Amino acid characteristics in protein native state structures

Tatjana Škrbić[1,2], Achille Giacometti[1,3], Trinh X. Hoang[4], Amos Maritan[5], and Jayanth R. Banavar[2]

[1] Ca' Foscari University of Venice, Department of Molecular Sciences and Nanosystems, Venice, Italy

[2] University of Oregon, Department of Physics and Institute for Fundamental Science, Eugene, Oregon, USA

[3] European Centre for Living Technology (ECLT), Ca' Bottacin, Dorsoduro 3911, Calle Crosera, 30123 Venice, Italy

[4] Vietnam Academy of Science and Technology, Institute of Physics, Hanoi, Vietnam

[5] University of Padua, Department of Physics and Astronomy, Padua, Italy

Correspondence

Tatjana Škrbić

Ca' Foscari University of Venice, Department of Molecular Sciences and Nanosystems, Via Torino 155, 30170 Venice, Italy.

Email: tatjana.skrbic@unive.it

Abstract

We present a geometrical analysis of the protrusion statistics of side chains in more than 4,000 high-resolution protein structures. We find that when the amino acids are grouped based on chemical similarity, the resulting groups equally well fit the geometrical attributes in a harmonious manner. Our analysis underscores the role of minimal frustration in the evolution of protein sequences that fit within a pre-sculpted menu of putative native state folds.

Keywords: local Frenet frame, amino-acid classes, side-chain protrusion, pre-sculpted landscape

I. INTRODUCTION

Proteins, the molecular machines of life, are relatively short linear chains of amino acids with a common backbone. There are twenty types of naturally occurring amino-acids, each possessing a distinct side chain attached to the main chain protein backbone [1-4]. The complexity of the protein problem stems from the myriad of degrees of freedom. A protein is surrounded by

water molecules within the cell. Each of the twenty side chains has its own chemical properties and geometry. Despite the complexity, small globular proteins share a great deal of properties because of their common backbone. They fold rapidly and reproducibly into their respective unique native state structures [5]. Protein native state structures are modular and comprise secondary structure building blocks: topologically one-dimensional α-helices and almost planar parallel and antiparallel β-sheets. Hydrogen bonds provide support to the building blocks [6-7]. Because of the modularity, the total number of distinct native folds is of the order of just several thousand [8-11]. Furthermore, the native state folds are evolutionarily conserved [12-13] and are immutable (Platonic). This surprising simplicity present in the complex protein problem can be rationalized through the notion of a free energy landscape of proteins pre-sculpted by symmetry and geometry [14-16]. Moreover, it has been recently shown that both the secondary and the tertiary protein structure can be derived without any empirical or chemical input by building on the common protein backbone and the notion of poking interactions [17-21] arising from the affinity of one part of a chain with another. It follows that the textbook knowledge that ‘sequence determines

structure' [4-5] ought to be replaced by a simpler scenario of 'sequence selects structure' from the set of the Platonic folds [14-16].

The side chains play a critical role in the selection process in two crucial ways. First, the chemistry of the interacting side chains [22-25] must be harmonious [26-28], maximizing favorable interactions (including water-mediated hydrophobic, van der Waals, electrostatic, and hydrogen bonding interactions). The net result is to create a protein hydrophobic core shielded from the surrounding water molecules thereby ensuring the stability and compactness of protein native structure. Second, the side chains must fill the space in the interior of the protein, packing tightly against each other, maximizing favorable self-interactions in the hydrophobic interior and minimizing empty space [29] (see Figure 1).

The specific arrangement of side chains in the protein interior is determined by two factors. The first is the primary protein sequence of amino acids that can roughly be classified as being hydrophobic (non-polar residues mainly buried in the protein interior and forming its hydrophobic core), hydrophilic (polar residues that readily interact with water molecules and tend to be positioned at the protein surface), or neutral (somewhere between

the two categories) [32]. The second is the overall folded geometry selected from the menu of available folds to provide an optimal, best possible fit to the sequence. The orientation of the side chain is flexible and the set of specific conformations and/or orientations constitute the so-called side chain rotamers [33-37].

Here we adopt a simplified coarse-grained description. We view a protein as a chain of C_α atoms. Our approach then consists in the determining the locations and orientations of the protruding side chain atoms. Because of the imperative need to fill space in the interior while assiduously avoiding steric clashes, our focus is on the heavy atom protruding furthest from the corresponding C_α atom. We seek to characterize the geometry of this protrusion in a coordinate frame relative to the portion of protein backbone corresponding to the given amino acid, which enables us to determine both the average side chain behavior as well as a specific behavior of different amino acid classes. Our longer-term goal is to set the stage for the more important step of understanding the role of side chains in tertiary structure

assembly. This will enable us to understand how a particular sequence chooses a specific fold geometry for its native state.

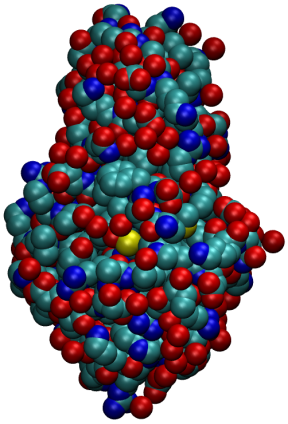

Figure 1: a) Native state of bacteriophage T4 lysozyme (PDB code: 2LZM) in the CPK representation [30-31] in which all heavy atoms of the protein backbone and its side chains are represented as spheres with radii proportional to their respective van der Waals atomic radii. Color code: carbon (cyan), oxygen (red), nitrogen (blue), and sulfur (yellow). The side chains in the protein interior are very well packed.

II. MATERIALS AND METHODS

A. Local Frenet coordinate system of an amino acid

We view a protein backbone as a chain of discrete points on which the consecutive C_α atoms are located. For the side chains, we account for all heavy atoms (excluding only the hydrogen atoms) to determine the maximally protruding atom from the protein backbone. This is the farthest heavy atom from the corresponding C_α atom and at a distance that we call R_{max}. To characterize the orientation of this maximally protruding sidechain atom, we employ a Frenet coordinate system [38] local to the portion of the backbone to which the side chain belongs. For the i-th amino acid in question, the origin of its local Frenet frame is located at the i-th C_α atom. The orthonormal set of axes are the tangent **t**, anti-normal **an**=**-n**, and binormal **b**. These basis vectors are defined from the positions of three consecutive C_α atoms associated with residues i-1,i, and i+1, as shown in Figure 2.

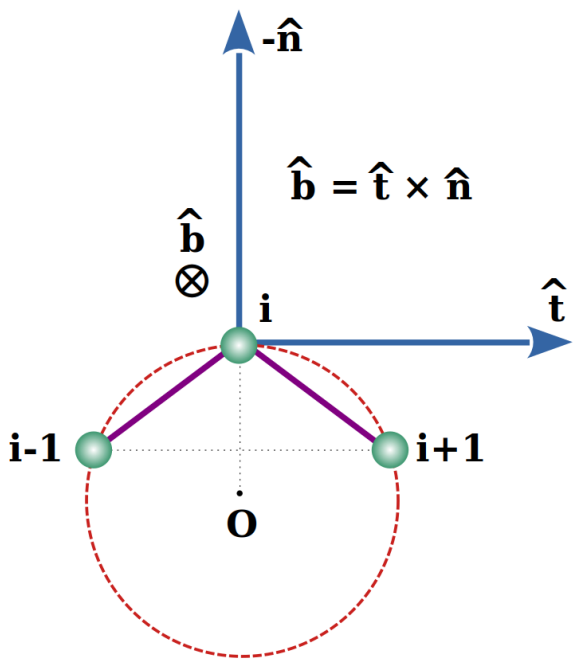

Figure 2: Local Frenet frame of the amino acid i. The three consecutive C_α atoms are at points i-1, i, and i+1 and lie in the plane of the paper. The point O is at the center of a circle passing through them. Please see text for a description of the orthonormal basis set.

99.7% of the bond lengths in proteins are, to a very good approximation, equal to 3.81Å [39], corresponding to the prevalent *trans* isomeric conformation of a peptide backbone group, where the two neighboring C_α atoms are on opposite sides of the peptide bond with the third Ramachandran angle of ω close to 180°. However, the remaining ~0.3% of protein bonds are

shorter, having a length around ~2.95Å [39] and correspond to the so-called *cis*-conformation [40]. The third Ramachandran angle ω ~ 0° when the two consecutive C_α atoms are placed on the same side of the connecting peptide bond. We define a local Frenet frame of a given amino acid in a manner that is robust to variations in the bond lengths. First, we draw a circle passing through points i-1, i, and i+1 and determine its center and the radius. The direction of the anti-normal (negative normal direction) **an**=**-n** is along the straight line joining the center of the circle to the C_α atom. The tangent vector **t** points along the direction (i-1, i+1), and finally the binormal vector **b** is found as a cross product of the unit vectors **t** x **n** (see Figure 2).

We then proceed to a characterization of the orientation of the maximal protruding heavy atom of the side chain from the C_α atom by means of three projections, along the unit vectors **t**, **b**, and **-n**, in the corresponding local Frenet system.

B. Chemistry of amino acids

For describing the chemical properties of the twenty naturally occurring amino acids we adopt a standard amino acid classification that is based on their so-called physico-chemical properties, the IMGT classification [41]. The nine amino acid classes adopted in this classification are presented in Table I.

Aliphatic A	*ALA, ILE, LEU, VAL*	*Hydrophobic, linear and b ranched chains*
Aliphatic B	*PRO (ring connected back to backbone)*	
Aromatic	*PHE, TRP, TYR*	*Hydrophobic, aromatic ring*
Acidic	*ASP, GLU*	*Negatively charged*
Basic	*ARG, LYS, HIS*	*Positively charged*
Hydroxylic	*SER, THR*	*Polar, Uncharged, -OH hydroxyl group*
Sulfur-containing A	*CYS(disulfide bond)*	*Contain sulfur atom*
Sulfur-containing B	*MET*	
Amidic	*ASN, GLN*	*Polar, uncharged, amidic -R group*

Table I: Summary of the side chains' distinct chemistries in the nine amino acid classes in the 'physico-chemical' IMGT classification [41].

C. Curation and data analysis

Our protein data set consists of 4,366 globular protein structures from the PDB, a subset of Richardsons' Top 8000 set [42] of high-resolution, quality-filtered protein chains (resolution < 2Å, 70% PDB homology level), that we

further distilled out to exclude structures with missing backbone and sidechain atoms, as well as amyloid-like structures. The program DSSP (CMBI version 2.0) [43] has been used to determine for each protein residue its location in an α-helix, in a β-strand or elsewhere.

Our data set comprises of a total of 940,812 residues (864,815 non-glycine and 75,997 glycine amino acids) in the native state structures of more than 4000 proteins. Their abundances and relative frequencies, in order of decreasing prevalence, in our data set, are shown in Table II.

Amino acid type	**Number**	**Frequency [%]**
LEU	83,875	8.91
ALA	82,068	8.72
GLY	75,997	8.08
VAL	68,515	7.28
GLU	59,537	6.33
ASP	56,080	5.96
ILE	53,051	5.64
SER	52,989	5.63
THR	52,789	5.61
LYS	52,101	5.54
ARG	44,589	4.74
PRO	44,119	4.69
ASN	41,247	4.38
PHE	38,689	4.11
TYR	34,457	3.66

GLN	33,185	3.53
HIS	22,065	2.35
MET	18,338	1.95
TRP	14,504	1.54
CYS	12,616	1.34

Table II: Number and relative frequency of twenty amino acid types in our data set comprised of over 4,000 protein native state structures.

III. RESULTS AND DISCUSSION

A. The average orientation of amino acids in globular proteins

To determine the direction of the overall dominant amino acid protrusion, we simply sum up all the unit vectors (directors), which connect the C_α atoms of a given residue with its maximally protruding atom. We then work in a plane perpendicular to the dominant direction to determine the second dominant direction and the third dominant direction forming a rotated orthonormal set of the Frenet frame (see Figure 3). The direction of the dominant axis lies predominantly in the (anti-normal-binormal) plane with a relatively small component in the tangent direction.

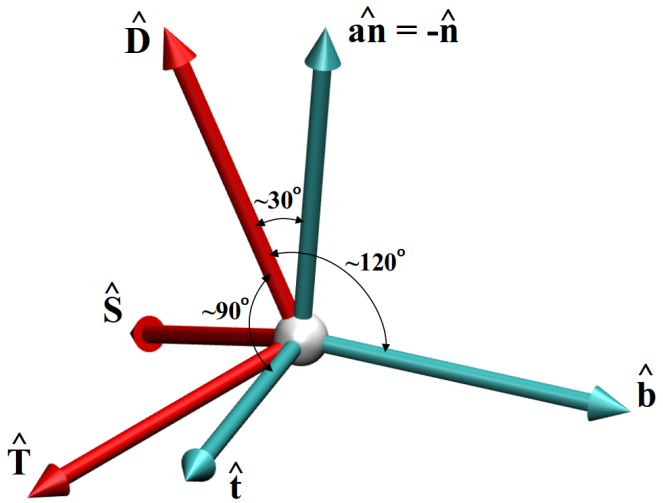

Figure 3: The figure shows the relative orientation of two orthonormal sets of axes: the Frenet frame and the rotated consensus frame showing the dominant (**D**), the second dominant (**S**), and the third dominant (**T**) axes. The dominant axis forms angles of 29.57°, 92.04°, and 119.49°, with the anti-normal, tangent and binormal vectors, respectively.

Figure 4(a) shows the distributions of the projections of the maximally protruding heavy atom along the three consensus axes. The dominant axis plays the key role. Figure 4(b) shows the frequency distribution along just the dominant axis but for amino acids separated into ‘helical’, ‘strand’, and

'loop' environments and elucidate the origins of the peaks marked 1-4 in both (a) and (b). The peaks numbered 1 and 2 arise from the α-helical backbone and β-strands respectively. Both the peaks numbered 3 and 4 originate from the amino acid PRO. The proline (PRO) residue consists of a five-atom ring connected back covalently to the protein backbone and introduces steric constraints that disrupt protein secondary structure elements and is primarily found in protein loops. PRO has the protrusion maximum of R_{max} = (2.42 +/- 0.07) Å of the order of the radius of the backbone tube Δ=2.63Å and is almost completely buried within the protein backbone tube and do not have steric problems despite its odd protruding direction.

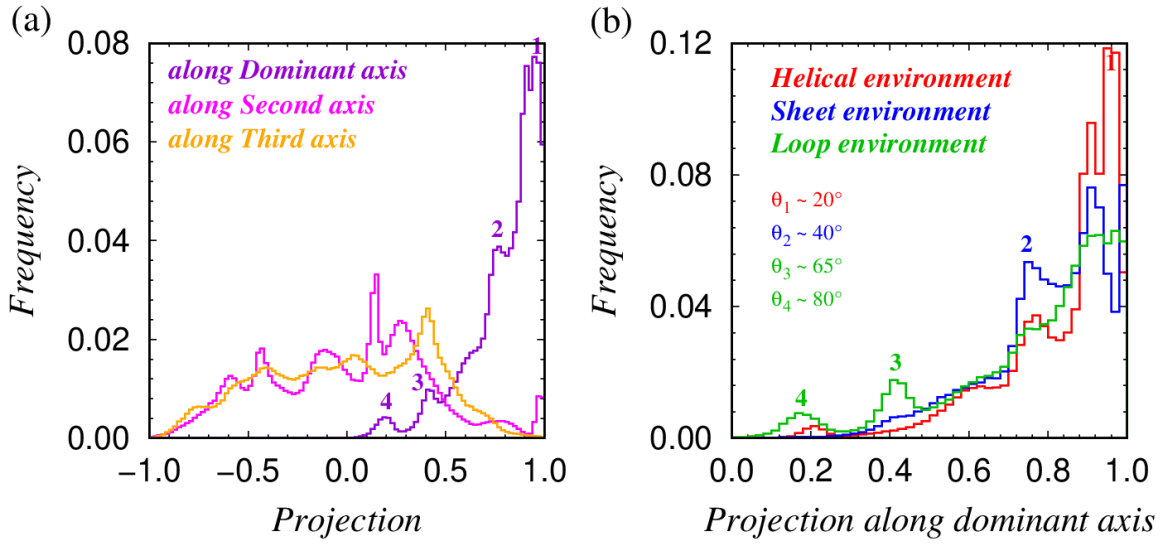

Figure 4: (a) Probability distribution of the projections along the dominant axis, the second axis, and the third axis of ~900,000 non-

glycine residues in more than 4,000 high-resolution structures of globular proteins. (b) Probability distribution of the cosine of the angle of the protrusion direction with the dominant axis for three subsets of all consecutive triplets of C_α atoms, those belonging to α-helical segments (red histogram), to β-strands (blue histogram), and those for which the consecutive triplets of C_α atoms are in protein loops.

B. The protruder atom type and the amount of amino acid protrusion

Figure 5 indicates for each of the nineteen amino acids (except for glycine (GLY) that does not possess any heavy atoms) the atom involved in maximum protrusion along with how often this is the case. We note that in most of the cases there is prevalently only one atom that protrudes the most (~90% of times or more) and this is the case for the thirteen amino acids: ALA, ASN, ASP, CYS, ILE, LYS, MET, SER, THR, TYR, TRP, PHE, and PRO. For the remaining six amino acids: ARG, GLN, GLU, HIS, LEU, and VAL there were two viable candidate atoms. We note that both hydrophilic and hydrophobic residues are present in both these classes showing that this result is largely chemistry independent.

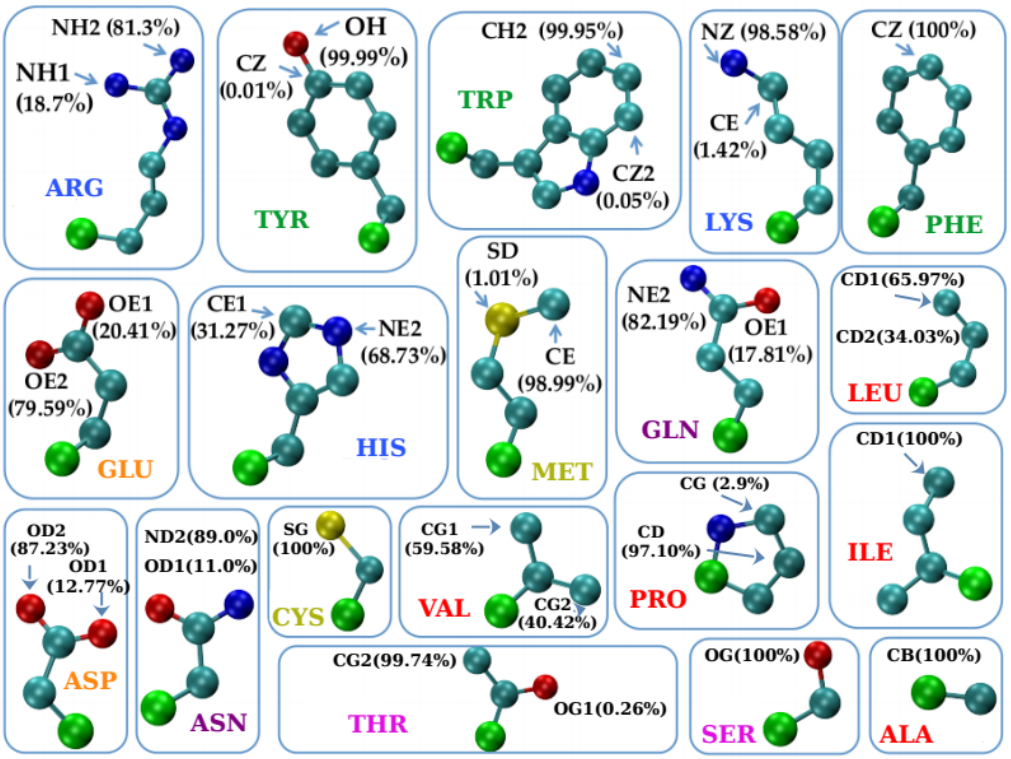

Figure 5: Gallery of nineteen amino acids (with glycine excluded). Three-letter amino acid codes are used. For each amino acid, the maximally protruding atom along with the frequency with which it occurred are shown. The color code of the atoms is: carbon C_α in green, carbon C atoms other than C_α in turquoise, oxygen O atoms in red, nitrogen N atoms in dark blue, and sulfur S atoms in yellow. Carbon C_α atoms (green spheres) are artificially represented as spheres with slightly larger radius than the rest of C atoms (turquoise

spheres) to enhance visibility. The measure of the degree of protrusion of a given side chain atom with respect to the backbone was defined to be the distance of the atom from the corresponding C_α atom.

Figure 6 displays the frequency distribution of the values of the maximal protrusion R_{max} observed in our data set. The observed peaks in this distribution can be readily assigned to specific amino acids because of their non-overlapping mean values and their relatively sharp widths. Additionally, we can conveniently divide the observed range of R_{max} into three distinct classes: 1) small with $R_{max} < 3$Å, comprised of [ALA, CYS, PRO, SER, VAL], 2) medium $R_{max} \sim (3\text{-}5)$Å, composed of [ASN, ASP, GLN, GLU, HIS, ILE, LEU, MET] and 3) large with $R_{max} > 5$Å, containing [ARG, LYS, PHE, TRP, TYR].

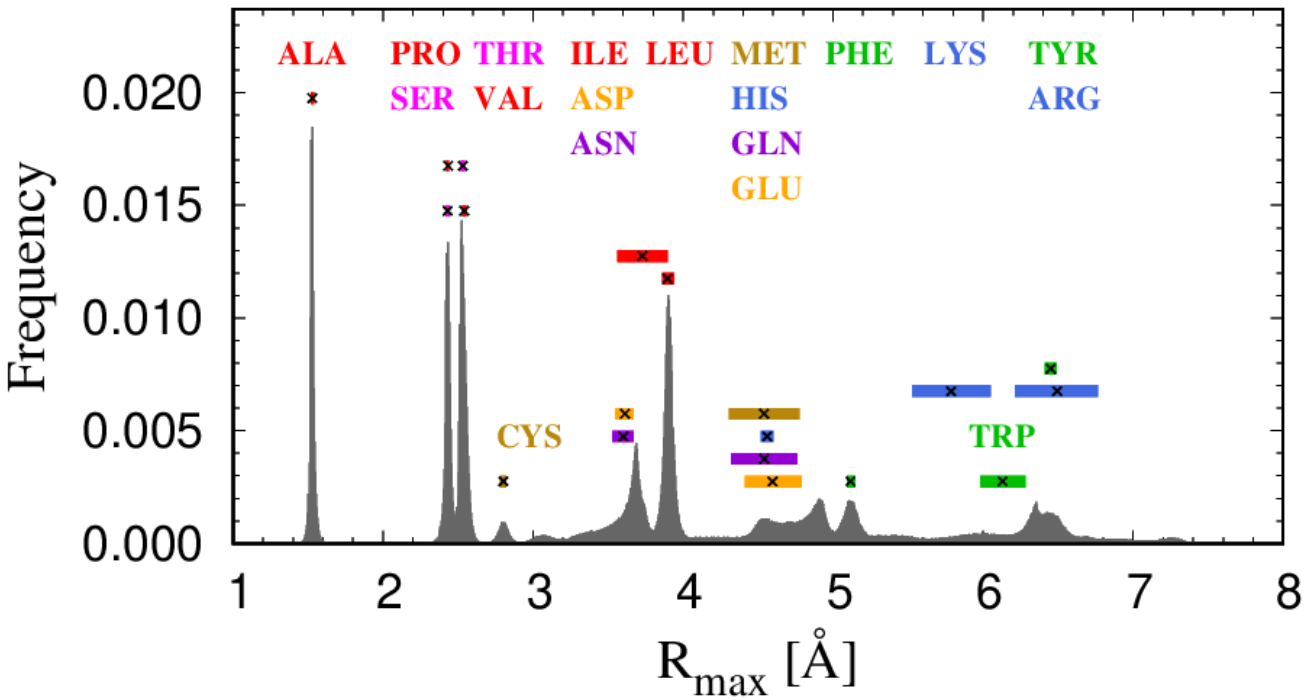

Figure 6: Histogram of the maximal protrusion R_{max} (with respect to the protein backbone) of amino acids in more than 4,000 high-resolution structures of globular proteins. The 19 amino acids (with glycine being excluded, having no side chain atoms) are denoted with three-letter amino acid code and are colored according to the amino acid classification summarized in Table I. The mean values of R_{max} for each of the amino acids are shown as black X symbols, while the colored rectangles have a width that corresponds to the standard deviation.

C. Geometry of amino acid protrusion from the backbone

For each of the nine IMGT amino acid classes [41], we present data below that shows the complete geometrical attributes of their respective protrusion, both in magnitude and orientation. We note the complementarity of the amino acids belonging to Aliphatic A class (Figure 7) and Aromatic class (Figure 9). On one hand, these classes share similar chemistry because they are both hydrophobic. From the geometrical point of view, they are, however, complementary. Aliphatic A group amino acids have 'small-to-medium' protrusion amounts (Figure 7, Panels (h)-(k)) and are very well aligned with the dominant axis, with a projection close to 1 with a small deviation for VAL [that could be related to the fact that is the only amino acid in this group that shows a sizable ambivalence of 60%-40% in the choice of the maximally protruding atom, see Figure 5] (see Panels (d)-(g) of Figure 7). The Aromatic group amino acids belong to the 'large' protrusion class (see Figure 9, Panels (g)-(i)) and have significantly smaller projections along the dominant axis, of the order of ~0.6 that are substantially spread out (see Panels (d)-(f) of Figure 9).

Let us compare the geometries of protrusion of Basic (Figure 10), Acidic (Figure 11), Amidic (Figure 12) and Sulfur-containing B (composed solely of MET, shown in Figure 15) amino acid classes. They have a variety of chemical properties, but their geometries are very similar: they have ‘medium-to-large’ protrusions, with significant values of protrusion components along the dominant axis, larger than ~0.7-0.8.

Finally, the three remaining amino acid classes, Aliphatic B (composed solely of PRO, see Figure 8), Hydroxylic (Figure 13), Sulfur-containing A (composed solely of CYS, see Figure 14), have largely different geometries of protrusion between themselves and with the previously discussed amino acid classes. We note that they significantly deviate from the dominant axis, and PRO is extreme in this regard.

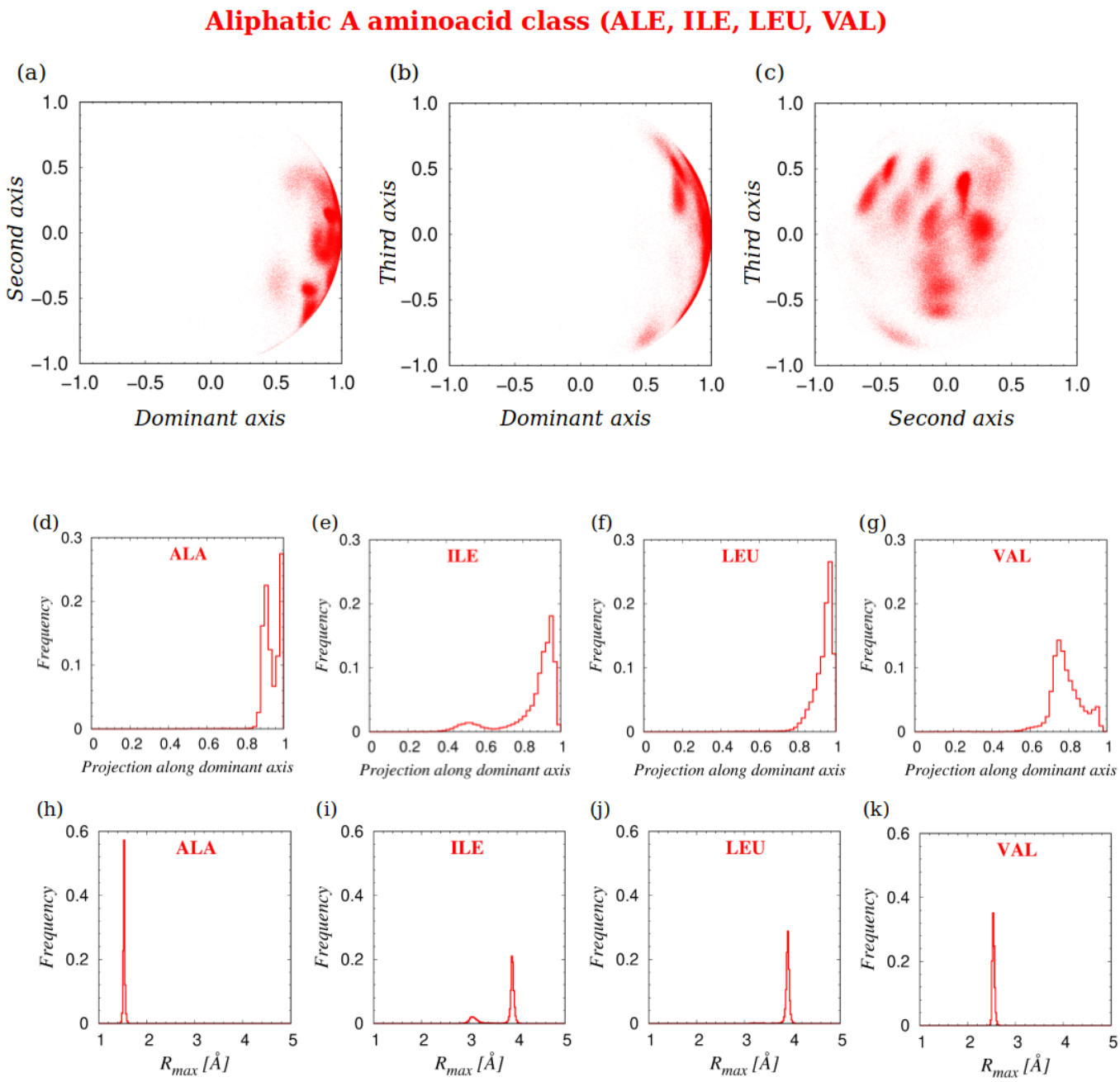

Figure 7: Aliphatic A amino acid class: geometrical attributes of protrusion. Panels (a), (b), and (c) are the three two-dimensional projection cross-plots: (a) in the (dominant axis, second axis) plane,

(b) in the (dominant axis, third axis) plane, and (c) in the (second axis, third axis) plane. Panels (d)-(g) show the frequency distributions of the projections along the dominant axis separately for the participant amino acids. Panels (h)-(k) show the frequency distributions of amount of the maximal protrusion R_{max} separately for the participant amino acids.

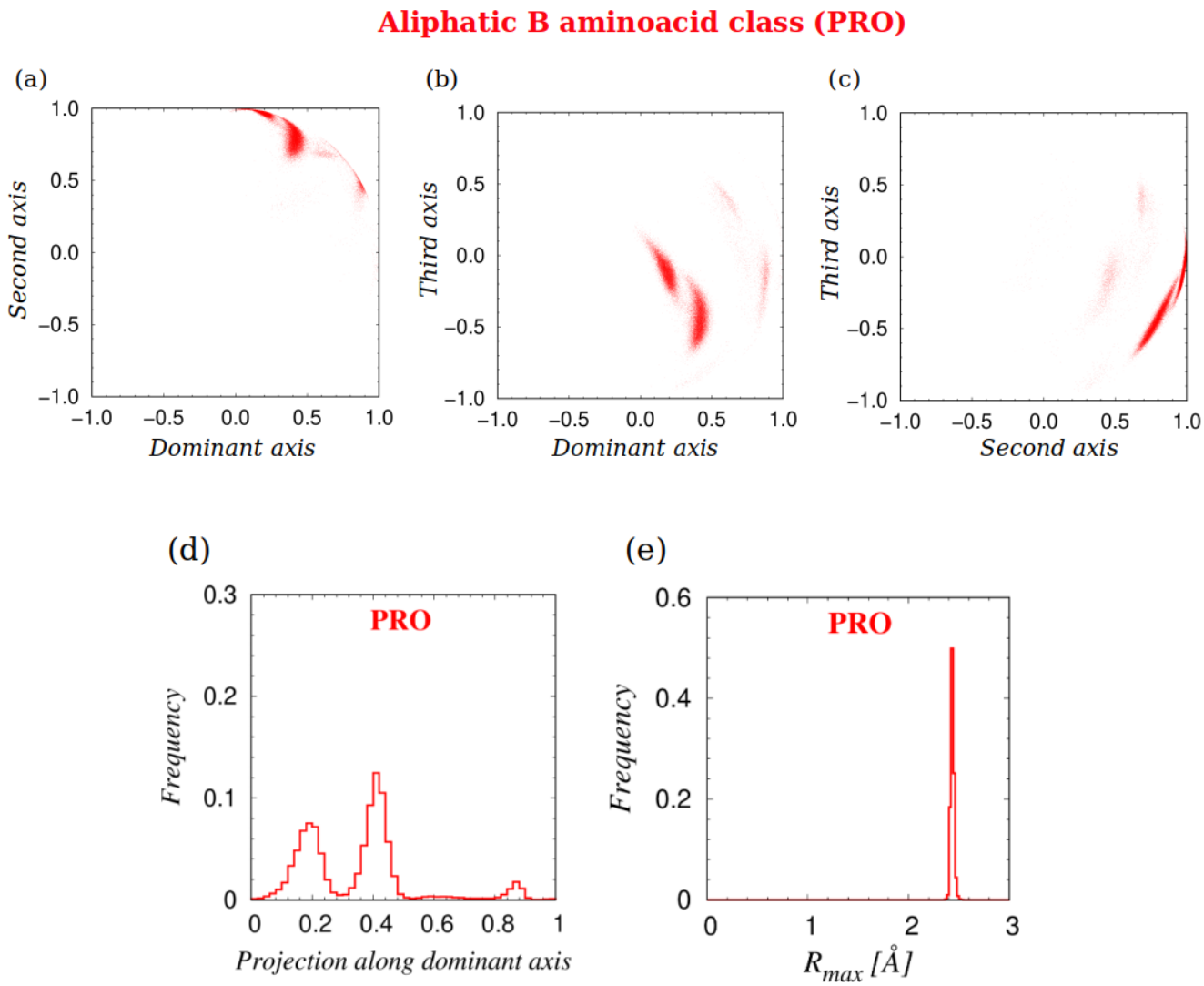

Figure 8: Aliphatic B amino acid class is composed only of PRO: geometrical attributes of protrusion. Panels (a), (b), and (c) are the three two-dimensional projection cross plots. Panel (d) shows the frequency distribution of the projections along the dominant axis for PRO. Panel (e) shows the frequency distributions of the amount of the maximal protrusion R_{max} for PRO.

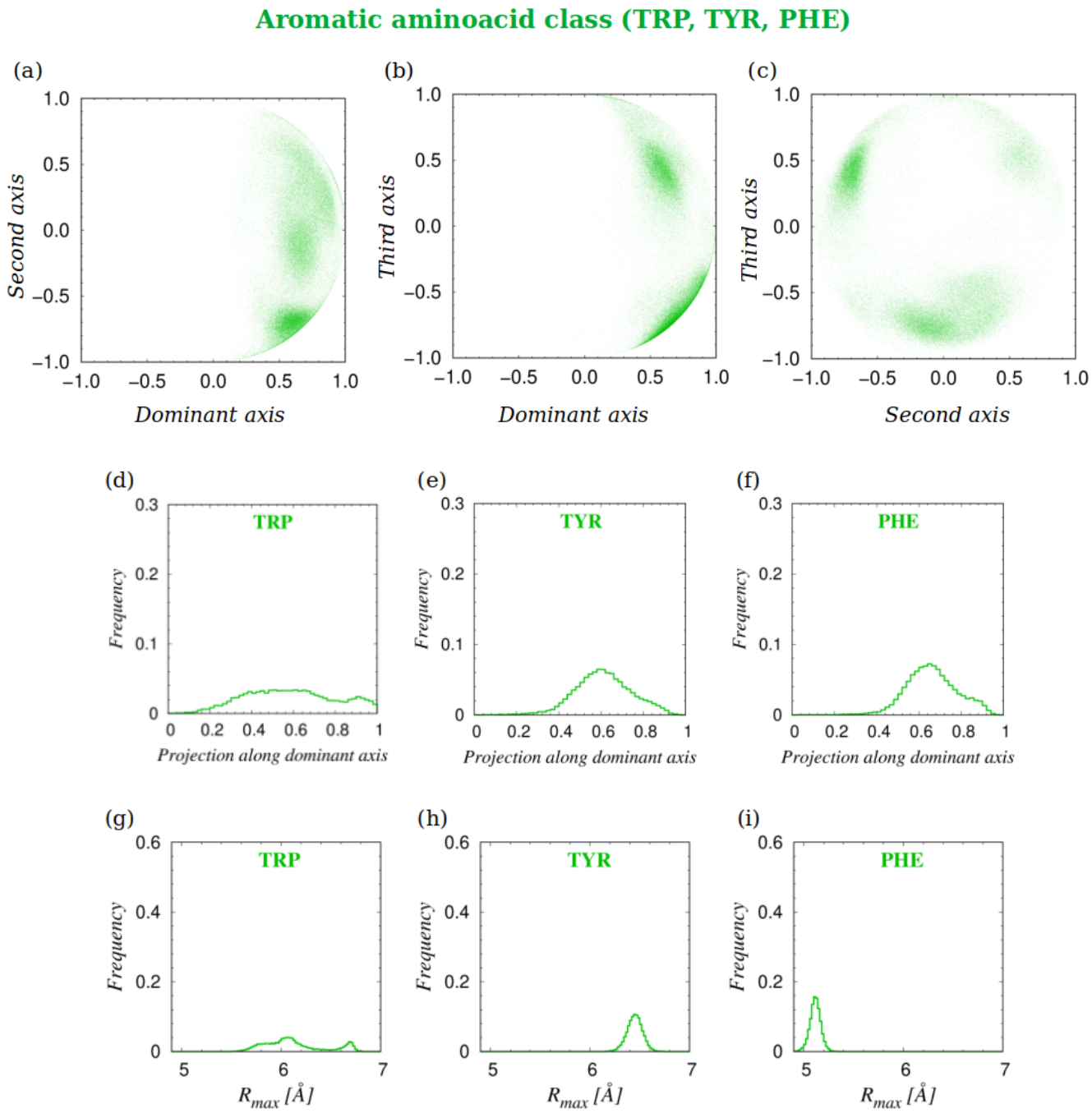

Figure 9: Aromatic amino acid class: geometrical attributes of protrusion. Panels (a), (b), and (c) are the three two-dimensional projection cross plots. Panels (d)-(f) shows the frequency distributions

of the projections along the dominant axis separately for the participant amino acids. Panels (g)-(i) show the frequency distributions of amount of the maximal protrusion R_{max} separately for the participant amino acids.

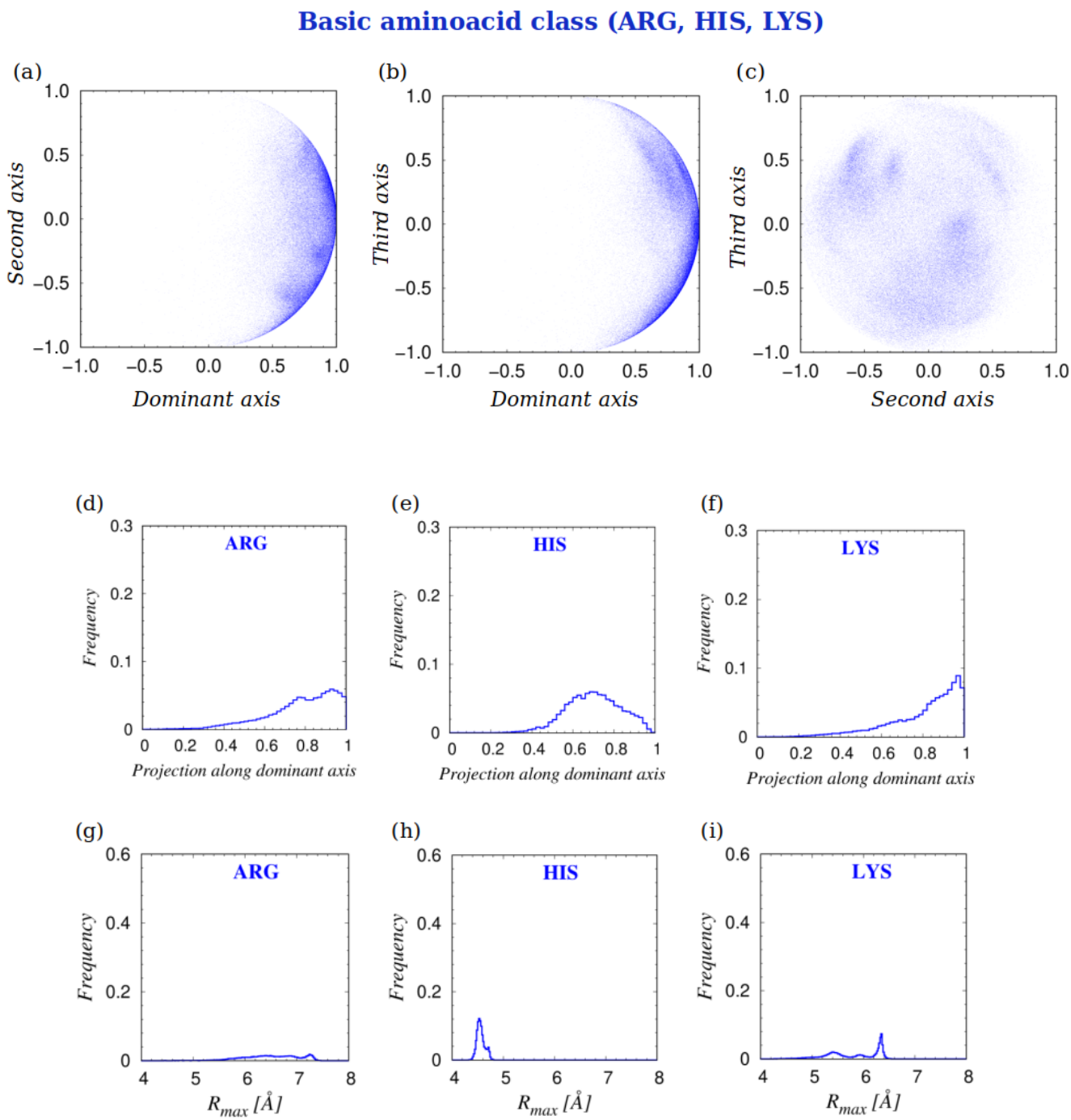

Figure 10: Basic amino acid class: geometrical attributes of protrusion. Panels (a), (b), and (c) are the three two-dimensional projection cross plots. Panels (d)-(f) shows the frequency distributions of the

projections along the dominant axis separately for the participant amino acids. Panels (g)-(i) show the frequency distributions of amount of the maximal protrusion R_{max} separately for the participant amino acids.

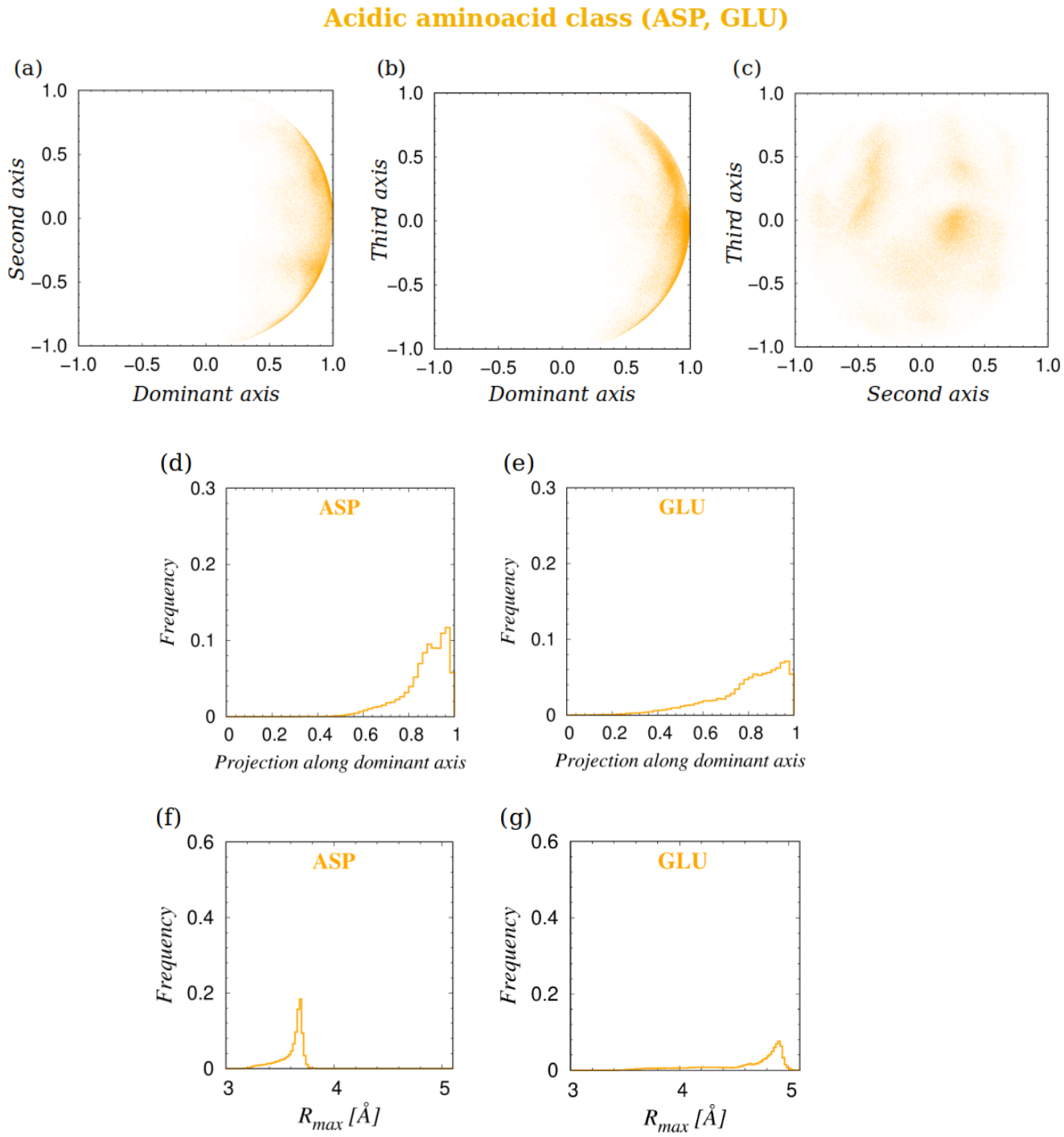

Figure 11: Acidic amino acid class: geometrical attributes of protrusion. Panels (a), (b), and (c) are the three two-dimensional projection cross plots. Panels (d)-(e) shows the frequency distributions of the projections along the dominant axis separately for the participant amino acids. Panels (f)-(g) show the frequency distributions of amount of the maximal protrusion R_{max} separately for the participant amino acids.

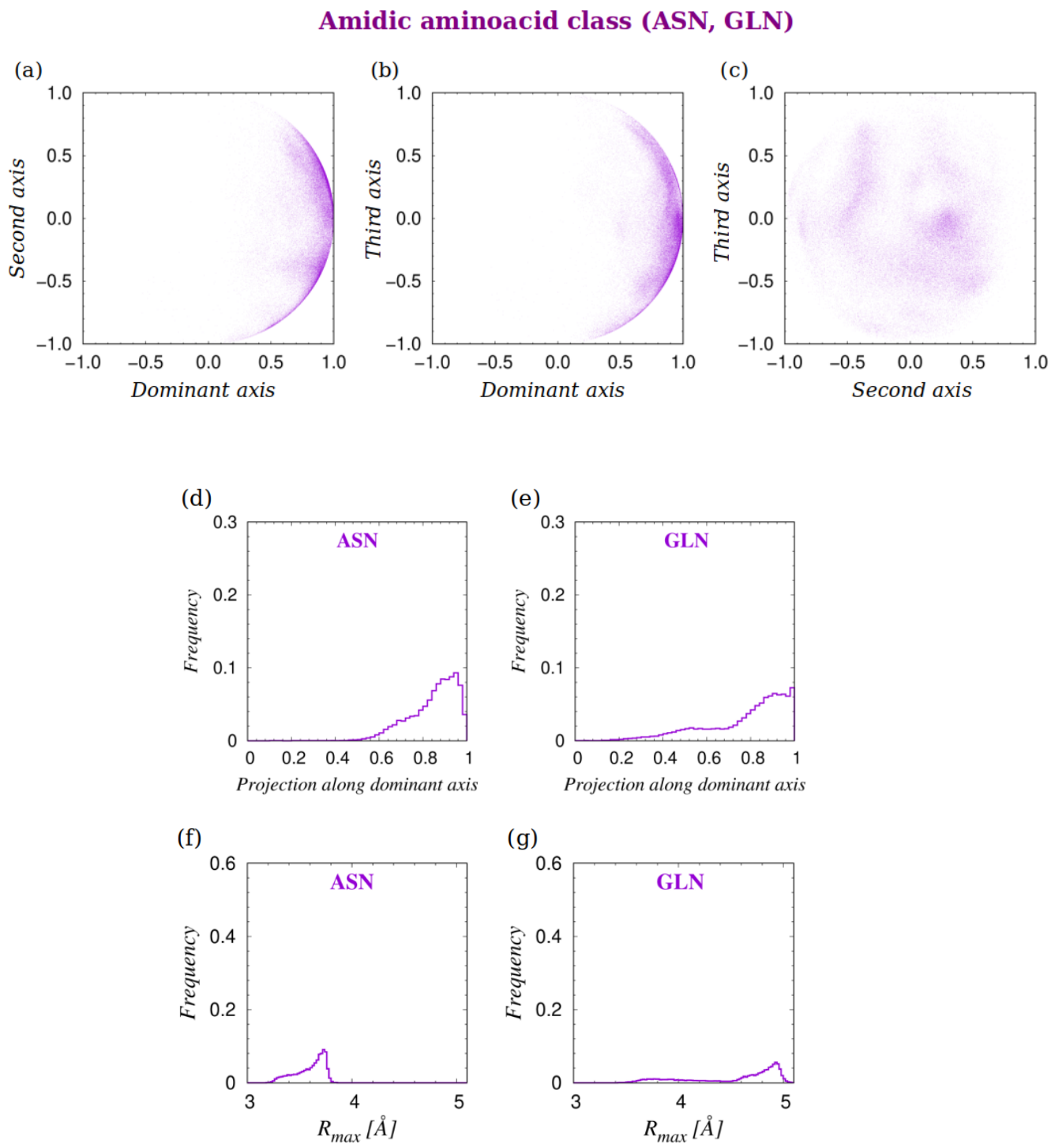

Figure 12: Amidic amino acid class: geometrical attributes of protrusion. Panels (a), (b), and (c) are the three two-dimensional

projection cross plots. Panels (d)-(e) shows the frequency distributions of the projections along the dominant axis separately for the participant amino acids. Panels (f)-(g) show the frequency distributions of amount of the maximal protrusion R_{max} separately for the participant amino acids.

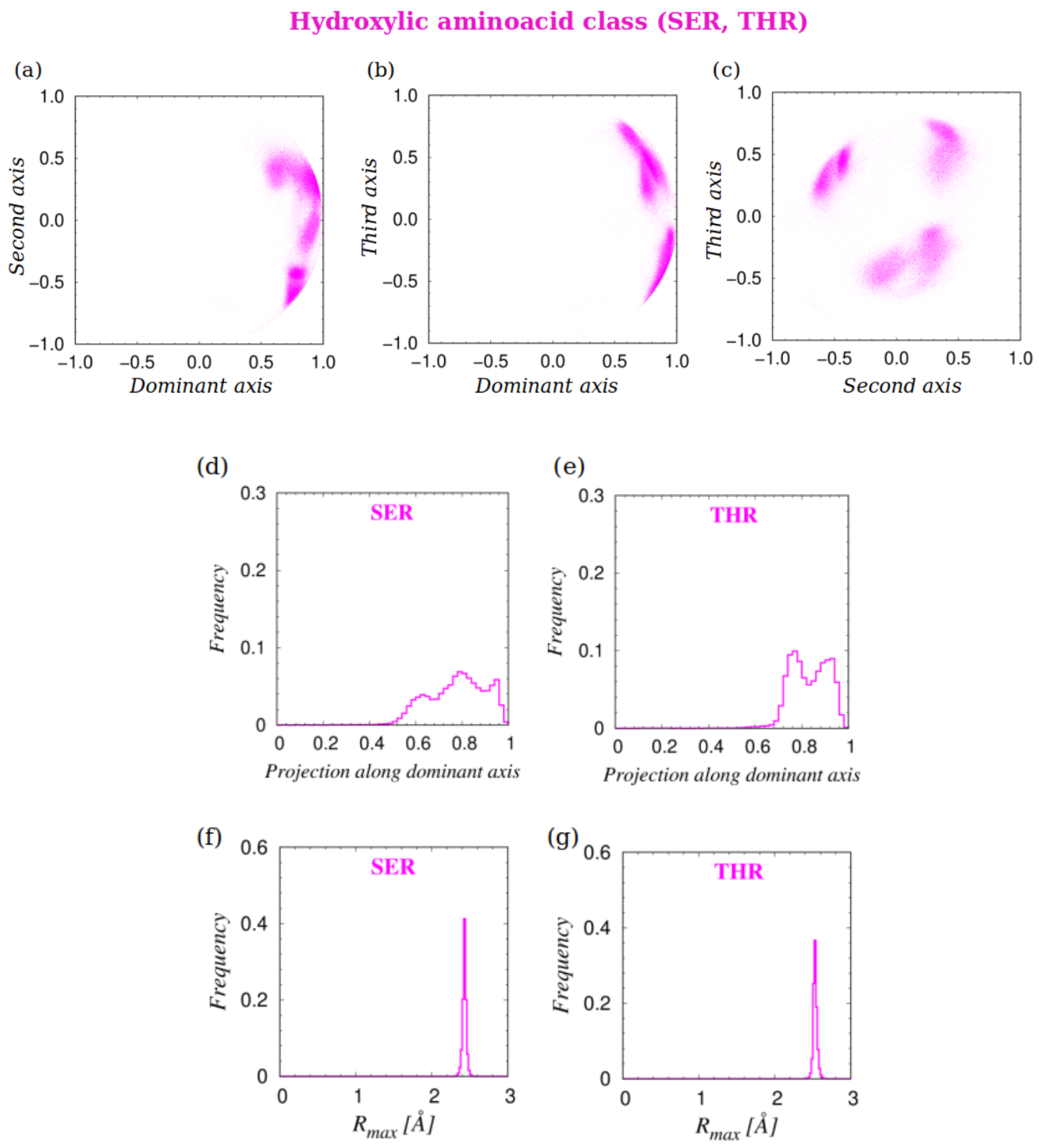

Figure 13: Hydroxylic amino acid class: geometrical attributes of protrusion. Panels (a), (b), and (c) are the three two-dimensional

projection cross plots. Panels (d)-(e) shows the frequency distributions of the projections along the dominant axis separately for the participant amino acids. Panels (f)-(g) show the frequency distributions of amount of the maximal protrusion R_{max} separately for the participant amino acids.

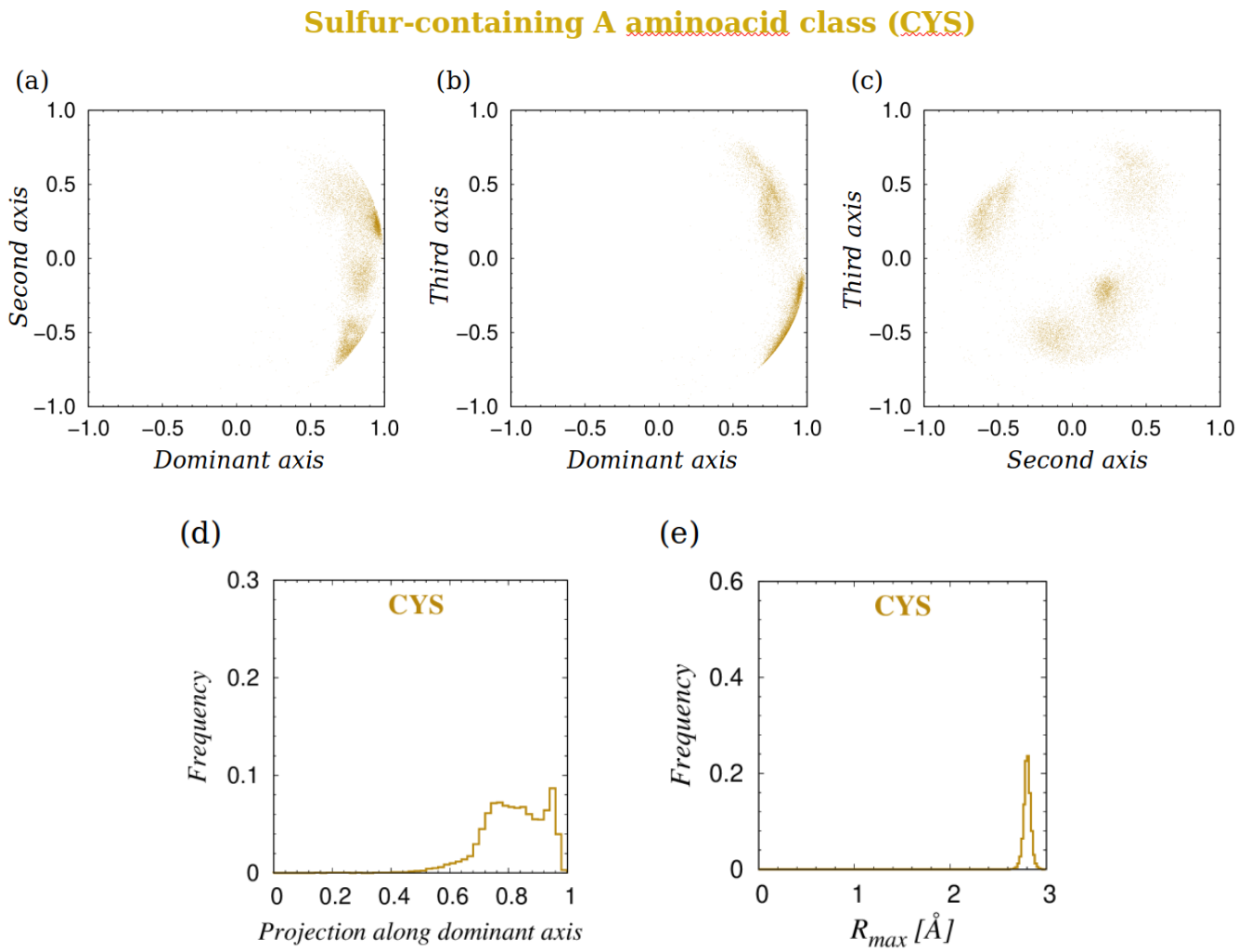

Figure 14: Sulfur-containing A amino acid class is composed only of CYS: geometrical attributes of protrusion. Panels (a), (b), and (c) are

the three two-dimensional projection cross plots. Panel (d) shows the frequency distribution of the projections along the dominant axis for CYS. Panel (e) shows the frequency distributions of amount of the maximal protrusion R_{max} for CYS.

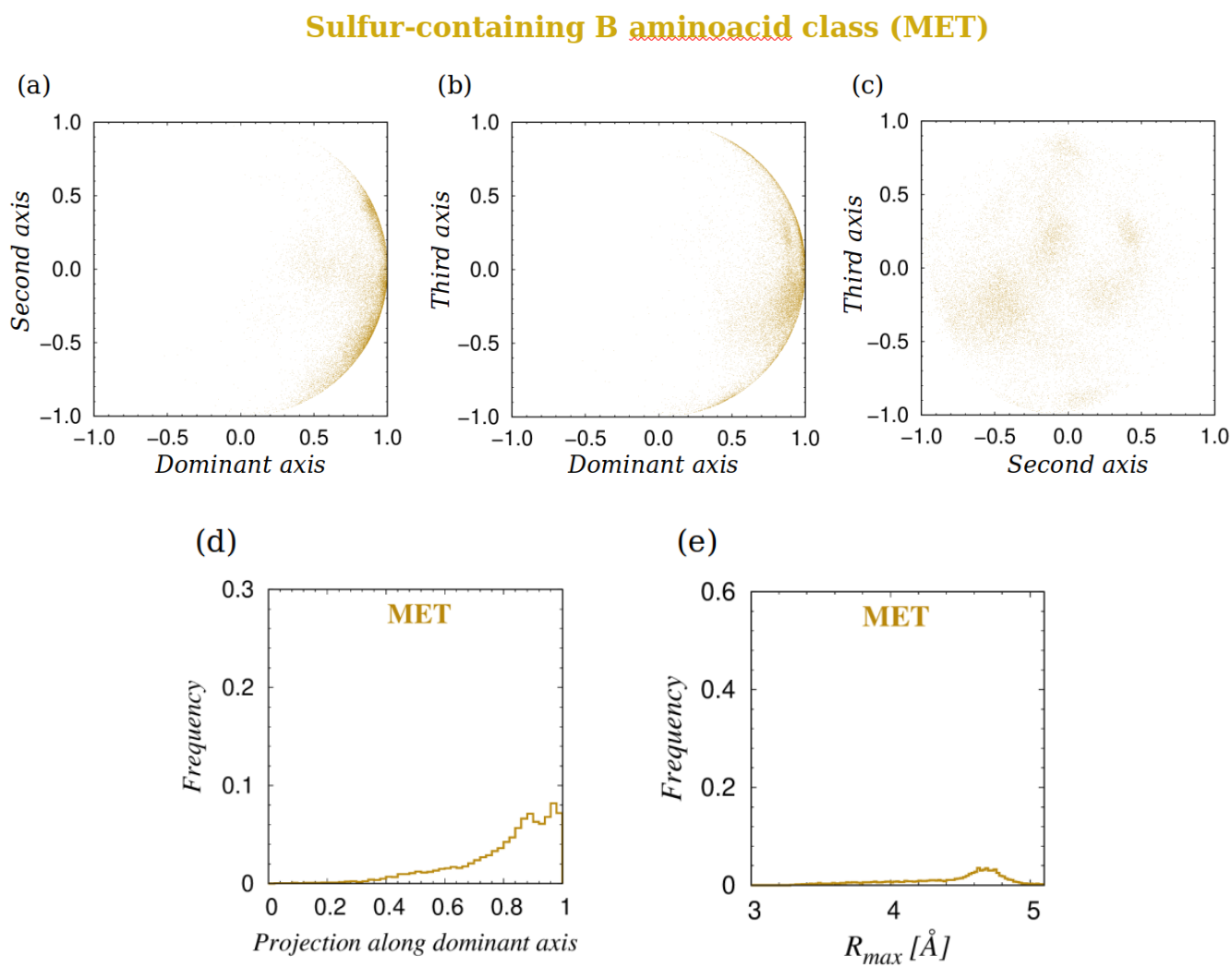

Figure 15: Sulfur-containing B amino acid class is composed only of MET: geometrical attributes of protrusion. Panels (a), (b), and (c) are

the three two-dimensional projection cross plots. Panel (d) shows the frequency distribution of the projections along the dominant axis for MET. Panel (e) shows the frequency distributions of amount of the maximal protrusion R_{max} for MET.

IV. CONCLUSIONS

Our results validate the adequacy of the nine IMGT 'physico-chemical' amino acid classes [41]. Not only do they capture the chemical behavior of the participant amino acids, but they also nicely fit the geometrical attributes of the protrusion of side chains. Amino acid participants in the same class display similar geometrical attributes of their amino acid protrusions, that makes them suitable for interchangeability in protein native state structures. Indeed, one ought to expect that protein structure, stability and function are not significantly affected by amino acid substitutions within a class. An atomistic study [44] considered the interactions between the side chains and different solvents and found that water was optimal (in comparison with cyclohexane (hydrophobic) and ethanol (polar)) in the ranges of free energies of solvation.

There are vastly more protein sequences than native state folds and sometimes unrelated sequences adopt the same native state conformation. Also, native state protein structures are usually robust to significant amino acid mutations [45] except at certain key locations [46-48] in the protein interior, when the sidechain packing in the hydrophobic core ceases to remain optimal, so that a sequence switches to a different fold topology that provides a better fit. The amino acid interchangeability is often observed during protein evolution, when amino acid substitutions occur over time, that allows for functional diversification. The availability of a rich repertoire of amino acids facilitates the optimization of sequence within the space of the thousands of putative native state folds, already pre-sculpted at the backbone level [14-16].

A codon is a triplet of nucleotides that encodes a specific amino acid residue in a protein. Interestingly, the genetic codes of different organisms select one of several choices for encoding a given amino acid. This is a subject of great interest (see, for instance, Refs. [49-52]) and there have been careful studies and modeling of distinct factors that may rationalize this. These factors include gene expression level and the speed of translation, the

content of the nucleotides, the stability of RNA, and conservation of amino acids in similarity groups like the ones we have studied here.

There are two competing drives during evolution and natural selection. First, when a good sequence is found that not only folds well but also is effective functionally, one would like to maintain the sequence in future generations. Equally, if not more importantly, one would seek mutations which are effective in exploring sequence space effectively even if it means that selection spurns many of the putative sequences that are deleterious. We now proceed to a brief look at the genetic code which programs for the groups of amino acids. Table III shows the number of triplet codons coding for each of 11 groups, this time adding glycine and the STOP as groups to the 9 IGMT classes [41]. In steady state, assuming equal number of amino acids for each codon (a condition that satisfies detailed balance), one obtains the propensities shown in the Table under neutral evolution without selection. There are some deviations in actual abundances, which may signal selection mechanisms at work. Notable among these selection pressures is a drive to equalize the abundances of the acidic and basic groups; a tendency to suppress the abundance of hydroxylic amino acids (SER and THR), which

are located predominantly in turns and loops; the reduction in the abundance of CYS, which plays a significantly beneficial role only in rare situations when a pair of CYS are near each other and form a disulfide bond; and an enhancement in the abundances of amino acids in the Aliphatic A group because they are hydrophobic and occupy the core of the protein structure.

Group name	**Number of codons**	**Expected propensity with no selection [%]**	**Observed propensity [%]**
Aliphatic A **ALA (GCA, GCC, GCG, GCT)** **ILE (ATA, ATC, ATT)** **LEU (TTA, TTG, CTT, CYC, CTA, CTG)** **VAL (GTT, GTC, GTA, GTG)**	17	26.6	30.6
Aliphatic B **PRO (CCT, CCC, CCA, CCG)**	4	6.2	4.7
Aromatic **PHE (TTT, TTC)** **TRP (TGG)** **TYR (TAT, TAC)**	5	7.8	9.3
Acidic **ASP (GAC, GAT)** **GLU (GAA, GAG)**	4	6.2	12.3

Basic **ARG (CGA, CGC, CGG, CGT, AGA, AGG)** **LYS (AAA, AAG)** **HIS (CAC, CAT)**	10	15.6	12.6
Hydroxylic **SER (TCA, TCC, TCG, TCT, AGC, AGT)** **THR (ACA, ACC, ACG, ACT)**	10	15.6	11.2
Amidic **ASN (AAC, AAT)** **GLN (CAA, CAG)**	4	6.2	7.9
Sulphur-containing A **CYS (TGC, TGT)**	2	3.1	1.3
Sulphur-containing B and 'Start' **MET (ATG)**	1	1.6	1.9
Glycine **GLY (GGA, GGC, GGG, GGT)**	4	6.2	8.1
Stop **TAA, TAG, TGA**	3	4.7	-

Table III: Table of codons for 11 groups along with expected propensity of occurrence and actual frequencies. See text for additional details.

The sensitivity-diversity balance is underscored in Table IV, where the number of single-step (just one nucleotide mutating) pseudo-silent mutations are shown for each of the 11 groups. The pseudo-silent mutations are a coarse-grained analog of silent mutations in which the mutations result in a given amino acid being replaced by another in the same group, which harmoniously fits both geometry and chemistry and thus is likely to be a successful substitution. The percentage of such pseudo-silent mutations is found to be between 10% and 55% with the upper number holding for the all-important Aliphatic A group.

Amino acid group	**Number of mutations**	**Number of *internal* mutations (percentage of all possible mutations)**
Aliphatic A	153	82 (~54%)
Aliphatic B	36	12 (~33%)
Aromatic	45	8 (~18%)

Acidic	36	12 (~33%)
Basic	90	30 (~33%)
Hydroxylic	90	38 (~42%)
Amidic	36	4 (~11%)
Sulphur-containing A	18	2 (~11%)
Sulphur-containing B and 'Start'	9	0 (~0%)
Glycine	36	12 (~33%)
Stop	27	4 (~15%)

Table IV: The table shows the number of possible single base mutations for each of 11 groups (nine IGMT classes [41] and glycine and 'Stop' class) along with the number and percentage of mutations that result in the new amino acid still belonging to the group.

We conclude with a discussion of some influential ways of thinking about proteins, Anfinsen [5] is credited with an idea that for a small globular

protein, the native state structure is determined by the amino acid sequence. This may appear as a logical conclusion of his observations that distinct sequences have different native state structures and, remarkably, a protein folds reproducibly and rapidly to its own native state structure. Each sequence must sculpt its own free energy landscape with a stable and kinetically accessible minimum corresponding to its native state structure. The principle of minimal frustration [26-28] highlights the importance of sequence selection (through evolution) to facilitate this sculpting of a very high dimensional funneled free energy landscape.

We [14-21] have suggested a simpler viewpoint. We observe that all small globular proteins share the same backbone. The poking interactions, both local and non-local, between the backbone atoms with appropriate geometrical constraints leads to the creation of the common building blocks of all proteins: helices and sheets. We have deduced the geometries of these building blocks with just a few geometrical assumptions. The modularity of protein native state structures then yields a set of putative native state folds, which number just several thousand [10]. This is in accord with observations that the number of distinct sequences vastly exceed the number of native state

conformations that they adopt. Unlike the idea credited to Anfinsen [4-5,26], an equally viable hypothesis is that a sequence does not sculpt its own free energy landscape but instead chooses the native state structure from among the library of modular folds. Thus, there are already a set of pre-sculpted (even before consideration of a sequence) free energy landscapes that a sequence can choose from. The principle of minimal frustration [26-28] emphasizes the important point that a sequence should be harmonious with the pre-sculpted landscape selected by it. Frustrating influences or a lack of compatibility or a lack of harmony will only serve to thwart the stability and/or kinetic accessibility of the native state fold.

A loose analogy is a township with a predetermined set of architectural styles for the homes. A person who wishes to live happily in the township ought to fit in harmoniously in one of the existing styles. Another analogy is a restaurant in which the items available are predetermined by the chef and listed on a menu. A customer who joyfully frequents the place ought to love an item from the menu in a manner free of conflicts or frustration.

Our work here underscores the remarkable harmony that exists both in geometry and in chemistry within the groups of similar amino acids. This

promotes the ideas of stability (many choices of amino acids that serve a purpose and can be harmlessly interchanged) and, at the same time, diversity allowing evolution to fully explore sequence space that yet fit within the pre-sculpted menu of native state structures, lending to stability and facilitating powerful functionality.

In future work, we intend to use the results of the present study as a starting point to tackle two important questions: What is the detailed role of side chains in selecting the optimal tertiary protein structure? And what are the relative orientations and geometrical and chemical fits of tightly packed sidechains that poke towards each other in the protein interior?

Acknowledgments

The computer calculations were performed on the Talapas cluster at the University of Oregon.

Funding information

This project received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie Grant Agree-

ment No. 894784 (TŠ). The contents reflect only the authors' view and not the views of the European Commission. JRB was supported by a Knight Chair at the University of Oregon. TXH is supported by The Vietnam Academy of Science and Technology under grant No. NCXS02.05/22-23.

Author Contributions

TŠ carried out all the calaculations under the guidance of JRB. JRB and TŠ wrote the paper. All authors participated in reviewing the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Creighton TE. Proteins: Structures and Molecular Properties. New York: W. H. Freeman; 1983.

2. Lesk AM. Introduction to Protein Science: Architecture, Function and Genomics. Oxford: Oxford University Press; 2004.

3. Bahar I, Jernigan RL, Dill KA. Protein Actions. New York: Garland Science; 2017.

4. Berg JM, Tymoczko JL, Gatto GJ Jr, Stryer L. Biochemistry. New York: Macmillan Learning; 2019.

5. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973; 181: 223.

6. Pauling L, Corey RB, Branson HR. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. USA 1951; 37: 205.

7. Pauling L, Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. Proc. Natl. Acad. Sci. USA 1951; 37: 251.

8. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976; 261: 552.

9. Chothia C. One thousand families for the molecular biologist. Nature 1992; 357: 543.

10. Przytycka T, Aurora R, Rose GD. A protein taxonomy based on secondary structure. Nat. Struct. Biol. 1999; 6: 672.

11. Taylor W. A ‘periodic table’ for protein structures. Nature 2002; 416: 657.

12. Bordin N, Sillitoe I, Lees JG, Orengo C. Tracing Evolution Through Protein Structures: Nature Captured in a Few Thousand Folds. Front. Mol. Biosci. 2021; 8: 668184.

13. Alvarez-Carreno C, Gupta RJ, Petrov AS, Williams LD. Creative destruction: New protein folds from old. Proc. Natl. Acad. Sci. USA 2022; 119: e2207897119.

14. Hoang TX, Trovato A, Seno F, Banavar JR, Maritan A. Geometry and symmetry presculpt the free-energy landscape of proteins. Proc. Natl. Acad. Sci. USA 2004; 101: 7960.

15. Rose GD, Fleming PJ, Banavar JR, Maritan A. A backbone-based theory of protein folding. Proc. Natl. Acad. Sci. USA 2006; 103: 16623.

16. Banavar JR, Maritan A. Physics of proteins. Ann. Rev. Biophys. Biomol. Struct. 2007; 36: 261.

17. Škrbić T, Maritan A, Giacometti A, Rose GD, Banavar JR. Building blocks of protein structures: Physics meets biology. Phys. Rev. E 2021; 104: 014402.

18. Škrbić T, Hoang TX, Giacometti A, Maritan A, Banavar JR. Proteins — a celebration of consilience. Intl. J. Mod. Phys. B 2022; 36: 07n08.

19. Banavar JR, Giacometti A, Hoang TX, Maritan A, Škrbić T. A geometrical framework for thinking about proteins. Proteins 2023, https://doi.org/10.1002/prot.26567

20. Škrbić T, Giacometti A, Hoang TX, Maritan A, Banavar JR. II. Geometrical framework for thinking about globular proteins: The power of poking. Proteins 2023, https://doi.org/10.1002/prot.26566

21. Škrbić T, Giacometti A, Hoang TX, Maritan A, Banavar JR. III. Geometrical framework for thinking about globular proteins: Turns in proteins. Submitted to Proteins.

22. Bhattacharyya M, Bhat CR, Vishveshwara S. An automated approach to network features of protein structure ensembles. Protein Sci. 2013; 22: 1399.

23. Bhattacharyya M, Ghosh S, Vishveshwara S. Protein Structure and Function: Looking through the Network of Side-Chain Interactions. Curr. Protein Pept. Sci. 2016; 17: 4.

24. Gadiyaram V, Vishveshwara Smitha, Vishveshwara Saraswathi. From Quantum Chemistry to Networks in Biology: A Graph Spectral Approach to Protein Structure Analyses. J. Chem. Inf. Model. 2019; 59:1715.

25. Rose GD. Ramachandran maps for side chains in globular proteins. Proteins 2019; 87:357.

26. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis, Proteins 1995; 21: 167.

27. Wolynes PG, Onuchic JN, Thirumalai D. Navigating the folding routes. Science 1995; 267: 1619.

28. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nature Struct. Biol. 1997; 4: 10.

29. Richards FM. Areas, volumes, packing, and protein structure. Annu. Rev. Biophys. Bioeng. 1977; 6: 151.

30. Corey RB, Pauling L. Molecular models of amino acids, peptides, and proteins. Review of Scientific Instruments 1953; 8: 621.

31. Koltun WL. Precision space-filling atomic models. Biopolymers 1965; 3: 665.

32. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 11982; 57: 105.

33. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins 2000; 40: 389.

34. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc. Natl. Acad. Sci. USA 2000; 97: 10383.

35. Dunbrack RL, Jr. Rotamer libraries in the 21st century. Curr. Opin. Struct. Biol. 2002; 12: 431.

36. Gaines JC, Acerbes S, Virrueta A, Butler M, Regan L, O'Hern CS. Comparing side chain packing in soluble proteins, protein-protein interfaces, and transmembrane proteins. Proteins 2018; 86: 581.

37. Huang X, Pearce R, Zhang Y. Toward the Accuracy and Speed of Protein Side-Chain Packing: A Systematic Study on Rotamer Libraries. J. Chem. Inf. Model 2020; 60: 410.

38. Kamien RD. The geometry of soft materials: a primer. Rev. Mod. Phys. 2002; 74: 953.

39. Škrbić T, Maritan A, Giacometti A, Banavar JR. Local sequence-structure relationships in proteins. Protein Sci 2021; 30: 818.

40. Ramachandran GN, Mitra AK. An explanation for the rare occurrence of cis peptide units in proteins and polypeptides. J. Mol. Biol. 1976; 107: 85.

41. Pommié C, Levadoux S, Sabatier R, Lefranc G, Lefranc M-P. IMGT (ImMunoGeneTics information system) standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J. Mol. Recognit. 2004; 17: 17.

42. 3D Macromolecule Analysis & Kinemage Home Page at Richardson Laboratory. http://kinemage.biochem.duke.edu/databases/top8000/

43. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983; 22: 2577.

44. Dongmo Foumthuim CJ, Carrer M, Houvet M, Škrbić T, Graziano G, Giacometti A. Can the roles of polar and non-polar moieties be reversed in non-polar solvents? Phys. Chem. Chem. Phys. 2020; 22: 25848.

45. Alber T, Bell JA, Dao-Pin S, Nicholson H, Wozniak JA, Cook S, Matthews BW. Replacements of Pro86 in Phage T4 Lysozyme Extend an α-helix But Do Not Alter Protein Stability. Science 1998; 239: 631.

46. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. The design and characterization of two proteins with 88% sequence identity but different structure and function. Proc. Acad. Sci. USA 2007; 104: 11963.

47. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. A minimal sequence code for switching protein structure and function. Proc. Acad. Sci. USA 2009; 106: 21149.

48. Porter LL, Looger LL. Extant fold-switching proteins are widespread. Proc. Acad. Sci. USA 2018; 115: 5968.

49. Saunders R, Deane CM. Synonymous codon usage influences the local protein structure observed. Nucleic Acids Research 2010; 38: 6719.

50. Gonzalez DL, Giannerini S, Rosa R. On the origin of degeneracy in the genetic code. Interface Focus 2019; 9: 20190038.

51. Liu Y. A code within the genetic code: codon usage regulates co-translational protein folding. Cell Commun. Signal. 2020; 18: 145.

52. Inouyea M, Takinoa R, Ishida Y, Inouyea K. Evolution of the genetic code; Evidence from serine codon use disparity in Escherichia coli. Proc. Acad. Sci. USA 2020; 117: 28572.