Supplementary material

for the manuscript

Geometry-based Distance for Clustering Amino Acids

by Samira F. Abushilah[1,2], Charles C. Taylor[1], and Arief Gusnanto[1]

[1]Department of Statistics, University of Leeds, Leeds LS2 9JT, United Kingdom; [2]Department of Mathematics, Faculty of Education for Girls, University of Kufa, Najaf, Iraq

1 Further descriptions on the datasets

In this section we describe further some summary statistics in for the datasets in this study.

1.1 First dataset

The first dataset is described in Lovell *et al.* (2003) and are available from

`http://kinemage.biochem.duke.edu/databases/top500.php.`

The dataset contains a selection of 500 proteins from the Protein Data Bank (PDB) are catalogued. The 500 proteins were selected because of their high quality, low homology and high resolution (1.8 Å or better) (Lovell *et al.*, 2003). The summary statistics are presented in Table S1. Note that, in angular data, the minimum and maximum *points* ($-180°$, $180°$) are arbitrary because the data are wrapped around a circle: both are the same. However, we present in the table the minimum and maximum values that we observe in the data to show that some amino acids' dihedral angles do not occupy the whole range of domain. Based on the values of mean and median for ϕ and ψ, some amino acids indicate that they are not concentrated symmetrically. This indication is illustrated in Figure 1 in the main manuscript.

1.2 Second dataset

The second dataset is also from the Protein Data Bank (PDB). We select *all* proteins from PDB that have less than 1.5Å in resolution, and published (uploaded to the repository) from 2015 to May 2019 (inclusive). In this dataset, we have 1,259 proteins and none of them are overlapping with the first dataset. The summary statistics are presented in Table S2. The table indicates that the summary statistics for amino acids in the second dataset are similar to those in Table S1. This is interesting since the proteins are different in both datasets. In a way, this can be considered as a validation dataset in terms of results. Further description of this dataset, and the results of analysis on this dataset, are available in Section 4 of this supplementary material.

Amino Acids	Number of dihendral angles[1]	The first angle ϕ						The second angle ψ					
		Min[2]	1st Qu.	Median	Mean	3rd Qu.	Max[2]	Min[2]	1st Qu.	Median	Mean	3rd Qu.	Max[2]
ASN	5029	-177.90	-107.40	-82.70	-73.98	-62.90	176.90	-179.90	-27.80	25.60	39.67	113.90	180.00
ILE	5850	-170.10	-116.80	-90.20	-91.64	-65.30	137.30	-179.70	-40.90	111.15	55.85	130.20	176.80
PHE	4375	-177.00	-121.30	-90.30	-92.26	-64.30	92.60	-179.90	-37.95	83.10	56.93	140.40	180.00
GLU	6501	-177.50	-91.90	-68.60	-77.38	-62.00	178.90	-179.30	-40.60	-26.50	22.48	123.90	179.80
MET	2153	-179.00	-106.90	-72.80	-84.60	-63.30	90.30	-180.00	-39.40	-11.20	37.49	134.50	178.70
LEU	9083	-175.70	-102.10	-74.60	-82.91	-63.60	121.80	-179.20	-39.80	-11.50	36.53	129.60	179.80
ARG	4866	-179.00	-106.50	-72.30	-82.28	-62.40	153.60	-179.60	-39.58	-11.25	35.61	133.00	179.70
ASP	6555	-176.60	-97.40	-75.80	-77.31	-62.90	177.20	-180.00	-36.40	5.10	30.88	112.50	180.00
GLY	8749	-180.00	-74.20	62.90	9.293	88.20	180.00	-180.00	-45.20	-3.30	-2.09	33.70	179.90
LYS	6298	-178.50	-102.00	-71.50	-79.29	-61.92	178.90	-179.90	-39.30	-12.70	34.63	131.38	180.00
TYR	3926	-176.80	-121.50	-92.50	-92.50	-64.50	179.80	-179.40	-35.60	95.50	58.97	140.30	179.50
THR	6280	-179.20	-121.20	-94.30	-95.14	-67.70	163.50	-180.00	-33.60	89.95	56.71	144.20	180.00
HIS	2525	-176.20	-119.60	-85.30	-86.93	-64.30	178.40	-177.40	-34.90	37.20	51.30	136.30	179.10
SER	6628	-178.30	-115.22	-77.90	-87.41	-64.30	176.60	-180.00	-31.30	11.85	51.36	147.00	179.90
PRO	5052	-111.70	-71.40	-63.40	-64.95	-57.60	-21.90	-179.70	-26.10	128.55	69.01	147.80	179.90
ALA	9427	-178.60	-87.10	-66.40	-77.79	-61.20	179.40	-179.90	-40.70	-27.20	25.83	133.20	179.60
VAL	7730	-175.70	-121.20	-96.90	-95.12	-66.60	103.80	-179.20	-39.10	117.10	63.15	133.60	179.00
GLN	4129	-176.00	-100.80	-70.90	-79.60	-62.70	168.40	-179.90	-39.70	-18.00	29.53	128.10	179.30
TRP	1624	-171.90	-112.40	-77.75	-87.40	-62.98	176.90	-179.80	-40.10	7.75	45.61	137.50	179.10
CYS	1737	-178.20	-119.20	-87.30	-89.66	-65.80	177.90	-180.00	-31.10	102.90	61.47	142.80	179.80

Table S1: *Summary statistics of the first dataset described in Lovell* et al. *(1996). The dataset contains a selection of 500 proteins from the Protein Data Bank (PDB) are catalogued. The 500 proteins were selected because of their high quality, low homology and high resolution (1.8 Å or better).* [1] *This is the number of pairs of dihedral angles ϕ and ψ of the different types of amino acids across the 500 proteins.* [2] *Note that, in angular data, the minimum and maximum points ($-180°$, $180°$) are arbitrary because the data are wrapped around a circle: the both are the same. We present the minimum and maximum values that we observe in the data to show that some amino acids' dihedral angles do not occupy the whole range of domain.*

Amino Acids	Number of dihendral angles[1]	The first angle ϕ						The second angle ψ					
		Min[2]	1st Qu.	Median	Mean	3rd Qu.	Max[2]	Min[2]	1st Qu.	Median	Mean	3rd Qu.	Max[2]
ASN	5943	-179.73	-111.93	-81.77	-73.14	-60.64	179.96	-179.98	-33.86	24.18	34.84	112.15	179.99
ILE	7989	-178.88	-115.11	-88.44	-89.91	-65.97	179.76	-179.92	-39.60	100.07	52.04	132.14	179.99
PHE	5496	-179.48	-120.67	-89.47	-89.29	-64.00	179.78	-179.90	-39.71	68.83	49.71	138.16	179.94
GLU	10710	-179.87	-101.18	-72.82	-77.26	-61.11	179.95	-179.79	-41.53	-23.88	20.07	113.64	179.95
MET	3019	-179.79	-108.45	-74.16	-79.29	-62.06	179.86	-179.64	-40.93	-14.00	30.76	124.52	179.80
LEU	12522	-179.85	-103.24	-76.41	-80.99	-62.34	179.89	-179.91	-41.39	-16.21	31.27	126.84	179.92
ARG	7561	-179.84	-111.55	-78.06	-81.15	-61.63	179.95	-179.95	-40.94	-9.50	31.26	126.75	179.99
ASP	8429	-179.97	-102.94	-76.59	-75.43	-61.0	179.88	-179.93	-39.42	-1.59	25.95	111.76	179.85
GLY	10351	-179.92	-86.08	53.29	6.087	95.58	179.93	-179.99	-48.23	-0.95	2.407	56.85	179.95
LYS	9952	-179.57	-111.37	-77.86	-79.99	-61.47	179.99	-179.93	-40.62	-8.86	31.60	125.59	179.96
TYR	4464	-179.23	-123.11	-91.69	-92.18	-65.82	177.85	-179.89	-38.45	91.51	53.52	139.26	179.87
THR	7827	-180.00	-123.06	-92.01	-91.20	-67.92	179.79	-179.94	-36.23	85.65	51.54	139.98	179.99
HIS	3772	-179.98	-121.04	-86.84	-85.25	-64.74	179.04	-179.98	-35.49	52.43	45.37	129.79	179.76
SER	9464	-179.99	-122.90	-83.40	-83.88	-63.19	179.95	-180.00	-35.10	36.81	44.97	138.16	180.00
PRO	6505	-154.34	-75.21	-69.13	-67.93	-59.17	157.50	-179.92	-28.73	71.18	50.31	140.77	179.95
ALA	10440	-179.94	-96.33	-70.08	-76.20	-60.25	179.91	-179.99	-42.39	-25.13	21.12	120.29	179.98
VAL	10126	-178.60	-121.75	-93.37	-93.30	-66.94	179.76	-179.91	-35.48	112.47	63.76	136.52	179.75
GLN	5652	-179.77	-105.62	-74.66	-78.97	-61.60	179.99	-179.83	-41.15	-21.02	23.36	119.63	179.93
TRP	1722	-177.25	-114.45	-86.65	-89.62	-66.54	176.46	-179.39	-37.42	76.63	49.54	135.70	177.84
CYS	3308	-179.85	-121.42	-90.62	-89.27	-66.26	179.44	-179.95	-31.02	104.85	59.80	140.80	179.97

Table S2: *Summary statistics of the second dataset from the Protein Data Bank (PDB). We select all proteins from PDB that have less than 1.5Å in resolution, and published (uploaded to the repository) from 2015 to May 2019 (inclusive). In this dataset, we have 1,259 proteins and none of them are overlapping with the first dataset.* [1] *This is the number of pairs of dihedral angles ϕ and ψ of the different types of amino acids across the 1,259 proteins.* [2] *Note that, in angular data, the minimum and maximum points ($-180°$, $180°$) are arbitrary because the data are wrapped around a circle: the both are the same. We present the minimum and maximum values that we observe in the data to show that some amino acids' dihedral angles do not occupy the whole range of domain.*

2 Dependencies between adjacent amino acids in a protein

Our proposed method to construct the dissimilarity matrix between amino acids involves a test of distribution equality of dihedral angles. When two amino acids are tested, then we include dihedral angles of each amino acid across proteins. For example, consider the amino acids GLY and PRO in the first dataset (illustrated in the bottom panel of Figure 1 in the main manuscript). When we are testing GLY and PRO in a two-sample setting, we include the dihedral angles of GLY across all proteins in one sample, and those of PRO in the other sample. It is possible that, within the amino acids GLY, there are two pairs of dihedral angles that are adjacent in a protein. I.e. in a protein, there is a sequence $\ldots -$ GLY $-$ GLY $- \ldots$. Similarly, for the PRO sample, it is possible that it contains dihedral angles that are adjacent; i.e. there is a sequence $\ldots -$ PRO $-$ PRO $- \ldots$ in a protein.

Adjacent amino acids are expected to have some sort of dependencies or correlation of dihedral angles. This is because, once an amino acid occupy a position in a three-dimensional structure of a protein, then the next amino acid can only occupy a certain position in the structure of protein. This is true regardless of the type of amino acids that are adjacent and is still true in our context when we focus on the same type of amino acids when performing a test. However, the correlation of adjacent amino acids are relatively weak as shown in Figure S1. The figure shows the distribution of correlations between (1) ϕ's of adjacent same-type amino acids in a protein (2) ψ's, of adjacent same-type amino acids in a protein, (3) ϕ and ψ of adjacent same-type amino acid in a protein and (4) ϕ and ψ of the same-type amino acid in a protein.

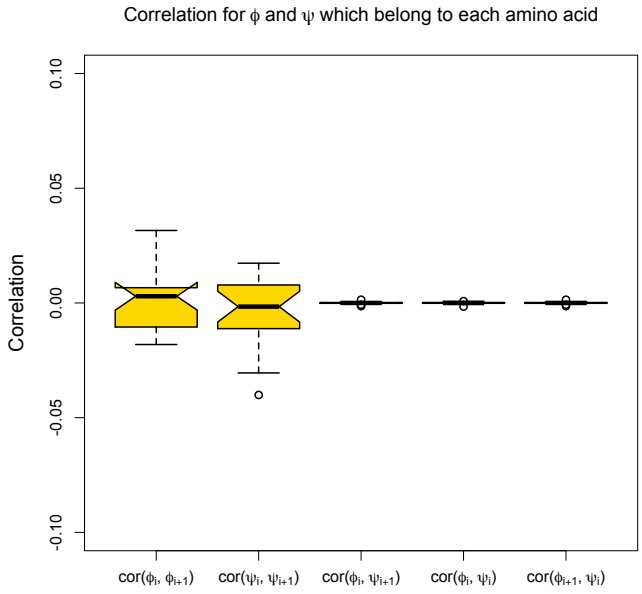

Figure S1: *Correlations between (1) ϕ's of adjacent same-type amino acids in a protein (2) ψ's, of adjacent same-type amino acids in a protein, (3) ϕ and ψ of adjacent same-type amino acid in a protein and (4) ϕ and ψ of the same-type amino acid in a protein.*

With a range of approximately between -0.04 to 0.04, Figure S1 indicates that the correlation between adjacent amino acids in the test are really week to have practical importance. Relative to the number of dihedral angles involves in the test that ranges from 1,624 to 9,427 (in the first dataset – Table S1), the number of adjacent amino acid (of the same type) is relative low. As an illustration, there are 12 pairs of dihedral angles (of GLY) that are adjacent in a protein, which represents a very small fraction of the total 8749 dihedral angles in the first dataset. Similarly for PRO, there are 6 pairs of dihedral angles that are adjacent in a protein, representing a small fraction of the total 5052 dihedral angles.

To check that these adjacent amino acids do not cause a problem, we have performed the analysis where we remove the dihedral angles of one of the adjacent amino acids. As expected, we end up with the same p-values of the test, hence the same dissimilarity matrix. As a result, the resulted clusters are also the same. Hence these dependencies are not a problem in our analysis.

As a note, the correlation between dihedral angles that we consider in this investigation is based on the Fisher-Lee (FL) circular coefficient (Fisher and Lee, 1983). Suppose that (ϕ_1, ψ_1) and (ϕ_2, ψ_2) are independently distributed as (Φ, Ψ). Then the Fisher-Lee (FL) circular correlation coefficient is defined by

$$\rho_{FL}(\Phi, \Psi) = \frac{E[\sin(\phi_1 - \phi_2)\sin(\psi_1 - \psi_2)]}{\sqrt{E[\sin^2(\phi_1 - \phi_2)]E[\sin^2(\psi_1 - \psi_2)]}}, \text{ where } \rho_{FL} \in [-1, 1]$$

3 Physicochemical properties of amino acids

Stanfel (1996) describe the physicochemical properties of amino acids that they use to construct dissimilarity matrix for clustering amino acids. The dissimilarity matrix is given in Table S3 below. The table shows the physicochemical properties of amino acids that they consider to construct the dissimilarity matrix.

Amino Acids		Volume	Area	Polarity	Charge	Shape
1	ASN	126	160	-4.8	0	5.1
2	ILE	164	175	3.1	0	1.45
3	PHE	193	210	3.7	-1	12
4	GLU	142	190	-8.2	0	5.2
5	MET	167	185	3.4	0	3.3
6	LEU	164	170	2.8	1	1.4
7	ARG	195	225	-12.3	-1	8.6
8	ASP	118	150	-9.2	0	5
9	GLY	64	75	1.0	1	1
10	LYS	170	200	-8.8	0	8.5
11	TYR	197	230	-0.7	0	12.05
12	THR	121	140	1.2	1	2.1
13	HIS	159	195	-3.0	0	7
14	SER	95	115	0.6	0	2
15	PRO	124	145	-0.2	0	1.25
16	ALA	90	115	1.6	0	1.1
17	VAL	139	155	2.6	0	1.3
18	GLN	142	180	-4.1	0	5.3
19	TRP	231	255	1.9	0	12.15
20	CYS	113	135	2.0	0	3

Table S3: *Physicochemical properties of amino acids as described in Stanfel (1996). The column 'volume' refers to the volume of amino acids in $Å^3$ as described in Gerstein* et al. *(1994). The column 'area' corresponds to the amino acids area in $Å^2$ as described by Chothia (1984). The column 'polarity' is in kcal/mol as described by Engelman* et al. *(1986). The column of 'charge' is measured just by -1, 0, 1 for negatively charged, uncharged, and positively charged, respectively. The column 'shape' corresponds to the shape of amino acids, by their contructive atoms involved as described by Stryer (1988).*

The physicochemical properties in the table are only indirectly related to the three-dimensional structure of protein. We argue that, when our interest is the three-dimensional structure of protein, then measures that are directly related to the structure would be more relevant for the clustering of

amino acids. Lovell *et al.* (2003) has described that the distribution of dihedral angles in protein are directly related to the protein structure. In a Ramachandran plot that describes the joint distribution of ϕ and ψ of dihedral angles, the occupation of certain regions of the plot are directly associated with the shape of secondary structure: α-helix, β-sheet, or others.

However, one might consider to take into account the above physicochemical properties in the construction of dissimilarity matrix for clustering amino acids, alongside the information from the distribution of dihedral angles. To do this, we can construct a composite dissimilarity matrix that is the result of weighting each dissimilarity matrix from the physicochemical and dihedral angle information. They are normalised first by taking the ratio to make the maximum value to be one.

Using the notation from the main manuscript, this can be done as follows. Let M^d be a 20×20 distance matrix between amino acids based on dihedral angle information. The entries of the matrix, denoted $m^d_{kk'}$ are defined as the negative log or p-value of the test described in Sections 3.2 and 3.3 in the main manuscript. Let M^p be a 20×20 distance matrix between amino acids, whose entries, denoted $m^p_{kk'}$, are defined as the Euclidean distance between amino acids from the physicochemical measurements in Table S3. We assume that the entries of the matrices have been divided by the maximum to get a maximum value of one. Let a be a weight, $0 \leq a \leq 1$. The composite distance matrix, denoted as M^c of size 20×20 with entries $m^c_{kk'}$ is calculated as

$$m^c_{kk'} = a m^p_{kk'} + (1 - a) m^d_{kk'}. \tag{1}$$

Based on this new composite matrix, we can investigate the clustering of amino acids as the weight a varies. The results of this investigation is shown in Figure S2.

Figure S2 indicates that the resulting clusters share some *common* and *different* (distinct) characteristics of each cluster based on the proposed method and physicochemical properties, depending on the weight. For example, Figure S2 indicates that GLY is its own cluster, whether the clustering is based on the proposed distance matrix or based on the physicochemical properties. However, ARG, for example, shows to be a separate cluster in the figure $a = 1$ (based on physicochemical properties) and its position starts to join other amino acids to form a bigger cluster as a decreases to zero (based on dihedral angle information). These types of indication can be inferred with other amino acids. It is therefore reasonable to expect the results in the main manuscript that there are some agreement between clustering based on dihedral angle information and that based on physicochemical properties as shown in the main manuscript (Section 4.4).

4 Results of analysis on the second dataset

In this section, we present our analysis with the second dataset. As we have discussed earlier in Section 1 of this supplementary material, the second dataset does not have any observation in the first dataset, and hence can be utilised to validate the results that we obtained from the first dataset. A comparison between the two datasets are illustrated in Figure S3. The figure shows the distribution of some amino acids across the different proteins in the first dataset (left column) and the second dataset (right column).

The figure indicates that the distributions of amino acids, as a comparison between the first and second dataset, are quite similar. There are some minor differences, but overall we consider them to be similar. This is a promising preliminary result because there is an indication, albeit in only some amino acids and the fact that there is no overlapping protein, that the second dataset may be able to confirm our previous results.

We performed the BAPT method to estimate the distance matrix of the amino acids in the second dataset. To simplify this distance matrix, we summarise it in Figure S4. An edge that connect a pair of amino acids indicate that the associated p-value of the test for the pair is greater than 0.05. The figure indicates that the natural clustering of amino acids in the second dataset is equal to that in the first dataset (Figure 4 in the main manuscript). For example, GLY and PRO are in their own cluster. There are some slight differences in the edges that connected the amino acids within a cluster. Some

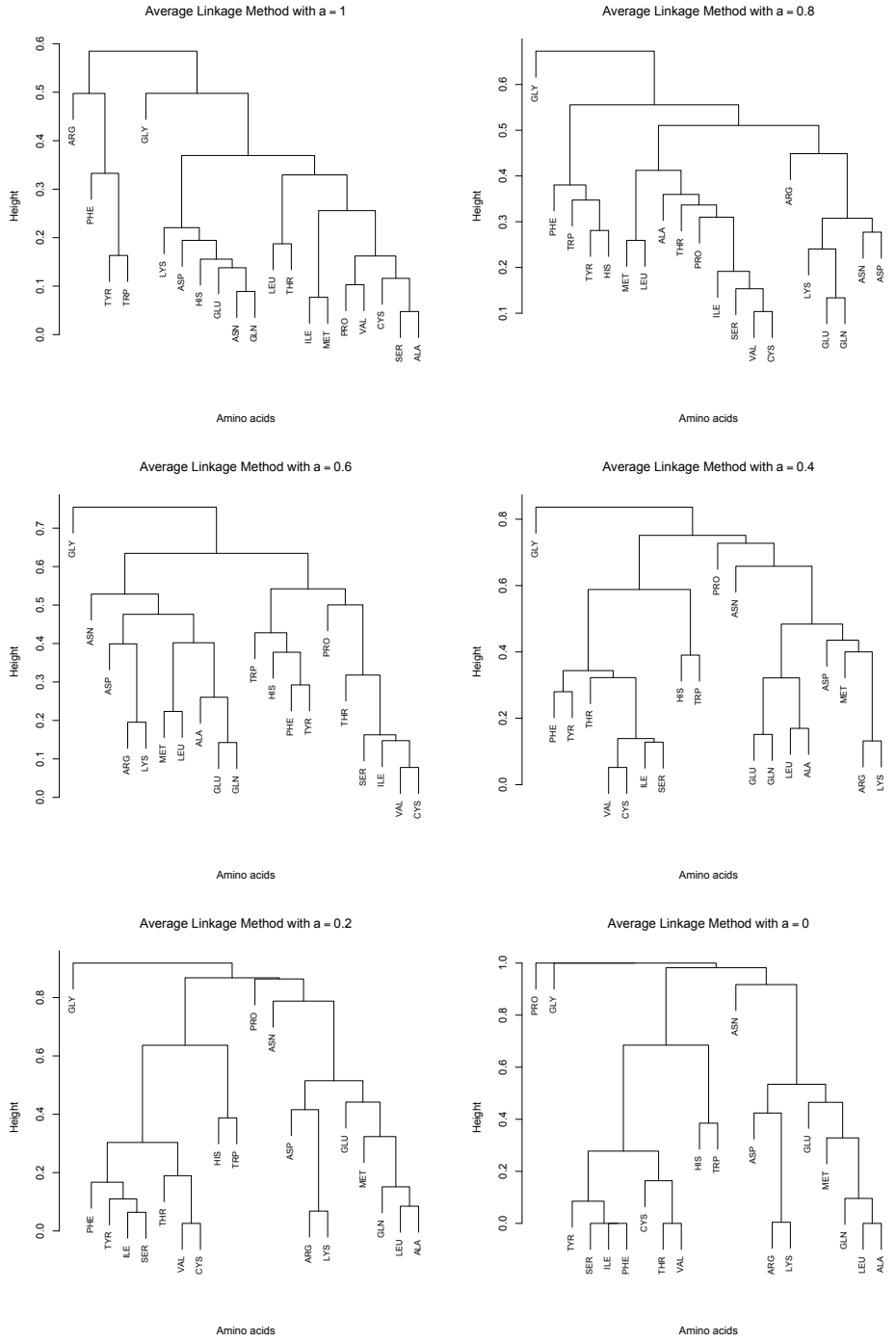

Figure S2: *The results of hierarchical clustering using average linkage based on a composite distance matrix (Equation 1) for different weights (a) from 1 (top left panel) to zero (bottom right panel). The composite distance matrix is constructed from distance matrix based in our proposed method (weight* $1 - a$*) and that based on the physicochemical properties (weight a) in Table S3.*

pairs of amino acids that are not significant at the 5% level in the first dataset are now significant in the second dataset. For example, in the first dataset, LYS is idenfified to have the same population distribution with GLU, ALA, and GLN (Figure 4 in the main manuscript). However, in the second dataset as shown in Figure S4, LYS is now identified to have diffrent population distribution from GLU, ALA, and GLN.

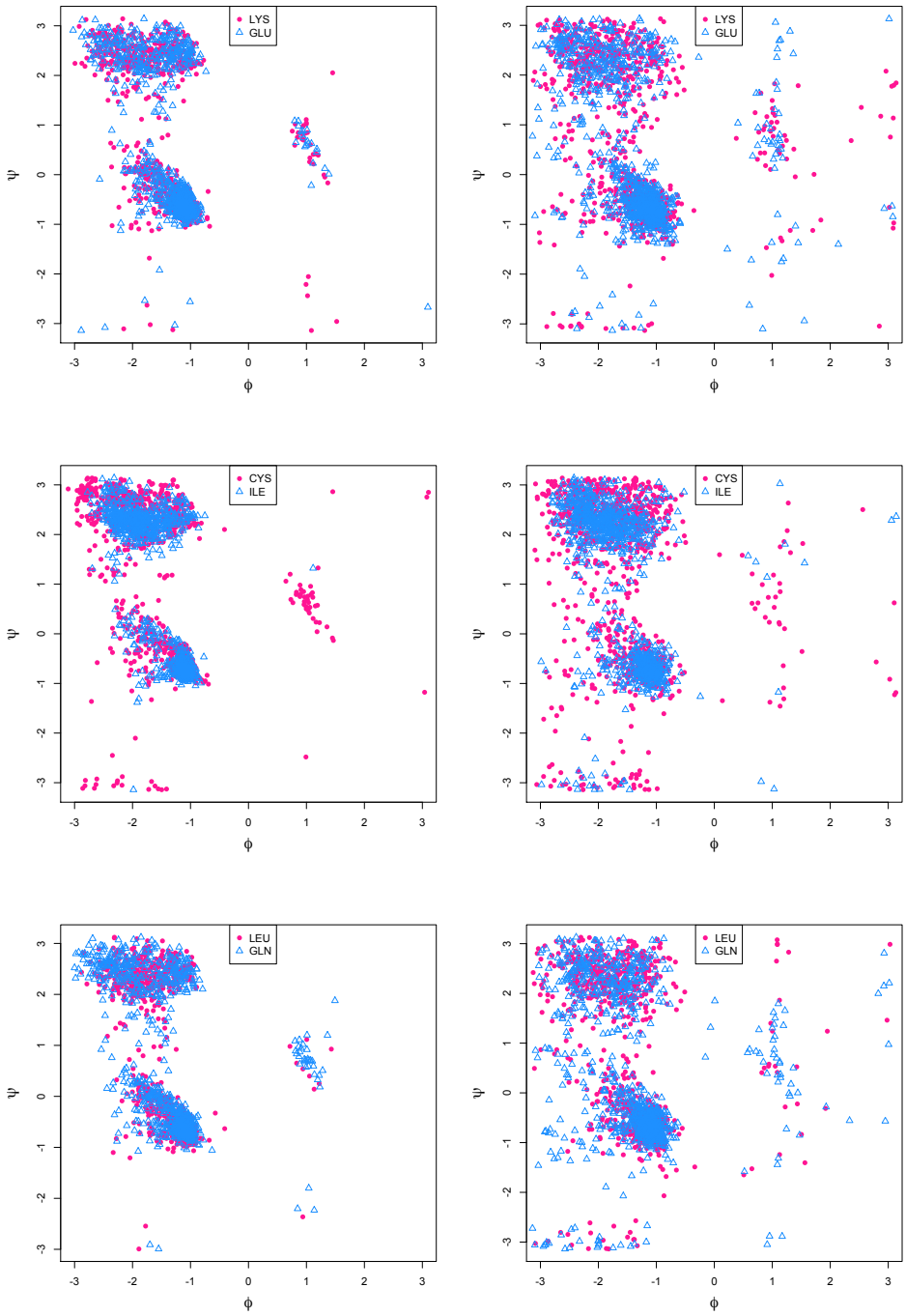

Figure S3: *Comparison between the first dataset (left column) and the second dataset (right column) in terms of the distribution of amino acids across proteins in their respective dataset: LYS and GLU (top row), CYS and ILE (middle row), and LEU and GLN (bottom row).*

From the distance matrix, we can now create dendogram under the single, average, and complete linkage, and the Ward method. The relevant dendograms are presented in Figure S5. The figure indicates that the resulting dendograms in the second dataset are very close to those in the first dataset (Figure 5 in the main manuscript). This is as expected since the result of the tests in the second dataset in Figure S4 shows the same natural clustering as the first dataset (Figure 4 in the main manuscript).

This result is able to confirm that the clustering we previously have seen in the first dataset is not

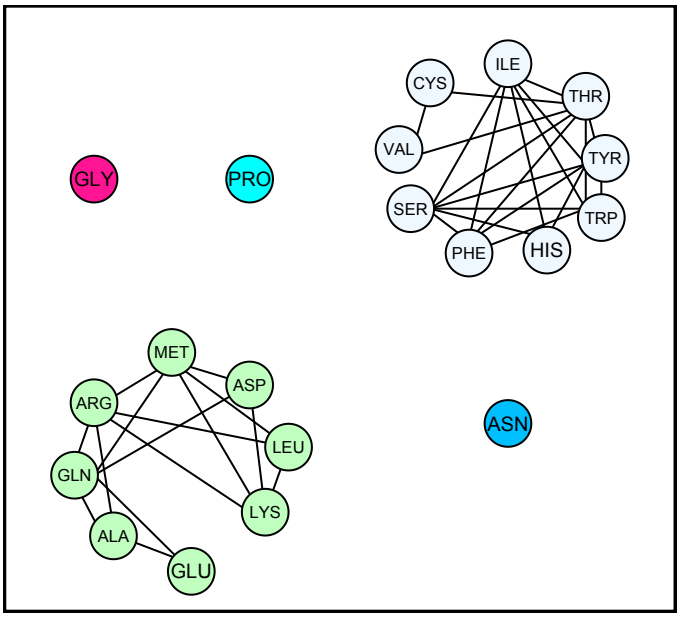

Figure S4: *The result of appling the BAPT method to estimate the distance matrix between amino acids in the second dataset. An edge that connect a pair of amino acids indicate that the associated p-value of the test for the pair is greater than 0.05.*

coincidence. The second dataset that does not contain any protein from the first dataset also suggests the same clustering of amino acids.

5 A note on type-I error in the simulation study

One of the most important aspects in the simulation study (Section 3.4 in the main manuscript) is the ability of the BPAT test to properly control type-I error. Figure 2 in the main manuscript has shown that the test is able to control properly type-I error when H_0 is true, for different configurations of n_1 and n_2. To show this in more detail, the numerical figures for the type-I error control achieved in the simulation are presented in Table S4. The figures in the table indicate that the test has a proper control of type-I error. All of the proportions of falsely rejected null hypotheses (out of 1000 replicates) are well within 1.96 standard errors of 0.05.

n_1	n_2	Prop.	St.Error
1000	1000	0.055	0.0072
500	1000	0.051	0.0069
200	1000	0.060	0.0075
100	1000	0.053	0.0071
1000	1000	0.055	0.0072
500	1000	0.057	0.0073
200	1000	0.060	0.0069
100	1000	0.054	0.0071

Table S4: *Proportion of p-values (out of 1000 replications) that are less than or equal to 0.05 in the simulation study under H_0 at different values of n_1. The first four rows are from the simulation study when we varied the mean and the last four rows are from that when we varied the variance.*

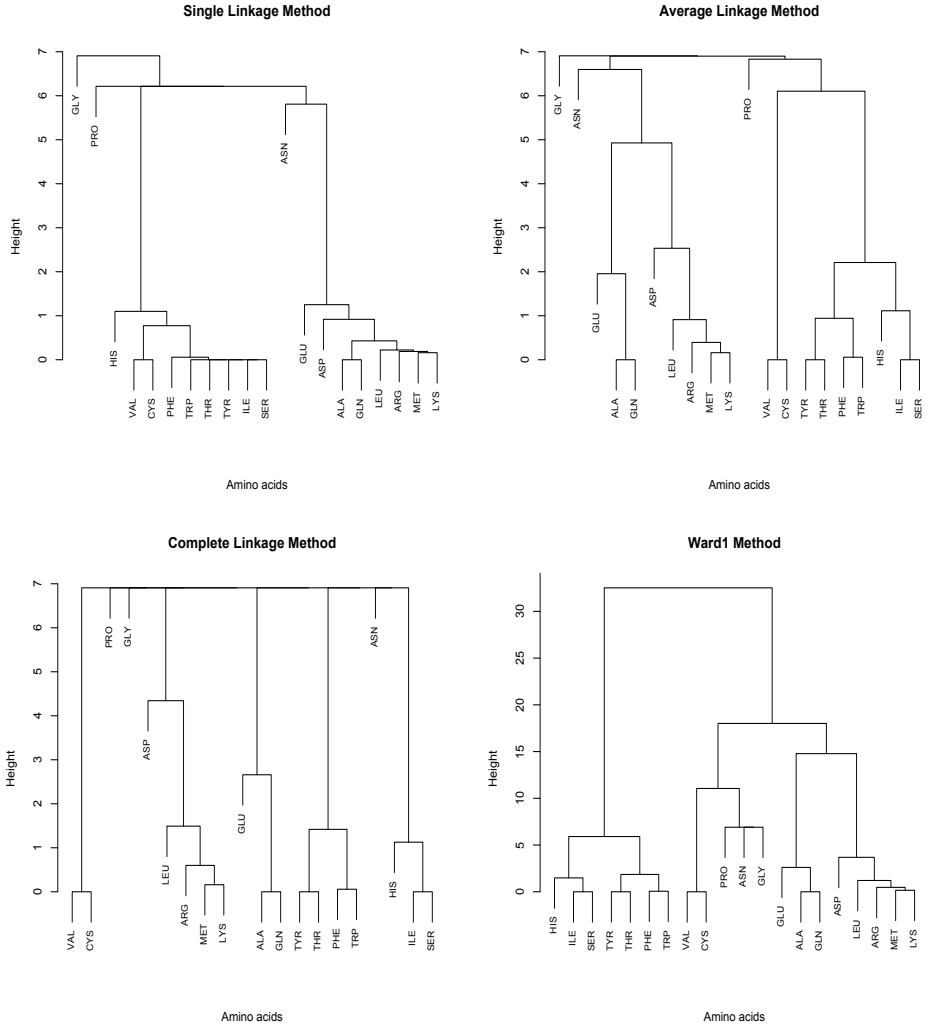

Figure S5: *Dendogram of hierarchical clustering of amino acids under single, average, and complete linkage, and Ward's clustering based on the dissimilarity matrix from p-values of test of equality of dihedral angle distributions between pairs of amino acids in the second dataset.*

References

Burley *et al.* (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy, *Nucleic Acids Research* **47**: D464–D474, doi: 10.1093/nar/gky1004

Carrascoza *et al.* Computational study of protein secondary structure elements: Ramachandran plots revisited, *Journal of Molecular Graphics and Modelling*, **50**:125–133, https://doi.org/10.1016/j.jmgm.2014.04.001

Chothia (1984), Principles that determine the structure of proteins, *Annual Review of Biochemistry*, **53**(1): 537-572

Engelman *et al.* (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, *Annual Review of Biophysics and Biophysical Chemistry*, **15**(1): 321-353

Fisher and Lee (1983) A correlation coefficient for circular data, *Biometrika*, **70**(2):327–332

Gerstein *et al.* (1994) Volume changes in protein evolution, *Journal of Molecular Biology*, **236**(4): 1067-1078, https://doi.org/10.1016/0022-2836(94)90012-4

Lovell *et al.* (2003) Structure validation by C_α geometry: ϕ, ψ and C_β deviation, *Proteins: Structure, Function, and Bioinformatics*, **50**:437-450

Stanfel (1996), A New Approach to Clustering the Amino Acid, *Journal of Theoretical Biology*, **183**(2):195-205, https://doi.org/10.1006/jtbi.1996.0213

Stryer (1988), *Biochemistry*, New York: H. Freeman