# The dependence of the amino acid backbone conformation on the translated synonymous codon is not statistically significant

Javier González-Delgado[1,2], Pablo Mier[3], Pau Bernadó[4],
Pierre Neuvial[2] and Juan Cortés[1]

[1] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France.*

[2] *Institut de Mathématiques de Toulouse, Université de Toulouse, CNRS, Toulouse, France.*

[3] *Institute of Organismic and Molecular Evolution, Faculty of Biology, Johannes Gutenberg University Mainz, Germany.*

[4] *Centre de Biologie Structurale, Université de Montpellier, INSERM, CNRS, France.*

In their recent work, Rosenberg *et al.* [1] studied the dependence between the identity of synonymous codons and the distribution of the backbone dihedral angles of the translated amino acids. In the past, it has been shown that the use of synonymous codons is highly relevant in multiple biological processes including, among others, mRNA splicing, translational rates and protein folding [2, 3]. While the correlation between synonymous codons and secondary structure in translated proteins has been widely studied [4–6], Rosenberg *et al.* evaluated the effect of codon identity on a finer scale, analyzing whether the distribution of $(\phi, \psi)$ dihedral angles within secondary structure elements is significantly altered when synonymous codons are used. Their conclusion, showing significant differences, particularly for amino acid residues involved in $\beta$-strands, would represent a new paradigm for the role played by synonymous codons in defining protein structure. However, the statistical methodology used in that study was formally incorrect, casting doubt on the obtained results. Besides, it is based on density estimates that might be imprecise for small sample sizes, yielding misleading comparisons. These methodological errors are described in the following section. Then, using an appropriate methodology, we reanalyzed the data presented in [1]. Our results show that the influence of the codon on the distribution of the dihedral angles is statistically non-significant for all types of secondary structures, contradicting the conclusion by Rosenberg *et al.*. These results were corroborated by repeating the analysis on structures extracted from the AlphaFold Database [7, 8] for the same set of proteins, and shown to be robust with respect to the definition secondary structural classes and also when considering the nature of the neighbor residues. Overall, our observations demonstrate that the influence of the synonymous codons on the backbone dihedral angles can not be inferred with current data.

## Limitations of the original methodology

### *Ill-defined p-values*

The goal of Rosenberg *et al.* was to assess the effect of synonymous codons on the distribution of $(\phi, \psi)$ dihedral angles by comparing codon-specific Ramachandran plots. Keeping the notation of [1], if $(c, c')$ denotes a pair of synonymous codons and $\mathcal{X}$ a type of secondary structure, they aimed at testing the null hypothesis $H_{0,(c,c')|\mathcal{X}}$ that both codon-specific distributions are the same. To do so, the authors introduced a metric to quantify differences between the distributions corresponding to different codons. Then, to assess the significance of such differences, Rosenberg *et al.* proposed to draw $B = 25$ pairs of bootstrapped samples, and to compare them with their synonymous codon counterparts using a permutation test procedure, with $K = 200$ permutations. For each bootstrap sample $b \in \{1, \ldots, B\}$, if $n_b$ denotes the number of permutations where the permuted metric is larger than the base metric (obtained from non-permuted data), they proposed the quantity

$$p_{(c,c'),\mathcal{X}} = \frac{1 + \sum_{b=1}^{B} n_b}{1 + BK} \tag{1}$$

as a $p$-value for $H_{0,(c,c')|\mathcal{X}}$. We can reformulate (1) in order to gain insight into its statistical behavior. First, let us define

$$p_b = \frac{1 + n_b}{1 + K}\,, \tag{2}$$

which is a well-defined $p$-value for the $b$-th permutation test. Letting

$$\overline{p}_B = \frac{1}{B}\sum_{b=1}^{B} p_b\,, \tag{3}$$

it can be shown (see SI) that

$$|p_{(c,c'),\mathcal{X}} - \overline{p}_B| \leq \frac{1}{K}\,. \tag{4}$$

That is, for sufficiently large $K$, $p_{(c,c'),\mathcal{X}}$ is approximately the empirical mean of the $B$ $p$-values associated to individual permutation tests.

However, $\overline{p}_B$ is not a valid $p$-value (see SI for a formal proof). Let us recall that a $p$-value $p$ is statistically valid if and only if its distribution under the null hypothesis is Super-Uniform. A random variable is said to be Super-Uniform if its cumulative distribution function (CDF) $F$ is upper bounded by that of the Uniform distribution (denoted by $\mathrm{U}[0,1]$ below), that is:

$$F(x) \leq x \text{ for all } x \text{ in } [0,1] \tag{5}$$

(see e.g. [9, Section 3.3]). Moreover, the closer the $p$-value distribution under the null hypothesis is to $\mathrm{U}[0,1]$, the more powerful the corresponding test is. Condition (5) is satisfied for classical permutation $p$-values such as $p_b$ (with the CDF getting closer to the $\mathrm{U}[0,1]$ distribution as $K$ increases), but not for averages of $p$-values like $\overline{p}_B$. Instead, all the $p_b$ could be correctly aggregated by taking their minimum and correcting the result for multiple testing (Bonferroni aggregation).

If the $p_b$ were independent, then, by the Central Limit Theorem (e.g. [10, Theorem 27.1]), the distribution of $\overline{p}_B$ would be asymptotically Gaussian $\mathcal{N}(1/2, 1/\sqrt{12B})$ as $B$ tends to infinity. This distribution does not verify (5), and therefore tests based on such a distribution are mathematically invalid. In the setting of [1], the $p_b$ are not independent since they have been computed by bootstrapping from one initial sample. However, for small values of $B$ (including the choice of $B = 25$ in [1]), the null distribution of (1) deviates only slightly from the asymptotic independence setting. This is illustrated in Figure 1, where the null distribution of (1) is simulated using the parameters chosen in [1]. Details on the simulation and further analyses of the effect of $K$ and $B$ are included in the SI.

The empirical distribution of $p_{(c,c'),\mathcal{X}}$ presented in Figure 1 does not satisfy Condition (5). Moreover, it is extremely conservative for large values of the statistic realization that is, low $p$-values, yielding an important number of false negatives and thus ignoring substantial differences appearing between the compared samples.

Finally, since the scores $p_{(c,c'),\mathcal{X}}$ are not valid $p$-values, they cannot be incorporated in a multiple testing procedure [11]. In particular, the Benjamini-Hochberg procedure [12] used in [1] needs the $p$-values to be Super-Uniform under the null hypothesis to control the False Discovery Rate (FDR). Consequently, using and adjusting (1) for multiplicity will yield misleading analyses of the overall behaviour of all the

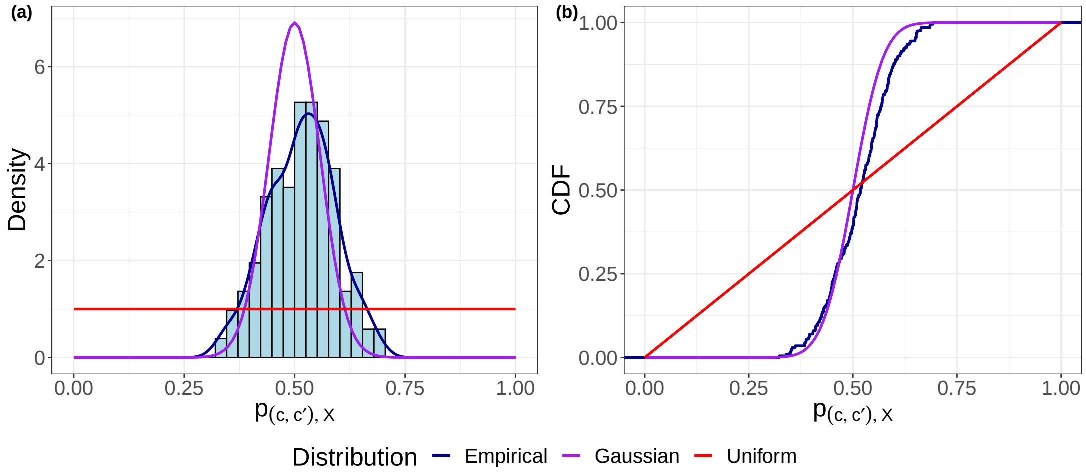

Figure 1: Simulation of the null distribution of $p_{(c,c'),\mathcal{X}}$ for $K = 200$ and $B = 25$, chosen in [1]. Left panel (a): histogram and kernel density estimate. Right panel (b): empirical Cumulative Distribution Function (CDF). Purple lines: asymptotic Gaussian distribution $\mathcal{N}(1/2, 1/\sqrt{12B})$; red lines: uniform distribution on $[0, 1]$.

null hypotheses and therefore, inaccurate results when the specificities of individual amino acids are studied *a posteriori*.

### *Density estimates and reduced sample sizes*

Beyond the above-mentioned methodological issues, the approach proposed in [1] presents several practical limitations. It needs, on the one hand, a substantial reduction of sample sizes, which may imply an important loss of information in some cases and thus a substantial power reduction. Indeed, the maximum sample size in [1] was set to $N_{\max} = 200$, whereas, for instance, the mean sample size for $\alpha$-helical conformations was 724 and only 18% of the samples had sizes below $N_{\max}$. On the other hand, it requires a prior parametric estimation of the underlying densities, whose parameters would need to be optimized. In [1], the authors opted to fix the same bandwidth for all comparisons. However, too small bandwidths can lead to density undersmoothing, especially for small sample sizes. This would yield biased kernel estimates whose comparison might lead to false positives.

## Goodness-of-fit between codon-specific $(\phi, \psi)$ distributions

In our recent work [13], we defined two two-sample goodness-of-fit tests for probability distributions supported on the two-dimensional flat torus, in order to study local changes on polypeptide backbone conformations. Both approaches are non-parametric and they use the information provided by entire datasets. The test statistic is based on the 2-Wasserstein distance, which integrates the geometry of the underlying space and provides strong theoretical guarantees and attractive empirical performance [14]. Here, we implemented the first of the testing procedures defined in [13], called $N_g$-geod, to detect differences between the codon-specific Ramachandran plots provided in [1]. For each amino acid, we tested all the pairwise differences between the $(\phi, \psi)$ distributions for all pairs of synonymous codons. As in [1], we kept data points where codons were unambiguously assigned (whose codon scores equal one). Similarly, we repeated the same redundancy filtering based on averaging $(\phi, \psi)$ points with identical Uniprot ID and sequence position. Note that an alternative filtering approach, which kept every first redundant point in

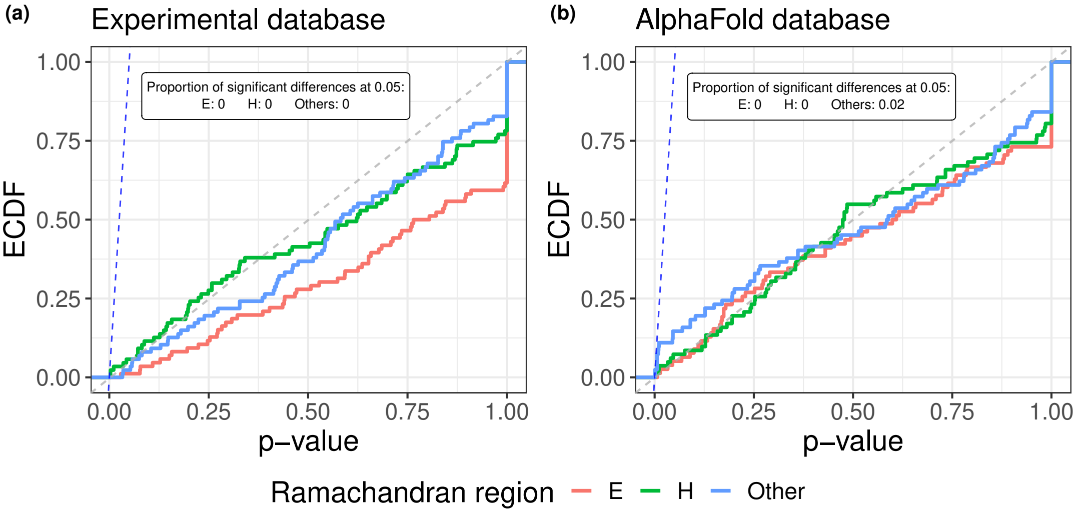

Figure 2: Empirical cumulative distribution function (ECDF) of corrected $p$-values corresponding to testing the equality of $(\phi, \psi)$ distribution pairs corresponding to different synonymous codons in the (a) experimental and (b) AlphaFold database, for conformations in extended strand (E, red), $\alpha$-helix (H, green) and other (*Others*, blue) secondary structures. The dashed blue line of slope $1/\alpha$ corresponds to a target FDR set to $\alpha = 0.05$ for the Benjamini-Hochberg (BH) correction, determining the proportion of rejections among each set of tested hypotheses. The dashed gray line represents the CDF of a Uniform distribution.

the dataset, produced similar results. To facilitate the comparison with the results in [1], we kept only pairs of samples with sizes $n, m \geq 30$ and we classified all conformations according to their secondary structure according to DSSP [15]: extended strand (E) and $\alpha$-helix (H). We also performed the analysis for all the conformations not belonging to any of these two classes, which we named *Others*. As in [1], the Benjamini-Hochberg multiplicity correction [12] was performed to the computed $p$-values. When representing the Empirical Cumulative Distribution Function (ECDF) of the $p$-values, points laying above the line of slope $1/\alpha$ are considered rejections for a target FDR of $\alpha$ according to the BH procedure. The results are presented in Figure 2(a).

The $p$-value distributions presented in Figure 2(a) indicate that, for all the tested hypotheses, differences between codon-specific Ramachandran plots are not significant at level $\alpha = 0.05$. Indeed, the three depicted ECDF lay below the line of slope $1/\alpha$ and therefore no rejections are produced for a target FDR of $\alpha$. Note that our results for $\alpha$-helical structures agree with those presented in [1], for which no significant discrepancies were retrieved. However, results for E structures strongly contradict those in [1], for which significant conformational differences were found for 66% of the synonymous codon pairs tested. Therefore, the main conclusion in [1] is firmly refuted when an appropriate statistical approach is used and, as a consequence, no significant effect of the translated codon identity on the amino-acid backbone conformation can be inferred from the current data.

The discrepancies between the two analyses are most probably due to the above-discussed incorrectness of the methods applied in the original study and, especially, from the potential use of biased density estimates to describe small $(\phi, \psi)$ samples. Indeed, when we looked at the codon pairs whose $(\phi, \psi)$ distributions were found significantly different in [1], we observed that their sample sizes concentrated around the smallest values in the dataset (see Figure 3). This correspondence between significant differences and small sample sizes is counterintuitive for a well-defined statistical test. When the sample size is small and there is limited information about the underlying distribution, the null hypothesis is not rejected unless

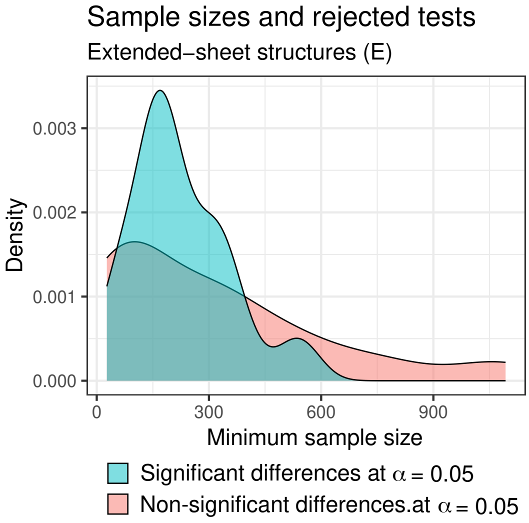

Figure 3: Distribution (kernel density estimates) of the minimum sample size for the codon pairs in the dataset provided in [1], after aggregating redundant data points and removing those with ambiguous codon assignment. Groups correspond to codon pairs for which differences between their codon-specific $(\phi, \psi)$ distributions were found significant (blue) and non-significant (red) at level $\alpha = 0.05$ in [1].

the evidence against it is very strong. Similarly, $p$-values closer to zero were often found for larger sample sizes, where the statistical power (the test's ability to detect differences) is higher. This phenomenon, which is depicted in Figure 3 for E structures, suggested that false positives might be appearing due to misleading comparisons of small $(\phi, \psi)$ samples. As we discussed in the previous section, this may be caused by undersmoothed kernel density estimates computed with too small bandwidths for that setting.

### Additional analyses confirm the robustness of our observations

To validate our conclusions, we performed additional analyses considering different settings. First, we used structures extracted from the AlphaFold Database [7, 8] for the same set of sequences. Note that only residues with pLDDT values larger than 90 were collected for the analysis. Results, depicted in Figure 2(b), qualitatively match those in Figure 2(a) and therefore support the aforementioned conclusions.

The results presented above, as those of Rosenberg *et al.*, were based on the structural classification provided by DSSP [15]. We performed the same analyses using a less restrictive classification of structural classes, only considering conformational regions on the Ramachandran space based on non-overlapping angular intervals and disregarding the formation of hydrogen bonds [16, 17]. The corresponding results (see SI) showed that differences between codon-specific Ramachandran plots were statistically non-significant.

Finally, we assessed whether the consideration of the nearest neighbors effects on $(\phi, \psi)$ distributions may alter our results. Indeed, the invalidity of Flory's Isolated Pair Hypothesis [18] and the interdependence of neighbor effects have been demonstrated in several experimental [19, 20] and theoretical/computational studies [21, 22]. The consideration of neighboring residues is particularly relevant here, as the dataset in [1] exhibits important differences in the proportion of left and right neighboring amino acid types among synonymous codons (see Figure S3). After repeating the same analysis but fixing the identities of left and right neighbors, the overall conclusions remain the same. Results, presented in the SI, show once again that differences between codon-specific Ramachandran plots are non-significant for all secondary structure types.

### Concluding remarks

The work of Rosenberg *et al.* introduced a new paradigm in biology: the nature of the codon influences the $(\phi, \psi)$ angles of protein secondary structures. While the correlation between synonymous codons and secondary structure in the translated proteins is a well known phenomenon [4–6], differences at the $(\phi, \psi)$ level for the most populated conformational states emerged as an intriguing and controversial observation [23]. The conclusions reached from their work could have major impact on one of the paradigms of structural biology, which should shift from protein-sequence to DNA-sequence structure encoding, an information that is not currently stored in structural databases.

With the present study, we have demonstrated the incorrectness of the statistical methodology proposed in [1] to compare probability distributions supported on the two-dimensional flat torus. This, together with the use of density estimates that are not appropriately tuned for small sample sizes, makes the approach in [1] unsuitable to correctly compare codon-specific Ramachandran plots. When applying our previously developed statistical methodology [13] to the same database, no statistically significant differences between the structure encoded by synonymous codons could be detected. Importantly, we demonstrated that this observation is robust with respect to the origin of the 3D structure, the definition of the structural classes and the type of the flanking residues. Therefore, the ensemble of our results unambiguously show that, based on available data, a significant influence of the codon usage in the distribution of backbone dihedral angles in proteins cannot be inferred.

It is worth mentioning, however, that our results have been derived from a limited set of *Escherichia coli* proteins for which the structure had been experimentally determined, and assuming that the gene used for the production of the protein was the same as in the original organism. We believe that a general understanding of the of codon-specific Ramachandran plots can only be achieved by using extensive structural databases, including the corresponding gene sequence, and applying robust statistical methods, such as the one presented here.

### Software availability

The code to reproduce the analyses presented here is available at https://github.com/gonzalez-delgado/synco. The two testing procedures defined in [13] for assessing differences between $(\phi, \psi)$ distributions are implemented in the `R` package `torustest`, available at https://github.com/gonzalez-delgado/torustest.

### Author contributions

All the authors designed the studies, interpreted the results and wrote the manuscript; J.G. developed all the computational methods and performed the analyses; J.G. and P.N. carried out the theoretical analyses.

### Competing interests

The authors declare no competing interests.

## References

[1] A. A. Rosenberg, A. Marx, and A. M. Bronstein. “Codon-specific Ramachandran plots show amino acid backbone conformation depends on identity of the translated codon”. *Nature Communications* 13.1 (2022).

[2] F. Pagani, M. Raponi, and F. E. Baralle. “Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution”. *Proceedings of the National Academy of Sciences* 102.18 (2005), pp. 6368–6372.

[3] F. Buhr, S. Jha, M. Thommen, J. Mittelstaet, F. Kutz, H. Schwalbe, M. V. Rodnina, and A. A. Komar. “Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations”. *Molecular Cell* 61.3 (2016), pp. 341–351.

[4] M. Orešič and D. Shalloway. “Specific correlations between relative synonymous codon usage and protein secondary structure”. *Journal of Molecular Biology* 281.1 (1998), pp. 31–48.

[5] R. Saunders and C. M. Deane. “Synonymous codon usage influences the local protein structure observed”. *Nucleic Acids Research* 38.19 (2010), pp. 6719–6728.

[6] S. Pechmann and J. Frydman. “Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding”. *Nature Structural & Molecular Biology* 20.2 (2013), pp. 237–243.

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

[8] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, and S. Velankar. “AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models”. *Nucleic Acids Research* 50.D1 (2021), pp. D439–D444.

[9] E. L. Lehmann, J. P. Romano, and G. Casella. *Testing statistical hypotheses*. Vol. 3. Springer, 2005.

[10] P. Billingsley. *Probability and measure*. John Wiley & Sons, 1995.

[11] E. Roquain. “Type I error rate control for testing many hypotheses: a survey with proofs”. *Journal de la Société Française de Statistique* 152.2 (2011), pp. 3–38.

[12] Y. Benjamini and Y. Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300.

[13] J. González-Delgado, A. González-Sanz, J. Cortés, and P. Neuvial. “Two-sample goodness-of-fit tests on the flat torus based on Wasserstein distance and their relevance to structural biology”. *Electronic Journal of Statistics* 17.1 (2023), pp. 1547–1586.

[14] G. Peyré and M. Cuturi. “Computational Optimal Transport: With Applications to Data Science”. *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.

[15] W. Kabsch and C. Sander. “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features”. *Biopolymers* 22.12 (1983), pp. 2577–2637.

[16] V. Ozenne, R. Schneider, M. Yao, J.-r. Huang, L. Salmon, M. Zweckstetter, M. R. Jensen, and M. Blackledge. “Mapping the Potential Energy Landscape of Intrinsically Disordered Proteins at Amino Acid Resolution”. *Journal of the American Chemical Society* 134.36 (2012). PMID: 22901047, pp. 15138–15148.

[17] A. Estaña, A. Barozet, A. Mouhand, M. Vaisset, C. Zanon, P. Fauret, N. Sibille, P. Bernadó, and J. Cortés. “Predicting Secondary Structure Propensities in IDPs Using Simple Statistics from Three-Residue Fragments”. *Journal of Molecular Biology* 432.19 (2020), pp. 5447–5459.

[18] P. J. Flory and M. Volkenstein. “Statistical mechanics of chain molecules”. *Biopolymers* 8.5 (1969), pp. 699–700.

[19] K.-I. Oh, Y.-S. Jung, G.-S. Hwang, and M. Cho. “Conformational distributions of denatured and unstructured proteins are similar to those of 20 x 20 blocked dipeptides”. *Journal of Biomolecular NMR* 53 (2012), pp. 25–41.

[20] R. Schweitzer-Stenner and S. E. Toal. “Anticooperative nearest-neighbor interactions between residues in unfolded peptides and proteins”. *Biophysical Journal* 114.5 (2018), pp. 1046–1057.

[21] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and R. Dunbrack. “Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model”. *PLoS Computational Biology* 6.4 (2010), e1000763.

[22] J. González-Delgado, P. Bernadó, P. Neuvial, and J. Cortés. “Statistical proofs of the interdependence between nearest neighbor effects on polypeptide backbone conformations”. *Journal of Structural Biology* 214.4 (2022), p. 107907.

[23] O. J. Akeju and A. L. Cope. “Re-examining correlations between synonymous codon usage and protein bond angles in E. coli”. *Genome Biology and Evolution* (2024), evae080.