Statistical proofs of the interdependence between nearest neighbor effects on polypeptide backbone conformations

Javier González-Delgado, Pau Bernadó, Pierre Neuvial, Juan Cortés

Statistical proofs of the interdependence between nearest neighbor effects on polypeptide backbone conformations

Javier González-Delgado[a,b], Pau Bernadó[c], Pierre Neuvial[b], Juan Cortés[a,*]

[a] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France*
[b] *Institut de Mathématiques de Toulouse, Université de Toulouse, CNRS, France*
[c] *Centre de Biologie Structurale, Université de Montpellier, INSERM, CNRS, France*

Abstract

Backbone dihedral angles ϕ and ψ are the main structural descriptors of proteins and peptides. The distribution of these angles has been investigated over decades as they are essential for the validation and refinement of experimental measurements, as well as for structure prediction and design methods. The dependence of these distributions, not only on the nature of each amino acid but also on that of the closest neighbors, has been the subject of numerous studies. Although neighbor-dependent distributions are nowadays generally accepted as a good model, there is still some controversy about the combined effects of left and right neighbors. We have investigated this question using rigorous methods based on recently-developed statistical techniques. Our results unambiguously demonstrate that the influence of left and right neighbors cannot be considered independently. Consequently, three-residue fragments should be considered as the minimal building blocks to investigate polypeptide sequence-structure relationships.

Keywords: Protein and peptide structure, Local conformational preferences, Sequence-structure relationship, Statistical tests, Wasserstein distance

*Corresponding author
Email address: juan.cortes@laas.fr (Juan Cortés)

1. Introduction

Proteins and peptides are essential molecules in all living organisms. Their numerous functions are closely related to their structural and dynamic properties. The main variables to investigate these properties are the backbone ϕ and ψ dihedral angles of each of the amino acid residues along the sequence (Brändén and Tooze, 1998; Liljas et al., 2009) (see Figure 1 for an illustration). The allowed values of this pair of angles and its statistical distribution have been studied over half a century, since the seminal work by (Ramachandran et al., 1963; Ramachandran and Sasisekharan, 1968). Several applications have motivated the detailed analysis of ϕ and ψ angles in polypeptide chains, such as the validation and refinement of structures determined from biophysical techniques (Morris et al., 1992; Lovell et al., 2003) and the development of models or scoring functions for protein structure prediction and design (Gibrat et al., 1991; Kang et al., 1993; Betancourt and Skolnick, 2004; Boomsma et al., 2008; Rata et al., 2010; Ting et al., 2010). The study of local structural preferences of polypeptides is also essential for the investigation of denatured states of globular proteins (Smith et al., 1996b; Jha et al., 2005a) and intrinsically disordered proteins (IDPs) (Shen et al., 2018; Estaña et al., 2019).

Each amino acid type has a particular distribution of the ϕ and ψ angles (Swindells et al., 1995; Serrano, 1995; Deane et al., 1999; Hovmöller et al., 2002; Anderson et al., 2005). These distributions are relatively similar for all natural amino acids, with the exception of glycine and proline. While glycine lacks a side chain, thus providing enhanced conformational variability, proline has a cyclic side chain, which severely restricts the accessible ϕ and ψ values (Ho and Brasseur, 2005). Some early work assumed that the distribution depends only on the amino acid type, independently of the context, which is usually referred to as Flory's isolated-pair hypothesis (Flory, 1969; Zimmerman et al., 1977). Despite its simplicity, Flory's isolated-pair hypothesis has been very useful to interpret Small Angle Scattering data reporting on the overall size of disordered and denatured proteins (Kohn et al., 2004). However, the availability of

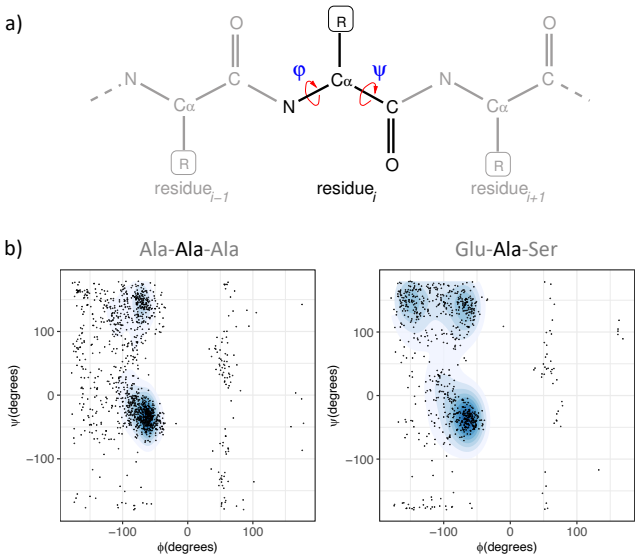


Figure 1: (a) Illustration of a three-residue fragment indicating the ϕ and ψ angles of the central residue. Only heavy atoms are represented, and the R corresponds to each amino acid side chain. (b) Distributions of the ϕ-ψ angles for an alanine residue with different neighbors.

residue-specific information provided my Nuclear Magnetic Resonance (NMR) measurements, such as residual/scalar couplings and chemical shifts, evidenced that the conformational preferences of individual amino acid residues is influenced by their nearest neighbors (Braun et al., 1994; Penkett et al., 1997). A wide variety of short peptides have been used in order to rationalize and quantify the effects exerted by the nearest neighbors (Dames et al., 2006; Oh et al., 2012a,b; Jung et al., 2014; Toal et al., 2015; Schweitzer-Stenner and Toal, 2018). The ensemble of these studies identified aromatic and β-branched amino acids as having the strongest influence on the structure of their neighbors due to their steric hindrance (Penkett et al., 1997; Jung et al., 2014), although the role of solvation has been also pointed out by some authors (Avbelj and Baldwin, 2004). Nearest neighbor effects were found particularly significant in several cases of repeated amino acids along the sequence (Milorey et al., 2021b,a).

Various theoretical/computational approaches, building on experimentally-determined protein structures, have also been developed to investigate sequence-

dependent structural preferences and to integrate them within predictive methods (Kabat and Wu, 1973; Gibrat et al., 1987; Betancourt and Skolnick, 2004; Boomsma et al., 2008; Rata et al., 2010; Ting et al., 2010). In addition, the inability of simple single-residue-based coil models to recapitulate NMR data supports the influence of the sequence context in defining conformational ensembles of disordered and denatured proteins (Smith et al., 1996a; Pappu et al., 2000; Jha et al., 2005a,b; Bernadó et al., 2005; Huang et al., 2013; Estaña et al., 2019). Finally mention that molecular dynamics simulations of simple tripeptides showed that the nearest neighbors influence the relative population between the regions of the Ramachandran space and the transitions rates between them (Zaman et al., 2003). Overall, these experimental and computational studies provide strong evidence for the effect of the nearest neighbors in defining the conformational preferences of a given amino acid residue. However, from a statistical perspective, no inference method (i.e., hypothesis testing methods, as described below) has yet been applied to formally test Flory's isolated pair hypothesis.

An additional question arises when investigating nearest neighbor dependence: is the influence of left and right neighbors interdependent? This is an important issue as it determines whether the influence exerted by both neighbors can be studied separately. Contradictory answers have been given to this question. For instance, Griffiths-Jones et al. (1998) postulated that electrostatic interactions between the left and right neighbors significantly affect the conformation of the central residue. Betancourt and Skolnick (2004) directly considered three-residue fragments for the analysis of neighbor dependence, thus implicitly assuming that left and right neighbors cannot be dissociated for this analysis. Results by Huang et al. (2013) also suggested that left and right residues have to be simultaneously taken into account in order to appropriately estimate NMR residual dipolar couplings (RDCs) measured in IDPs. Conversely, independence of adjacent residues has been asserted by other authors, although only descriptive or statistically vague methods have been applied in this regard. Rata et al. (2010) stated independence after visual comparison of two density estimations,

and Shen et al. (2018) based their analysis on an amino acid clustering approach, valid only under the independence hypothesis.

We have implemented statistical hypothesis testing methods (Lehmann and Romano, 2005) to investigate the interdependence between nearest neighbor effects on backbone dihedral angles. These statistical tests make it possible to: **(i)** reject Flory's isolated-pair hypothesis by finding significant differences between dihedral angle distributions when both neighbors are or not taken into account (Section 2.1), and **(ii)** prove the interdependence of neighbor effects, which is the main contribution of this work, by assessing the independence of two categorical variables (Sections 2.2-2.4). Details on the implemented statistical tests are provided in Section 4.2.

Data for our analyses were extracted from a non-redundant set of experimentally-determined high-resolution protein structures. We constructed two datasets from three-residue fragments (called tripeptides from now on) depending on the structural context: considering all tripeptides in all structures, and considering only fragments from coil regions (i.e. tripeptides not contained in α-helices or β-strands). In the following, we will refer to these datasets as *All* and *Coil*, respectively. We would like to clarify here that although it is well known that statistical models to investigate disordered or unfolded proteins are in general more accurate when they are built from restricted structural datasets that do not contain secondary structure elements (Swindells et al., 1995; Smith et al., 1996a; Jha et al., 2005a), we decided to perform the statistical tests for restricted and unrestricted datasets with the aim of analyzing differences. Details about the tripeptide datasets are provided in Section 4.1.

The software and the data used in this work are freely available (see detail in section "Software and data availability" at the end of the manuscript).

2. Results and discussion

2.1. Left and right neighbors significantly affect dihedral angle distributions

First, we evaluated Flory's isolated-pair hypothesis. To do this, we performed a recently-developed two-sample goodness-of-fit test (González-Delgado et al., 2021), defined to assess the equality of two probability distributions supported on the two-dimensional flat torus (which is the space corresponding to two angles, ϕ and ψ). This test is based on the Wasserstein distance, which is a suitable metric to compare distributions on non-euclidean spaces (Villani, 2008). Details about the method are provided in Section 4.2.1. We found highly significant differences (at significance level $\alpha = 0.05$) between dihedral angle distributions when both neighbors are or are not taken into account. Results are represented through p-values, which quantify the plausibility of the observed data assuming that Flory's hypothesis is true. Hence, small p-values provide strong quantitative evidence against this hypothesis. Results were consistent for the *All* and *Coil* datasets, and similar for all amino acid types, including glycine and proline. This is shown in Table 1, where a p-value is presented for each central amino acid type. Note that p-values for the *All* and *Coil* datasets should not be compared, because in general rejection strength is higher when more information is available (as in the *All* dataset). However, within-dataset p-value comparisons are reasonable, since the sample sizes are comparable. For instance, for the *All* database, the effect of neighbors on the (ϕ, ψ) distribution is higher for alanine than for asparagine. Overall, these analyses represent statistically strong evidence to reject Flory's isolated-pair hypothesis and confirm the results of many previous studies mentioned in Section 1 that show the importance of nearest neighbor effects.

2.2. Influence of the left and right neighbors are statistically interdependent

Next, we assessed the significance of combined effects exerted by the nearest neighbors. From formulations developed in previous studies (Rata et al., 2010; Ting et al., 2010), we can derive that assessing the independence of left

Dataset	Ala	Arg	Asn	Asp	Cys	Gln
All	0.00603	0.00582	0.01316	0.01284	0.00351	0.01071
Coil	0.01809	0.01746	0.01880	0.01926	0.00390	0.01530
Glu	Gly	His	Ile	Leu	Lys	Met
0.00852	0.01570	0.00522	0.00456	0.00426	0.00872	0.00168
0.01704	0.02016	0.00580	0.01368	0.01704	0.01962	0.00210
Phe	Pro	Ser	Thr	Trp	Tyr	Val
0.00756	0.01425	0.01589	0.01393	0.01500	0.00648	0.00540
0.01260	0.01250	0.02043	0.01990	0.00030	0.01080	0.01620

Table 1: Amino-acid-specific p-values for each statistical test of the Flory's isolated-pair hypothesis performed for the *All* and *Coil* datasets. All p-values are highly significant at level $\alpha = 0.05$ after correction for multiple testing.

and right neighbor influences is mathematically equivalent to evaluating the independence of the left and right amino acid identities given a central residue in a defined conformation (ϕ,ψ). In practice, this can be assessed via a χ^2 (chi-square) test of independence per central amino acid residue and value of (ϕ,ψ). To do so, we discretized the Ramachandran space and carried out one test per subdivision (see Section 4.2.2 for details). Then, the results of all tests across the discretization grid were summarized by a p-value, which quantifies the plausibility of the observed data assuming independence for each amino-acid type. Hence, p-values close to zero provide strong statistical evidence for the interdependence of the influence of the left and right nearest neighbors. The obtained p-values were lower than 10^{-10} for all amino acid types, from both *All* and *Coil* datasets. This implies the interdependence of the left and right nearest neighbors in determining the (ϕ, ψ) angles of the central residue. Note that the difference between the scale of the p-values in Sections 2.1 and 2.2 is due to methodological differences and not to the nature of the contrasted hypotheses. Indeed, p-values in Section 2.1 were computed by simulating the null distribution of the test statistic, and therefore cannot be lower than $1/N_{\text{sim}}$, where N_{sim} is the number of simulations (Phipson and Smyth, 2010), whereas

the null distribution of the χ^2 statistic is exact, and consequently p-values can take any value in $(0, 1)$.

2.3. The physicochemical properties of the nearest neighbors affect the magnitude of the interdependence

We assessed whether properties such as the polarity and the size of the neighbors affect the strength of the interdependence. Note that some previous studies on nearest neighbor effects (not dealing with interdependence) divided amino acid types into only two classes (Penkett et al., 1997): those involving aromatic and beta-branched side chains (FHITVWY) and the others, with the exception of glycine and proline. The large size of our databases enabled a finer classification. Consequently, we chose six representative amino acids for each of the following groups:

- Polar (P): Arg, Lys, Asp, Glu, Asn, Gln.
- Hydrophobic (H): Ala, Ile, Leu, Met, Phe, Val.
- Small (S): Ala, Ser, Thr, Asp, Asn, Cys.
- Large (L): Phe, Tyr, Trp, Arg, Ile, Lys.

To facilitate the comparison between these categories, we defined a score measuring the strength of the interdependence effect for a given neighbor setting (a fixed group for the left residue and one for the right residue). This score corresponds to the Area Under the Curve (AUC) of the empirical cumulative distribution function of p-values (see Section 4.2.3 for details). When comparing two different neighbors settings, a higher AUC means p-values being closer to zero and thus a higher interdependence.

Figure 2 shows radar plots with AUC values for possible combinations of neighbors depending on their polarity (left plot) and size (right plot), and averaged for all amino acid types at the central position. General trends can be observed from this representation: (i) Interdependence is stronger when both

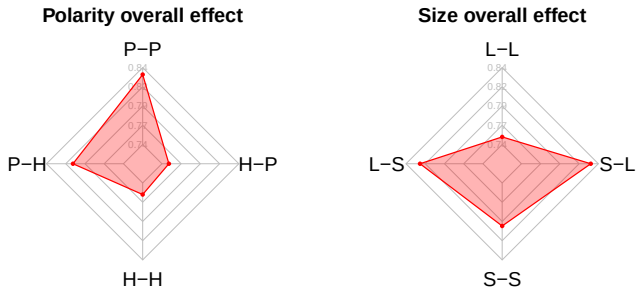


Figure 2: Radar plots showing the interdependence score (AUC) between neighbors with different physicochemical properties: (left) polarity/hydrophobicity, (right) size. P, H, L and S stand for polar, hydrophobic, large and small, respectively.

neighbors are polar. This can be justified by the presence of attractive or repulsive electrostatic interactions between them, which may constrain the conformational space for the central residue, and that strongly depend on the specific pair of neighbors (Milorey et al., 2021b). Adjacent charged amino acids can also modify the solvation energy and perturb the central residue (Avbelj and Baldwin, 2004). (ii) Interdependence is weaker when both neighbors are large. This observation is less intuitive. A possible explanation would be that when both neighbors are large (regardless of their type), the conformational space of the central residue is more constrained (Cho et al., 2007), and thus, other effects due to the nature of each neighbor are less visible. The contrary occurs when at least one of the neighbors is small, as the central residue exhibits a less constrained conformational space.

Nevertheless, exceptions to these above-described general trends emerged when the strength of the interdependence was analyzed individually for each amino acid. Case-by-case results comparing the four settings for each central residue are shown in Figures 3 (for polarity) and 4 (for size). The AUC values for each group were also compared with the "free" setting, for which no physicochemical properties are imposed to neighbors (i.e. considering all possible neighbors). They illustrate the very diverse degrees of interdependence depending on the central amino acid and the properties of the nearest neighbors, which highlights the need to take into account (at least) three-residue fragments to

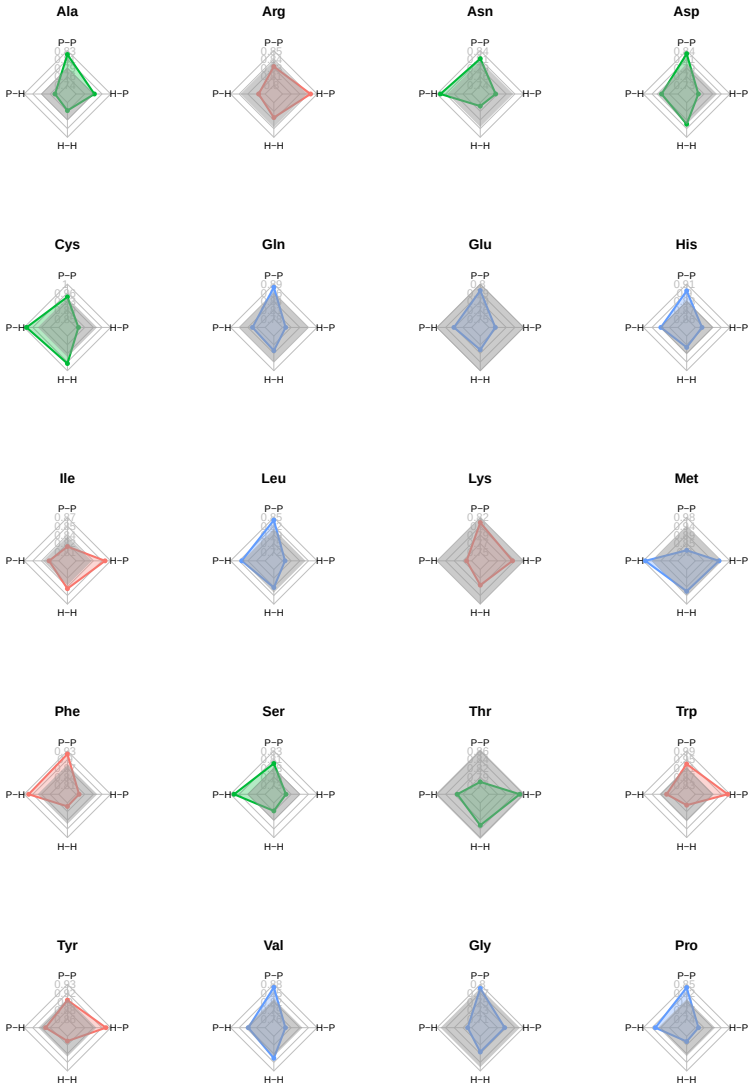


Figure 3: For each central amino acid type, comparison of the interdependence score (AUC) between the four possible polarity combinations for neighbors (where P stands for polar and H for hydrophobic). In gray, the same score when no physicochemical properties are imposed. Polar and hydrophobic central residues correspond to red and green plots respectively. Blue plots correspond to central residues not belonging to any of both categories.

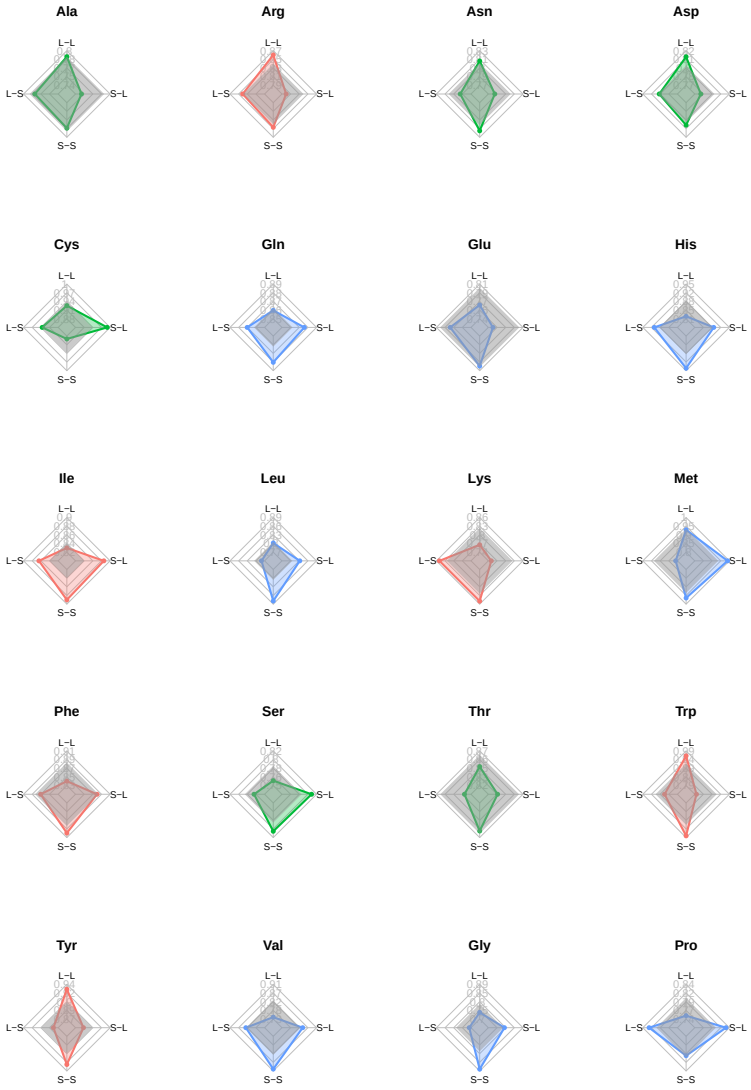


Figure 4: For each central amino acid type, comparison of the interdependence score (AUC) between the four possible size combinations for neighbors (where L stands for large and S for small). In gray, the same score when no physicochemical properties are imposed. Large and small central residues correspond to red and green plots respectively. Blue plots correspond to central residues not belonging to any of both categories.

locally describe backbone conformational preferences. Due to how the score has been defined, one must not compare AUC values between different individual plots, but only inside each plot. All the differences between AUC scores were statistically significant.

Our analyses performed on the *Coil* dataset showed that, for 14 out of the 20 amino acids, interdependence is stronger when both neighbors are polar than when they are both hydrophobic. With respect to size effect, for 16 out of 20 central residues interdependence was found stronger when both neighbors were small than when they were large. No relationship was found between amino acids not following both general trends. However, more detailed analyses showed that amino acids that did not follow the general trend were among those for which the amount of data was more limited. This may suggest that with additional data, the general trend would probably be more widely satisfied. With respect to mixed neighbors settings, no clear general trend was found among all central amino acids. In all cases, all the corresponding AUC scores were significantly different to the "free" setting ones, showing that both polarity and size do affect interdependence also when neighbors have mixed properties. Moreover, all plots in Figures 3 and 4 were strongly asymmetrical with respect to the vertical axis, which evidences that polarity and size effects have a non-negligible directional component.

2.4. Combined neighbor effects are stronger in coil regions

We implemented two approaches to quantitatively assess whether neighbor interdependence is influenced by the structural origin of the datasets. The first approach lies in comparing the computed p-value distributions for each dataset (*All* and *Coil*). Here, a distribution of p-values is associated to each central amino acid (one test is performed at each point of its discretized Ramachandran space, see Section 4.2.2 for details). Moreover, sample sizes are fixed for both datasets and thus p-values for *All* and *Coil* are now quantitatively comparable. Consequently, we can compare each pair of distributions and evaluate whether the interdependence of neighbors is stronger in one of the two datasets.

Three representative examples of this comparison are shown in Figure 5(a,c,e) for alanine, glutamic acid and leucine as central residues. p-value density estimates show that the independence hypothesis is more significantly rejected (i.e. interdependence is stronger) for the *Coil* dataset (Kolmogorov-Smirnov test states highly significant discrepancies). Note that the scales in Figure 5 vary between the three amino acids in order to better reflect the different behaviour of *All* and *Coil* distributions. Comparisons between different plots are not really relevant.

The second approach simulates the proportion of statistically non-significant tests for both datasets (the lower this proportion, the more interdependent the left and right neighbors). This is explained in detail in Section 4.2.2. Figure 5(b,d,f) exemplifies this approach using the same central residues. These figures show that the simulated distribution corresponding to the *Coil* dataset is significantly closer to zero than that of the *All* dataset, substantiating our aforementioned conclusion. This observation contradicts the statements of Rata et al. (2010), who suggested that the correlated effects of left and right neighbors were weak, especially in coil regions. Nevertheless, their statements about the lack of interdependence were based on vague statistical analyses compared to the rigorous statistical approach presented above.

3. Conclusions

We have investigated local sequence effects on the distribution of the ϕ-ψ angles, which are the main descriptors of polypeptide conformations, using rigorous statistical methods on datasets built form experimentally-determined high-resolution protein structures. Results of our analyses corroborate the large amount of experimental and computational studies describing the influence of the nearest neighbors, thus providing additional evidence for the rejection of Flory's isolated-pair hypothesis, even in disordered regions. Furthermore, our results unambiguously demonstrate coupled effects of the left and right neighbors, which cannot be considered independently of each other. This observation

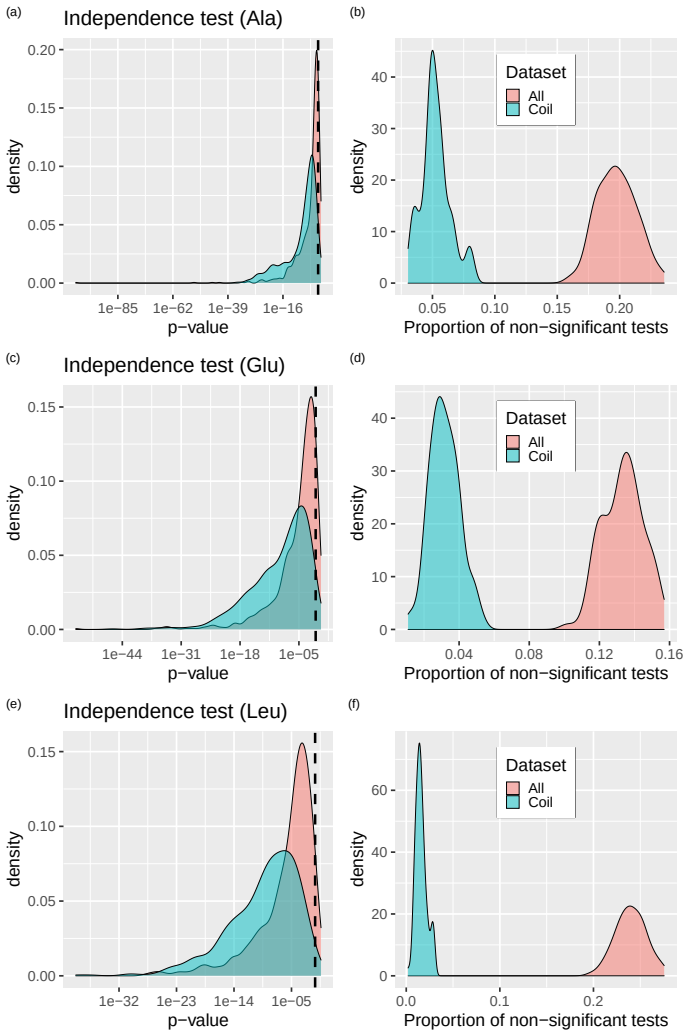


Figure 5: (a,c,e) Distribution of (-log10-scaled) p-values for the independence hypothesis tests performed on the *All* (red) and *Coil* (blue) datasets, for a fixed central residue. Dashed line indicates a level of significance of 0.05. (b,d,f) Distribution of the proportion of non-significant tests for a fixed central residue for the *All* and *Coil* datasets.

clarifies questions still open on this subject, and represents a fundamental step to understand sequence-structure relationships in peptides and proteins.

These results also have several direct implications for methodological de-
260 velopments in the context of molecular modeling and protein design. The

most obvious one concerns sampling algorithms that use ϕ-ψ distributions to model flexible regions in proteins, such as loops or intrinsically disordered regions (Bernadó et al., 2005; Ozenne et al., 2012; Estaña et al., 2019; Barozet et al., 2021). More accurate conformational ensemble models will be obtained when explicitly considering coupled neighbor dependencies. The parameterization of the constants associated with backbone torsion angles in force-fields used for molecular dynamics simulations, and more particularly in the case of coarse-grained models, would also benefit from protocols that consider the local sequence context (i.e. going beyond residue-specific parameterization). Regarding structure prediction algorithms applied to globular proteins, although modern machine-learning-based algorithms mostly exploit evolutionary-conserved pairwise residue contacts, the incorporation of local structural constraints and preferences are crucial to obtain accurate solutions (Jumper et al., 2021). Thus, our observations suggest that the performance of these algorithms could be improved by explicitly considering triplets of consecutive amino acids for the conception of the neural network architecture.

This work focused on studying the interdependent effects of the nearest neighbors along the sequence (i.e. residues $i \pm 1$). It would be very interesting to extend the analysis to more distant neighbors ($i \pm n$, with $n = 2, 3, 4, \dots$). Unfortunately, the amount of experimental data currently available does not allow such an analysis. With the increase of available data from experimental techniques and/or high quality models generated by simulation or structural prediction methods, such an analysis seems feasible in the near future. It should be noted, however, that non-trivial mathematical challenges would also arise in addressing this question, which would require new methodological developments.

4. Methods

4.1. Data collection

A database of three-residue fragments (tripeptides) was built from a curated database of experimentally-determined high-resolution protein structures. More

precisely, we used protein domains from the SCOPe 2.07 release (Chandonia et al., 2018). In order to remove highly-redundant sequences, we used the 95% sequence-identity-filtered subset of these domains. In addition to structures determined by X-ray crystallography (with a resolution below 3Å), SCOPe also contains structures from NMR experiments. For each input file from NMR experiments containing more than one model, a distance filter was applied to corresponding tripeptides in each model to avoid repetitions in the database. A tripeptide structure was considered sufficiently different from another one already extracted from the same file, and was thus added to the database, if it met at least one of the two following criteria: the RMSD on ϕ and ψ angles was above 0.2 radians, or one of the dihedral angles differed by more than 0.6 radians. In total, 6,740,433 tripeptide structures were extracted. Tripeptides were classified by sequence (i.e. 8,000 tripeptide classes) and the backbone dihedral angles were collected in a dataset called *All*, since no additional structural criteria were considered for filtering.

A structurally filtered dataset, called *Coil*, was generated by removing tripeptides contained in α-helices or β-strands. For this, DSSP (Kabsch and Sander, 1983; Touw et al., 2014) was employed to assign secondary structure labels to each amino acid residue in the structures extracted from the SCOPe database. A tripeptide was included in the filtered subset if none of its three residues had a DSSP code of H or E. Note that π-helices or 3-10-helices, which are relatively rare in our database, were not filtered out because they are usually small and can be observed inserted into coil regions. The secondary structure filtering reduced the number of tripeptide structures to 3,141,877, which is less than half the size of the *All* dataset.

For the analyses in this work, for both *All* and *Coil* datasets, we considered only tripeptides involving peptide bonds in *trans* conformation, which corresponds to the vast majority of the instances. Therefore, tripeptides involving at least one peptide bond in *cis* conformation were removed. We treated the cases of glycine and proline separately. We excluded from the datasets tripeptide sequences for which the number of available structures was very low, and

thus not statistically interpretable. The number of required structures depends on each test, and is detailed in Sections 4.2.1 and 4.2.2.

It should be noted that, in order to collect enough data for the analyses, we were less restrictive in the construction of the datasets compared to previous studies (e.g. Rata et al. (2010); Ting et al. (2010)). Nevertheless, this is acceptable in our case since our aim is not to develop a (differentiable) statistical potential, but to perform statistical tests, and because our implementation of these tests is reasonably resilient to noise. For the sake of rigor, we generated more restrictive (supposedly higher-quality) datasets considering only structures determined by X-ray crystallography with a resolution below 2Å. We performed the same analyses using these datasets, but considering only tripeptides for which the amount of data allowed a correct implementation of the statistical tests. Overall, the analyses (restricted to a small number or tripeptides sufficiently represented in the so filtered datasets) led to the same conclusions regarding the rejection of the Flory's isolated-pair hypothesis and the interdependent effects of left and right neighbors. These results are not presented here.

4.2. Statistical methodology

4.2.1. Rejecting Flory's isolated-pair hypothesis

For a given central amino acid residue C, the associated dihedral angles (ϕ, ψ) describing its conformation follow a certain distribution F_C, which is supported on the 2-dimensional flat torus $\mathbb{T}^2$. If left and right neighbors are taken into account, this distribution may also depend on their identities, L and R, respectively, it is noted as $G_C^{L,R}$, and it is also supported on $\mathbb{T}^2$. This can be rewritten as follows:

$$(\phi, \psi \,|\, C) \rightsquigarrow F_C \qquad (\phi, \psi \,|\, C, L, R) \rightsquigarrow G_C^{L,R}, \tag{1}$$

where, for each value of C, L and R, $(\phi, \psi \,|\, C)$ and $(\phi, \psi \,|\, C, L, R)$ are two-dimensional random variables following the above-mentioned distributions. In order to evaluate the dependence of the backbone dihedral angles corresponding

to residue C on the identity of its nearest neighbors, the following statistical test has to be performed:

$$H_0 \,:\, F_C = G_C^{L,R} \quad \forall\, C, L, R \text{ amino acids.} \tag{2}$$

This corresponds to a goodness-of-fit (GoF) test between two continuous distributions on $\mathbb{T}^2$. This test has to be implemented for each combination of all possibles C, L and R. Note that it suffices to reject the null hypothesis $F_C = G_C^{L,R}$ for one single tripeptide C, L, R in order to reject (2). The statistical analysis of protein conformations was the motivation of the recent work by González-Delgado et al. (2021) to explore the definition of two-sample goodness-of-fit tests for measures on $\mathbb{T}^2$. In that paper, two approaches were proved efficient both on synthetic data and experimental protein conformation data. The presented techniques use the p-Wasserstein distance as metric between distributions, which corresponds to the p-th root of the minimum transportation cost between two probability laws. This metric takes the geometry of the underlying space into account, and therefore accounts for the periodicity of the Ramachandran space. We refer to Peyré and Cuturi (2019) for an introduction to Optimal Transportation. Here, the first of the approaches presented by González-Delgado et al. (2021) has been implemented, consisting on testing the equality of the projected laws on the torus geodesics, which are circles $\mathbb{T} = \mathbb{R}/\mathbb{Z}$.

Due to the large sizes of the available protein conformation datasets, further practical considerations had to be taken into account. The first one was the occurrence of ties when projecting data on the torus to the marginal one-dimensional space. As the test is built under the assumption of continuity of distributions, ties can bias the statistic realizations and therefore the resulting p-values can distort the test results. To break ties, a uniform background noise was added to the data. The second one is intrinsic to the mathematical procedure and to the exacting null hypothesis (2). Indeed, when the number of data points increases, the test becomes more sensitive to small discrepancies, and rejects the equality of distributions when those differences may not be relevant for practical purposes. Several approaches can be considered to deal with this.

Here, we decreased sampling resolution and searched for significant differences at a coarser scale. If, for a given low resolution, hypothesis (2) is rejected, the conclusion will be the same when all data points are taken into account. When analyzing the corresponding results, note that p-values for this test are computed via a Monte Carlo simulation, so they are lower-bounded by $1/N_{\mathrm{sim}}$, being N_{sim} the number of Monte Carlo iterations.

We carried out the GoF test assessing Flory's hypothesis (2) considering tripeptides having more than 200 conformations in the corresponding dataset. For each tripeptide, we performed one test comparing the available sample drawn from $G_C^{L,R}$ (all the points for a given tripeptide) with an equally sized subsample drawn from F_C. This comparison was repeated for 10 subsamples, leading to 10 p-values $(p_1, \ldots, p_{10})$, and we used the minimum adjusted p-value among these tests ($p = 10 \min_{i=1}^{10} p_i$) as a global p-value for each tripeptide. These p-values were then adjusted for multiplicity (Bonferroni, 1936) across tripeptides having the same central residue, whose minimum is depicted in Table 1, for each central amino acid type.

4.2.2. Assessing interdependence between left and right neighbors

We aimed at assessing whether the distribution of (ϕ, ψ), which depends on the three amino acids L, R and C, can be separately inferred from the information of L-C and C-R dipeptides, or the information on the tripeptide L-C-R is unavoidably required. Ting et al. (2010) stated that, under the hypothesis

$$L \text{ and } R \text{ independent given } C \text{ and } (\phi, \psi), \qquad \text{(indep)}$$

the probability density of (ϕ, ψ) given the whole tripeptide, $f(\phi, \psi \,|\, L, C, R)$, can be obtained from the information of the densities given by the left and right dipeptides as

$$f(\phi, \psi \,|\, L, C, R) = \frac{f(\phi, \psi \,|\, L, C)\, f(\phi, \psi \,|\, C, R)}{S\, f(\phi, \psi \,|\, C)}, \qquad (3)$$

where S is a normalization constant. Moreover, Rata et al. (2010) proved the reciprocal implication. We have thus the following equivalence:

$$\begin{gathered} L \text{ and } R \text{ independent given } C \text{ and } (\phi,\psi) \\ \Leftrightarrow \\ f(\phi,\psi \,|\, L,C,R) = \frac{f(\phi,\psi \,|\, L,C)\, f(\phi,\psi \,|\, C,R)}{S\, f(\phi,\psi \,|\, C)} \end{gathered} \tag{4}$$

The proof of this equivalence is stated in the Appendix for completeness.

If Equation (3) is false, then, the probability density of (ϕ,ψ) of a central residue for a given tripeptide cannot be inferred from the information on the corresponding dipeptides (at least via the functional form stated by Ting et al. (2010)). According to the equivalence (4), disproving hypothesis (indep) is enough to disprove (3). In order to test (indep), one can perform a χ^2 independence test between the categorical variables L and R for each fixed value of C and (ϕ,ψ). This requires a proper discretization of the space $\mathbb{T}^2$, in order to obtain a set of values for (ϕ,ψ) that accurately represent the bidimensional random variable and that allow the implementation of the statistical test. Generally, a finer or coarser discretization entails a more or less faithful representation of the angular variable (ϕ,ψ), which ideally is continuous on $\mathbb{T}^2$. Consequently, an optimal discretization procedure will be the thinnest one allowing contingency matrices of the maximum dimension and with a number of points sufficiently large for the independence tests to be performed correctly. We propose three different discretization methods, whose parameters should be optimized. The three methods are based on:

(D1) The choice of a representative set

$$\mathcal{R} = \{(\phi_i,\psi_i)\}_{i\in 1,\ldots,N_{\text{rep}}} \subset \mathbb{T}^2\,.$$

(D2) For each representative value $(\phi_i,\psi_i) \in \mathcal{R}$, the choice of the set of points $\mathcal{R}_i = \{(\phi_{ij},\psi_{ij})\}_{j\in 1,\ldots,J_i}$ for which $(\phi_{ij},\psi_{ij}) \equiv (\phi_i,\psi_i) \quad \forall\, j \in 1,\ldots,J_i$, where $a \equiv b$ means that, in terms of the discretization, a and b belong to the same space subdivision.

The three proposed methods were built as follows and are illustrated in Figure (6):

(I) $\mathcal{R}$ is a homogeneous square grid and $\mathcal{R}_i$ are the points belonging to the i-th cell.

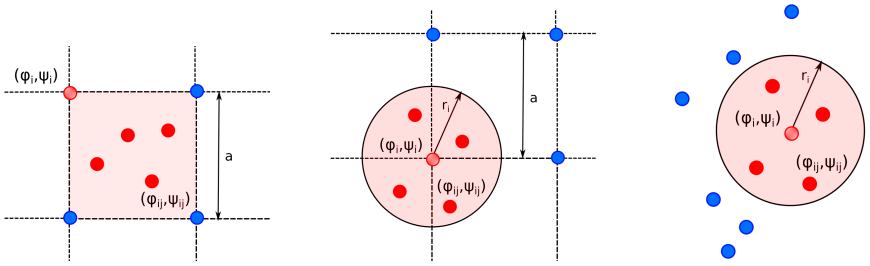
(II) $\mathcal{R}$ is a homogeneous square grid and $\mathcal{R}_i$ are the points in the vertex-centered ball $B_{\mathbb{T}^2}\left((\phi_i, \psi_i), r_i\right)$.

(III) $\mathcal{R}$ is a subset of the dataset sampled uniformly and without replacement, and the $\mathcal{R}_i$ are the points in the ball $B_{\mathbb{T}^2}\left((\phi_i, \psi_i), r_i\right)$.

For method I, the only parameter is the size $a = 2\pi/\sqrt{N_{\text{rep}}}$ of the square grid. It was chosen as the smallest value allowing maximum dimension contingency matrices with a large enough number of points. Due to physicochemical constraints, the whole $\mathbb{T}^2$ space is not accessible, and thus we limited ourselves to regions with non negligible density. To do so, a grid cell was kept only if it contained a minimum number of data points (i.e. if $J_i \geq N_{\min}$). For the analyses presented here, we chose $N_{\min} = 500$ and $N_{\text{rep}} = 30$.

For methods II and III, the radius r_i of each ball depends on (ϕ_i, ψ_i), and was determined in order to include a specific number $J_i = J$ of points in the ball, the same for all partitions. This allowed a discretization for which each subdivision had the same number of data points, and thus for which all the tests performed were comparable. In order to maintain a certain control on how (ϕ, ψ) values are identified together, a maximum radius R was established and only balls with $r_i < R$ were kept. The number of points J at each ball was chosen to guarantee contingency matrices with maximum dimension while providing a thin and reliable discretization. For method III, the number of representative points N_{rep} was also chosen according to the same considerations. Here, we chose $J = N_{\text{rep}} = 1000$ and $R = 0.1$.

It should be recalled that we do not need to perform tests over the whole $\mathbb{T}^2$ space. As the hypothesis (indep) is *conditional* to C and the continuous random



(a) Method I. (b) Method II. (c) Method III.

Figure 6: The three proposed discretization methods.

variable (ϕ, ψ), it is equivalent to the hypothesis

$$L \text{ and } R \text{ independent given } C = c \text{ and } (\phi, \psi) = (\phi_0, \psi_0)$$
$$\text{for all amino-acids } c \text{ and all } \phi_0, \psi_0 \in [-\pi, \pi].$$

Therefore, rejecting (indep) means rejecting the independence of L and R for any fixed values of C and (ϕ, ψ). Consequently, implementing tests for a subset of the discretized space will properly answer our question if a significant result is retrieved.

The independence test was performed for the two aforementioned datasets, *All* and *Coil*, using the three proposed discretization methods, whose corresponding parameters were chosen according to the previously specified considerations. Given a central amino acid, one test was performed per point (ϕ_0, ψ_0) of the chosen grid, associating a distribution of p-values to each central residue. For methods II and III, sample sizes were fixed and therefore p-values can be quantitatively compared (as it is illustrated in Section 2.4). Note that results follow the same general trend for all amino acid types (for the central residue and the neighbors) and discretization methods. Note also that, unlike the test described in Section 4.2.1, p-values are no longer lower-bounded and can take any real positive value.

As a large number of test was performed, a multiplicity adjustment was implemented (Bonferroni, 1936). Finally, an overall p-value for each amino acid was defined as the minimum adjusted p-value across the discretization.

Simulation of non-rejecting tests: The intrinsic randomness of discretization method III allows to simulate the proportion of non-rejecting tests for a given central amino acid. For $s = 1, \ldots, N_{\text{sim}} = 100$, we sample a representative set $\mathcal{R}_s$, perform the independence test across $\mathcal{R}_s$ and compute the proportion $\tilde{p}_s$ of p-values higher than a fixed threshold $\alpha = 0.05$. The set of all $\tilde{p}_1, \ldots, \tilde{p}_{N_{\text{sim}}}$ constitute a sample of the proportion of non-rejecting tests for the given amino acid. As p-values are quantitatively comparable, so are the proportions $\tilde{p}_s$. This corresponds to comparisons presented in Figure 5(b,d,f), between the two datasets.

4.2.3. Polarity and size effect on interdependence: AUC score

In order to assess whether the nearest neighbors' polarity and size have a significant effect on their interdependence, we chose six representative amino-acids for each one of the groups defined in Section 2.3. The strategy was to repeat the independence test for all central amino acid types, but restricting the admissible settings of neighbors identities to those in these groups. For polarity (resp. size) we computed (indep) p-values when left and right neighbors belonged to the settings P-P, P-H, H-P and H-H (resp. L-L, L-S, S-L and S-S). However, reducing the number of classes that the categorical variables L and R induces a power loss. In other words, if the information about the variables whose independence we want to assess is trimmed-down, the test will have less information to state any result with the same evidence. Nevertheless, relative comparisons between two groups of p-values for the same number of classes are allowed, and statistically informative.

To facilitate a more direct comparison between settings, we defined a score representing the strength of the interdependence of neighbors in a given configuration. For a given setting C_L-X-C_R, where $C_L, C_R \in \{P, H\}$ (for polarity) or $C_L, C_R \in \{L, S\}$ (for size), let $F_{N_{\text{rep}}}^{C_L, C_R}$ denote the empirical cumulative distribution function (ECDF) of the p-values retrieved after testing hypothesis (indep) across a fixed discretization of size N_{rep}. Then, the Area Under the Curve

(AUC) of $F_{N_{\text{rep}}}^{C_L,C_R}$ is defined as

$$\text{AUC}(C_L, C_R) = \sum_{i=1}^{N_{\text{rep}}} (p_{(i+1)} - p_{(i)})\, F_{N_{\text{rep}}}^{C_L,C_R}(p_{(i)}), \tag{5}$$

where $p_{(i)}$ is the i-th smallest p-value, for $i = 1, \ldots, N_{\text{rep}}$, and $p_{(N_{\text{rep}}+1)} = 1$. If the AUC for a given setting is close to 1, then the corresponding p-values are concentrated towards zero and, therefore, the statistical evidence that (indep) has to be rejected is high.

Appendix

Proof of (4)

Letting

$$\frac{1}{S} = \frac{P(L,C)\,\mathrm{P}(C,R)}{\mathrm{P}(C)\,\mathrm{P}(L,C,R)}, \tag{6}$$

we have

$$\begin{aligned}
&\frac{\mathrm{P}(\varphi,\psi \,|\, L,C,R)\,\mathrm{P}(\varphi,\psi \,|\, C)}{\mathrm{P}(\varphi,\psi \,|\, L,C)\,\mathrm{P}(\varphi,\psi \,|\, C,R)} \\
&= \frac{\mathrm{P}(\varphi,\psi,L,C,R)\,\mathrm{P}(\varphi,\psi,C)}{\mathrm{P}(\varphi,\psi,L,C)\,\mathrm{P}(\varphi,\psi,C,R)}\,\frac{\mathrm{P}(L,C)\,\mathrm{P}(C,R)}{\mathrm{P}(L,C,R)\,\mathrm{P}(C)} \\
&= \frac{\mathrm{P}(L,R,\varphi,\psi\,,C)}{\mathrm{P}(\varphi,\psi,C)}\,\frac{\mathrm{P}(\varphi,\psi,C)}{\mathrm{P}(L,\varphi,\psi,C)}\,\frac{\mathrm{P}(\varphi,\psi,C)}{\mathrm{P}(R,\varphi,\psi,C)}\,\frac{1}{S} \\
&= \frac{\mathrm{P}(L,R \,|\, \varphi,\psi,C)}{S\,\mathrm{P}(L \,|\, \varphi,\psi,C)\,\mathrm{P}(R \,|\, \varphi,\psi,C)}\,,
\end{aligned}$$

so that the conditional independence of L and R given C and (φ,ψ) is indeed equivalent to (3).

Software and data availability

The code implementing the statistical tests described in this work as well as the datasets are freely available:

- Software: https://gitlab.laas.fr/moma/STINA
- Data: https://moma.laas.fr/static/data/tripeptide_angles_data.tar

Acknowledgements

This work has been partially supported by the French National Research Agency (ANR) through grant ANR-19-P3IA-0004, the LabEx CIMI (ANR-11-LABX-0040) and EpiGenMed (ANR-10-LABX-12-01) within the French State Programme "Investissements d'Avenir", and by the European Research Council under the European Union's H2020 Framework Programme (2014-2020) / ERC Grant agreement n° [648030] awarded to PB.

References

Anderson, R.J., Weng, Z., Campbell, R.K., Jiang, X., 2005. Main-chain conformational tendencies of amino acids. Proteins 60, 679–689.

Avbelj, F., Baldwin, R.L., 2004. Origin of the neighboring residue effect on peptide backbone conformation. Proc. Natl. Acad. Sci. U.S.A 101, 10967–10972.

Barozet, A., Chacón, P., Cortés, J., 2021. Current approaches to flexible loop modeling. Curr. Res. Struct. Biol. 3, 187–191.

Bernadó, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R.W.H., Blackledge, M., 2005. A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. Proc. Natl. Acad. Sci. U.S.A 102, 17002–17007.

Betancourt, M.R., Skolnick, J., 2004. Local propensities and statistical potentials of backbone dihedral angles in proteins. J. Mol. Biol. 342, 635–649.

Bonferroni, C.E., 1936. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze, Libreria internazionale Seeber.

Boomsma, W., Mardia, K.V., Taylor, C.C., Ferkinghoff-Borg, J., Krogh, A., Hamelryck, T., 2008. A generative, probabilistic model of local protein structure. Proc. Natl. Acad. Sci. U.S.A 105, 8932–8937.

Brändén, C., Tooze, J., 1998. Introduction to Protein Structure (2nd ed.). Garland Science, New York.

Braun, D., Wider, G., Wuethrich, K., 1994. Sequence-corrected ^{15}N "random coil" chemical shifts. J. Am. Chem. Soc. 116, 8466–8469.

Chandonia, J.M., Fox, N.K., Brenner, S.E., 2018. SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. Nucleic Acids Res. 47, D475–D481.

Cho, M.K., Kim, H.Y., Bernado, P., Fernandez, C.O., Blackledge, M., Zweckstetter, M., 2007. Amino acid bulkiness defines the local conformations and dynamics of natively unfolded α-synuclein and tau. J. Am. Chem. Soc. 129, 3032–3033.

Dames, S.A., Aregger, R., Vajpai, N., Bernado, P., Blackledge, M., Grzesiek, S., 2006. Residual dipolar couplings in short peptides reveal systematic conformational preferences of individual amino acids. J. Am. Chem. Soc. 128, 13508–13514.

Deane, C.M., Allen, F.H., Taylor, R., Blundell, T.L., 1999. Carbonyl–carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid. Protein Eng. Des. Sel. 12, 1025–1028.

Estaña, A., Sibille, N., Delaforge, E., Vaisset, M., Cortés, J., Bernadó, P., 2019. Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database. Structure 27, 381–391.e2.

Flory, P.J., 1969. Statistical Mechanics of Chain Molecules. Wiley, New York.

Gibrat, J.F., Garnier, J., Robson, B., 1987. Further developments of protein secondary structure prediction using information theory: New parameters and consideration of residue pairs. J. Mol. Biol. 198, 425–443.

Gibrat, J.F., Robson, B., Garnier, J., 1991. Influence of the local amino acid sequence upon the zones of the torsional angles ϕ and ψ adopted by residues in proteins. Biochemistry 30, 1578–1586.

González-Delgado, J., González-Sanz, A., Cortés, J., Neuvial, P., 2021. Two-sample goodness-of-fit tests on the flat torus based on wasserstein distance and their relevance to structural biology. ArXiv:2108.00165.

Griffiths-Jones, S.R., Sharman, G.J., Maynard, A.J., Searle, M.S., 1998. Modulation of intrinsic ϕ,ψ propensities of amino acids by neighbouring residues in the coil regions of protein structures: NMR analysis and dissection of a β-hairpin peptide. J. Mol. Biol. 284, 1597–1609.

Ho, B.K., Brasseur, R., 2005. The ramachandran plots of glycine and preproline. BMC Struct. Biol. 5, 14.

Hovmöller, S., Zhou, T., Ohlson, T., 2002. Conformations of amino acids in proteins. Acta Crystallogr. D 58, 768–776.

Huang, J.R., Ozenne, V., Jensen, M.R., Blackledge, M., 2013. Direct prediction of nmr residual dipolar couplings from the primary sequence of unfolded proteins. Angew. Chem. Int. Ed. 52, 687–690.

Jha, A.K., Colubri, A., Freed, K.F., Sosnick, T.R., 2005a. Statistical coil model of the unfolded state: Resolving the reconciliation problem. Proc. Natl. Acad. Sci. U.S.A 102, 13099–13104.

Jha, A.K., Colubri, A., Zaman, M.H., Koide, S., Sosnick, T.R., Freed, K.F., 2005b. Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library. Biochemistry 44, 9691–9702.

Jumper, J.M., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D.A., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with alphafold. Nature 596, 583 – 589.

Jung, Y.S., Oh, K.I., Hwang, G.S., Cho, M., 2014. Neighboring residue effects in terminally blocked dipeptides: Implications for residual secondary structures in intrinsically unfolded/disordered proteins. Chirality 26, 443–452.

Kabat, E.A., Wu, T.T., 1973. The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in proteins: Comparison of predicted and experimental determination of β-sheets in concanavalin a. Proc. Natl. Acad. Sci. U.S.A 70, 1473–1477.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.

Kang, H.S., Kurochkina, N.A., Lee, B., 1993. Estimation and use of protein backbone angle probabilities. J. Mol. Biol. 229, 448–460.

Kohn, J.E., Millett, I.S., Jacob, J., Zagrovic, B., Dillon, T.M., Cingel, N., Dothager, R.S., Seifert, S., Thiyagarajan, P., Sosnick, T.R., Hasan, M.Z., Pande, V.S., Ruczinski, I., Doniach, S., Plaxco, K.W., 2004. Random-coil behavior and the dimensions of chemically unfolded proteins. Proc. Natl. Acad. Sci. U.S.A 101, 12491–12496.

Lehmann, E.L., Romano, J.P., 2005. Testing statistical hypotheses. volume 3. Springer, New York.

Liljas, A., Liljas, L., Piskur, J., Lindblom, G., Nissen, P., Kjeldgaard, M., 2009. Textbook Of Structural Biology. World Scientific Publishing, Singapore.

Lovell, S.C., Davis, I.W., Arendall III, W.B., de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., Richardson, D.C., 2003. Structure validation by cα geometry: ϕ, ψ and cβ deviation. Proteins 50, 437–450.

Milorey, B., Schwalbe, H., O'Neill, N., Schweitzer-Stenner, R., 2021a. Repeating aspartic acid residues prefer turn-like conformations in the unfolded state: Implications for early protein folding. J. Phys. Chem. B 125, 11392–11407.

Milorey, B., Schweitzer-Stenner, R., Andrews, B., Schwalbe, H., Urbanc, B., 2021b. Short peptides as predictors for the structure of polyarginine sequences in disordered proteins. Biophys. J. 120, 662–676.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G., Thornton, J.M., 1992. Stereochemical quality of protein structure coordinates. Proteins 12, 345–364.

Oh, K.I., Jung, Y.S., Hwang, G.S., Cho, M., 2012a. Conformational distributions of denatured and unstructured proteins are similar to those of 20 x 20 blocked dipeptides. J. Biomol. NMR 53, 25–41.

Oh, K.I., Lee, K.K., Park, E.K., Jung, Y., Hwang, G.S., Cho, M., 2012b. A comprehensive library of blocked dipeptides reveals intrinsic backbone conformational propensities of unfolded proteins. Proteins 80, 977–990.

Ozenne, V., Bauer, F., Salmon, L., Huang, J.r., Jensen, M.R., Segard, S., Bernadó, P., Charavay, C., Blackledge, M., 2012. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics 28, 1463–1470.

Pappu, R.V., Srinivasan, R., Rose, G.D., 2000. The flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding. Proc. Natl. Acad. Sci. U.S.A 97, 12565–12570.

Penkett, C.J., Redfield, C., Dodd, I., Hubbard, J., McBay, D.L., Mossakowska, D.E., Smith, R.A., Dobson, C.M., Smith, L.J., 1997. NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein. J. Mol. Biol. 274, 152–159.

Peyré, G., Cuturi, M., 2019. Computational optimal transport: With applications to data science. Foundations and Trends in Machine Learning 11, 355–607.

Phipson, B., Smyth, G.K., 2010. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. Stat. Appl. Genet. Mol. Biol. 9.

Ramachandran, G., Ramakrishnan, C., Sasisekharan, V., 1963. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 7, 95–99.

Ramachandran, G., Sasisekharan, V., 1968. Conformation of polypeptides and proteins. Adv. Protein Chem. 23, 283–437.

Rata, I.A., Li, Y., Jakobsson, E., 2010. Backbone statistical potential from local sequence-structure interactions in protein loops. J. Phys. Chem. B 114, 1859–1869.

Schweitzer-Stenner, R., Toal, S.E., 2018. Anticooperative nearest-neighbor interactions between residues in unfolded peptides and proteins. Biophys. J. 114, 1046–1057.

Serrano, L., 1995. Comparison between the ψ distribution of the amino acids in the protein database and nmr data indicates that amino acids have various ψ propensities in the random coil conformation. J. Mol. Biol. 254, 322–333.

Shen, Y., Roche, J., Grishaev, A., Bax, A., 2018. Prediction of nearest neighbor effects on backbone torsion angles and nmr scalar coupling constants in disordered proteins. Protein Sci. 27, 146–158.

Smith, L.J., Bolin, K.A., Schwalbe, H., MacArthur, M.W., Thornton, J.M., Dobson, C.M., 1996a. Analysis of main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil conformations. J. Mol. Biol. 255, 494–506.

Smith, L.J., Fiebig, K.M., Schwalbe, H., Dobson, C.M., 1996b. The concept of a random coil: Residual structure in peptides and denatured proteins. Fold. Des. 1, R95–R106.

Swindells, M.B., MacArthur, M.W., Thornton, J.M., 1995. Intrinsic ϕ and ψ propensities of amino acids, derived from the coil regions of known structures. Nat. Struct. Mol. 2, 596–603.

Ting, D., Wang, G., Shapovalov, M., Mitra, R., Jordan, M.I., Dunbrack, R., 2010. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. PLoS Comput. Biol. 6, e1000763.

Toal, S.E., Kubatova, N., Richter, C., Linhard, V., Schwalbe, H., Schweitzer-Stenner, R., 2015. Randomizing the unfolded state of peptides (and proteins) by nearest neighbor interactions between unlike residues. Chem. Eur. J. 21, 5173–5192.

Touw, W.G., Baakman, C., Black, J., te Beek, T.A.H., Krieger, E., Joosten, R.P., Vriend, G., 2014. A series of PDB-related databanks for everyday needs. Nucleic Acids Res. 43, D364–D368.

Villani, C., 2008. Optimal Transport: Old and New. Springer-Verlag, Berlin, Heidelberg.

Zaman, M.H., Shen, M.Y., Berry, R., Freed, K.F., Sosnick, T.R., 2003. Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the Flory isolated-pair hypothesis for peptides. J. Mol. Biol. 331, 693–711.

Zimmerman, S.S., Pottle, M.S., Némethy, G., Scheraga, H.A., 1977. Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP. Macromolecules 10, 1–9.