). A multivariate clustering of AAindex database for protein numerical representation. *In: Iranian Conference on Signal Processing and Intelligent Systems*, pp. 1–4. https://doi.org/10.1109/ICSPIS.2017.8311579.
16. MacQueen, J., (1967). Some methods for classification and analysis of multivariate observations. *Berkeley Symp. Math. Stat. Probab.* **5**, 281–297.
17. Calinski, T., Harabasz, J., (1974). A dendrite method for cluster analysis. *Commun. Stat.* **3**, 1–27.
18. Rousseeuw, P.J., (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65.
19. Bard, J.B.L., Rhee, S.Y., (2004). Ontologies in biology: Design, applications and future challenges. *Nature Rev. Genet.* **5**, 213–222.
20. van Rees, R., (2003). Clarity in the usage of the terms ontology, taxonomy and classification. *Comput. Sci.* **1–8**
21. Shen, Y., Tang, J., Guo, F., (2019). Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. *J. Theor. Biol.* **462**, 230–239.
22. Li, F. et al, (2019). Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: A comprehensive revisit and benchmarking of existing methods. *Brief. Bioinform.* **20**, 2150–2166.
23. Tang, Y.-J., Pang, Y.-H., Liu, B., (2020). IDP-Seq2Seq: Identification of intrinsically disordered regions based on sequence to sequence learning. *Bioinformatics* **36**, 5177–5186.
24. Teng, Z., Zhang, Z., Tian, Z., Li, Y., Wang, G., (2021). ReRF-Pred: Predicting amyloidogenic regions of proteins based on their pseudo amino acid composition and tripeptide composition. *BMC Bioinf.* **22**, 1–18.
25. Wright, P.E., Dyson, H.J., (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature Rev. Mol. Cell Biol.* **16**, 18–29.
26. Theillet, F.-X. et al, (2014). Physicochemical properties of cells and their effects on intrinsically disordered proteins (IDPs). *Chem. Rev.* **114**, 6661–6714.

27. Hessa, T. et al, (2005). Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433**, 377–381.
28. MacCallum, J.L., Tieleman, D.P., (2011). Hydrophobicity scales: A thermodynamic looking glass into lipid-protein interactions. *Trends Biochem. Sci.* **36**, 653–662.
29. Anoosha, P., Sakthivel, R., Michael Gromiha, M., (2016). Exploring preferred amino acid mutations in cancer genes: Applications to identify potential drug targets. *Biochim. Biophys. Acta* **1862**, 155–165.
30. Szpiech, Z.A. et al, (2017). Prominent features of the amino acid mutation landscape in cancer. *PLoS One* **12**, 1–12.
31. Liu, J.-J. et al, (2021). The structure-based cancer-related single amino acid variation prediction. *Sci. Rep.* **11**, 1–17.
32. Iadanza, M.G., Jackson, M.P., Hewitt, E.W., Ranson, N. A., Radford, S.E., (2018). A new era for understanding amyloid structures and disease. *Nature Rev. Mol. Cell Biol.* **19**, 755–773.
33. Eisenberg, D., Jucker, M., (2012). The amyloid state of proteins in human diseases. *Cell* **148**, 1188–1203.
34. Chiti, F., Dobson, C.M., (2017). Protein misfolding, amyloid formation, and human disease: A summary of progress over the last decade. *Annu. Rev. Biochem.* **86**, 27–68.
35. Stone, E.A., Sidow, A., (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15**, 978–986.
36. Serohijos, A.W.R., Shakhnovich, E.I., (2014). Merging molecular mechanism and evolution: Theory and computation at the interface of biophysics and evolutionary population genetics. *Curr. Opin. Struct. Biol.* **26**, 84–91.
37. Starr, T.N., Thornton, J.W., (2016). Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218.
38. Pandurangan, A.P., Blundell, T.L., (2020). Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning. *Protein Sci.* **29**, 247–257.
39. Iqbal, S. et al, (2020). Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc. Natl. Acad. Sci.* **117**, 28201–28211.
40. Rodrigues, C.H.M., Pires, D.E.V., Ascher, D.B., (2021). DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci.* **30**, 60–69.
41. Du, X. et al, (2012). Mapping of H3N2 influenza antigenic evolution in China reveals a strategy for vaccine strain recommendation. *Nature Commun.* **3**, 1–9.
42. Hebditch, M., Warwicker, J., (2019). Charge and hydrophobicity are key features in sequence-trained machine learning models for predicting the biophysical properties of clinical-stage antibodies. *PeerJ*. https://doi.org/10.7717/peerj.8199.
43. Koehler, J., Woetzel, N., Staritzbichler, R., Sanders, C.R., Meiler, J., (2009). A unified hydrophobicity scale for multi-span membrane proteins. *Proteins: Struct. Funct.* **76**, 13–29.
44. Lins, L., Thomas, A., Brasseur, R., (2003). Analysis of accessible surface of residues in proteins. *Protein Sci.* **12**, 1406–1417.
45. Breimann, S., Frishman, D., (2024). AAclust: k-optimized clustering for selecting redundancy-reduced sets of amino acid scales. *bioRxiv*.
46. Chothia, C., (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1–12.
47. Janin, J., Wodak, S., Levitt, M., Maigret, B., (1978). Conformation of amino acid side-chains in proteins. *J. Mol. Biol.* **125**, 357–386.
48. Tsai, C.J., Lin, S.L., Wolfson, H.J., Nussinov, R., (1996). Protein-protein interfaces: Architectures and interactions in protein- protein interfaces and in protein cores. Their similarities and differences. *Crit. Rev. Biochem. Mol. Biol.* **31**, 127–152.
49. Gromiha, M.M., Selvaraj, S., (1999). Importance of long-range interactions in protein folding. *Biophys. Chem.* **77**, 49–68.
50. Cantor, E.J. et al, (1997). Effects of amino acid side-chain volume on chain packing in genetically engineered periodic polypeptides. *J. Biochem.* **122**, 217–225.
51. Gromiha, M.M., Oobatake, M., Kono, H., Uedaira, H., Sarai, A., (2002). Importance of mutant position in ramachandran plot for predicting protein stability of surface mutations. *Biopolymers* **64**, 210–220.
52. Bigelow, C.C., (1967). On the average hydrophobicity of proteins and the relation between it and protein structure. *J. Theor. Biol.* **16**, 187–211.
53. Murphy, L.R., Matubayasi, N., Payne, V.A., Levy, R.M., (1998). Protein hydration and unfolding – insights from experimental partial specific volumes and unfolded protein models. *Fold. Des.* **3**, 105–118.
54. Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M., (2005). Prinicipal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins: Struct. Funct.* **58**, 22–30.
55. Zimmerman, J.M., Eliezer, N., Simha, R., (1968). The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **21**, 170–201.
56. Brosnan, J.T., Brosnan, M.E., (2006). Branched-chain amino acids: metabolism, physiological function, and application. *J. Nutr.* **136**, 269–273.
57. Bull, H.B., Breese, K., (1974). Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues. *Arch. Biochem. Biophys.* **161**, 665–670.
58. Dayhoff, M.O., Hunt, L.T., Hurst-Calderone, S., (1978). Amino acid composition. *Atlas Protein Seq. Struct.* **5**
59. Jones, D.T., Taylor, W.R., Thornton, J.M., (1992). The rapid generation of mutation data matrices. *Bioinformatics* **8**, 275–282.
60. Nakashima, H., Nishikawa, K., (1992). The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. *FEBS Lett.* **303**, 141–146.
61. Fukuchi, S., Nishikawa, K., (2001). Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *J. Mol. Biol.* **309**, 835–843.
62. Nakashima, H., Nishikawa, K., Ooi, T., (1990). Distinct character in hydrophobicity of amino acid compositions of mitochondria1 proteins. *Proteins: Struct. Funct.* **178**, 173–178.
63. Cedano, J., Aloy, P., Perez-Pons, J.A., Querol, E., (1997). Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.* **266**, 594–600.

64. Killian, J.A., Von Heijne, G., (2000). How proteins adapt to a membrane-water interface. *Trends Biochem. Sci.* **25**, 429–434.
65. Guy, H.R., (1985). Amino acid side-chain partition energies and distribution of residues in soluble proteins. *Biophys. J.* **47**, 61–70.
66. Baker, J.A., Wong, W.-C., Eisenhaber, B., Warwicker, J., Eisenhaber, F., (2017). Charged residues next to transmembrane regions revisited: 'Positive-inside rule' is complemented by the 'negative inside depletion/outside enrichment rule'. *BMC Biol.* **15**, 1–29.
67. Aurora, R., Rose, G.D., (1998). Helix capping. *Protein Sci.* **240**, 21–38.
68. von Heijne, G., Gavel, Y., (1988). Topogenic signals in integral membrane proteins. *Eur. J. Biochem.* **174**, 671–678.
69. Punta, M., Maritan, A., (2003). A knowledge-based scale for amino acid membrane propensity. *Proteins: Struct. Funct.* **50**, 114–121.
70. de Brevern, A.G., (2022). A perspective on the (rise and fall of) protein β-turns. *Int. J. Mol. Sci.* **23**, 12–16.
71. Fasman, G.D., (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.*, 455–468.
72. Chou, P.Y., Fasman, G.D., (1978). Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **47**, 251–276.
73. Richardson, J.S., (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34**, 167–339.
74. Richardson, J.S., Richardson, D.C., (1988). Amino acid preferences for specific locations at the ends of α-helices. *Science* **240**, 1648–1652.
75. Qian, N., Sejnowski, T.J., (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865–884.
76. Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V., (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99.
77. Kabsch, W., Sander, C., (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-nonded and heometrical features. *Biopolymers* **22**, 2577–2637.
78. Fitzkee, N.C., Rose, G.D., (2004). Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci.* **101**, 12497–12502.
79. Nguyen, H.D., Marchut, A.J., Hall, C.K., (2004). Solvent effects on the conformational transition of a model polyalanine peptide. *Protein Sci.* **13**, 2909–2924.
80. Grigsby, J.J., Blanch, H.W., Prausnitz, J.M., (2002). Effect of secondary structure on the potential of mean force for poly-L-lysine in the α-helix and β-sheet conformations. *Biophys. Chem.* **99**, 107–116.
81. Cerpa, R., Cohen, F.E., Kuntz, I.D., (1996). Conformational switching in designed peptides: The helix/sheet transition. *Fold. Des.* **1**, 91–101.
82. Richardson, J.S., Richardson, D.C., (2002). Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl. Acad. Sci.* **99**, 2754–2759.
83. Imai, K., Mitaku, S., (2005). Mechanisms of secondary structure breakers in soluble proteins. *Biophysics (Oxf)* **1**, 55–65.
84. Narwani, T.J. et al, (2018). Dynamics and deformability of α-, 310- and π-helices. *Arch. Biol. Sci.* **70**, 21–31.
85. Fodje, M.N., Al-Karadaghi, S., (2002). Occurrence, conformational features and amino acid propensities for the π-helix. *Protein Eng.* **15**, 353–358.
86. Weaver, T.M., (2000). The π-helix translates structure into function. *Protein Sci.* **9**, 201–206.
87. Cooley, R.B., Arp, D.J., Karplus, P.A., (2010). Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices enhancing protein functionality. *J. Mol. Biol.* **404**, 232–246.
88. Palau, J., Puigdoménech, P., (1974). The structural code for proteins: Zonal distribution of amino acid residues and stabilization of helices by hydrophobic triplets. *J. Mol. Biol.* **88**, 457–469.
89. Shi, Z., Olson, C.A., Bell, A.J., Kallenbach, N.R., (2001). Stabilization of α-helix structure by polar side-chain interactions: Complex salt bridges, cation-π interactions, and C-H···O H-bonds. *Biopolymers* **60**, 366–380.
90. Butterfield, S.M., Patel, P.R., Waters, M.L., (2002). Contribution of aromatic interactions to α-helix stability. *J. Am. Chem. Soc.* **124**, 9751–9755.
91. Finkelstein, A.V., Badretdinov, A.Y., Ptitsyn, O.B., (1991). Physical reasons for secondary structure stability: α-Helices in short peptides. *Proteins: Struct. Funct.* **10**, 287–299.
92. Geisow, M.J., Roberts, R.D.B., (1980). Amino acid preferences for secondary structure vary with protein class. *Int. J. Biol. Macromol.* **2**, 387–389.
93. Novotny, M., Kleywegt, G.J., (2005). A survey of left-handed helices in protein structures. *J. Mol. Biol.* **347**, 231–241.
94. Tanaka, S., Scheraga, H.A., (1977). Statistical mechanical treatment of protein conformation. 5. A multistate model for specific sequence copolymers of amino acids. *Macromolecules* **10**, 9–20.
95. Chen, S.-Y., Feilen, L.P., Chávez-Gutiérrez, L., Steiner, H., Zacharias, M., (2023). Enzyme-substrate hybrid β-sheet controls geometry and water access to the γ-secretase active site. *Commun. Biol.* **6**
96. Remaut, H., Waksman, G., (2006). Protein-protein interaction through β-strand addition. *Trends Biochem. Sci.* **31**, 436–444.
97. Stanfield, R.L., Wilson, I.A., (1995). Protein-peptide interactions. *Curr. Opin. Struct. Biol.* **5**, 103–113.
98. Lifson, S., Sander, C., (1979). Antiparallel and parallel β-strands differ in amino acid residue preferences. *Nature* **282**, 109–111.
99. Nowick, J.S., (2008). Exploring β-sheet structure and interactions with chemical model systems. *Acc. Chem. Res.* **23**, 1–7.
100. FarzadFard, F., Gharaei, N., Pezeshk, H., Marashi, S.-A., (2008). β-Sheet capping: Signals that initiate and terminate β-sheet formation. *J. Struct. Biol.* **161**, 101–110.
101. Porter, L.L., Rose, G.D., (2011). Redrawing the Ramachandran plot after inclusion of hydrogen-bonding constraints. *Proc. Natl. Acad. Sci.* **108**, 109–113.
102. Zhou, A.Q., O'Hern, C.S., Regan, L., (2011). Revisiting the Ramachandran plot from a new angle. *Protein Sci.* **20**, 1166–1171.
103. Caballero, D. et al, (2014). Intrinsic α-helical and β-sheet conformational preferences: A computational case study of Alanine. *Protein Sci.* **23**, 970–980.

104. Burgess, A.W., Ponnuswamy, P.K., Scheraga, H.A., (1974). Analysis of conformations of amino acid residues and prediction of backbone tropography in proteins. *Isr. J. Chem.* **12**, 239–286.

105. Maxfield, F.R., Scheraga, H.A., (1976). Status of empirical methods for the prediction of protein backbone topography. *Biochemistry* **15**, 5138–5153.

106. de Brevern, A.G., (2016). Extension of the classical classification of β-turns. *Sci. Rep.* **6**, 1–15.

107. Robson, B., Suzuki, E., (1976). Conformational properties of amino acid residues in globular proteins. *J. Mol. Biol.* **107**, 327–356.

108. Monné, M., Hermansson, M., Von Heijne, G., (1999). A turn propensity scale for transmembrane helices. *J. Mol. Biol.* **288**, 141–145.

109. Leszczynski, J.F., Rose, G.D., (1986). Loops in globular proteins: A novel category of secondary structure. *Science* **234**, 849–855.

110. Uversky, V.N., Dunker, A.K., (2010). Understanding protein non-folding. *Biochim. Biophyisca Acta* **1804**, 1231–1264.

111. Mészáros, B., Simon, I., Dosztányi, Z., (2011). The expanding view of protein–protein interactions: Complexes involving. *Phys. Biol.* **8**, 1–10.

112. Thornton, J.M., Sibanda, B.L., Edwards, M.S., Barlow, D. J., (1988). Analysis, design, and modiciation of loop regions in proteins. *Bioessays* **8**, 63–69.

113. Ring, C.S., Kneller, D.G., Langridge, R., Cohen, F.E., (1992). Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol.* **224**, 685–699.

114. George, R.A., Heringa, J., (2003). An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng.* **15**, 871–879.

115. Charton, M., Charton, B.I., (1983). The dependence of the Chou-Fasman parameters on amino acid side chain structure. *J. Theor. Biol.* **102**, 121–134.

116. Klein, P., Kanehisa, M., DeLisi, C., (1984). Prediction of protein function from sequence properties. Discriminant analysis of a data base. *Biochim. Biophys. Acta* **787**, 221–226.

117. Fauchère, J.-L., Charton, M., Kier, L.B., Verloop, A., Pliska, V., (1988). Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **32**, 269–278.

118. Ramanadham, M., Jakkal, V.S., Chidambaram, R., (1993). Carboxyl group hydrogen bonding in X-ray protein structures analysed using neutron studies on amino acids. *FEBS Lett.* **323**, 203–206.

119. Doig, A.J., Sternberg, M.J.E., (1995). Side-chain conformational entropy in protein folding. *Protein Sci.* **4**, 2247–2251.

120. Hu, X., Kuhlman, B., (2006). Protein design simulations suggest that side-chain conformational entropy is not a strong determinant of amino acid environmental preferences. *Proteins: Struct. Funct.* **62**, 739–748.

121. Tzeng, S.-R., Kalodimos, C.G., (2012). Protein activity regulation by conformational entropy. *Nature* **488**, 236–240.

122. Hutchers, J.O., (1970). Handbook of Biochemistry B60–B61. Chemical Rubber Co., Cleaveland, Ohio.

123. Yutani, K., Ogasahara, K., Tsujita, T., Sugino, Y., (1987). Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase a subunit. *Proc. Natl. Acad. Sci.* **84**, 4441–4444.

124. Radzicka, A., Wolfenden, R., (1988). Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1 - octano1, and neutral aqueous solution. *Biochemistry* **27**, 1664–1670.

125. Muñoz, V., Serrano, L., (1994). Elucidating the folding problem of helical peptides using empirical paramters. *Nature Struct. Mol. Biol.* **1**, 399–409.

126. Muñoz, V., Serrano, L., (1994). Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. *Proteins: Struct. Funct.* **20**, 301–311.

127. Laurence, C., Brameld, K.A., Graton, J., Le Questel, J.-Y., Renault, E., (2009). The pKBHX database: Toward a better understanding of hydrogen-bond basicity for medicinal chemists. *J. Med. Chem.* **52**, 4073–4086.

128. Chowdhury, N., Bagchi, A., (2015). An overview of DNA-protein interactions. *Curr. Chem. Biol.* **9**, 73–83.

129. Cosic, I., (1994). Macromolecular bioactivity: Is it resonant interaction between macromolecules?—theory and applications. *IEEE Trans. Biomed. Eng.* **41**, 1101–1114.

130. Oobatake, M., Ooi, T., (1977). An analysis of non-bonded energy of proteins. *J. Theor. Biol.* **67**, 567–584.

131. Kyte, J., Doolittle, R.F., (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.

132. Eisenberg, D., McLachlan, A.D., (1986). Solvation energy in protein folding and stability. *Nature* **319**, 199–203.

133. Ponnuswamy, P.K., Prabhakaran, M., Manavalan, P., (1980). Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim. Biophyisca Acta* **623**, 301–316.

134. White, S.H., Wimley, W.C., (1998). Hydrophobic interactions of peptides with membrane interfaces. *Biochim. Biophys. Acta* **1376**, 339–352.

135. Mitaku, S., Hirokawa, T., Tsuji, T., (2002). Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics* **18**, 608–616.

136. Giménez-Andrés, M., Čopič, A., Antonny, B., (2018). The many faces of amphipathic helices. *Biomolecules* **8**, 1–14.

137. Segrest, J.P., De Loof, H., Dohlman, J.G., Brouillette, C. G., Anantharamaiah, G.M., (1990). Amphipathic helix motif: Classes and properties. *Proteins: Struct. Funct.* **8**, 103–117.

138. Argos, P., Rao, J.K.M., Hargrave, P.A., (1982). Structural prediction of membrane-bound proteins. *Eur. J. Biochem.* **128**, 565–575.

139. Drin, G., Antonny, B., (2010). Amphipathic helices and membrane curvature. *FEBS Lett.* **584**, 1840–1847.

140. Cornette, J.L. et al, (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **195**, 659–685.

141. Kakraba, S., Knisley, D., (2016). A graph-theoretic model of single point mutations in the cystic fibrosis transmembrane conductance regulator. *J. Adv. Biotechnol.* **6**, 780–786.

142. Rackovsky, S., Scheraga, H.A., (1977). Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins. *Proc. Natl. Acad. Sci.* **74**, 5248–5251.

143. Rhodes, G., (1993). Other diffraction methods. *Crystallogr. Made Cryst. Clear*.
144. Prabhakaran, M., Ponnuswamy, P.K., (1982). Shape and surface features of globular proteins. *Macromolecules* **15**, 314–320.
145. Levitt, M., (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.
146. Mandell, D.J., Kortemme, T., (2009). Backbone flexibility in computational protein design. *Curr. Opin. Biotechnol.* **20**, 420–428.
147. Halle, B., (2002). Flexibility and packing in proteins. *Proc. Natl. Acad. Sci.* **99**, 1274–1279.
148. Radivojac, P. et al, (2004). Protein flexibility and intrinsic disorder. *Protein Sci.* **13**, 71–80.
149. Zavodszky, M.I., Leslie, A.K., (2005). Side-chain flexibility in protein-ligand binding: The minimal rotation hypothesis. *Protein Sci.* **14**, 1104–1114.
150. Karplus, P.A., Schulz, G.E., (1985). Prediction of chain flexibility in proteins. *Naturwissenschaften* **72**, 212–213.
151. Krigbaum, W.R., Komoriya, A., (1979). Local interactions as a structure determinat for protein molecules: II. *Biochim. Biophys. Acta* **576**, 204–228.
152. Vihinen, M., Torkkila, E., Riikonen, P., (1994). Accuracy of protein flexibility predictions. *Proteins: Struct. Funct.* **19**, 141–149.
153. Parthasarathy, S., Murthy, M.R.N., (2000). Protein thermal stability: Insights from atomic displacement parameters (B values). *Protein Eng.* **13**, 9–13.
154. Ptitsyn, O.B., Finkelstein, A.V., (1983). Theory of protein secondary structure and algorithm of its prediction. *Biopolymers* **22**, 15–25.
155. Zhou, H., Zhou, Y., (2004). Quantifying the effect of burial of amino acid residues on protein stability. *Proteins: Struct. Funct.* **54**, 315–322.
156. Sueki, M. et al, (1984). Helix-coil stability constants for the naturally occurring amino acids in water. 22. histidine parameters from random poly[(hydroxybutyl)glutamine-co-L-histidine]. *Macromolecules* **17**, 148–155.
157. Berjanskii, M.V., Wishart, D.S., (2005). A simple method to predict protein flexibility using secondary chemical shifts. *J. Am. Chem. Soc.* **127**, 14970–14971.
158. Bundi, A., Wüthrich, K., (1979). 1H-nmr parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH. *Biopolymers* **18**, 285–297.
159. Andersen, N.H., Cao, B., Chen, C., (1992). Peptide/protein structure analysis using the chemical shift index method: Upfield α-CH values reveal dynamic helices and αL sites. *Biochem. Biophys. Res. Commun.* **184**, 1008–1014.
160. Sneath, P.H.A., (1966). Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* **12**, 157–195.
161. Wold, S., Esbensen, K., Geladi, P., (1987). Principal component analysis. *Chemom. Intell. Lab. Syst.* **2**, 37–52.
162. Takahashi, S., Ihara, S., Ooi, T., (1978). C-terminal side of α-helix is more stable than N-terminal side. *Nature* **276**, 735–736.
163. Santiveri, C.M., Santoro, J., Rico, M., Jiménez, M.A., (2004). Factors involved in the stability of isolated β-sheets: Turn sequence, β-sheet twisting, and hydrophobic surface burial. *Protein Sci.* **13**, 1134–1147.
164. Yang, W.Y., Larios, E., Gruebele, M., (2003). On the extended β-conformation propensity of polypeptides at high temperature. *J. Am. Chem. Soc.* **125**, 16220–16227.
165. Boonyaratanakornkit, B.B., Park, C.B., Clark, D.S., (2002). Pressure effects on intra- and intermolecular interactions within proteins. *Biochim. Biophys. Acta* **1595**, 235–249.
166. Nelson, D.L., Cox, M., (2017). Lehninger Principles of Biochemistry. WH Freeman.
167. Raimondi, D., Orlando, G., Vranken, W.F., Moreau, Y., (2019). Exploring the limitations of biophysical propensity scales coupled with machine learning for protein sequence analysis. *Sci. Rep.* **9**, 1–11.
168. Yang, K.K., Wu, Z., Bedbrook, C.N., Arnold, F.H., (2018). Learned protein embeddings for machine learning. *Bioinformatics* **34**, 2642–2648.
169. Cui, F., Zhang, Z., Zou, Q., (2021). Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. *Brief. Funct. Genomics* **20**, 61–73.
170. Villegas-Morcillo, A. et al, (2021). Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. *Bioinformatics* **37**, 162–170.
171. Bernhofer, M. et al, (2021). PredictProtein – predicting protein structure and function for 29 years. *Nucleic Acids Res.* **49**, W535–W540.
172. Higgins, D.G., Sharp, P.M., (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244.
173. Zvelebil, M.J., Barton, G.J., Taylor, W.R., Sternberg, M.J., (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.*, 957–961.
174. Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., Thornton, J.M., (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326**, 347–1252.
175. LeCun, Y. et al, (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551.
176. LeCun, Y., Bengio, Y., Hinton, G., (2015). Deep learning. *Nature* **521**, 436–444.
177. Hochreiter, S., Schmidhuber, J., (1997). Long short-term memory. *Neural Comput.* **9**, 1–32.
178. Vaswani, A. et al, (2017). Attention is all you need. *Neural Inf. Process. Syst.* **31**, 1–15.
179. Jumper, J. et al, (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.
180. Elnaggar, A. et al, (2021). ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127.
181. Weißenow, K., Heinzinger, M., Rost, B., (2022). Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure* **30**, 1169–1177.e4.
182. Heinzinger, M. et al, (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinf.* **20**, 1–17.
183. Iman, M., Arabnia, H.R., Rasheed, K., (2023). A review of deep transfer learning and recent advancements. *Technologies* **11**, 1–14.

184. Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., Church, G.M., (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* **16**, 1315–1322.

185. Taylor, W.R., Chelliah, V., Hollup, S.M., MacDonald, J.T., Jonassen, I., (2009). Probing the 'dark matter' of protein fold space. *Structure* **17**, 1244–1252.

186. Perdigão, N. et al, (2015). Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci.* **112**, 15898–15903.

187. Pechlaner, M., Oostenbrink, C., van Gunsteren, W.F., (2021). On the use of multiple-time-step algorithms to save computing effort in molecular dynamics simulations of proteins. *J. Comput. Chem.* **42**, 1263–1282.

188. Audagnotto, M., Dal Peraro, M., (2017). Protein post-translational modifications: In silico prediction tools and molecular modeling. *Comput. Struct. Biotechnol. J.* **15**, 307–319.

189. Fairlie, D.P. et al, (2000). Conformational selection of inhibitors and substrates by proteolytic enzymes: Implications for drug design and polypeptide processing. *J. Med. Chem.* **43**, 1271–1281.

190. Madala, P.K., Tyndall, J.D.A., Nall, T., Fairlie, D.P., (2010). Update 1 of: Proteases universally recognize β strands in their active sites. *Chem. Rev.* **110**, PR1–PR31.

191. Zhou, R. et al, (2019). Recognition of the amyloid precursor protein by human γ-secretase. *Science* **363**, 708–716.

192. Yang, G. et al, (2019). Structural basis of Notch recognition by human γ-secretase. *Nature* **565**, 192–197.

193. Jespersen, M.C., Mahajan, S., Peters, B., Nielsen, M., Marcatili, P., (2019). Antibody specific B-cell epitope predictions: Leveraging information from antibody-antigen protein complexes. *Front. Immunol.* **10**, 1–17.

194. Ruffolo, J.A., Sulam, J., Gray, J.J., (2022). Antibody structure prediction using interpretable deep learning. *Patterns* **3**, 1–12.

195. Kosugi, T., Ohue, M., (2021). Quantitative estimate index for early-stage screening of compounds targeting protein-protein interactions. *Int. J. Mol. Sci.* **22**, 1–15.

196. Sudha, G., Nussinov, R., Srinivasan, N., (2014). An overview of recent advances in structural bioinformatics of protein-protein interactions and a guide to their principles. *Prog. Biophys. Mol. Biol.* **116**, 141–150.

197. Dyson, H.J., Wright, P.E., (2002). Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **12**, 54–60.

198. Bloemeke, N. et al, (2022). Intramembrane client recognition potentiates the chaperone functions of calnexin. *EMBO J.* **41**, 1–21.

199. Chiu, T.-P., Rao, S., Rohs, R., (2022). Physicochemical models of protein–DNA binding with standard and modified base pairs. *Proc. Natl. Acad. Sci.* **120**, 1–9.

200. Fischer, G., Rossmann, M., Hyvönen, M., (2015). Alternative modulation of protein-protein interactions by small molecules. *Curr. Opin. Biotechnol.* **35**, 78–85.

201. Qing, R. et al, (2022). Protein design: From the aspect of water solubility and stability. *Chem. Rev.* **122**, 14085–14179.

202. de Bree, F.M., (2000). Trafficking of the vasopressin and oxytocin prohormone through the regulated secretory pathway. *J. Neuroendocrinol.* **12**, 589–594.

203. Liu, X. et al, (2020). Phase separation drives decision making in cell division. *J. Biol. Chem.* **295**, 13419–13431.

204. Fang, Y., Gao, S., Tai, D., Middaugh, C.R., Fang, J., (2013). Identification of properties important to protein aggregation using feature selection. *BMC Bioinf.* **14**, 1–9.

205. Bouziane, H., Chouarfia, A., (2020). Sequence- and structure-based prediction of amyloidogenic regions in proteins. *Soft Comput.* **24**, 3285–3308.

206. Guggino, W.B., Stanton, B.A., (2006). New insights into cystic fibrosis: Molecular switches that regulate CFTR. *Nature Rev. Mol. Cell Biol.* **7**, 426–436.

207. Menezes, L.F.S., Sabiá Júnior, E.F., Tibery, D.V., dos Carneiro, L. dos A., Schwartz, E.F., (2020). Epilepsy-related voltage-gated sodium channelopathies: A review. *Front. Pharmacol.* **11**, 1–32.

208. Moore, J.R., Leinwand, L., Warshaw, D.M., (2012). Understanding cardiomyopathy phenotypes based on the functional impact of mutations in the myosin motor. *Circ. Res.* **111**, 375–385.

209. Cozzetto, D., Minneci, F., Currant, H., Jones, D.T., (2016). FFPred 3: Feature-based function prediction for all Gene Ontology domains. *Sci. Rep.* **6**, 1–11.

210. Pazos, F., (2021). Prediction of protein sites and physicochemical properties related to functional specificity. *Bioengineering* **8**, 1–10.

211. Vu, T.T.D., Jung, J., (2021). Protein function prediction with Gene Ontology: From traditional to deep learning models. *PeerJ* **9**, 1–24.

212. Pinazo, A., Pons, R., Pérez, L., Infante, M.R., (2011). Amino acids as raw material for biocompatible surfactants. *Ind. Eng. Chem. Res.* **50**, 4805–4817.

213. Tripathy, D.B., Mishra, A., Clark, J., Farmer, T., (2017). Synthesis, chemistry, physicochemical properties and industrial applications of amino acid surfactants: A review. *Comptes Rendus Chim.* **21**, 112–130.

214. Feehan, R., Montezano, D., Slusky, J.S.G., (2021). Machine learning for enzyme engineering, selection and design. *Protein Eng.* **34**, 1–10.

215. Chiangjong, W., Chutipongtanate, S., Hongeng, S., (2020). Anticancer peptide: Physicochemical property, functional aspect and trend in clinical application (review). *Int. J. Oncol.* **57**, 678–696.

216. Fosgerau, K., Hoffmann, T., (2015). Peptide therapeutics: Current status and future directions. *Drug Discov. Today* **20**, 122–128.

217. Shin, W.-H., Kumazawa, K., Imai, K., Hirokawa, T., Kihara, D., (2020). Current challenges and opportunities in designing protein–protein interaction targeted drugs. *Adv. Appl. Bioinforma. Chem.* **13**, 11–25.

218. Caradonna, T.M., Schmidt, A.G., (2021). Protein engineering strategies for rational immunogen design. *npj Vaccines* **6**, 1–11.

219. Lynn, G.M. et al, (2015). In vivo characterization of the physicochemical properties of polymer-linked TLR agonists that enhance vaccine immunogenicity. *Nature Biotechnol.* **33**, 1201–1210.

220. Beck, A., Goetsch, L., Dumontet, C., Corvaïa, N., (2017). Strategies and challenges for the next generation of antibody-drug conjugates. *Nature Rev. Drug Discov.* **16**, 315–337.

221. Buecheler, J.W., Winzer, M., Weber, C., Gieseler, H., (2020). Alteration of physicochemical properties for antibody-drug conjugates and their impact on stability. *J. Pharm. Sci.* **109**, 161–168.

222. Leung, D. et al, (2020). Antibody conjugates-recent advances and future innovations. *Antibodies* **9**, 1–27.
223. Gao, B., Han, J., Reddy, S.T., (2022). Learning what not to select for in antibody drug discovery. *Cell Rep. Methods* **2**, 1–3.
224. Savojardo, C., Manfredi, M., Martelli, P.L., Casadio, R., (2021). Solvent accessibility of residues undergoing pathogenic variations in humans: from protein structures to protein sequences. *Front. Mol. Biosci.* **7**, 1–9.
225. Quint, S. et al, (2010). Residue-specific side-chain packing determines the backbone dynamics of transmembrane model helices. *Biophys. J.* **99**, 2541–2549.
226. Lan, M., Tan, C.L., Su, J., (2009). Feature generation and representations for protein-protein interaction classification. *J. Biomed. Inform.* **42**, 866–872.
227. Garla, V.N., Brandt, C., (2012). Ontology-guided feature engineering for clinical text classification. *J. Biomed. Inform.* **45**, 992–998.
228. Ronan, T., Qi, Z., Naegle, K.M., (2016). Avoiding common pitfalls when clustering biological data. *Sci. Signal.* **9**, 1–13.
229. Pedregosa, F. et al, (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2826–2830.
230. Schwarz, G., (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464.