

### III. Geometrical framework for thinking about globular proteins: turns in proteins

Short running title (max. 40 characters): Turns in globular proteins

5 Tatjana Škrbić<sup>1,2</sup>, Achille Giacometti<sup>1,3</sup>, Trinh X. Hoang<sup>4</sup>, Amos Maritan<sup>5</sup>, and  
Jayanth R. Banavar<sup>2</sup>

<sup>1</sup> Ca' Foscari University of Venice, Department of Molecular Sciences and  
Nanosystems, Venice, Italy

10

<sup>2</sup> University of Oregon, Department of Physics and Institute for Fundamental  
Science, Eugene, Oregon, USA

15 <sup>3</sup> European Centre for Living Technology (ECLT), Ca' Bottacin, Dorsoduro  
3911, Calle Crosera, 30123 Venice, Italy

<sup>4</sup> Vietnam Academy of Science and Technology, Institute of Physics, Hanoi,  
Vietnam

20

<sup>5</sup> University of Padua, Department of Physics and Astronomy, Padua, Italy

#### 25 **Correspondence**

Tatjana Škrbić

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

30 **Email: [tatjana.skrbic@unive.it](mailto:tatjana.skrbic@unive.it)**

## Abstract

We have shown recently that the notion of poking pairwise interactions along a chain provides a unifying framework for understanding the formation of both secondary and the tertiary protein structure based on symmetry and geometry.  $\alpha$ -helices and  $\beta$ -sheets are found to be special geometries that have systematic poking contacts in a repetitive manner with the contacts being local along the  $\alpha$ -helix and non-local along a pair of adjacent strands within a  $\beta$ -sheet. Pairwise poking interactions also govern tertiary structure formation, but they are weaker and there are no special geometrical constraints as in secondary structure formation. Here we demonstrate that protein turns, the most prevalent non-repetitive structural element in proteins, are instances of local (as in  $\alpha$ -helices) and isolated (non-repetitive) poking pairwise contacts for which the geometrical constraints are partially relaxed. This simple and purely geometrical definition of protein turns (also sometimes known as reverse turns,  $\beta$ -turns,  $\beta$ -bends, hairpin bends,  $3_{10}$  bends, kinks, widgets, ...) provides a simple framework for unifying them. We present the results of a

systematic analysis and identify their structural classes as well as their respective amino acid preferences.

50

*Keywords:* geometry, symmetry, poking, backbone, amino acids

### **Statement for broader audience**

55 Poking interaction along a discrete chain signals the affinity of one part of the chain with another. It has been shown that a simple and purely geometrical model based on poking interactions is able to capture the essential features of complex protein chains, especially their building blocks and their assembly. We demonstrate that poking interactions also provide a simple framework to  
60 describe protein turns as isolated (not coordinated) local poking interactions. This framework allows one to comprehend the structural variety of protein turns that have been extensively studied in the past.

## 1. Introduction

The building blocks of the structure of globular proteins, topologically one-  
65 dimensional  $\alpha$ -helices [1] and almost planar  $\beta$ -sheets [2] are lower dimen-  
sional manifolds and thus not subject to compactification in three-dimen-  
sional space. Furthermore, both a helix and a strand are iso-directional. For  
the formation of a reasonably compact tertiary structure of a globular protein,  
one requires a sharp turn allowing for a change in the chain direction [3-28].  
70 Indeed, after helices and sheets, protein turns are the most common structural  
feature recognized in proteins [8,14]. In contrast to helices and sheets, they  
are non-repetitive structural elements with highly heterogeneous structures,  
that have made them, despite their ubiquity, much more challenging to define  
and categorize [27]. Protein turns are prevalently located at the surface of a  
75 protein and facilitate interactions with other molecules and often serve as  
sites for high specificity ligand binding [20,27]. Having a high preference for  
hydrophilic residues [13,15], because of their location at the protein surface  
and being exposed to water, turns help with the overall stabilization of the  
three-dimensional native state of a globular protein [21,25]. Protein turns are  
80 thus special in their ability to provide for the combined structural and chemi-

cal backdrop [26]: tight turns bring non-local parts of a protein chain together and at the same time have highly specific amino acid composition [8,13-15,20,27]. In this way, they allow for an efficient navigation of complex folding pathways and play an important role in protein folding and function [4,10,19,22,28].

Protein turns go under various names as reverse turns,  $\beta$ -turns,  $\beta$ -bends, hairpin bends,  $3_{10}$  bends, kinks, widgets, ... and they are found to display an enormous structural variety [8,9,14,16]. Consequently, there are many distinct definitions of protein turns present in the literature. Here we seek a physically motivated unified perspective for understanding turns.

The structural characterization and classification of protein turns was presented in the seminal work of Venkatachalam [3], who explored the conformations available to a system of three linked peptide units (i.e., to a consecutive amino acid quartet) that could be stabilized by a backbone hydrogen bond between the -CO carboxyl group of the  $i$ -th residue and the -NH amino group of the  $(i+3)$ -th residue. Venkatachalam [3] suggested three turn types, named Type I, Type II and Type III (G-helix or  $3_{10}$ -turn), as well as their mirror images (Types I', II' and III', respectively) and determined their respec-

tive backbone conformations, in terms of the Ramachandran angles [29]

100  $(\varphi_{i+1}, \Psi_{i+1})$  and  $(\varphi_{i+2}, \Psi_{i+2})$  of the two inner residues  $i+1$  and  $i+2$  of the protein quartet  $(i, i+1, i+2, i+3)$ . Types I and II (as well as their mirror images I' and II') are the most common turn types and are known as  $\beta$ -turns or U-turns, because they provide for the sharpest protein turns, as in  $\beta$ -hairpins where the polypeptide chain folds back on itself in a U-shape. Types III and III' are  
105 much milder turns and correspond to the turn in a G-helix and are thus called G-turns or  $3_{10}$ -turns. Two mirror image turns are transformed from one to the other by inversion symmetry with respect to the  $(0,0)$  point in the Ramachandran plot [29].

After extensive surveys of protein structures, it became increasingly ev-  
110 ident that hydrogen bonding between residues  $i$  and  $i+3$  was not a necessary condition to have a good turn. Furthermore, many turns were found to have the distance between the  $C_{\alpha}$  atoms of the  $i$  and  $i+3$  residues of up to  $\sim 7-7.5$  Å lying further away than the plausible hydrogen bonding distance [6]. Thus, in addition to the three turn types of Ventakachalam and their mirror images [3],  
115 five additional turn types were identified that seemed to account for all observed cases [7,11,18].

Apart from the turn classification relying on Ramachandran angles and detailed atomistic conformational analysis, many authors have successfully adopted a coarse-grained approach considering only the  $C_\alpha$  atoms. Kuntz [5] considered turns as all non-helical segments that effectively change the direction of the protein chain by more than  $90^\circ$ , while Levitt and Greer [12] assigned turns as non- $\alpha$ , non- $\beta$  segments for which the dihedral angle defined by four successive  $C_\alpha$  atoms is between  $-90^\circ$  and  $90^\circ$ . In addition, Rose and Seltzer [9] introduced a structurally unbiased, purely geometrical definition of protein turns, and identified them as local minima in the radius of curvature calculated from the  $i-2$ ,  $i$ , and  $i+2$   $C_\alpha$  atoms and in which both the ‘inlet’ and the ‘outlet’ of the presumed turn cannot both be fitted within a single pipe of diameter  $\sim 5.2\text{\AA}$ . Interestingly this estimate of Rose and Selzer is close to our recent theoretical prediction for the thickness of a discrete protein backbone of  $2\Delta \sim 5.26\text{\AA}$ , as measured by the diameter of a Kepler coin or twice the radius of curvature of a Kepler helix [30,31]. The Selzer-Rose pipe criterion is effective in eliminating spurious local minima within a single helix.

Recently, relying entirely on symmetry and geometry, and the notion of  
135 poking pairwise interactions within a discrete chain, we presented a unifying  
framework to understand the formation of both secondary and tertiary protein  
structure [30,31]. A poking interaction signals the affinity of one part of a  
chain to another and is often associated with a backbone hydrogen bond. The  
protein  $\alpha$ -helix geometry is the unique way in which a pair of Kepler coins of  
140 a discrete chain at the  $i$ -th and  $(i+3)$ -th positions locally poke towards each  
other in a repetitive manner and just touch each other [30,31]. The turn angle  
of a quartet  $\tau$  is defined as the angle between the directions  $(i,i+1)$  and  
 $(i+2,i+3)$ . A cautionary note is that these two lines are skew and do not inter-  
sect. The  $\alpha$ -helix is iso-directional consisting of successive quartets compris-  
145 ing tight ‘Kepler turns’ that are necessarily non-planar to avert steric clashes  
and have local turn angles of around  $130^\circ$  (see Figure 1).

150



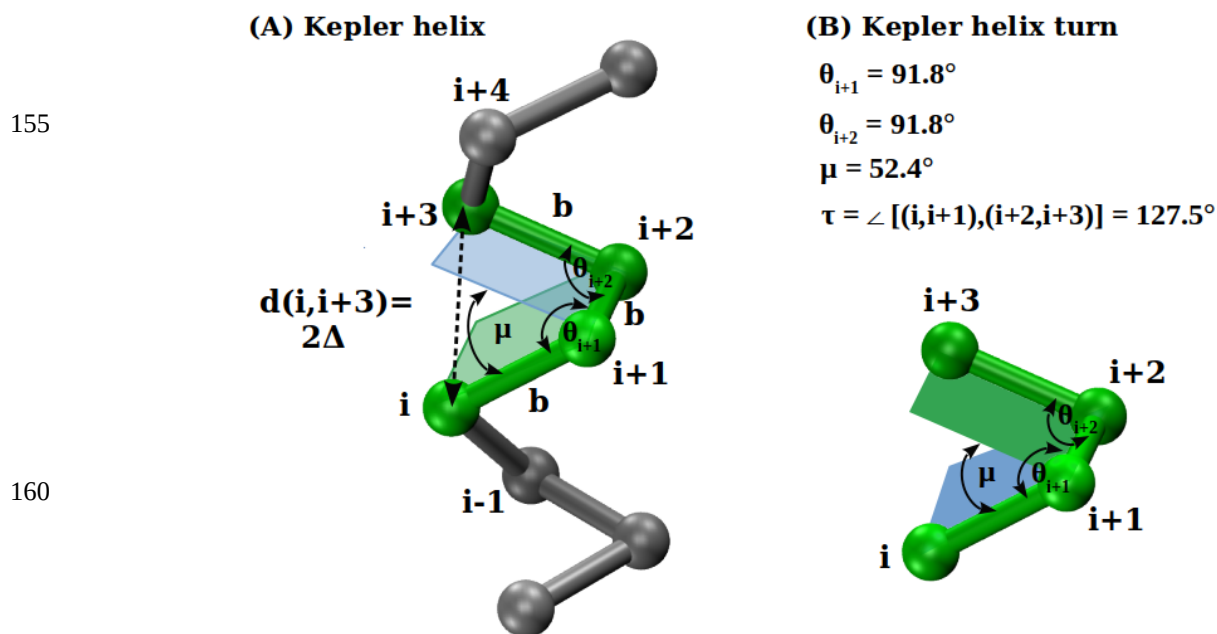


Figure 1: Panel (A) shows Kepler helix that consists of the consecutive local turns (Kepler helix turns), where one such turn is formed by quartet of points  $(i, i+1, i+2, i+3)$  and highlighted in green color. The bond length  $b=3.81\text{\AA}$  sets the length scale, and the distance  $d(i, i+3)$  is equal to twice the value of the local radius of curvature of Kepler helix  $2\Delta$ . The geometrical characteristics of the Kepler turn, two bond bending angles  $\theta_{i+1}$  and  $\theta_{i+2}$  and dihedral angle  $\mu$  are indicated. Panel (B) depicts enlarged Kepler turn, with the theoretically predicted values of its geometrical characteristics. Protein helices are characterized by a mean bond bending angle of  $91.7\pm 2.8^\circ$ , a mean dihedral angle of

49.5±6.7°, and a mean local turn angle of 130.2±5.8°. These mean values and their standard deviations have been calculated over

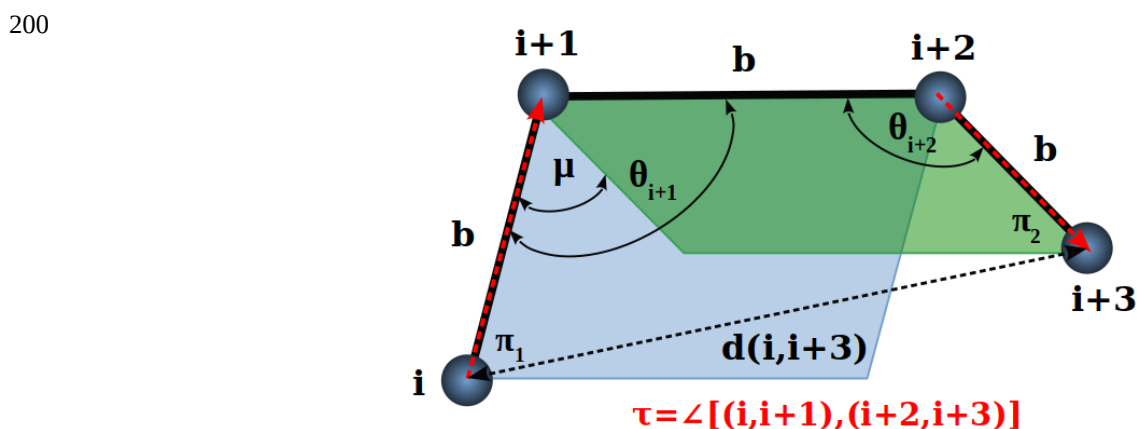
175 226,478  $\alpha$ -helical quartets (for definition of an  $\alpha$ -helical quartet see Materials and Methods, Section 2.4) in our data set (defined in Materials and Methods, Section 2.3).

Instead of strictly repetitive turns, as in the Kepler helix, let us consider  
180 a completely isolated poking contact between  $i$  and  $i+3$  with the absence of a poking contact between  $i-1$  and  $i+2$  and  $i+1$  and  $i+4$ . The absence of any constraint pertaining to the presence of a succession of pairwise poking contacts leads to a relaxation of the rigid constraint with the possibility of a close to planar configuration and thence a larger turn angle. Our main goal is to demon-  
185 strate that protein turns are indeed instances of such local and isolated poking interactions.

## 2. Materials and Methods

### 2.1 Local protein geometry: protein quartets calculation

190 We view the protein backbone as a sequence of equidistant points, where the  $C_\alpha$  atoms are located, separated by a constant bond length  $b$  of  $3.81\text{\AA}$  [32]. The shortest portion of protein local structure that can effectively ‘turn’ the chain consists of four consecutive points (quartet), see Figure 2. A truly perfect turn angle corresponds to  $\tau = 180^\circ$ . A turn angle associated with just  
195 three amino acids is given by  $\tau = 180^\circ - \mu$  (the local bond bending angle). The latter angle is typically not much smaller than  $90^\circ$ , so that a typical sharp triplet turn is around  $90^\circ$  (a right-angle turn) and is not effective in turning a chain back. This naturally leads to a quartet of beads, as in earlier studies, as the basic portion of protein local structure that serve as local turn candidates.



205

Figure 2: Definition of the local turn consisting of four successive points  $i$ ,  $i+1$ ,  $i+2$ , and  $i+3$  along a chain where the  $C_\alpha$  atoms are lo-

cated. The bond length  $b$  is the distance between the consecutive beads. The bond bending angle  $\theta_{i+1}$  is the angle subtended at point  
210  $(i+1)$  by points  $i$  and  $i+2$ , while the bond bending angle  $\theta_{i+2}$  is the angle subtended at point  $i+2$  by points  $i+1$  and  $i+3$ . The dihedral angle  $\mu$  is the angle between the planes  $\pi_1$  and  $\pi_2$  formed by  $[i, i+1, i+2]$  and  $[i+1, i+2, i+3]$ , respectively. The distance between the first bead in the turn  $i$  and the last bead in the turn  $i+3$ ,  $d(i, i+3)$ , is indicated. The turn  
215 angle  $\tau$  is defined as the angle between the directions  $(i, i+1)$  and  $(i+2, i+3)$ . Thus, there are several variables that characterize a local quartet of amino acids with constant bond length  $b$ :  $\theta_{i+1}$ ,  $\theta_{i+2}$ ,  $\mu$ ,  $d(i, i+3)$ , and  $\tau$ , of which just three are independent.

220 Figure 2 depicts the quartet of beads and defines its relevant geometrical attributes. Under the conditions of constant bond length  $b$ , the geometry of the quartet is defined by exactly three independent variables. These three variables can be chosen from five convenient variables: the two bond bending angles  $\theta_{i+1}$  and  $\theta_{i+2}$ , the dihedral angle  $\mu$ , the distance  $d(i, i+3)$ , and the turn  
225 angle  $\tau$  (see Figure 2).

An important issue is the difference between unidirectionality and uniaxiality and its repercussions for the turn analysis we present in this paper. A protein chain is unidirectional at the atomic level because the C-terminal and N-terminal unambiguously denote the protein beginning and end. Uniaxiality, on the other hand, does not distinguish the direction in which the chain is going, as is the case when the geometries of the quartets  $[i, i+1, i+2, i+3]$  and  $[i+3, i+2, i+1, i]$  are exactly the same and the atomic structure is not considered. Because of this difference, we will be careful in all that follows to distinguish between the positions of the two inner quartet beads,  $(i+1)$  and  $(i+2)$ , and consequently the bond bending angles  $\theta_{i+1}$  and  $\theta_{i+2}$ , respecting the chain directionality.

We are after quartets of beads that form large turn angles  $\tau$ . Ideally, the maximum possible turn angle is  $\tau_{\max} = 180^\circ$ . To understand the general circumstances under which large turn angles are to be expected, we perform calculations in which we create all possible quartet geometries that satisfy steric constraints. In other words, we generate sequences of four points  $(i, i+1, i+2, i+3)$  with a constant bond length  $b=3.81\text{\AA}$ , that yield a fixed distance  $d(i, i+3)$  between the ends of the quartet. In our calculations, for sim-

245 plicity, we place the second point of the quartet ( $i+1$ ) at the origin of the coordinate system resulting in mathematical expressions for the quartet coordinates:

	<b>i=1</b>	<b>i=2</b>	<b>i=3</b>	<b>i=4</b>
$\mathbf{x}_i$	$b \cos \theta_{i+1}$	0	b	$b-b \cos \theta_{i+2}$
$\mathbf{y}_i$	$b \sin \theta_{i+1}$	0	0	$b \sin \theta_{i+2} \cos \mu$
$\mathbf{z}_i$	0	0	0	$b \sin \theta_{i+2} \sin \mu$

250 The bond bending angles  $\theta_{i+1}$  and  $\theta_{i+2}$ , are chosen in the restricted range  $[85^\circ, 180^\circ]$ , while the dihedral angle  $\mu$  is unrestricted and can be in the range  $[-180^\circ, 180^\circ]$  as in protein structures [32]. We find that there is a correlation between the degree of the planarity of the quartet making a turn (equivalent to the dihedral angle  $\mu$  approaching  $0^\circ$ ) and its turn angle  $\tau$ , as illustrated in Figure 3. For a given value of the dihedral angle  $\mu$ , there are distinct combinations of the two bond bending angles ( $\theta_{i+1}, \theta_{i+2}$ ) that yield the same  $d(i, i+3)$  distance yet have a range of turn angles  $\tau$ . For larger  $d(i, i+3)$  distances the number of such quartets increases, as reflected in the wider colored regions. The general tendency observed is increased planarity of the turn ( $\mu \rightarrow 0^\circ$ ) leads to bigger turn angles, as one might intuitively expect.

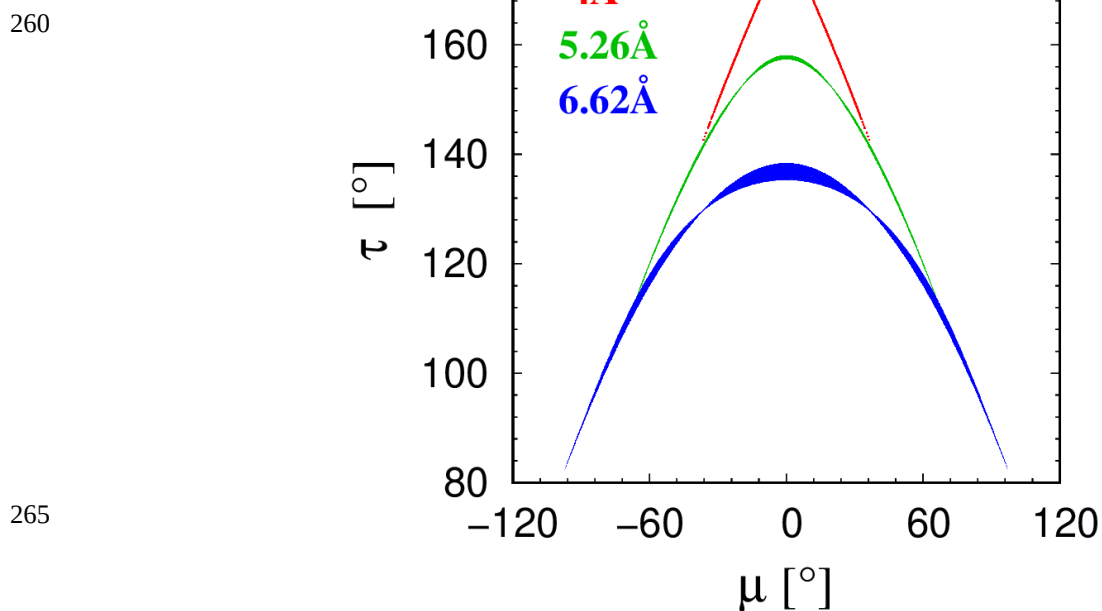


Figure 3: Cross plots of the turn angle  $\tau$  versus the dihedral angle  $\mu$  of quartets having a given fixed distance  $d(i,i+3)$  between the pair of beads  $(i,i+3)$ .

270

## 2.2 Local protein contacts: symmetric poking, one-way poking and no poking

We have shown recently [30,31] that the notion of poking pairwise interactions is a simple yet powerful concept allowing for the development of a unifying framework to understand the formation of both secondary and tertiary protein structure. Here we aim to show that a pairwise poking contact is a key

275

concept in the analysis of protein turns as well. Figure 4 illustrates an arrangement of two pieces of a discrete chain such that a pairwise poking contact is established between two  $C_{\alpha}$  atoms,  $i$  and  $j$ , located at the positions  $\mathbf{r}_i$  and  $\mathbf{r}_j$ , respectively. The point  $i$  is closer to  $j$  than its two neighbors along the chain, and likewise for  $j$ , so that the pieces of the chain effectively ‘poke’ towards each other:

$$d(i,j) < d(i,j-1) \quad \text{Eq. (1a)}$$

$$d(i,j) < d(i,j+1) \quad \text{Eq. (1b)}$$

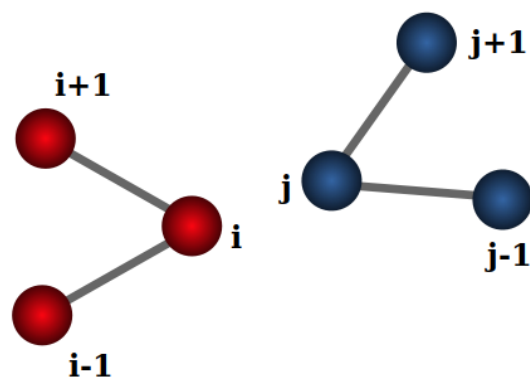
$$d(i,j) < d(i-1,j) \quad \text{Eq. (2a)}$$

$$d(i,j) < d(i+1,j) \quad \text{Eq. (2b)}$$

The term ‘poking contact’ is symmetric in nature. However, one can introduce the concept of asymmetric or ‘one-way’ poking contacts as well. ‘One-way’ poking contacts correspond to just the top two (Eq. (1a) and Eq. (1b)) or the bottom two (Eq. (2a) and Eq. (2b)) holding. Finally, there is ‘no-poking’ contact when at least one of the inequalities Eq. (1a) or Eq. (1b) is not satisfied and at the same time at least one of the two bottom inequalities Eq. (2a) or Eq. (2b), is also not satisfied.



295



300

305

Figure 4: Schematic representation of the relative position of two pieces of a chain  $(i-1,i,i+1)$  and  $(j-1,j,j+1)$  each comprising of three consecutive beads that result in a poking pairwise contact between  $i$  and  $j$ .  $i$  is closer to  $j$  than to the two neighbors of  $j$  and  $j$  is closer to  $i$  than to the two neighbors of  $i$ .

310

The final concept we introduce here is the notion of a ‘good’ contact when a contact is not poking but is close. A good contact is one that is on the verge of becoming a symmetric poking contact. If a decrease of the distance between  $(i,j)$  within  $0.263\text{\AA}$  (or 10% of  $\Delta$ ) is sufficient to make  $(i,j)$  a genuine poking contact, we denote it as a good contact. Otherwise, we call it a bad

contact to indicate that it is not close to being a genuine symmetric poking contact.

315

## 2.3 Data set of native structures of globular proteins

Our protein data set consists of 4,391 globular protein structures from the PDB, a subset of Richardsons' Top 8000 set [33] of high-resolution, quality-filtered protein chains (resolution  $< 2\text{\AA}$ , 70% PDB homology level), that we  
320 further distilled out to exclude structures with missing backbone atoms, as well as amyloid-like structures. Program DSSP (CMBI version 2.0) [34] was used to identify the presence of hydrogen bonds, as well as to determine for each protein residue if it was located within an  $\alpha$ -helical, or  $\beta$ -strand environment or elsewhere.

325

## 2.4 Protein quartets analysis

In this work we are after an objective definition of protein turns, the smallest portions of a protein chain that can effectively change the chain direction. To this end, we have analyzed all 940, 538 quartets present in our data set of  
330 more than 4,000 globular proteins. We used DSSP (CMBI version 2.0) [34] to

divide all these quartets into three disjunctive groups: purely  $\alpha$ -helical quartets (in which all four beads were assigned the flag ‘H’ by DSSP), purely extended  $\beta$ -strand quartets (for which all four beads were assigned the flag ‘E’ by DSSP) and all ‘other’ quartets. In this way we have identified 226,478  $\alpha$ -  
335 helical quartets, 98,648  $\beta$ -strand quartets and 615,412 ‘other’ quartets. The largest class of ‘other’ quartets includes those that are in the beginnings and the endings of the secondary structure elements, as well as those that are in protein loops.

## 340 **2.5 Identification of structural templates in protein turns**

We will demonstrate later in the Results section that the purest turns correspond to isolated poking contacts (meant in the full symmetric sense). We first consider all 21,571 isolated (i,i+3) poking contacts located in 59,464 loops of 4,391 globular protein chains of our data set. We further ensured that  
345 all quartets distilled in this manner have standard (or ‘long’) bonds (whose values are equal to  $3.81\text{\AA}$  within  $\pm 2\%$  [32]). We note that isolated (i,i+3) poking contacts with at least one ‘short’ bond length of  $\sim 3\text{\AA}$  are very few in number, only 228 of them, and will be analyzed separately. The occurrence of

these ‘short’  $C_\alpha - C_\alpha$  pseudo-bonds are due to the rare *cis*-configuration of the  
350 protein backbone (third Ramachandran angle being  $\omega \approx 0^\circ$ , so that the neigh-  
boring  $C_\alpha$  atoms along the chain are on the same side of the planar peptide  
bond), as opposed to the common *trans*-configuration (when  $\omega \approx 180^\circ$ , with  
the neighboring  $C_\alpha$  atoms along the chain on the opposite sides of the planar  
peptide bond, and thus further away) [35].

355 In order to systematically assess the structural variety of protein turns,  
we have analyzed the distribution of 21,571 quartets all having standard  
bonds in the  $(\theta_{i+1}, \theta_{i+2}, \mu)$  three-dimensional space. The distribution of the den-  
sity of quartets in the three-dimensional space can be characterized by identi-  
fying significant local maxima to create structural templates. We separately  
360 treat 18,096 ‘right-handed’ quartets ( $\mu \geq 0^\circ$ ) and 3,475 ‘left-handed’ quartets  
( $\mu < 0^\circ$ ), because it is not possible to superimpose an object on its mirror im-  
age by mere rotations and translations. The optimal binning of the three-di-  
mensional  $(\theta_{i+1}, \theta_{i+2}, \mu)$  structural space of quartets was found to be  $2.5^\circ$  for the  
two bond bending angles  $\theta_{i+1}$  and  $\theta_{i+2}$  and  $5^\circ$  in the dihedral angle  $\mu$ . Three  
365 templates were identified for the right-handed class populated with 394, 132  
and 92 quartets (corresponding to Regions I, III and II, respectively, see the

Results section). For the left-handed class, two local maxima were identified containing 76 and 56 quartets (corresponding to Regions V and IV, respectively). We then worked out the geometries of the five structural templates using the mean values of three angles  $\theta_{i+1}$ ,  $\theta_{i+2}$ , and  $\mu$  in the bins (see Table II in the Results section).

## 2.6 Mapping protein turns

Each of the 21,571 isolated (i,i+3) poking contacts with standard bond lengths was assigned to one of the five basins associated with the five structural templates. For a given protein quartet, the actual values of the three bond lengths (i,i+1), (i+1,i+2) and (i+2,i+3) were measured and five templates were constructed with those values of the three bond lengths, as well as the values of the three mean angles  $\theta_{i+1}$ ,  $\theta_{i+2}$ , and  $\mu$  calculated for a given template in a manner described above. Then a given protein quartet and each template geometry were superimposed and the value of the root-mean-square deviation (RMSD) was calculated using the Visual Molecular Dynamics (VMD) Software package [36]. Finally, by identifying the structural template to which a given protein quartet was closest to (i.e., having the smallest

385 RMSD), the protein quartet was assigned as belonging to that template basin.  
In this way, the map of the protein turns was determined.

## 3 Results

### 3.1 Theoretical considerations

390 In our recent work [30,31], we introduced a concept of poking pairwise  
interactions for a discrete chain, a simple yet powerful concept that allowed  
understanding of the formation of both secondary and tertiary protein  
structure. The geometry of  $\alpha$ -helices is found to be the unique arrangement in  
which local  $(i,i+3)$  poking contacts with special geometrical constraints are  
395 accommodated in a repetitive manner. On the other hand,  $\beta$ -sheets are  
geometries that have pairwise poking non-local  $(i,j)$  contacts with  $(j-i) \geq 4$   
along a pair of coupled  $\beta$ -strands. It was argued that pairwise poking  
interactions also govern, in a harmonious and compatible manner, tertiary  
structure formation, but they are non-repetitive and weaker with no special  
400 geometrical constraints.

Here we will demonstrate that protein turns, the most prevalent non-repetitive structural element in proteins, are also instances of local and isolated (non-repetitive) poking pairwise contacts for which the geometrical constraints are partially relaxed. Indeed, it is intuitively obvious that when  
405 tight turns are not constrained to periodically repeat as in protein  $\alpha$ -helices, they can become more planar and offer a simple rationalization for their observed structural variety and versatility. That this is indeed the case can be seen in the Supplementary Information (Figure SI 1), which shows a detailed comparison between the constraints required for repetitive (completely  
410 embedded) and completely isolated poking contacts in proteins.

### **3.2 What is an objective definition of protein turns?**

We ask the basic question: which type of a consecutive quartet of beads, the minimal portion of the local structure that effectively can turn the chain, can  
415 be thought of as belonging to the ‘true turn’ category? Can one come up with an objective definition of a protein turn? George Rose [28] has pointed out that a  $\beta$ -turn quartet is solvent accessible, it often has a hydrogen bond

between residues 1 and 4 (or equivalently  $i$  and  $i+3$ ) and lies at the protein surface because of the inability of the two middle residues to satisfy their backbone polar groups within the turn without a steric clash. Unlike the building blocks of helices and sheets,  $\beta$ -turns are neither iso-directional nor completely solvent-shielded. Turns can form autonomously and potentially initiate the assembly of scaffold elements by bringing them together with a desired relative placement. Here we observe that the formation of a small piece of a helix is necessarily the only way to nucleate a helix [37-38]. In our way of thinking, this is an isolated local ( $i, i+3$ ) poking contact that, similarly to within helices, is a tight turn but without the baggage of geometrical constraints necessitated by a properly repeating helix. Such pieces are candidates for growing into a fully formed helix. We will argue that when such growth is thwarted for one reason or another, they can also serve, in a flexible manner by tuning their degree of planarity, as one of the ubiquitous classes of turns that allow the protein to become overall compact and break the iso-directionality of a strand or a helix.

We have studied all 940, 538 quartets that are present in our data set of more than 4,000 globular proteins. We use the DSSP program [34] to divide



all quartets into three disjunctive groups: purely  $\alpha$ -helical quartets, purely extended  $\beta$ -strand quartets, and ‘other’ quartets, as explained in the Materials and Methods section. Figure 5 shows the ( $\tau$ - $\mu$ ) cross-plot for these three groups of quartets (in Panels (A), (B) and (C)). Panel (D) shows the same  
440 cross-plot but this time for *all* the quartets, which have an isolated poking contact of (i,i+3). Figure 6 shows the corresponding frequency distributions of the quartet turn angles  $\tau$ . As expected, quartets with an isolated (i,i+3) poking contact do indeed display large turn angles  $\tau$  and they are primarily located in the ‘other’ class of quartets, that is mainly in protein loops [17,23].

445

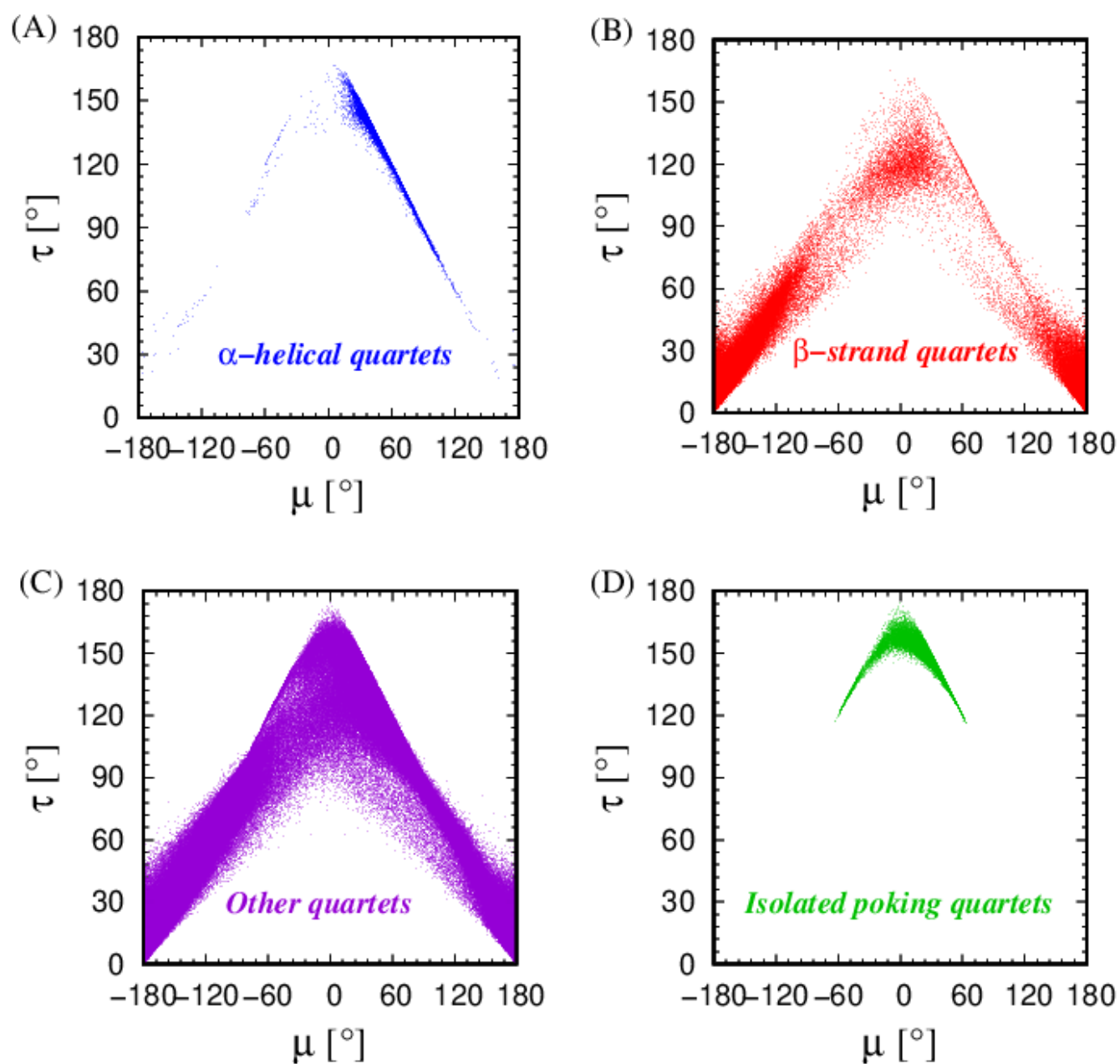
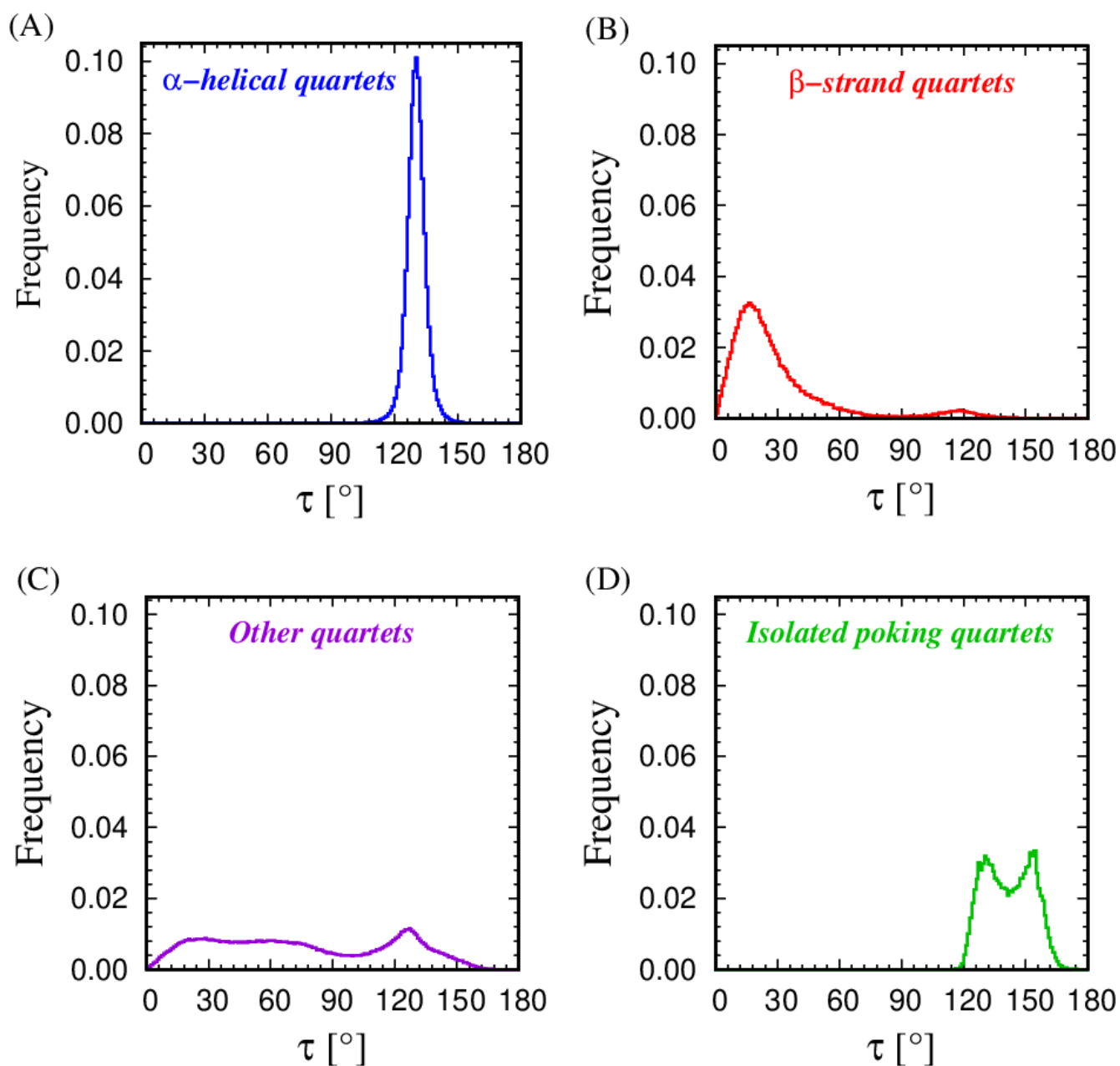


Figure 5: Cross-plots of the turn angles  $\tau$  vs dihedral angles  $\mu$  for the four classes of protein quartets: Panel (A) shows the frequency

450 distribution of the turn angles  $\tau$  in the 226,478  $\alpha$ -helical quartets in our data set (blue histogram), Panel (B) shows the frequency distribution

of the turn angles  $\tau$  in the 98,648  $\beta$ -strand quartets in our data set (red histogram), Panel (C) shows the frequency distribution of the turn angles  $\tau$  in the 615,412 ‘other’ quartets in our data set (purple histogram), and finally in Panel (D) the frequency distribution of the turn angles  $\tau$  in the 25,618 isolated poking quartets in our data set is shown (green histogram). We note that more than 95% of the (i,i+3) isolated poking contacts in our data set (24,349 in all) are in the class of ‘other’ quartets. There are 1,158 (i,i+3) isolated poking contacts in  $\alpha$ -helical quartets and a mere 111 in the  $\beta$ -strand quartets.

465



470 Figure 6: Distribution of the turn angles for the four classes of protein quartets: Panel (A) shows the frequency distribution of the turn angles  $\tau$  in the 226,478  $\alpha$ -helical quartets in our data set (blue histogram),

Panel (B) shows the frequency distribution of the turn angles  $\tau$  in the 98,648  $\beta$ -strand quartets in our data set (red histogram), Panel (C) 475 shows the frequency distribution of the turn angles  $\tau$  in the 615,412 ‘other’ quartets in our data set (purple histogram), and finally in Panel (D) the frequency distribution of the turn angles  $\tau$  in the 25,618 isolated poking quartets in our data set is shown (green histogram).

480 We inspect the distance between quartet ends,  $d(i,i+3)$ , and measure its frequency distributions across different quartet groups, and the result is shown in Figure 7. We see that helical quartets and quartets with isolated  $(i,i+3)$  poking contacts are similar (as expected) and have a  $d(i,i+3)$  distance around the value of the ‘Kepler coin’ diameter  $2\Delta$  [30,31]. The distances for 485 the extended  $\beta$ -strand quartets are understandably the largest among all groups with the ‘other’ being a mixture of the other three behaviors [17,23].

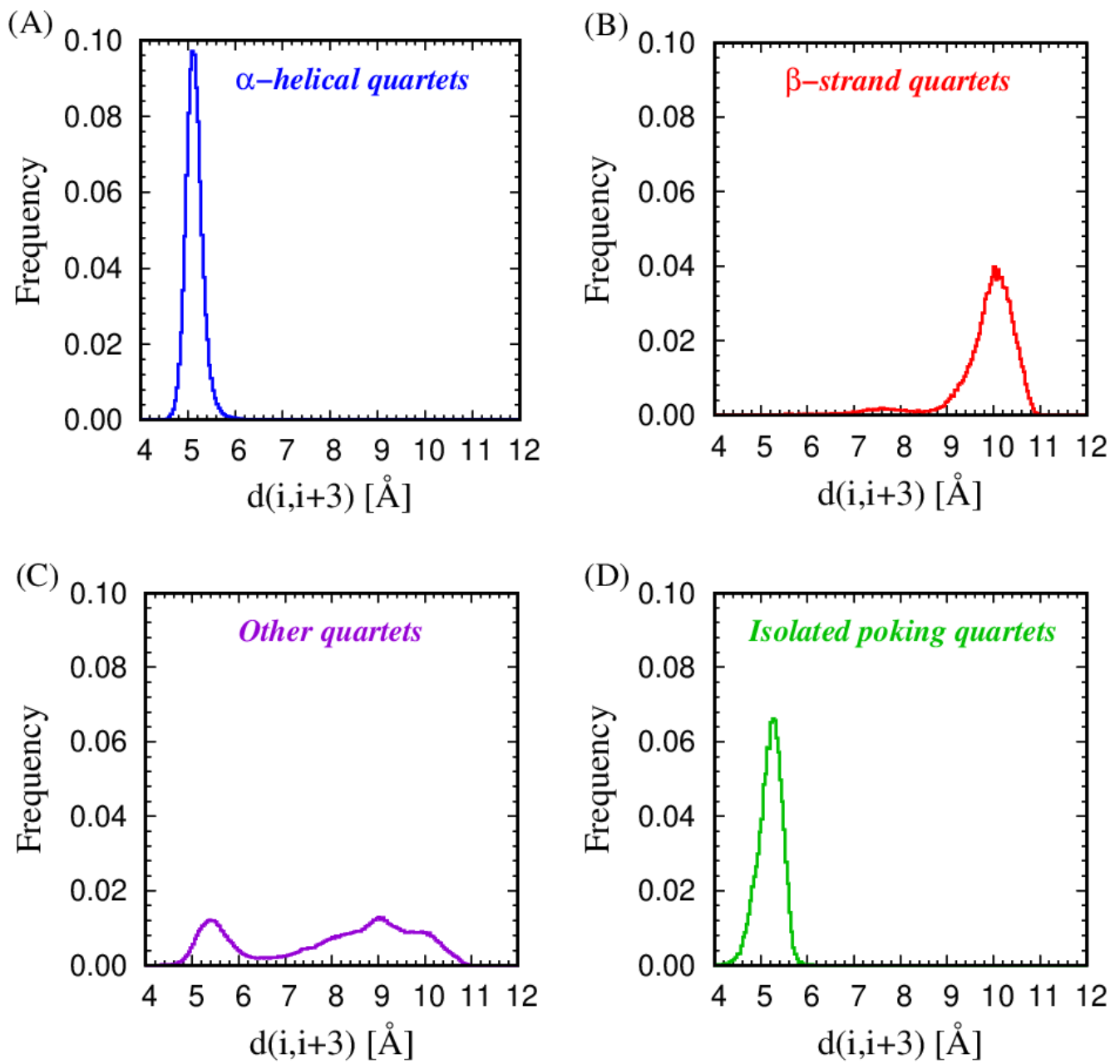


Figure 7: Same as in Figure 6, but for the distribution of frequencies of the  $(i, i+3)$  distances for the four classes of quartets.

The planarity of protein quartets is probed by analyzing the frequency distributions of their dihedral angles  $\mu$ , shown in Figure 8. The local and repetitive helical quartets are decisively non-planar with dihedral angles around  $\sim 50^\circ$ , while extended strand quartets are close to planar with  $|\mu| \sim$   
495  $180^\circ$ . The planar quartets with  $\mu \sim 0^\circ$  are located in the quartets with isolated (i,i+3) poking contacts and are predominantly a subset of the group of ‘other’ quartets. We have verified that the small weight around  $\mu \sim 0^\circ$  even in the group of  $\beta$ -strand quartets is present because the ‘E’ flag assigned by DSSP [34] sometimes includes hairpin loops.

500 Figure 9 shows the distribution of frequencies of the bond bending angles  $\theta_{i+1}$  and  $\theta_{i+2}$  in the four classes of quartets analyzed together. We see that with the exception of extended  $\beta$ -strand quartets, all the rest contain a significant amount of tight helix-like bond bending angles of around  $\sim 90^\circ$ . Our principal finding from the rudimentary analysis is that quartets with  
505 isolated (i,i+3) poking contact are in fact plausible turn candidates and are predominantly located in protein loops [17,23].

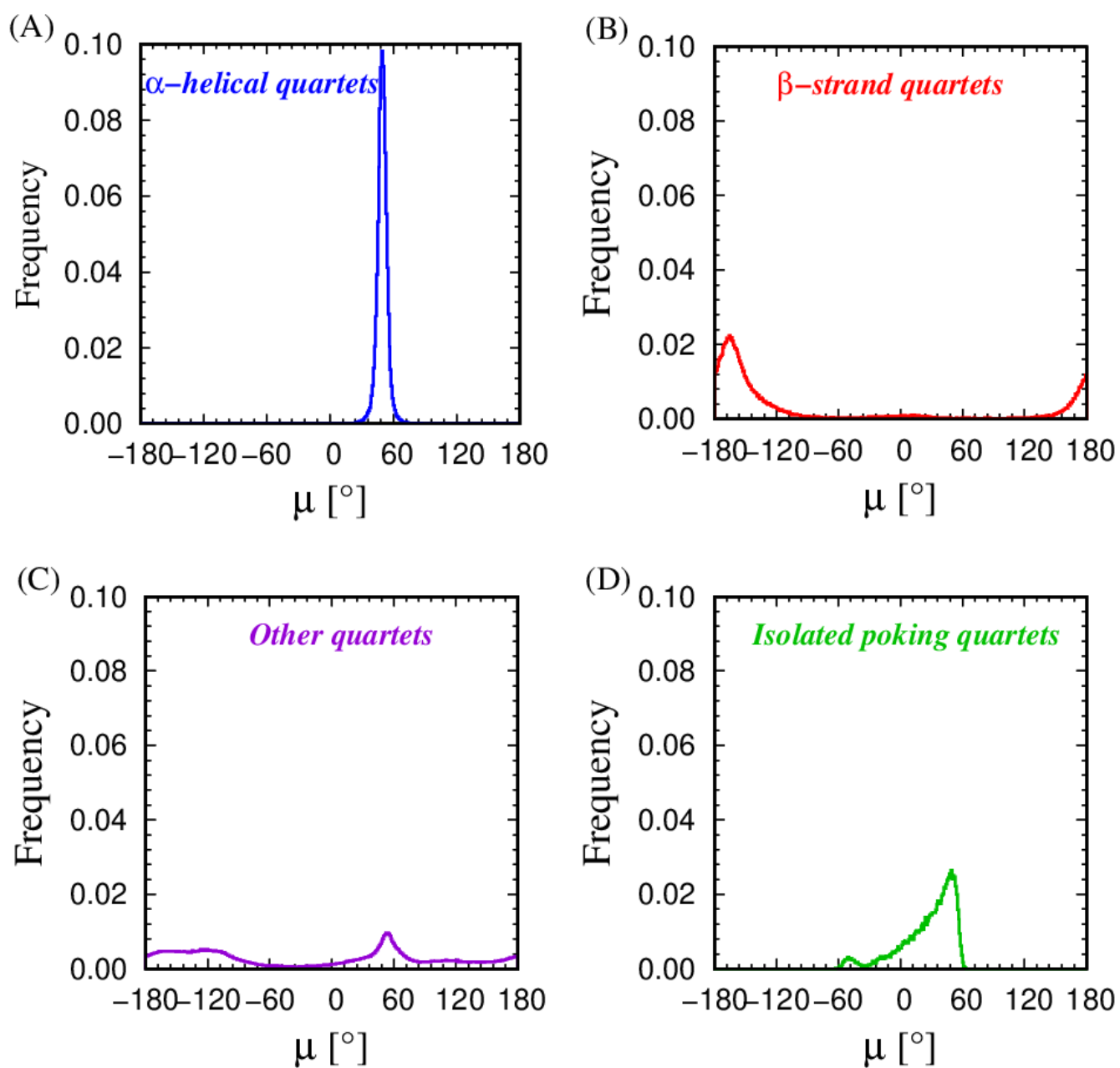
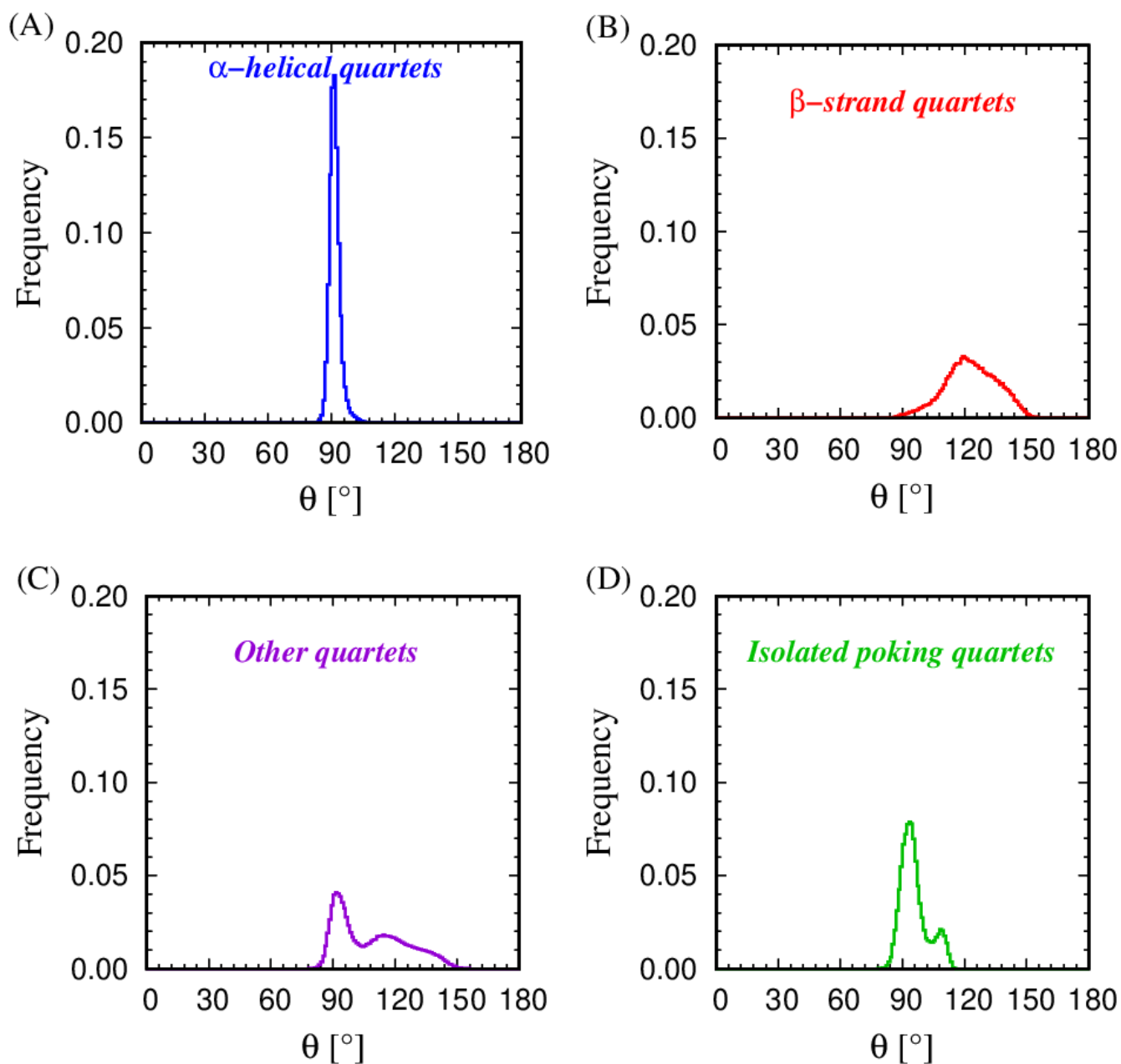


Figure 8: Same as in Figure 7, but for the distribution of frequencies of the dihedral angles  $\mu$  for the four classes of quartets.

510





515 Figure 9: Same as in Figure 8, but for the distribution of frequencies of the bond bending angles in the four classes of quartets. There are two bond-bending angles in a quartet and they are both treated together in this analysis.

We now proceed to study the isolated (i,i+3) poking contacts located  
520 within protein loops [17,23], thereby excluding those that are in the  
beginning and endings of helices and sheets. We define a protein loop as a  
segment of the protein that connects two elements of protein structure along  
the chain, namely  $\alpha$ - $\alpha$ ,  $\beta$ - $\beta$ ,  $\alpha$ - $\beta$ , or  $\beta$ - $\alpha$ . Approximately 85% of isolated  
(i,i+3) poking contacts are located in protein loops. The rest occur at the very  
525 beginning or ending of a protein, as well at the edges of secondary structure  
elements. In total, there are 59,464 loops in our proteins out of which 8,675  
(14.6%) have length of four (implying that there are two ‘O’ (beads other  
than ‘E’ or ‘H’, located in strands and helices, respectively) beads at positions  
2 and 3). There are 7,557 loops of length five (12.7%), 8,395 loops of length  
530 six (14.1%), 7,829 loops of length 7 (13.2%), 5,379 loops of length eight  
(9.0%), 4,071 loops of length nine (6.8%), and 3,245 loops of length 10  
(5.5%). Thus, we find that the number of loops decreases approximately  
monotonically with the length of the loop, a trend noted previously in Ref.  
[17,23].

535 We are now ready to take the next step in our search for the objective  
definition of protein turns. To this end, we discuss the detailed topology of  
the (i,i+3) contacts in protein quartets located in protein loops, in terms of  
whether they are symmetric poking contacts and, if not, whether they are  
good or bad (as per the definition introduced in Materials and Methods  
540 Section 2.2). Figure 10 shows the frequency distributions of the turn angles  $\tau$   
in protein quartets in protein loops, divided in different classes according to  
their topological characteristic. Panel (A) of Figure 10 underscores the sharp  
difference in the turn angle distributions for good and bad contacts with no  
poking at all -- the good ones decisively have a larger turn angle. Panel (B) of  
545 the same Figure shows that even for 'one-way' poking contacts, the good  
ones have a larger weight for bigger turn angles. Finally, Panel (C) shows that  
among symmetric poking contacts, the isolated ones turn the chain more  
sharply than the non-isolated ones.

550

555

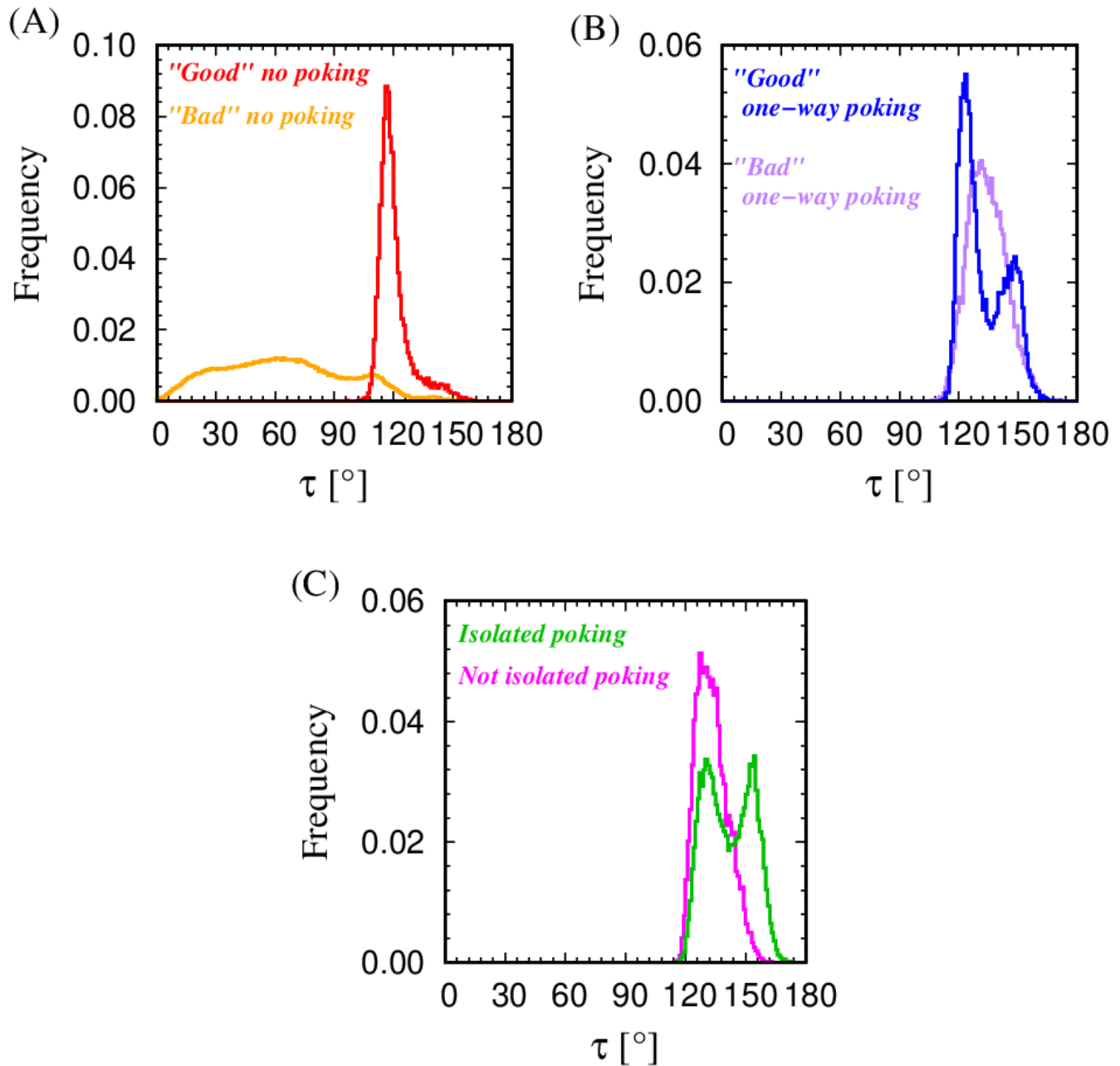


Figure 10: Turn angle distributions. Panel (A) shows the frequency distribution of the turn angle  $\tau$  for the classes of 11,954 'good' and

282,813 ‘bad’ (i,i+3) contacts in protein loops with no poking at all (in red and orange). The contact is classified as ‘good’ if its effective  
560 ‘distance’ from the symmetric poking condition is less than 10% of  $\Delta$  (0.263Å). Panel (B) shows the frequency distribution of the turn angle  $\tau$  for the class of 17,721 ‘good’ and 11,954 ‘bad’ (i,i+3) contacts in protein loops that have a one-way poking contact (in blue and violet respectively). Finally, Panel (C) shows the frequency distribution of the  
565 turn angle  $\tau$  for the 21,571 (i,i+3) isolated poking contacts (fully symmetric) as well as 6,886 (i,i+3) isolated poking contacts that are not fully isolated (‘embedded from one side’) in protein loops (in green and magenta).

570 Table I shows the number and percentages of turns belong to the ‘topological’ classes with respect to their (i,i+3) ‘poking’ status. In this way, we find that the purest class of sharp protein turns is indeed the one making isolated (i,i+3) poking contacts.

575

<b>Type of (i,i+3) contacts</b>	<b>All</b>	<b>Loops</b>
	<b><math>\tau &gt; 160^\circ</math></b>	<b><math>\tau &gt; 160^\circ</math></b>
<b>Isolated poking</b>	1,042	959
<b>Not isolated poking</b>	3	1
<b>‘Good’ one-way poking</b>	96	74
<b>‘Bad’ one-way poking</b>	85	53
<b>‘Good’ no poking</b>	34	7
<b>‘Bad’ no poking</b>	14	2
	<b>Total: 1,274</b>	<b>Total: 1,096</b>

580 Table I: Number of quartets with large turn angles,  $\tau > 160^\circ$ , divided  
into classes of (i,i+3) contacts. ‘All’ considered all 940,538 quartets in  
all our proteins whereas ‘Loops’ was restricted to 357,989 protein  
quartets just within loops (see Materials and Methods Section 2.4 for a  
description). The clear result is that large turn angles are  
predominantly associated with isolated poking contacts, the nuclei for  
585 the putative formation of protein helices.

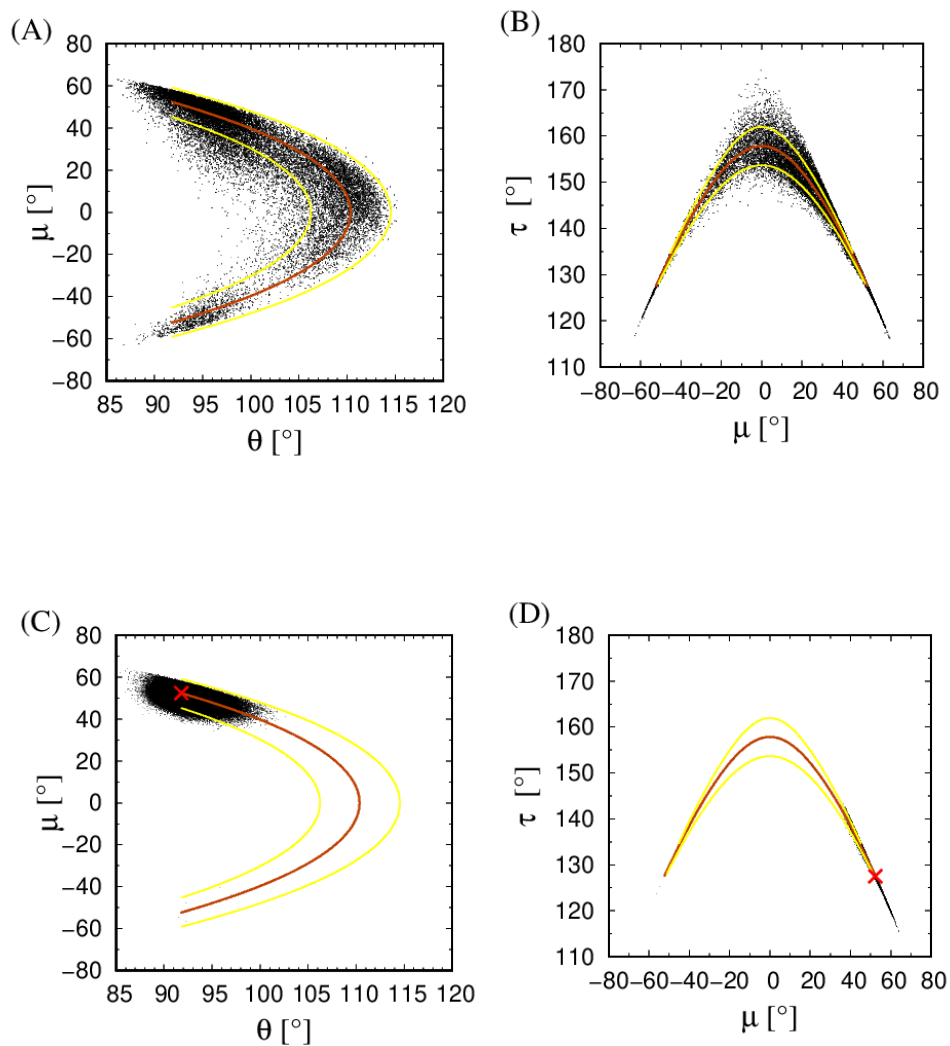
### 3.3 Geometrical classes of protein turns

Figure 11 depicts various cross-plots of geometrical attributes of protein turns defined as quartets of beads with  $(i,i+3)$  isolated poking contact (in Panels (A) and (B)), and contrast them with the geometrical attributes of quartets in the class of embedded (repetitive)  $(i,i+3)$  poking contacts (Panels (C) and (D)). The results of theoretical estimates are also shown alongside (see Figure caption for a description). There is a clear contrast between the two classes: quartets with ‘isolated’  $(i,i+3)$  poking contacts explore a much wider region of phase space and are much less constrained than the embedded helical contacts, which are restricted by the need for repetition.

600

605

610



615

620

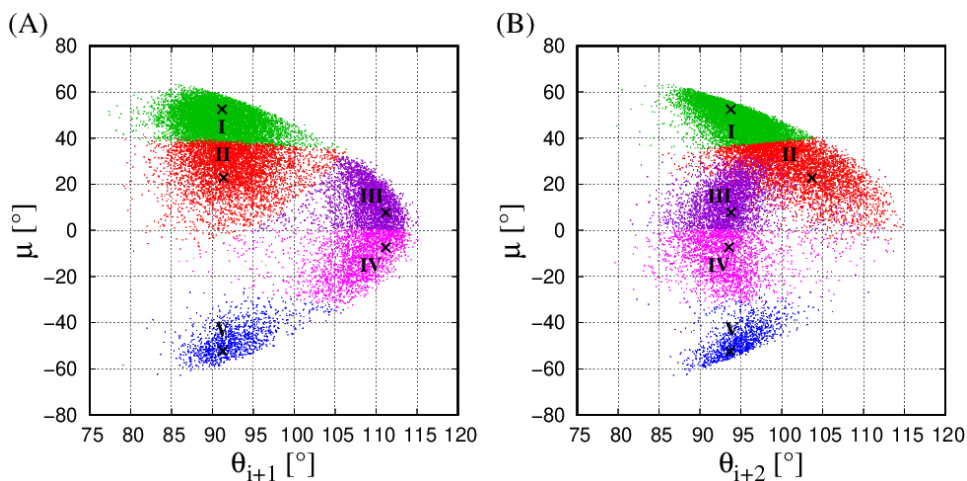
Figure 11: Panel (A) shows the  $(\theta, \mu)$  cross plot of the full set of 21,571 isolated  $(i, i+3)$  symmetric poking contacts located in the 59,464 loops of 4,391 globular protein chains of our data set, along with the



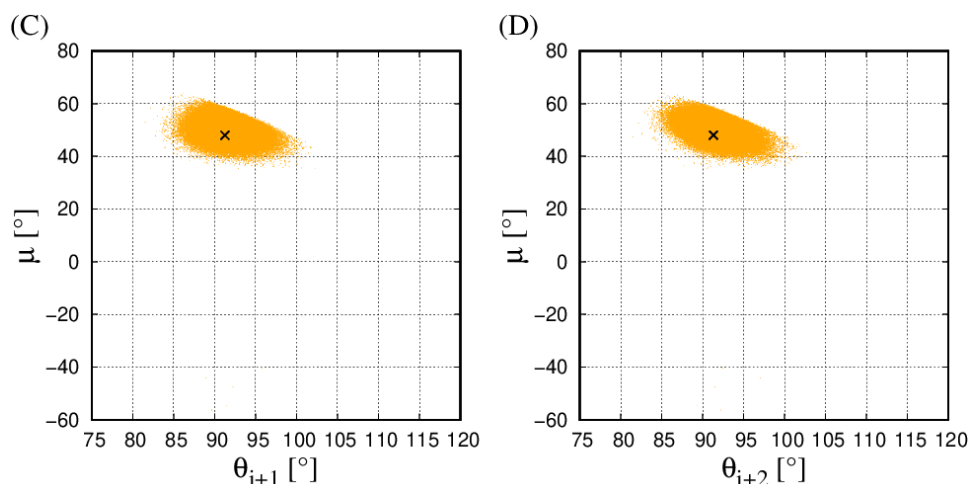
theoretical prediction curve (in dark orange color) for the dependence of the dihedral angle  $\mu$  of the turn and the larger of the two bond bending angles  $\theta$ , [that is  $\theta = \theta_1$ , if  $\theta_1 > \theta_2$ , and  $\theta = \theta_2$ , if  $\theta_2 > \theta_1$ ]. Panel (B) shows the  $(\mu, \tau)$  cross plot of the full set of 21,571 isolated  $(i, i+3)$  symmetric poking contacts located in the 59,464 loops of 4,391 globular protein chains of our data set. Panels (C) and (D) depict the same quantities as (A) and (B), but for the 176,711 embedded  $(i, i+3)$  symmetric poking helical contacts in 4,391 globular protein chains of our data set. The red X symbols in both panels (C) and (D) denote the corresponding values for the Kepler helix. The theoretical predictions are derived from the condition that the distance between the first and the fourth bead in the turn  $(i, i+3)$  equals  $2\Delta = 5.26\text{\AA}$  (the ideal distance for a Kepler helix), when one bond bending angle is fixed to correspond to the turn angle of a Kepler helix,  $91.8^\circ$ . The yellow curves in both panels are theoretical predictions when a  $\pm 5\%$  difference is allowed in the theoretical  $2\Delta$  distance  $d(i, i+3)$ .

As explained in Sections 2.5 and 2.6 of Materials and Methods, a  
640 quartet with standard bonds is assigned one of five representative structural  
templates. The five-colored map obtained in this way is shown in Panels (A)  
and (B) of Figure 12. Panels (C) and (D) show the case of quartets with  
embedded (i,i+3) poking contacts (helical turns) that are described with just  
one local maxima and one structural template. Region I (green) of protein  
645 turns corresponds to the unrelaxed helical turn. Figure 13 (A) shows the  
unique geometrical template for quartets containing embedded (i,i+3) poking  
contacts and correspond to the black X symbols in Panels (C) and (D) of  
Figure 12. The five structural templates of protein turns are shown in Panels  
(B-F) of Figure 13.

650



655



660

665

670

Figure 12: Panel (A) shows the map of the two dimensional projection in the  $(\theta_{i+1}, \mu)$  plane of the full set of 21,571 isolated  $(i,i+3)$  symmetric poking contacts located in the 59,464 loops of 4,391 globular protein chains of our data set. The black X symbols denote the locations of the five local maxima in the three-dimensional space of the three angles  $(\theta_{i+1}, \theta_{i+2}, \mu)$  that have the largest density of points with the binning of  $2.5^\circ, 2.5^\circ$ , and  $5^\circ$  in the three angles, respectively. Each of the 21,571 isolated  $(i,i+3)$  symmetric poking contacts is assigned to one of the five templates depicted in Panels (B)-(F) of Figure 13, based on the criterion of the smallest RMSD value from the winning template. The basin represented with Region I is found to contain

9,604 quartets (green points), Region II is found to contain 5,271  
675 quartets (red points), Region III is found to contain 3,237 quartets  
(purple points), Region IV is found to contain 2,298 quartets (magenta  
points), and finally Region V is found to contain 1,161 quartets (blue  
points). Panel (B) shows the cartography in a complementary two-  
dimensional plot in the  $(\theta_{i+2}, \mu)$  plane. Panels (C) and (D) depict the  
680 data for 176,711 embedded  $(i, i+3)$  symmetric poking contacts in 4,391  
globular protein chains of our data set (points in orange color). The  
black X symbols in Panels (C) and (D) denote the locations of the  
local maxima in the three-dimensional space of the three angles  $(\theta_{i+1},$   
 $\theta_{i+2}, \mu)$  with the binning of  $2.5^\circ, 2.5^\circ$ , and  $5^\circ$  in the three angles,  
685 respectively.

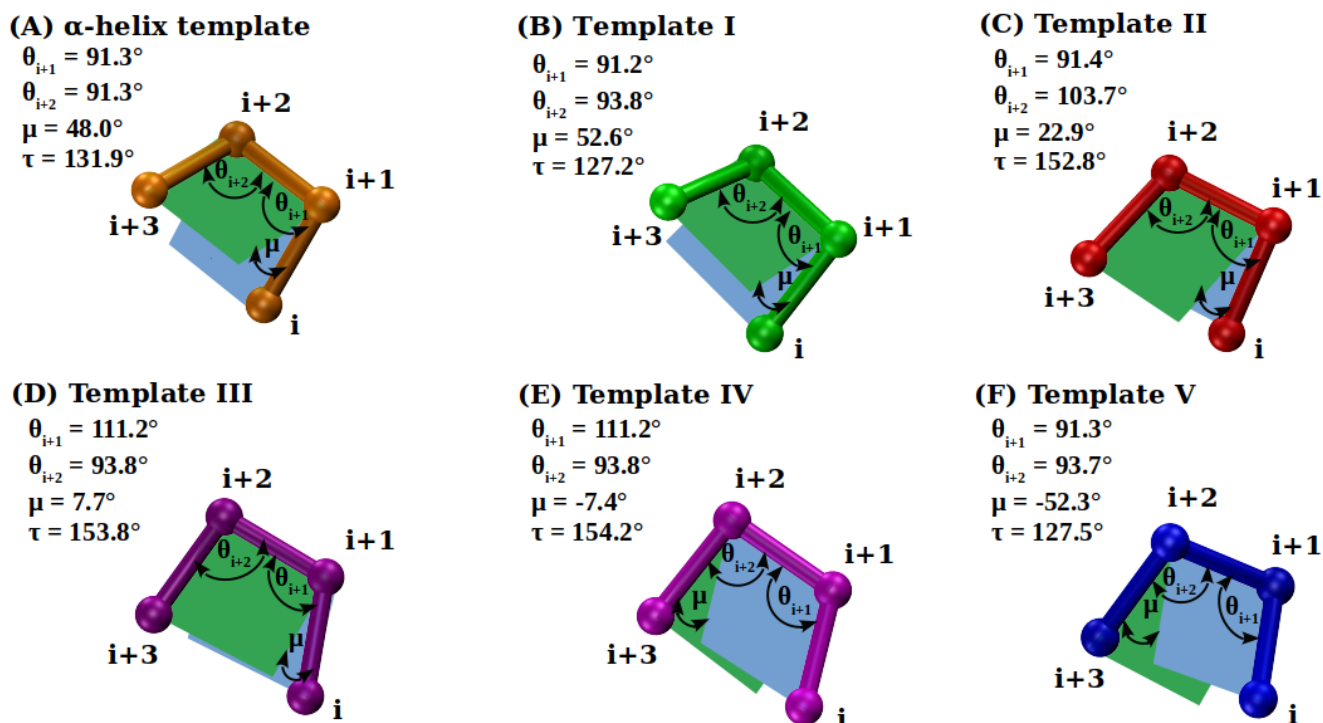


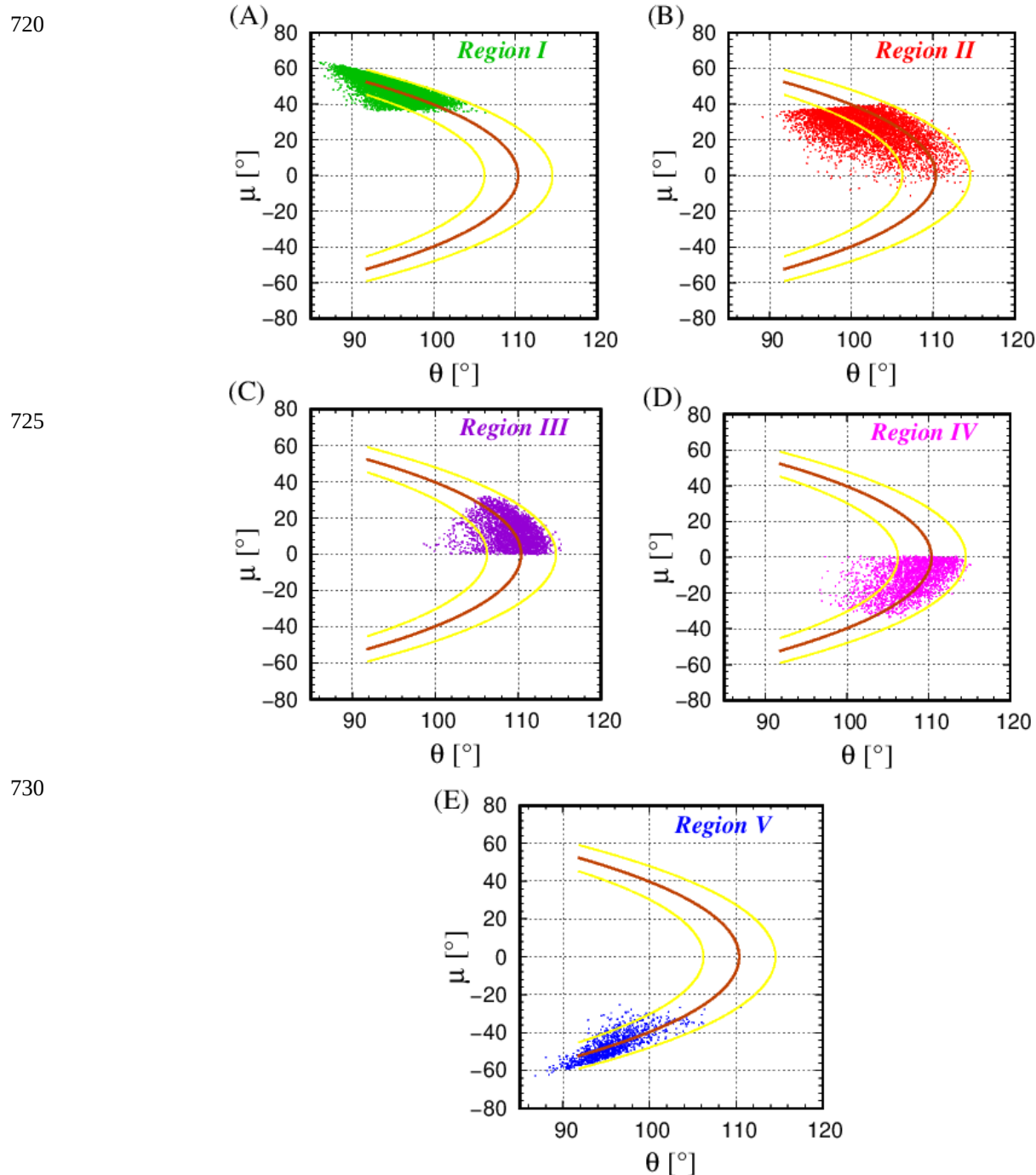
Figure 13: Panels (A) shows the single structural template for the embedded symmetric (i,i+3) poking contact characteristic of protein  $\alpha$ -helices. Panels (B-F) show five structural templates determined for the local turns with isolated symmetric (i,i+3) poking contacts. The spheres represent the locations of four consecutive  $C_\alpha$  atoms. The values of the bond bending angles  $\theta_{i+1}$ ,  $\theta_{i+2}$ , dihedral angle  $\mu$ , and the turn angle  $\tau$ , for each of the six templates reported in the figure correspond to the mean values calculated over the three-dimensional bin associated with each of the identified five local maxima. The three bond lengths are all drawn with length  $b$  of  $3.81\text{\AA}$ . The planes

containing points  $(i,i+1,i+2)$  are shown in blue, while the planes  
containing points  $(i+1,i+2,i+3)$  are shown in green. When the green  
plane is above the blue plane, the quartet of points is right-handed (it  
700 has a positive value of the dihedral angle  $\mu$ ), whereas when the blue  
plane is above the green plane the quartet of points is left-handed (it  
has a negative value of the dihedral angle  $\mu$ ).

Figures 14 and 15 show the  $(\theta,\mu)$  and  $(\mu,\tau)$  cross plots, respectively, of  
705 the five structural templates of protein turns alongside the theoretical  
estimates. Interestingly, most turn geometries lie within a  $\pm 5\%$  tolerance  
around the curve corresponding to the theoretical distance  $2\Delta$  between the  
quartet ends,  $d(i,i+3)$ . The exception to this behavior is in Region II (red  
points) that represents a more planar version of the helical turn having a  
710 smaller value of the end-to-end quartet distance of  $5.07\text{\AA}$  (see Table II). This  
turn type has a lower percentage of hydrogen bonds [24] associated with its  
ends.

Figures 16 shows Ramachandran plots  $(\varphi_{i+1}, \Psi_{i+1})$  and  $(\varphi_{i+2}, \Psi_{i+2})$  [29]  
for quartets with the embedded  $(i,i+3)$  contacts, with black X symbols

715 showing their mean values [3]. Figure 17, Panels (A-E) shows Ramachandran  
plots  $(\phi_{i+1}, \Psi_{i+1})$  and  $(\phi_{i+2}, \Psi_{i+2})$  for the five geometrical regions identified in  
protein turns, along with the black X symbols, at the positions of the  
conventional mean values for Type I and Type II turns, along with their  
mirror images (Types I' and II') predicted by Ventakachalam [3].

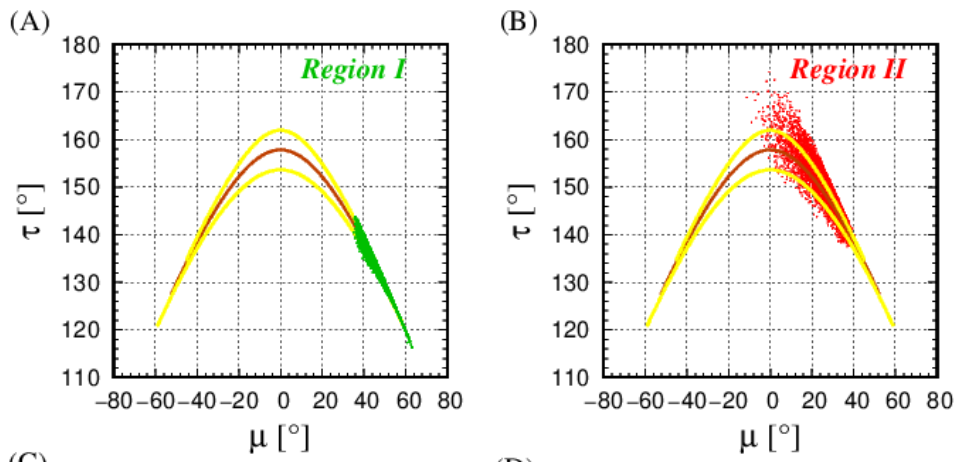


735 Figure 14: Panels (A-E) show separate  $(\theta, \mu)$  cross plots for the five  
regions for 21,571 total isolated  $(i, i+3)$  symmetric poking contacts  
located in the 59,464 loops of 4,391 globular protein chains of our  
data set, along with the theoretical prediction curve (in dark orange  
color) for the dependence of the dihedral angle  $\mu$  of the turn and the  
740 larger of the two bond bending angles  $\theta$ , [that is  $\theta = \theta_1$ , if  $\theta_1 > \theta_2$ , and  
 $\theta = \theta_2$ , if  $\theta_2 > \theta_1$ ]. The theoretical predictions are derived from the  
condition that the distance between the first and the fourth bead in the  
turn  $(i, i+3)$  equals  $2\Delta = 5.26\text{\AA}$  (the ideal distance for a Kepler helix),  
when one bond bending angle is fixed to correspond to the tight turn  
745 angle of a Kepler helix,  $91.8^\circ$ . The yellow curves in both panels are  
theoretical predictions when a  $\pm 5\%$  difference is allowed in the  
theoretical  $2\Delta$  distance  $d(i, i+3)$ .

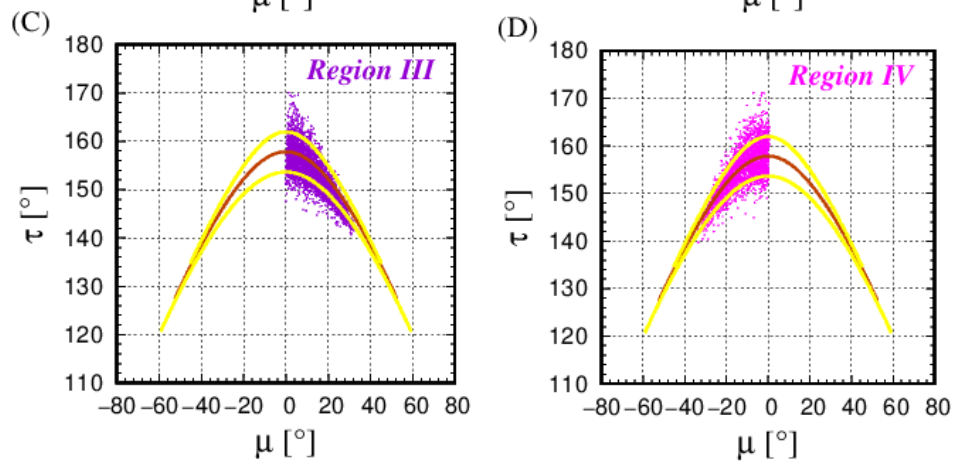
750



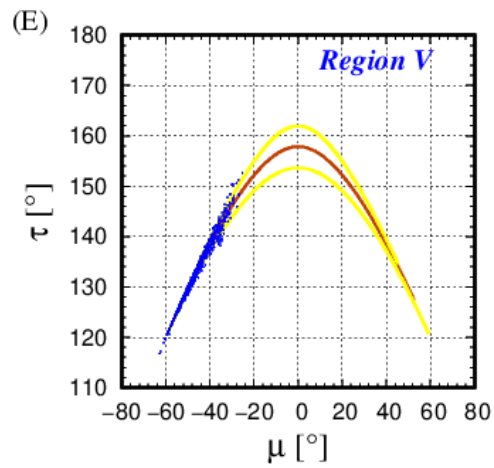
755



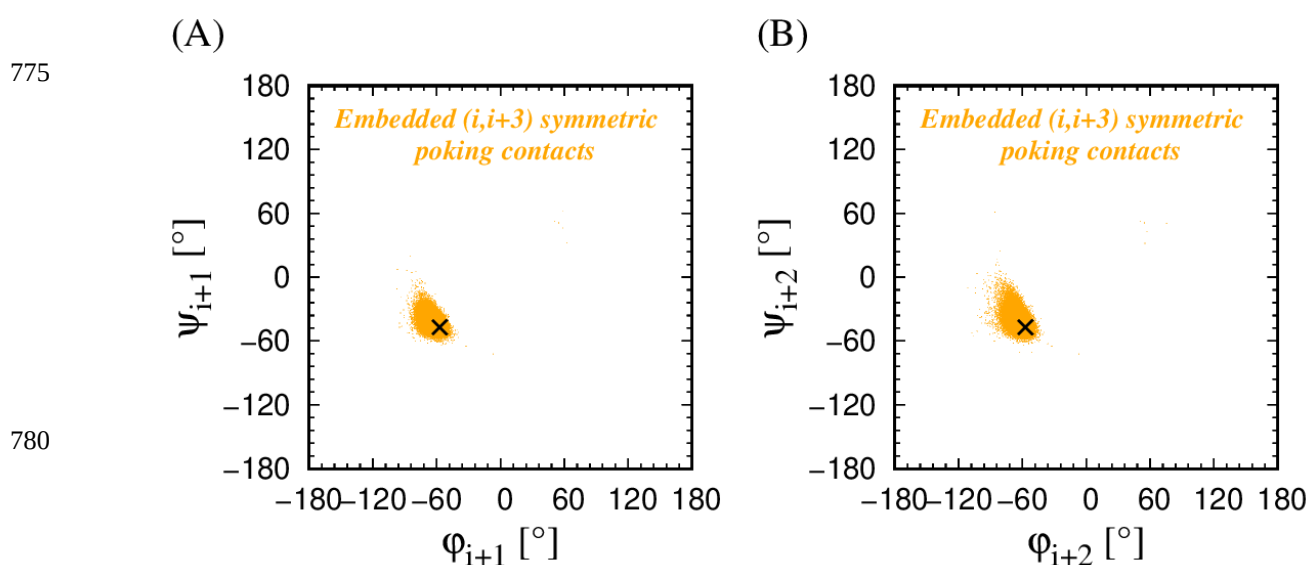
760



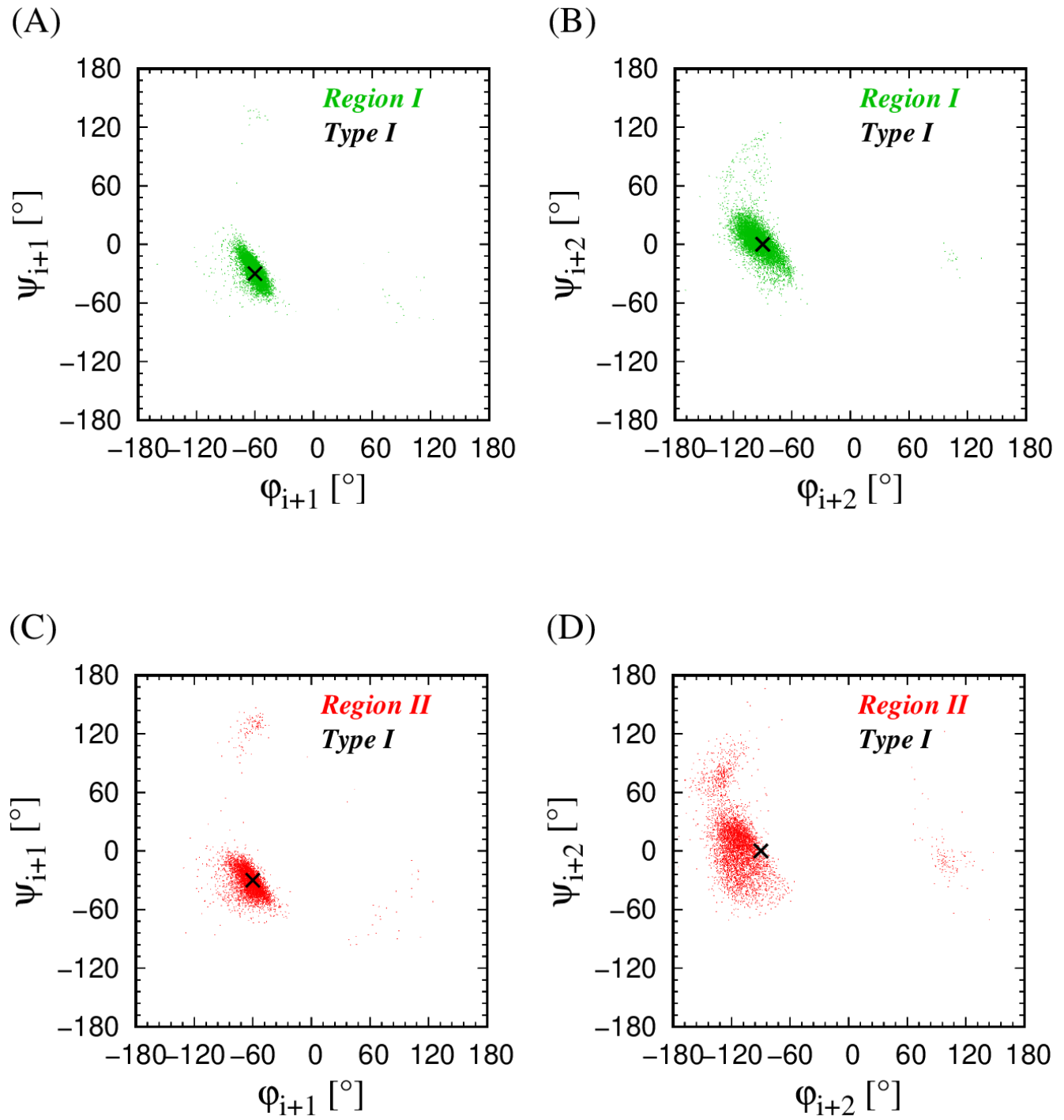
765



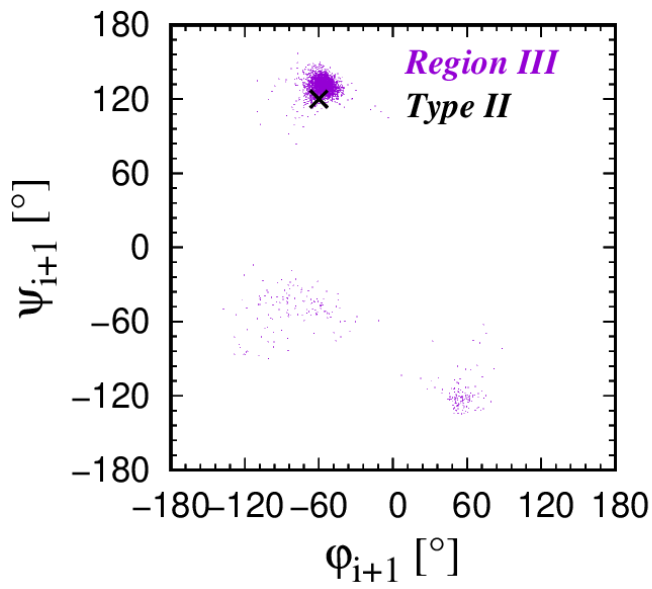
770 Figure 15: Panels (A-E) show the  $(\mu, \tau)$  cross plots for each of the five regions of the full set of 21,571 isolated  $(i, i+3)$  symmetric poking contacts located in 59,464 loops of 4,391 globular protein chains of our data set, along with the theoretical estimate curves (in dark orange color).



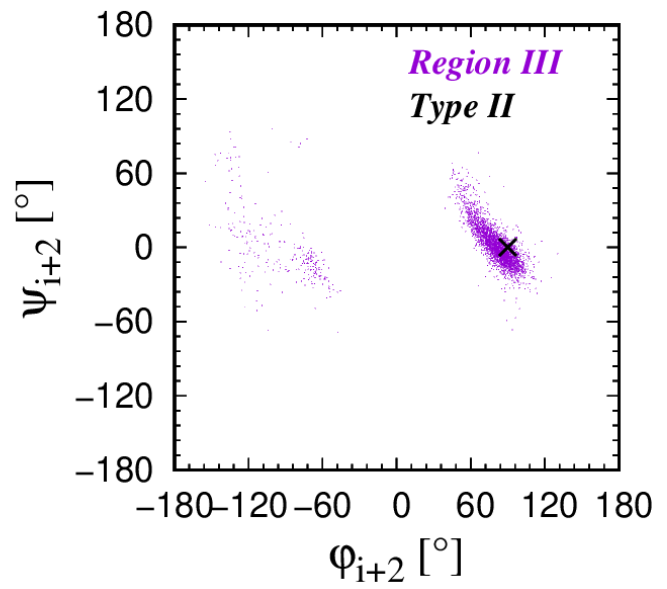
785 Figure 16: Ramachandran plots [29] for the beads  $(i+1)$  and  $(i+2)$  in the middle of the quartet  $(i, i+1, i+2, i+3)$ , when  $(i, i+3)$  is embedded symmetric poking contact,  $(\phi_{i+1}, \Psi_{i+1})$  and  $(\phi_{i+2}, \Psi_{i+2})$ , are shown in panels (A) and (B), respectively. The black X symbols indicate typical average values for protein  $\alpha$ -helices indicated in the literature.



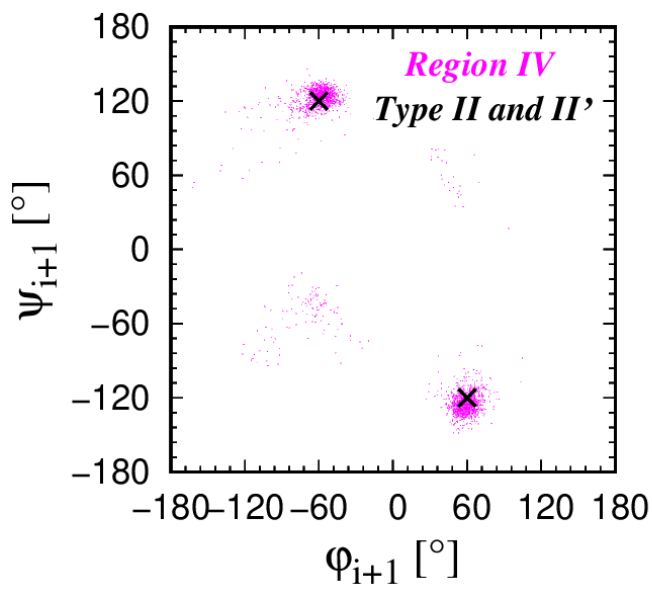
(E)



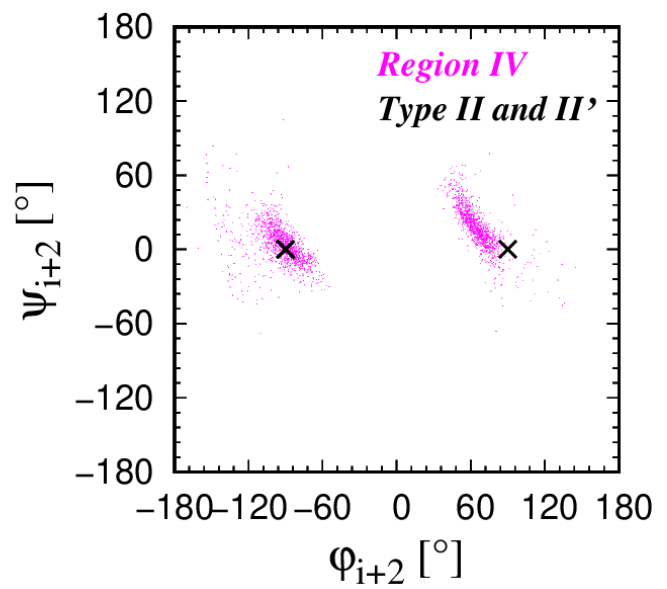
(F)

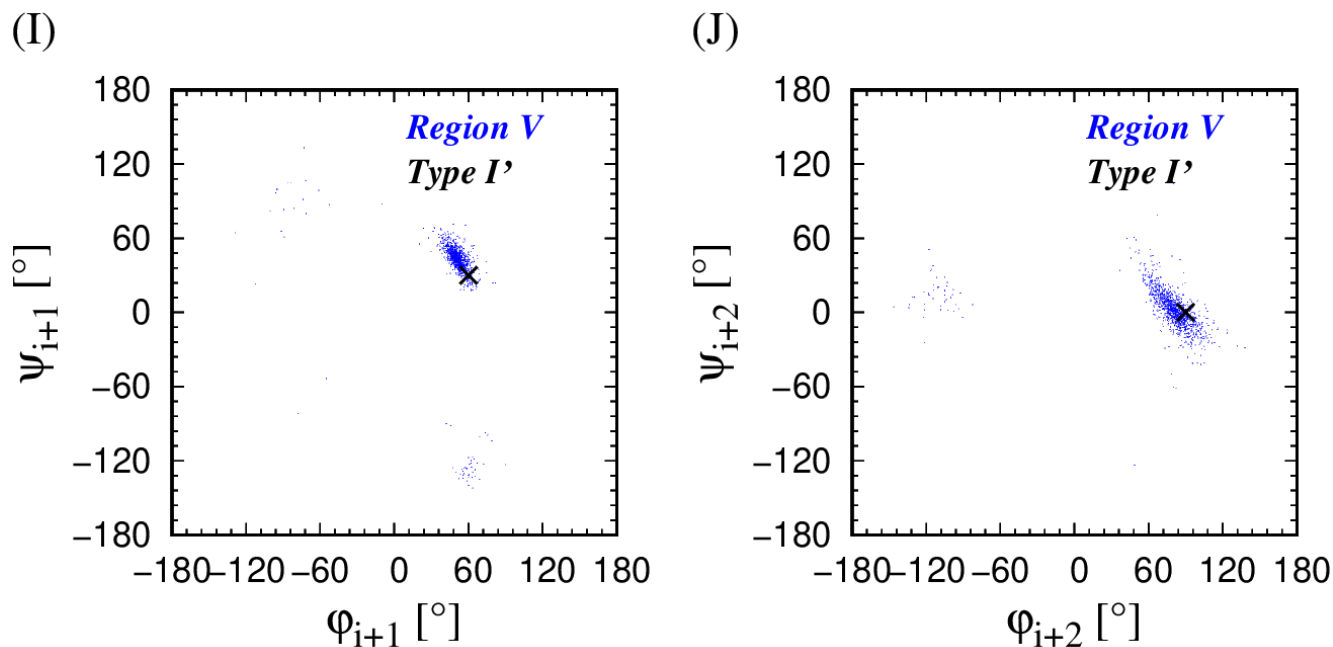


(G)



(H)





800 Figure 17: Panels (A-J) show the Ramachandran plots [29] for the  
beads (i+1) and (i+2) in the middle of the quartet (i,i+1,i+2,i+3) for the  
five map regions of the (i,i+3) isolated symmetric poking contacts  
within protein loops. The black X symbols indicate typical average val-  
ues [3].

805

	<b>Embedded poking contacts</b>	<b>Region I</b>	<b>Region II</b>	<b>Region III</b>	<b>Region IV</b>	<b>Region V</b>
<b>d(i,i+3) [Å]</b>	5.12 +/- 0.14	5.22 ± 0.20	5.07 ± 0.29	5.36 ± 0.22	5.25 ± 0.24	5.27 ± 0.17
<b>θ<sub>i+1</sub> [°]</b>	91.35 +/- 2.05	90.91± 3.52	92.49 ± 4.21	108.89±2.75	107.39 ± 4.06	92.57 ± 3.62
<b>θ<sub>i+2</sub> [°]</b>	91.33 +/- 2.11	94.63 ± 3.16	101.42±4.81	93.60 ± 3.46	93.21 ± 4.20	94.75 ± 2.95
<b>μ [°]</b>	49.78 +/- 3.52	48.07 ± 5.62	26.22 ± 8.85	12.66 ± 7.66	-12.03 ± 8.34	-46.96 ± 6.86
<b>τ [°]</b>	130.07 +/- 3.37	131.42 ± 5.24	149.35±6.16	153.42±4.70	155.00 ± 4.83	132.19 ± 6.09
<b>RMSD [Å]</b>	0.082 +/- 0.039	0.135 ± 0.066	0.183±0.071	0.144±0.074	0.172 ± 0.088	0.135 ± 0.087
<b>Hydrogen bonded [%]</b>	99.99	97.6	86.09	96.57	96.56	98.54

810

Table II: Mean values with standard deviations of various geometrical characteristics calculated for the embedded (i,i+3) symmetric poking contacts, as well as for the five regions of local (i,i+3) isolated symmetric poking contacts located within protein loops. (i,i+3) contact is considered to be hydrogen bonded if at least one hydrogen bond of the type (i,j), (i-1,j),(i+1,j),(i,j-1), or (i,j+1) is present.

815

### 3.4 Amino acid specificities

We studied the occurrence of 20 amino acid types in the inner positions (i+1) and (i+2) of 176,711 (i,i+3) *embedded symmetric poking contacts*, as well as the most over-expressed (i+1,i+2) amino acid pairs. We find that there are in total eight amino acids that are over-expressed with the normalized frequencies ranging from 1.12 to 1.56. We find that there is an equal participation of

820

hydrophilic and hydrophobic amino acids: hydrophilic (GLU, GLN, ARG,  
825 LYS) and hydrophobic (ALA, LEU, MET, ILE). More specifically, ALA,  
LEU, and ILE belong to the aliphatic group (non-polar amino acids with lin-  
ear or branched sidechains), GLU is negatively charged, ARG and LYS are  
positively charged, GLN is amidic (polar, uncharged, with R amino group)  
and MET is a sulfur-containing hydrophobic residue. The same eight amino  
830 acids are over-expressed at both positions (i+2) and (i+3). The over-ex-  
pressed amino acid pairs at positions (i+1,i+2) are made up uniformly from  
the above over-expressed individual amino acids. We direct an interested  
reader to Supplementary Information for more details.

Our analysis of the amino acid specificities in the *isolated poking con-*  
835 *tacts* in protein loops reveals certain trends: **In Region I (green in the map),**  
the predominant choice for the (i+1) position is PRO with normalized fre-  
quency > 3, whereas for the (i+2) position, the preferred amino acids are  
ASN and ASP both with normalized frequency of ~3, and the over expressed  
(i+1,i+2) amino acids pairs are naturally (PRO,ASP) and (PRO,ASN). **Re-**  
840 **gion II (red in the map)** has the same preferences as Region I, with the only  
difference being a smaller over-expression normalized frequency of the PRO

residue at the position (i+1) dropping to  $\sim 2$ . **In Region III (purple color in the map)**, the predominant choice for the (i+1) position is again PRO with an even larger normalized frequency of  $\sim 4$ , whereas GLY is *highly* preferred in position (i+2) with normalized frequency of  $\sim 9$ . The most over-expressed (i+1,i+2) amino acids pairs all have GLY at the position (i+2). **In Region IV (magenta color in the map)**, the predominant choice for the (i+1) position is GLY with normalized frequency of  $\sim 4$  and PRO follows with a normalized frequency of  $\sim 2$ . In the (i+2) position ASN, GLY, and ASP amino acids are favored with normalized frequencies of  $\sim 3$ . Note the ‘reversed’ preferences in positions (i+1) and (i+2) for Regions III and IV, that are roughly the geometrical mirror images of each other. **In Region V (blue color in the map)**, the mirror image of Region I, the PRO residue is significantly *under-expressed* in the position (i+1) with normalized frequency of only 0.02 [in Region I it was *over-expressed* with normalized frequency  $> 3$ ]. The predominant choices for the (i+1) position in **Region V (blue color in the map)** are ASN with normalized frequency of  $\sim 5.5$ , as well as ASP with normalized frequency of  $\sim 3.5$ , while the *highly preferred* amino acid in the position (i+2) is predominantly GLY with normalized frequency of  $\sim 10$ . In all five regions identified



860 in protein turns, all over-expressed amino acids are hydrophilic. For more detailed information we direct an interested reader to Supplementary Information (Tables SI II to Table SI VII).

We have also studied the amino acid preferences in the 228 local (i,i+3) isolated symmetric poking contacts within protein loops with at least one  
865 short bond. We find that in the position (i+1) PRO and ASN residues are preferred having a normalized frequency of  $\sim 4$ , while, for the (i+2) position, we find *overwhelming preference* for a PRO residue with a normalized frequency of  $\sim 16$ . In 75% of all cases, the short bond is the (i+1,i+2) bond.

## 870 **4 Conclusions**

While protein secondary structure elements are repetitive and thus iso-directional, the basic role of protein turns is to provide an effective and advantageous way to turn the chain back to itself, changing its direction and thus defining protein boundaries. Turns assist in ensuring the compactness of the  
875 globular protein structure. They can be viewed as the structural feature that literally unifies the structure of a globular protein as a whole: helices are made of tight turns locally repeating periodically; non-local  $\beta$ -strands are

coupled by tight turns in hairpin loops; turns delineate the surface of a globular protein and being predominantly placed at the protein surface necessarily  
880 have a specific amino acid composition; and these turn surface sites are often implicated in binding and protein function. Protein turns, after the repetitive protein secondary structure elements,  $\alpha$ -helices, and  $\beta$ -sheets, are the most abundant structural features in proteins. Here we have demonstrated that purest protein turns are instances of isolated, local (i,i+3) poking contacts,  
885 that naturally provide a link with our recently presented theoretical framework to understand proteins [30,31]. Our studies presented here allow us to systematically access, characterize and classify turns in proteins, comprehending and unifying numerous approaches and definitions present in the literature. This unifying perspective of protein turns provides an explanation for  
890 their large structural flexibility and versatility, which allows them to fulfill successfully all their roles. Even though our analysis is a simple coarse-grained approach at the level of  $C_{\alpha}$  atoms, it is compatible with the pioneering studies of turns in the context of Ramachandran maps [29]. Finally, our studies confirm the notion that turns are emergent snippets of local structures that  
895 are aborted nuclei of helices. They accommodate larger turn angles by pro-

gressively becoming more and more planar. Furthermore, the strict constraint of right-handed helices required by steric clashes for a helix containing many turns is relaxed allowing for turns which can be left-handed and yet carry out the desired goal of turning iso-directional structural entities.

900 Poking contacts, a measure of the affinity of two parts of a chain, are a vivid example of nature using the same theme in different contexts. Poking contacts with rigid geometrical constraints are local and repetitive in a helix. Distinct geometrical constraints are in play in the formation of sheets assembled from iso-directional strands and, this time, the poking contacts are non-  
905 local. The assembly of these building blocks occurs through side-chain mediated interactions, which can be captured in an approximate but harmonious manner by invoking poking contacts between backbone  $C_{\alpha}$  atoms that do not have any geometrical constraints. There is significant variation in the geometries of the interacting side chains unlike the backbone atoms responsible for  
910 the creation of the building blocks. Here we have demonstrated that isolated poking contacts of the backbone atoms play a pivotal role in determining the geometries and types of turns in globular proteins. The recurrent role of poking contacts, a very natural physical manifestation of attraction in the context

of a chain, is an elegant example of nature using every avenue available to  
915 her to full effect. Poking contacts are implicated in the formation of helices,  
sheets, turns, and the assembly of the tertiary structures of globular proteins.  
From a dynamic point of view, in early folding, there ought to be a profusion  
of poking ( $i-i+3$ ) contacts between the backbone atoms, common to all pro-  
teins. They ought to be able to form readily under folding conditions in the  
920 absence of restrictions on too tight a bending angle or clashes between  
sidechains because they are entirely local in character. Some of these poking  
contacts aggregate and become part of a helix, one of the building blocks of  
proteins. Those that are unable to become part of a coordinated helix can re-  
main as isolated turns fulfilling a useful role in protein structure and function.  
925 Thus, protein turns, be they eventually part of a helix or isolated, are heavily  
involved in the key first steps of the dynamics of folding.

We refer the reader to Supplementary Information for more detailed informa-  
tion.

## 930 **Acknowledgments**

We are indebted to George Rose for his inspiration and invaluable advice.

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

### 935 **Funding information**

This project received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie Grant Agreement No. 894784 (TŠ). The contents reflect only the authors' view and not the views of the European Commission. JRB was supported by a Knight  
940 Chair at the University of Oregon. TXH is supported by The Vietnam Academy of Science and Technology under grant No. NCXS02.05/22-23.

### **Conflict of interest**

The authors declare that there is no conflict of interest.

945

## References

1. 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.  
950
2. Pauling L, Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. Proc. Natl. Acad. Sci. USA 1951; 37: 251.
3. Venkatachalam CM. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers  
955 1968; 6: 1425.
4. Lewis PN, Momany FA, Scheraga HA. Folding of polypeptide chains in proteins: a proposed mechanism for folding. *Proc. Acad. Sci. USA* 1971; 68: 2293.
5. Kuntz ID. Protein Folding. *J. Am. Chem. Soc.* 1972; 94: 4009.
- 960 6. Crawford JL, Lipscomb WN, Schellman CG. The reverse turn as a polypeptide conformation in globular proteins. Proc. Nat. Acad. Sci. USA 1973; 70: 538.

7. Lewis PN, Momany FA, Scheraga HA. Chain reversals in proteins. Biochim. Biophys. Acta 1973; 303: 211.
- 965 8. Rose GD, Wetlaufer DB. The number of turns in globular proteins. Nature 1977; 268: 769.
9. Rose GD, Seltzer JP. A new algorithm for finding the peptide chain turns in a globular protein. J. Mol. Biol. 1977; 113: 153.
10. Zimmerman SS, Scheraga HA. Local interactions in bends of proteins. 970 Proc. Acad. Sci. USA 1977; 74: 4126.
11. Chou PY, Fasman GD. Beta-turns in proteins. J. Mol. Biol. 1977; 115: 135.
12. Levitt M, Greer J. Automatic identification of secondary structure in globular proteins. J. Mol. Biol. 1977; 114: 181.
- 975 13. Rose GD. Prediction of chain turns in globular proteins on a hydrophobic basis. Nature 1978; 272: 586.
14. Richardson JS. The anatomy and taxonomy of protein structure. Adv. Prot. Chem. 1981; 34: 167.
15. Rose GD, Young WB, Gierasch LM. Interior turns in globular proteins. 980 Nature 1983; 304: 654.

16. Rose GD, Gierasch LM, Smith JA. Turns in peptides and proteins. *Adv. Protein Chem.* 1985; 37: 1.
17. Leszczynski JF, Rose GD. Loops in Globular Proteins: A Novel Category of Secondary Structure. *Science* 1986; 234: 849.
- 985 18. Wilmot CM, Thornton JM. Analysis and prediction of the different types of beta-turn in proteins. *J. Mol. Biol.* 1988; 203: 221.
19. Dyson HJ, Wright PE. Peptide conformation and protein folding. *Curr. Opin. Struct. Biol.* 1993; 3: 60.
20. Ybe JA, Hecht MH. Sequence replacements in the central beta-turn of  
990 plastocyanin. *Protein Sci.* 1996; 5: 814.
21. Zhou HX, Hoess RH, DeGrado WF. *In vitro* evolution of thermodynamically stable turns. *Nat. Struct. Biol.* 1996; 3: 446.
22. Kim DE, Fisher C, Baker D. A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* 2000; 298: 971.
- 995 23. Fitzkee NC, Fleming PJ, Rose GD. The protein coil library: A structural database of nonhelix, nonstrand fragments derived from the PDB. *Proteins* 2005; 58: 852.



24. Panasik Nick Jr, Fleming PJ, Rose GD. Hydrogen-bonded turns in proteins: The case for a recount. *Protein Sci.* 2005; 14: 2910.
- 1000 25. Jager M, Zhang Y, Bieschke J, Nguyen H, Dendle M, Bowman ME, Noel JP, Gruebele M, Kelly JW. Structure–function–folding relationship in a WW domain. *Proc. Natl. Acad. Sci. USA* 2006; 103:10648.
26. Street TO, Fitzkee NC, Perskie LL, Rose GD. Physical-chemical determinants of turn conformations in globular proteins. *Protein Sci.* 2007; 16: 1720.
- 1005 27. Marcelino A-M C, Gierasch LM. Roles of  $\beta$ -Turns in Protein Folding: From Peptide Models to Protein Engineering. *Biopolymers* 2008; 89: 380.
28. Rose GD. From propensities to patterns to principles in protein folding. *Proteins* 2023, <https://doi.org/10.1002/prot.26540>
- 1010 29. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. *Adv. Prot. Chem.* 1968; 23: 283.
30. 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>

1015 31. 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>

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

- 1020 33. 3D Macromolecule Analysis & Kinemage Home Page at Richardson  
Laboratory. <http://kinemage.biochem.duke.edu/databases/top8000/>
34. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern  
recognition of hydrogen-bonded and geometrical features. *Biopolymers*  
1983; 22: 2577.
- 1025 35. Ramachandran GN, Mitra AK. An explanation for the rare occurrence of  
cis peptide units in proteins and polypeptides. *J. Mol. Biol.* 1976; 107: 85.
36. Humphrey W, Dalke A, Schulten, K. VMD - Visual Molecular Dynamics.  
*J. Molec. Graphics* 1996; 14: 33.
37. Badasyan AV, Giacometti A, Mamasakhlisov YS, Morozov VF, Benight  
1030 AS. Microscopic formulation of the Zimm-Bragg model for the helix-coil  
transition. *Phys. Rev. E* 2010; 81: 021921.
38. Badasyan AV, Tonoyan S, Giacometti A, Podgornik R, Parsegian VA,  
Mamasakhlisov Y, Morozov V. Osmotic Pressure Induced Coupling between  
Cooperativity and Stability of a Helix-Coil Transition. *Phys. Rev. Lett.* 2012;  
1035 109: 068101.