Theoretical aspects regarding the generation of the Bio-macro MDs for proteins

1. Protein structural representations for geometric information extraction

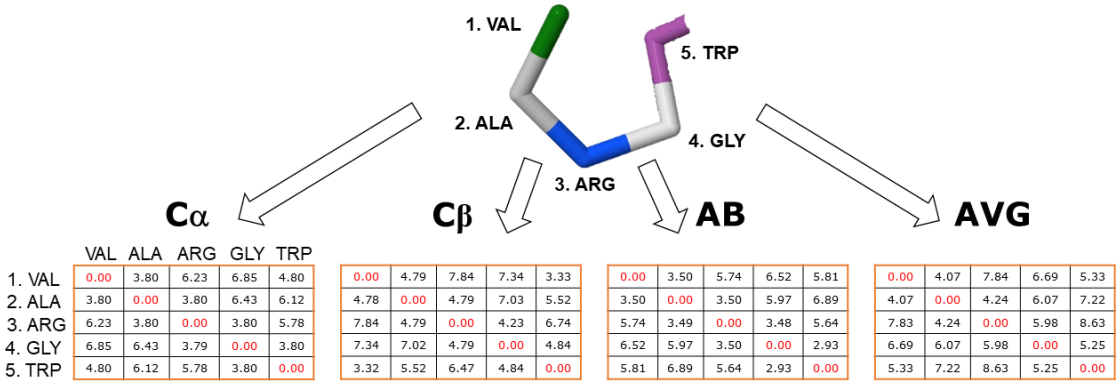
Spatial protein representations indicate the distribution of all amino acids present on the structure and allow the evaluation of their interactions.[1] The goal of these graphical structures is the extraction of valuable information for explaining experimental observations and bulk behavior.[2,3] Regarding the case of proteins, each amino acid from the spatial structure can be considered as a pseudo-vertex, which has spatial coordinates (x, y, z) defined by a carbon-atom representation. The importance of this pseudo-vertex approach is related to the structural simplification of the amino acid.[4]

Alpha carbon representation (C_α) has been the most used representation for protein geometrical/topological studies, without providing evidence of why this representation was chosen[3,5–7]. Moreover, Beta carbon representation (C_β) was considered as a simple atom(pseudo-node)-based representation in an article.[8] Considering these evidences, we have proposed two additional representations for protein spatial information extraction (Amide Carbon (AB) and the average of the coordinates of all atoms in the amino acid (AVG)) to observe the behavior and information content that these representations could bring respect the other existing representations (See Figure **1**).

2. Macromolecular vectors as weighting scheme

The transformation of chemical structures in molecular vectors has been explained on detail by several authors.[9–12] Regarding the case of proteins, this concept can be adapted by considering every amino acid on the structure as a compound element. Each

component of this macromolecular vector is a numerical value defined by a physicochemical property corresponding for each amino acid on the structure.[13,14]



	VAL	ALA	ARG	GLY	TRP
1. VAL	0.00	3.80	6.23	6.85	4.80
2. ALA	3.80	0.00	3.80	6.43	6.12
3. ARG	6.23	3.80	0.00	3.80	5.78
4. GLY	6.85	6.43	3.79	0.00	3.80
5. TRP	4.80	6.12	5.78	3.80	0.00

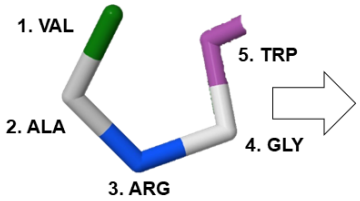
0.00	4.79	7.84	7.34	3.33
4.78	0.00	4.79	7.03	5.52
7.84	4.79	0.00	4.23	6.74
7.34	7.02	4.79	0.00	4.84
3.32	5.52	6.47	4.84	0.00

0.00	3.50	5.74	6.52	5.81
3.50	0.00	3.50	5.97	6.89
5.74	3.49	0.00	3.48	5.64
6.52	5.97	3.50	0.00	2.93
5.81	6.89	5.64	2.93	0.00

0.00	4.07	7.84	6.69	5.33
4.07	0.00	4.24	6.07	7.22
7.83	4.24	0.00	5.98	8.63
6.69	6.07	5.98	0.00	5.25
5.33	7.22	8.63	5.25	0.00

Figure 1. Numerical calculation example considering all protein representations proposed here, namely 1) alpha carbon (**Cα**), 2) beta carbon (**Cβ**), 3) amide carbon (**AB**) and 4) average of all atoms in the amino acid (**AVG**), employing the truncated peptide VG13P (pdb code: 5WRX). The two-tuple tensor (**D-SDST**) was calculated for every representation considering a Euclidean metric (**M5**), Non-Stochastic tensor (**NS**), *k*=**1** and distance to the center was not considered (z_{ii}=**0**).

These properties can be divided on 3 groups: steric, hydrophobic and electronic. Several examples are cited as follows: Isotropic contact area (ISA)[15], Kyte-Doolittle hydropathy index (KDS)[16], Hopp-Woods hydropathy index (HWS)[17], electronic charge index (ECI)[15], isoelectric point (PIE) [18], z parameters (Z_1, Z_2, Z_3)[19], molecular volume (MV)[20], alpha helix and beta sheet probability (PAH y PBS)[1]. All numerical values of these properties for every amino acid is shown in Table **1** (See Figure **2**).



	Position on the Primary Structure				
	1	*2*	*3*	*4*	*5*
Physicochemical Properties/ Amino acid (aa) code	***VAL***	***ALA***	***ARG***	***GLY***	***TRP***
Molecular Volume (MV)	140	88.6	174	60.1	227
Hopp-Woods Hydropathy Index (HWS)	-1.5	-0.5	3.0	0.0	-3.0
Electronic Charge Index (ECI)	0.07	0.05	1.69	0.02	1.08

Steric properties-based Vector $X = [140 \quad 88.6 \quad 174 \quad 60.1 \quad 227]$

Hydrophobic properties-based Vector $Y = [-1.5 \quad -0.5 \quad 3.0 \quad -0.0 \quad -3.0]$

Electronic properties-based Vector $P = [0.07 \quad 0.05 \quad 1.69 \quad 0.02 \quad 1.08]$

Figure 2. Numerical calculation example for the definition of macro-molecular vectors considering three different groups physicochemical properties (steric, hydrophobic and electronic), employing the truncated peptide 5WRX. All numerical values of these properties for every amino acid is shown in Table **1**.

Table 1. Physicochemical properties for the 20 elemental amino acids used for the computation of macromolecular vectors

Amino acid	Code	z-scale[a]			ISA[b]	ECI[c]	PIE[d]	HWS[e]	KDS[f]
		z_1	z_2	z_3					
Alanine	ALA	0.01	-1.73	0.09	62.9	0.05	6.01	-0.5	1.8
Arginine	ARG	2.88	2.52	-3.44	52.98	1.69	10.76	3	-4.5
Asparagine	ASN	3.22	1.45	0.84	17.87	1.31	5.41	0.2	-3.5
Aspartate	ASP	3.64	1.13	2.36	18.46	1.25	2.77	3	-3.5
Cysteine	CYS	0.71	-0.97	4.13	78.51	0.15	5.07	-1	2.5
Glutamate	GLU	3.08	0.39	-0.07	30.19	1.31	3.22	0.2	-3.5
Glutamine	GLN	2.18	0.53	-1.14	19.53	1.36	5.65	3	-3.5
Glycine	GLY	2.23	-5.36	0.3	19.93	0.02	5.97	0	-0.4
Histidine	HIS	2.41	1.74	1.11	87.38	0.56	7.59	-0.5	-3.2
Isoleucine	ILE	-4.44	-1.68	-1.03	149.77	0.09	6.02	-1.8	4.5
Leucine	LEU	-4.19	-1.03	-0.98	154.35	0.01	5.98	-1.8	3.8
Lysine	LYS	2.84	1.41	-3.14	102.78	0.53	9.74	3	-3.9
Methionine	MET	-2.49	-0.27	-0.41	132.22	0.34	5.74	-1.3	1.9
Phenylalanine	PHE	-4.92	1.3	0.45	189.42	0.14	5.48	-2.5	2.8
Proline	PRO	-1.22	0.88	2.23	122.35	0.16	6.48	0	-1.6
Serine	SER	1.96	-1.63	0.57	19.75	0.56	5.68	0.3	-0.8
Threonine	THR	0.92	-2.09	-1.4	59.44	0.65	5.87	-0.4	-0.7
Tryptophan	TRP	-4.75	3.65	0.85	179.16	1.08	5.89	-3.4	-0.9
Tyrosine	TYR	-1.39	2.32	0.01	132.16	0.72	5.66	-2.3	-1.3
Valine	VAL	-2.69	-2.53	-1.29	120.91	0.07	5.97	-1.5	4.2

[a]Z-scale (Hellberg et al., 1987), [b]Side-chain isotropic surface area (Collantes and Dunn III, 1995), [c]Atomic charge (Collantes and Dunn III, 1995), [d]Isoelectric point (Hellberg et al., 1987), [e]Hoop-Woods hydropathy index (Hopp and Woods, 1981), [f]Kyte-Doolittle hydropathy index (Kyte and Doolittle, 1982).

3. Spatial-Dis-Similarity Tensor generation

The spatial information extraction from a protein could be achieved by using a geometrical matrix which is defined by transforming the existing relationships between every element (amino acid) and its neighbors into numbers that represent such relation. [5,21,22].

An order two geometrical tensor, is a generalized spatial matrix that extracts the information of a 3D protein structure by using a metric as a function for the distance

between two amino acids.[23] For the application of this mathematical concept, a defined protein structural representation (Section **1**) is required.

Table 1. Physicochemical properties for the 20 elemental amino acids used for the computation of macromolecular vectors (*continued*)

Amino acid	Code	MV[g]	PAH[h]	PBS[i]
Alanine	ALA	88.6	1.29	0.9
Arginine	ARG	173.4	0.96	0.99
Asparagine	ASN	114.1	0.9	0.76
Aspartate	ASP	111.1	1.04	0.72
Cysteine	CYS	108.5	1.11	0.74
Glutamate	GLU	143.8	1.44	0.75
Glutamine	GLN	138.4	1.27	0.8
Glycine	GLY	60.1	0.56	0.92
Histidine	HIS	153.2	1.22	1.08
Isoleucine	ILE	166.7	0.97	1.45
Leucine	LEU	166.7	1.3	1.02
Lysine	LYS	168.6	1.23	0.77
Methionine	MET	162.9	1.47	0.97
Phenylalanine	PHE	189.9	1.07	1.32
Proline	PRO	112.7	0.52	0.64
Serine	SER	89	0.82	0.95
Threonine	THR	116.1	0.82	1.21
Tryptophan	TRP	227.8	0.99	1.14
Tyrosine	TYR	193.6	0.72	1.25
Valine	VAL	140	0.91	1.49

[g]Side-chain amino acid volume (Zamyatnin, 1972), [h,i] Relative frequencies with which an amino acid appear forming α-helices, and β-sheets, respectively (Mathews et al., 2000)

An order three geometrical tensor is a generalized set of order two tensors that extract the information of a 3D protein structure by the use of multi-metrics as a strategy to define relationships between 3 amino acids.[24] The detailed definitions of metric and multi-metric will be discussed in section **3.1**.

These geometrical tensors consider additional mathematical tools for generalization such as interaction ponderation considering a Haddamard matrix product (section **3.2.**), normalization procedures for tensor standardization (simple stochastic and mutual probability) (section **3.3.**) and topological and geometrical cut-offs which look to define the effect of certain proteins regions on the information extracted (section

3.4.). These operations result on obtaining the two-tuple Spatial-Dis-Similarity Tensor (D-SDST) ($^{D}\mathbb{Z}$) and three-tuple Spatial-Dis-Similarity Tensor (T-SDST) ($^{T}\mathbb{Z}$) (See Figure **3**).

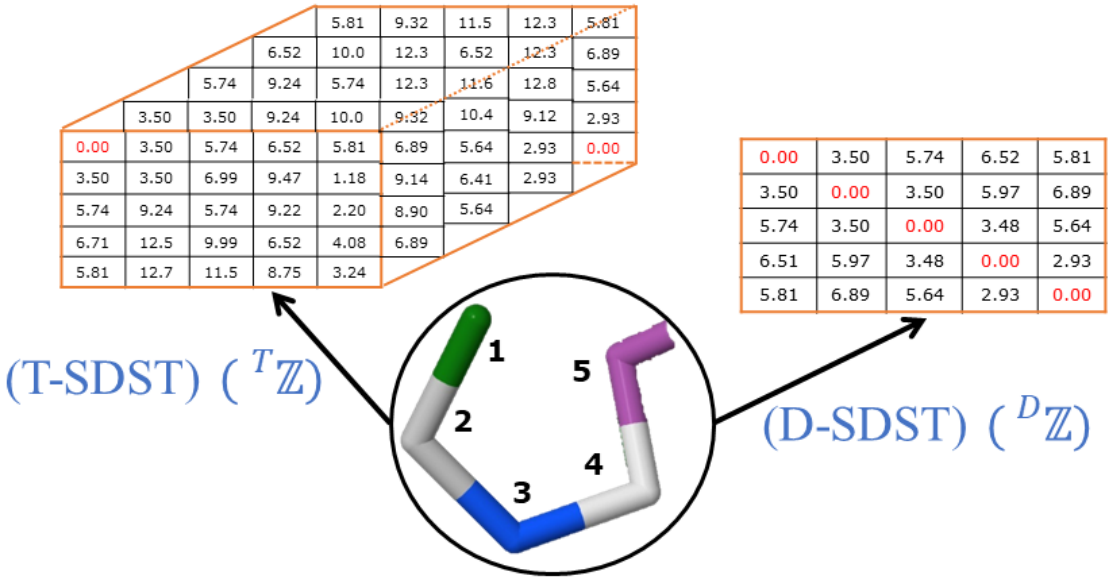


Figure 3. Numerical calculation example considering a two-tuple spatial dis similarity tensor (**D-SDST**) and a three-tuple spatial dis similarity tensor (**T-SDST**) employing the truncated peptide 5WRX. The two-tuple tensor was calculated considering an amide carbon protein representation (**AB**), a Euclidean metric (**M5**), Non-Stochastic tensor (**NS**), *k*=**1** and distance to the center was not considered. The three-tuple tensor was calculated considering an amide carbon protein representation (**AB**), a Perimeter multi-metric (**M37**), Non-Stochastic tensor (**NS**), *k*=**1** and distance to the center was not considered.

3.1. Metrics and Multi-metrics for geometric matrix inter-amino acid interaction generalization

3.1.1. Metric

A metric or a distance function is a mathematical expression that defines a distance between two elements (a,b) from a defined set. A metric has to fulfill the following conditions.[25]

i.) $d(a,b) \geq 0$ (it has to be positive)

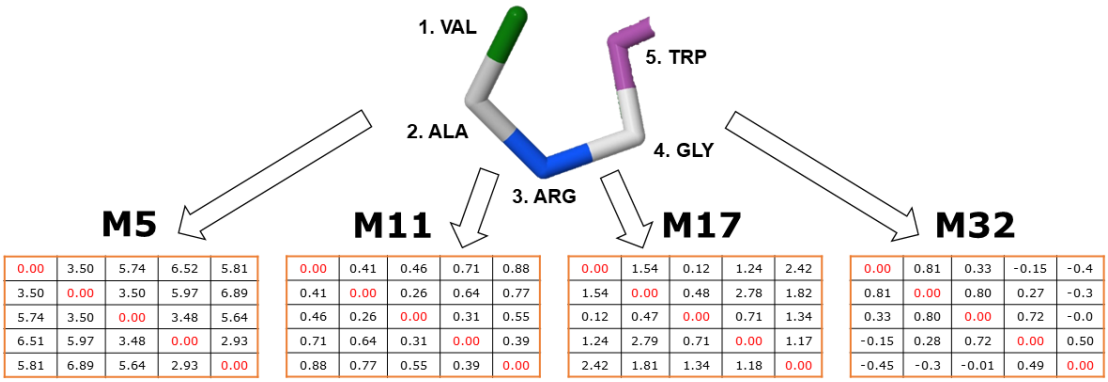
ii.) $d(a,b) = d(b,a)$ (it has to be symmetric)

iii.) d (a,b) ≤ d(a,c)+d(c,b) (it has to fulfill the triangle inequality) **(1)**

iv.) d (a,b) = 0 if a=b (it has to fulfill the identity axiom)

Metrics are essential elements in a variety of areas in science such as graph theory, molecular biology, among others.[23,26].

Table **2** presents all the metrics available for calculation of the proposed MDs on the software (See Figure **4**).



0.00	3.50	5.74	6.52	5.81
3.50	0.00	3.50	5.97	6.89
5.74	3.50	0.00	3.48	5.64
6.51	5.97	3.48	0.00	2.93
5.81	6.89	5.64	2.93	0.00

0.00	0.41	0.46	0.71	0.88
0.41	0.00	0.26	0.64	0.77
0.46	0.26	0.00	0.31	0.55
0.71	0.64	0.31	0.00	0.39
0.88	0.77	0.55	0.39	0.00

0.00	1.54	0.12	1.24	2.42
1.54	0.00	0.48	2.78	1.82
0.12	0.47	0.00	0.71	1.34
1.24	2.79	0.71	0.00	1.17
2.42	1.81	1.34	1.18	0.00

0.00	0.81	0.33	-0.15	-0.4
0.81	0.00	0.80	0.27	-0.3
0.33	0.80	0.00	0.72	-0.0
-0.15	0.28	0.72	0.00	0.50
-0.45	-0.3	-0.01	0.49	0.00

Figure 4. Numerical calculation example considering 4 different types of metrics applied to obtain a two-tuple spatial dis similarity tensor (**D-SDST**) employing the truncated peptide 5WRX. The two-tuple tensor was calculated considering an amide carbon protein representation (**AB**), Non-Stochastic tensor (**NS**), *k*=**1** and distance to the center not considered. The metrics considered for this example were: Euclidean metric (**M5**), Lance-Williams (**M11**), SL-Like (**M17**) and Proportionality corrected (**M32**)

3.1.2. Multi-metric

A multi-metric is a generalization of the metric concept since it seeks for relationships between two or more elements. The mathematical definitions for proposing an element as multi-metric are shown below, considering the notation proposed by Warrens [24]:

Let be, $x_{1,k} = (x_1, x_2, x_3, \dots, x_k)$, a k-uple and let be, $x_{1,k}^{-i} = (x_1, x_2, x_{i-1}, x_{i+1} \dots, x_k)$, the (k-1)-uple, where the minus sign on the index $x_{1,k}^{-i}$ is used to indicate that this object has been removed from the k-uple. From this point, a multi-metric can be defined as a dis-similarity measure that satisfies the following conditions:

$$
\begin{aligned}
&i.)\ (k-1)d_k\left(x_{1,k}\right) \le \sum_{i=1}^{k} d_k\left(x_{1,k+1}^{-i}\right) \text{ (polyhedral inequality)} \\
&ii.) d_k\left(x_1, x_{1,k-1}\right) = d_k\left(x_{1,2}, x_{2,k-1},\right) = \cdots = d_k\left(x_{1,k-1}, x_{k-1}\right) \text{ (invariant condition)} \\
&iii.)\ d_{k-1}\left(x_{1,k-1}\right) = \frac{1}{p} d_k\left(x_1, x_{1,k-1}\right) \\
&iv.)\ d_k\left(x_1, x_{1,k-1}\right) \le d_k\left(x_1, x_{2,k}\right)
\end{aligned}
\tag{2}
$$

Table 3, **Table 4**, **Table 5** and **Table 6** presents all the groups of multi-metrics available for calculation (See Figure **5**). These tables are shown below.

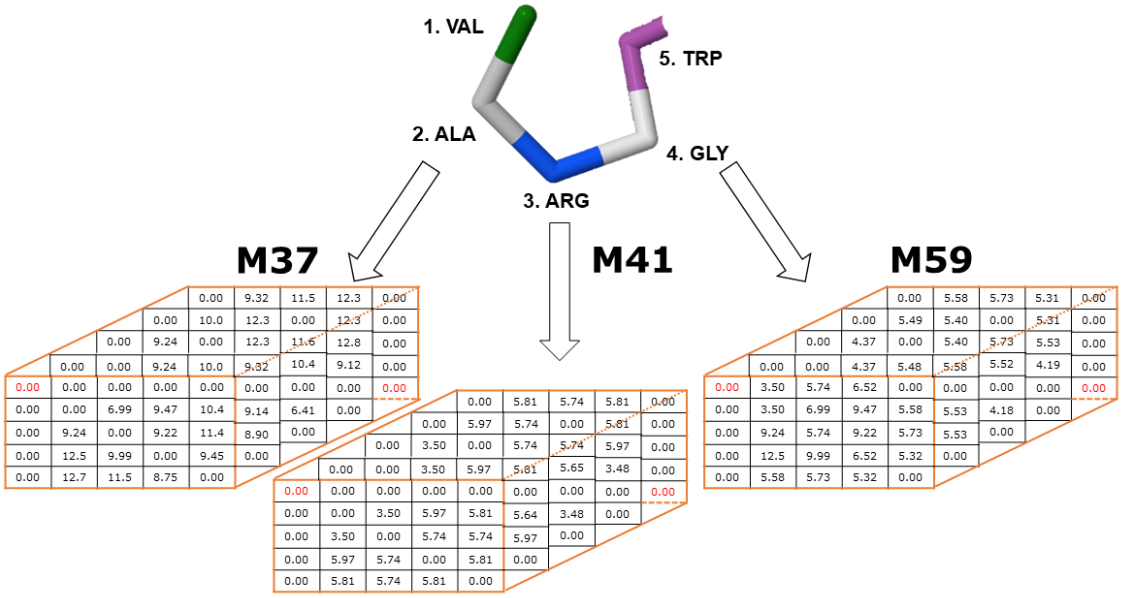


Figure 5. Numerical calculation example considering 3 different types of multi-metrics applied on a three-tuple spatial dis similarity tensor (**T-SDST**) employing the truncated peptide 5WRX. The three-tuple tensor was calculated considering an amide carbon protein representation (**AB**), Non-stochastic tensor (**NS**), *k*=**1**, and distance to the center was not considered. The multi-metrics considered for this example were: Summation Sides (**M37**), Min-Rule (**M41**) and Sum-Rule (**M59**).

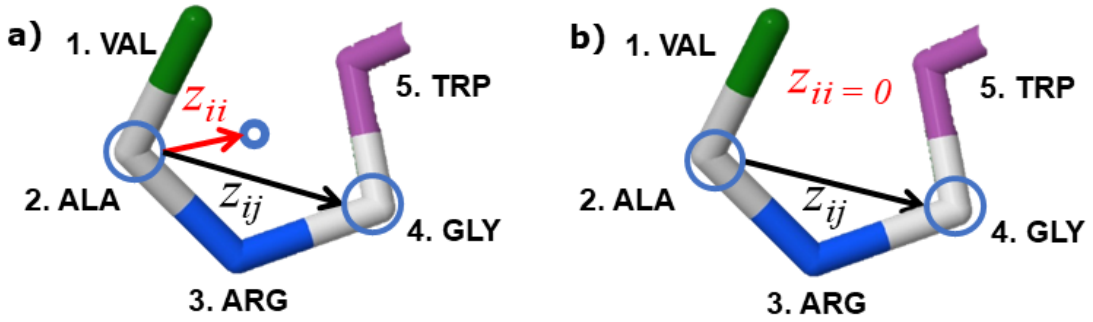
Table 2. Metrics available for the calculation of the novel 3D algebraic MDs for proteins. In bold, the software ID number of the multi-metric is indicated.

Coefficient	Code	Range	Formula
Minkowski r = 0.25, 0.5, 1, 1.5, 2, 2.5 and 3 (where, when r = 1 is general metric is Hamming distance (also known as Manhattan, city-block or taxi distance) and r = 2 is Euclidean distance)	**M1-M7**	∞ to 0	$d_{st} = \langle \sum_{j=1}^{P} \langle \lvert X_{sj} - X_{st} \rvert \rangle^{r} \rangle^{\frac{1}{r}}$
Chebyshev/Lagrange (Minkowski formula when r = ∞)	**M8**	∞ to 0	$d_{st} = \max \lvert X_{sj} - X_{tj} \rvert$
Minkowski (also known as power distance) r value can be defined by user	**M9**	∞ to 0	$d_{st} = \langle \sum_{j=1}^{P} \langle \lvert X_{sj} - X_{st} \rvert \rangle^{r} \rangle^{\frac{1}{r}}$
Canberra	**M10**		$d_{st} = \sum_{j=1}^{P} \frac{\lvert X_{sj} - X_{tj} \rvert}{\lvert X_{sj} \rvert + \lvert X_{tj} \rvert}$
Lance-Williams/Bray-Curtis	**M11**		$d_{st} = \frac{\sum_{j=1}^{P} \lvert X_{sj} - X_{tj} \rvert}{\sum_{j=1}^{P} \langle \lvert X_{sj} \rvert + \lvert X_{tj} \rvert \rangle}$
Clark	**M12**		$d_{st} = \sqrt{\sum_{j=1}^{P} \langle \frac{X_{sj} - X_{tj}}{\lvert X_{sj} \rvert + \lvert X_{tj} \rvert} \rangle^{2}}$
Soergel	**M13**		$d_{st} = \lvert \frac{\sum_{j=1}^{P} \lvert X_{sj} - X_{tj} \rvert}{\sum_{j=1}^{P} \max(X_{sj}, X_{tj})} \rvert$
Bhattacharyya	**M14**		$d_{st} = \sqrt{\sum_{j=1}^{P} (\sqrt{\lvert X_{sj} \rvert} - \sqrt{\lvert X_{tj} \rvert})^{2}}$
Wave-Edges	**M15**		$d_{st} = \lvert \sum_{j=1}^{P} (1 - \frac{\min(X_{sj}, X_{tj})}{\max(X_{sj}, X_{tj})}) \rvert$
Angular Separation/ [1-Cosine (**Ochiai**)]	**M16**	0 to 1	$d_{st} = 1 - \frac{\sum_{j=1}^{P} (X_{sj} * X_{tj})}{\sqrt{\sum_{j=1}^{P} {X_{sj}}^{2}} * \sqrt{\sum_{j=1}^{P} {X_{tj}}^{2}}}$
SL-Like	**M17**	0 to 1	$d_{st} = \frac{1}{p} \sum_{j=1}^{p} \frac{\lvert x_{sj} - x_{tj} \rvert}{(x_{sj} + x_{tj})}$
Average Euclidean	**M18**		$EM_{XY} = \frac{\sqrt{\sum_{j=1}^{n} \lvert x_j - y_j \rvert^{2}}}{n}$
Squared Euclidean coefficient	**M19**		$EC_{XY} = \sum_{j=1}^{n} \lvert x_j - y_j \rvert^{2}$

Table 2. Multi-metrics available for the calculation of the novel 3D algebraic MDs for proteins. In bold, the software ID number of the multi-metric is indicated. (*continued*)

Coefficient	Code	Range	Formula
Pearson correlation	**M20**	-1 to 1	$r_{XY} = \frac{\sum_{j=1}^{n}(x_j - \overline{X})(y_j - \overline{Y})}{\sqrt{\sum_{j=1}^{n}(x_j - \overline{X})^2 \sum_{j=1}^{n}(y_j - \overline{Y})^2}}$
Cosine coeficient/ **Ochiai coefficient** (essentially equivalent to the **Carbo index** for overlap of electron density functions.	**M21**	-1 to 1	$Cos_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sqrt{\sum_{j=1}^{n} x_j^2 \times \sum_{j=1}^{n} y_j^2}}$
Fossum	**M23**	0 to ∞	$F_{XY} = \frac{n\left(\sum_{j=1}^{n} x_j y_j - \frac{1}{2}\right)^2}{\sum_{j=1}^{n} x_j^2 \sum_{j=1}^{n} y_j^2}$
Jaccard/Tanimoto	**M24**	-1/3 to 1	$T_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2 - \sum_{j=1}^{n} x_j y_j}$
Kulczynski	**M25**	0 to ∞	$Kul1_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2 - 2\sum_{j=1}^{n} x_j y_j}$
Sokal/Sneath	**M26**	0 to 1	$SS1_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{2\sum_{j=1}^{n} x_j^2 + 2\sum_{j=1}^{n} y_j^2 - 3\sum_{j=1}^{n} x_j y_j}$
Simpson	**M27**	0 to 1	$Sim_{XY} = \frac{\sum_{j=1}^{n} min(x_j, y_j)}{min(\sum_{j=1}^{n} x_j, \sum_{j=1}^{n} y_j)}$
Ruzicka's dissimilarity	**M28**		$d_3(X,Y) = 1 - 2\left[\frac{\sum_i min\{x_i, y_i\}}{\sum_i max\{x_i, y_i\}}\right]$ Where, *X* and *Y* represent the compared molecular vectors, x_i and y_i their corresponding vector components.
Dice (also known as **Czekanowski** or **Sørenson** coefficient.)	**M29**	-1 to 1	$D_{XY} = \frac{2\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{n} y_j^2}$
Cosine coeficient/ **identity corrected**	**M30**	-1 to 1	$Cos_{XY} = \frac{\sum_{j=1}^{n} x_j y_j}{\sqrt{\sum_{j=1}^{n} x_j^2 \times \sum_{j=1}^{n} y_j^2}}$
Additivity	**M31**		$a_{XY} = \frac{2s_{XY}}{s_X^2 + s_Y^2}$
Proportionality corrected	**M32**		$L_{st} = \frac{1}{N}\sum_{k=1}^{N}\left(1 - \frac{\lvert P_{sk} - P_{tk}\rvert}{max(\lvert P_{sk}\rvert, \lvert P_{tk}\rvert)}\right)$

There are two possibilities regarding the application of multi-metrics or metrics on the protein structure, these could be amino acid-based, or protein mass center-based. In the first option, the multi-metric or the metric is calculated considering the distance functions against every *aa*, consequently, the elements z_{ij} of the D-SDST or the elements z_{ijl} of the T-SDST when $i = j$ *or* $i = j = l$, respectively, are zero. For the second case, the metric or multi-metric is calculated considering the distance functions against the mass center of the protein, and all elements z_{ij} on the D-SDST or all elements z_{ijl} on the T-SDST are different from zero; this approach may offer a better discrimination among protein spatial structures given that it provides information about the centrality of *aa* residues (See Figure **6**).



3.68	3.50	5.74	6.52	5.81
3.50	3.28	3.50	5.97	6.89
5.74	3.49	2.66	3.48	5.64
6.52	5.97	3.50	2.99	2.93
5.81	6.89	5.64	2.93	3.68

0.00	3.50	5.74	6.52	5.81
3.50	0.00	3.50	5.97	6.89
5.74	3.49	0.00	3.48	5.64
6.52	5.97	3.50	0.00	2.93
5.81	6.89	5.64	2.93	0.00

Figure 6. Numerical calculation example considering the difference between the distance to the center configuration applied on a two-tuple spatial dis similarity tensor (**D-SDST**) employing the truncated peptide 5WRX. The two-tuple tensor was calculated considering an amide carbon protein representation (**AB**), Non-Stochastic tensor (**NS**), k=**1** and two different options: **a)** distance to the center was considered and **b)** distance to the center was not considered.

Table 3. Multi-metrics (geometric-based) available for the calculation of the novel 3D algebraic MDs for proteins. In bold, the software ID number of the multi-metric is indicated.

Measure	Formula	Symmetry
Triangle Area (**M33-M34**)	$T_{XYZ} = \sqrt{s(s-d_{XY})(s-d_{YZ})(s-d_{ZX})}$ $s = \frac{d_{XY}+d_{YZ}+d_{ZX}}{2}$	S
Triangle's Incircle Area (**M35-M36**)	$T_{XYZ} = \pi\left(\frac{2\sqrt{s(s-d_{XY})(s-d_{YZ})(s-d_{ZX})}}{d_{XY}+d_{YZ}+d_{ZX}}\right)^2$	S
Summation Sides (**M37-M38**)	$T_{XYZ} = d_{XY} + d_{YZ}$	A
Bond angle (Angle between sides) (**M39-M40**)	$A_X, A_Y, A_Z \; coordinates \; of \; three \; aminoacids \; of \; a \; protein$ $U = A_X - A_Y, V = A_Z - A_Y$ $T_{XYZ} = \alpha = arccos\left(\frac{U*V}{\lvert U\rvert * \lvert V\rvert}\right)$	A

Table 4. Multi-metrics (cluster-similarity-based) available for the calculation of the novel 3D algebraic MDs for proteins. In bold, the software ID number of the multi-metric is indicated.

Measure	Formula	Symmetry
MIN-RULE [1-Nearest neighbor (NN)] (**M41-M42**)	$T_{1_{XYZ}} = min(d_{XZ}, d_{YZ})$ $V_2 = \begin{cases} Y, d_{XY} < d_{XZ} \\ Z, otherwise \end{cases}$ $V_3(V_2) = \begin{cases} Y, Y \neq V_2 \\ Z, otherwise \end{cases}$ $T_{2_{XYZ}} = min\left(d_{XV_3}, d_{V_2V_3}\right)$	A
JOIN-RULE (2-NN) **(M43-M44)**	$d\,min(d_{XY}, d_{YZ}, d_{ZX})_{min} \; d\,max(d_{XY}, d_{YZ}, d_{ZX})_{max}$ $join\left(d_{XY}, d_{YZ}, d_{ZX}, d_{\min}, d_{\max}\right) = \begin{cases} d_{XY}, d_{\min} > d_{XY} < d_{\max} \\ d_{YZ}, d_{\min} > d_{YZ} < d_{\max} \\ d_{ZX}, d_{\min} > d_{ZX} < d_{\max} \end{cases}$ $T_{XYZ} = join(d_{XY}, d_{YZ}, d_{ZX}, dmax_{min})$	S
MAX-RULE (Furthest neighbor) (**M45-M46**)	$T_{XYZ} = max(d_{XZ}, d_{YZ})$	A
AVE-RULE (Average-link) **(M47-M48)**	$T_{XYZ} = \frac{d_{XZ}+d_{YZ}}{2}$	A
MED-RULE (**M49-M50**)	$T_{XYZ} = \frac{d_{XZ}+d_{YZ}}{2} - \frac{d_{XY}}{4}$	A
WARD-RULE (**M51-M52**)	$T_{XYZ} = {d_{X\overline{C}}}^2 + {d_{Y\overline{C}}}^2 + {d_{Z\overline{C}}}^2 - {d_{X\overline{C}_{XY}}}^2 - {d_{Y\overline{C}_{XY}}}^2$	A
ADJ-RULE (**M53-M54**)	$T_{XYZ} = max(d_{XY}, d_{YZ}, d_{ZX}) - d_{XY}$	A
MAH-RULE (**M55-M56**)	$T_{XYZ} = {d^M}_{X\overline{C}}{}^2 + {d^M}_{Y\overline{C}}{}^2 + {d^M}_{Z\overline{C}}{}^2 - {d^M}_{X\overline{C}_{XY}}{}^2 - {d^M}_{Y\overline{C}_{XY}}{}^2$	

$\overline{C}_{XYZ}(\overline{C}_{XY})$ are the mean centroids for the amino acids X,Y,Z (XY) in the protein, respectively, d^M is the Mahalanobis distance.

3.2. k^{th} power operation for interaction account

Inter amino acid interactions do not occur only among near located amino acids, but also with amino acids located far in sequence. As a strategy for accounting these interactions on the proposed tensor, a Haddamard matrix product can be performed.[5] This procedure completes the power operation in every element of the spatial-(dis)similarity tensors. The exponent *k* is a real number whose values can be positive or negative; when parameter *k* is negative, the reciprocal operation is computed. The range of values to evaluate this product could be from -12 to 12, e.g. *k*=**-1** is related to the gravitational potential, *k*= **-2** is related to the Coulomb potential (See Figure **7** and Figure **8** for an illustration).

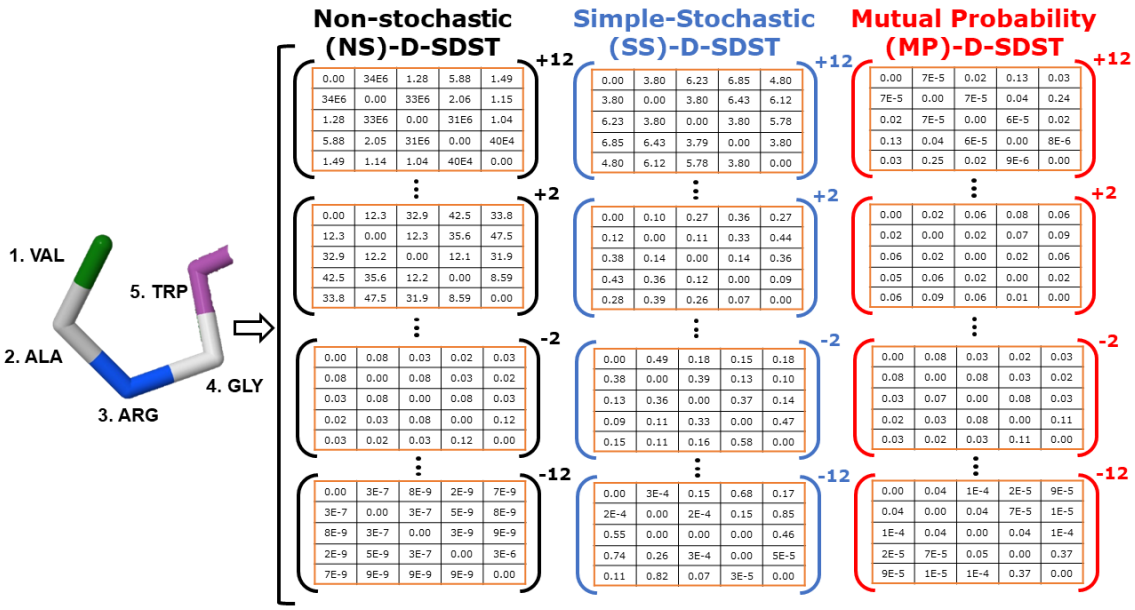


Figure 7. Numerical calculation example considering 2 different types of tensor normalization procedures and several exponents (**k** order) for the Haddamard product on a two-tuple spatial dis similarity tensor (**D-SDST**) employing the truncated peptide 5WRX. The two-tuple tensor was calculated considering an amide carbon protein representation (**AB**) and distance to the center was not considered. The normalization procedures considered for this example were: simple stochastic (**SS**) (see Equation **3**) and mutual probability (**MP**) (see Equation **5**). The exponents (**k**) considered were: -12, -2, +2 and +12.

Table 5. Multi-metrics (classic, data-fusion and statistics-based) available for the calculation of the novel 3D algebraic MDs for proteins. In bold, the software ID number of the multi-metric is indicated.

Measure	Formula	Symmetry
ADD-RULE (Average D/D degree) (**M57-M58**)	$T_{XYZ} = \frac{1}{3}\left(\frac{d_{XY}}{p_{XY}} + \frac{d_{YZ}}{p_{YZ}} + \frac{d_{ZX}}{p_{ZX}}\right)$	S
SUM-RULE (Wiener index) (**M59-M60**)	$T_{XYZ} = d_{XY} + d_{YZ} + d_{ZX}$	S
PRO-RULE (**M61-M62**)	$T_{XYZ} = d_{XY} \cdot d_{YZ} \cdot d_{ZX}$	S
QUA-RULE (**M63-M64**)	$T_{XYZ} = \left(\frac{{d_{XY}}^2 + \cdot\, {d_{YZ}}^2 + {d_{ZX}}^2}{3}\right)^{\frac{1}{2}}$	S
GEO-RULE (**M65-M66**)	$T_{XYZ} = \left(\frac{{d_{XY}}^3 + \cdot\, {d_{YZ}}^3 + {d_{ZX}}^3}{3}\right)^{\frac{1}{3}}$	S
RAN-RULE (**M67-M68**)	$T_{XYZ} = max(d_{XY}, d_{YZ}, d_{ZX}) - min(d_{XY}, d_{YZ}, d_{ZX})$	S

p_{XY} is the topological distance between the amino acids (X and Y)

Table 6. Multi-metrics (agreement coefficients-based) available for the calculation of the novel 3D algebraic MDs for proteins. In bold, the software ID number of the multi-metric is indicated.

Measure	Formula	Symmetry
IC-RULE Identity-corrected (**M69-M70**)	$T_{XYZ} = \frac{2(S_{XY} + S_{XZ} + S_{YZ})}{2\left({S_X}^2 + {S_Y}^2 + {S_Z}^2\right) + \left(\overline{X} - \overline{Y}\right)^2 + \left(\overline{X} - \overline{Z}\right)^2 + \left(\overline{Y} - \overline{Z}\right)^2}$	A
AC-RULE Additivity-corrected (**M71-M72**)	$T_{XYZ} = \frac{S_{XY} + S_{XZ} + S_{YZ}}{{S_X}^2 + {S_Y}^2 + {S_Z}^2}$	S
PC-RULE Proortionality-corrected (**M73-M74**)	$T_{XYZ} = \sum_{i<j}^{k} \frac{\left(\sum_t^n U_{it}U_{jt} - n\overline{U_i U_t}\right)\Big/ A}{\frac{k}{2}(k-1) - n\sum_{i<j}^{k}\left[\frac{\overline{U_i U_j}}{A}\right]}$ $A = \left(\sum_t^n {U_{it}}^2 \sum_t^n {U_{jt}}^2\right)^{\frac{1}{2}}$	S
LC-RULE Linearity-corrected (**M75-M76**)	$T_{XYZ} = \frac{r_{XY} + r_{YZ} + r_{ZX}}{3}$	S

n is the dimension (3), *k* is the number of combinations (i,j), when i<j [(1,2) (1,3) and (2,3)], $\overline{U}$is the arithmetic mean of the the variable U. The values of the subscript "***i*** *"* (1,2,3) stands for the amino acids (X,Y,Z), respectively (**e.g** for the combination (1,2) U_1 and U_2 represent the amino acids X and Y) and r_{XY} is the Pearson correlation between variables X and Y.

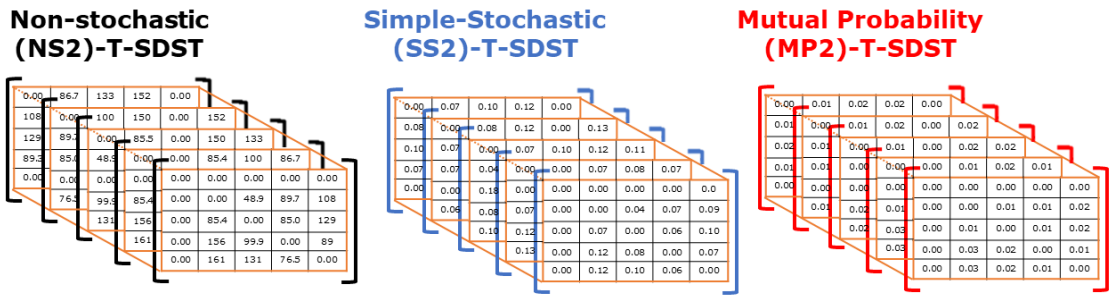


Figure 8. Numerical calculation example considering 2 different types of tensor normalization procedures and an order two exponent (*k*=**2**) for the Haddamard product on a three-tuple spatial dis similarity tensor (**T-SDST**) employing the truncated peptide 5WRX. The three-tuple tensor was calculated considering an amide carbon protein representation (**AB**) and distance to the center was not considered. The normalization procedures considered for this example were: simple stochastic (**SS**) (See Equation **4**) and mutual probability (MP) (See Equation **6**).

3.3. Normalization procedures

One of the main advantages of using normalization procedures is providing an information standardization when applied on a mathematical object.[27] These procedures have not been normally employed on geometrical matrices, however, this strategy has been used for several MDs definition.[28–32]

To apply probabilistic transformations for the two-tuple and three-tuple tensor approaches as a generalization, 2 normalization schemes will be evaluated: a) simple stochastic (**SS**) and b) mutual probability (**MP**); regarding the double stochastic approach, since it was proven that the aforementioned approach contains collinear information with **SS** and the calculation time for this scheme is considerably larger than for the other two approaches, this probability scheme was ruled out[33].

The kth simple-stochastic for two and three-tuple-(dis)similarity tensors ${}^{D}_{ss}\mathbb{Z}^{k} and {}^{T}_{ss}\mathbb{Z}^{k}$ (SS-D-SDST and SS-T-SDST) and kth mutual probability for two and three-tuple-(dis)similarity tensors ${}^{D}_{mp}\mathbb{Z}^{k} and {}^{T}_{mp}\mathbb{Z}^{k}$ (MP-D-SDST and MP-T-SDST), can be defined by applying the following equations:

$$ {}^{D}_{ss}z^{k}_{ij} = \frac{{}^{D}_{ns}z^{k}_{ij}}{S_i} = \frac{{}^{D}_{ns}z^{k}_{ij}}{\sum_{j=1}^{n} {}^{D}_{ns}z^{k}_{i}} \quad \textbf{(3)} $$

$$ {}^{T}_{ss}z^{k}_{ijl} = \frac{{}^{T}_{ns}z^{k}_{ijl}}{S_{jl}} = \frac{{}^{T}_{ns}z^{k}_{ijl}}{\sum_{j=1}^{n}\sum_{k=1}^{n} {}^{T}_{ns}z^{k}_{ijl}} \quad \textbf{(4)} $$

$$ {}^{D}_{mp}z^{k}_{ij} = \frac{{}^{D}_{ns}z^{k}_{ij}}{S_{ij}} = \frac{{}^{D}_{ns}z^{k}_{ijl}}{\sum_{i=1}^{n}\sum_{j=1}^{n} {}^{D}_{ns}z^{k}_{ij}} \quad \textbf{(5)} $$

$$ {}^{T}_{mp}z^{k}_{ijl} = \frac{{}^{T}_{ns}zt^{k}_{ijl}}{S_{ijl}} = \frac{{}^{T}_{ns}z^{k}_{ijl}}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} {}^{T}_{ns}z^{k}_{ijl}} \quad \textbf{(6)} $$

where, ${}^{D}_{ns}z^{k}_{ijl}$, ${}^{T}_{ns}z^{k}_{ijl}$ are the elements of the k[th] non-stochastic two and three-tuple-spatial (dis)similarity tensors. S_i is the summation of all elements on a row on the two-tuple tensor, S_{jl} is the summation of all entries of the two-tuple tensor corresponding to each aa i in a three-tuple matrix for the simple stochastic case. Considering the mutual probability scheme, S_{ij} is the summation of all elements on the two-tuple tensor, S_{ijl} is the summation of all elements of the three-tuple tensor (see Figure **7** and Figure **8**)

3.4. Topological and Geometrical Cut-offs for fragment evaluation

Non-covalent interactions have a central effect on the final structure of macromolecules, their specific binding modes and the self-organizing process of cellular structures and macromolecular as well other functions.[1] The relationship between the distance and the magnitude of the non-covalent interactions of diverse nature (functional groups and steric conditions) demonstrates their contribution to the maintenance of the 3D protein structure. On other hand, the relationship between the topology and the folding of biopolymers has been elucidated in diverse studies, where significant correlation between simple structural parameters and the speed of protein folding has been found. [7,34–36] In this way, sometimes it may be useful to build tensors with information on the interaction between amino acid residues found at a certain distance (or distance range) in the sequence with the objective of studying possible

relations between a specific property and topological features of the native state of the protein.

For the purpose of considering solely some types of non-covalent interactions in global or local indices, two different approaches are applied:

1) Geometric cut-off (*l*), based on Euclidean distance at lag *l*, termed as "*length cut-off"*.

2) Graph-theoretical cut-off (*p*) based on topological distance at lag *p*, designated as "*path cut-off"*.

The application of one or both cut-offs over ${}^{D,T}_{\boldsymbol{ns}}\mathbb{Z}^{\boldsymbol{k}}$ generates the geometric tensor at the lags *l* and/or *p* and their entries are calculated as follows:

$$\begin{aligned} {}^{(p/l)}_{ns}z^{k}_{ij} &= {}^{(p/l)}_{ns}z^{k}_{ij} \times \delta_{ij} \qquad (\delta_{ij} = 1 \text{ if } p_{min} \leq p_{ij} \leq p_{max} \text{ and/or } l_{min} \leq l_{ij} \leq l_{max}) \\ {}^{(p/l)}_{ns}z^{k}_{ij} &= 0 \qquad \text{otherwise} \end{aligned} \tag{7}$$

where, l_{min} and l_{max} are the lower and upper bounds for the metric dependent distance between amino acids, respectively, and l_{ij} is the metric dependent distance between the amino acids *i* and *j*; p_{min} and p_{max} are the pre-defined topological distance thresholds, p_{ij} is the topological distance between the amino acids *i* and *j*. It is important to note that when the length and/or path thresholds are applied to the computation of the ${}^{(\boldsymbol{p/l})}_{\boldsymbol{ns}}\mathbb{Z}^{\boldsymbol{k}}$, a sparse tensor (a tensor with relatively few nonzero elements) is obtained, where each entry ${}^{(\boldsymbol{p/l})}_{\boldsymbol{ns}}\boldsymbol{z}^{\boldsymbol{k}}_{\boldsymbol{ij}}$ coincides with its original definition (the term $\boldsymbol{\delta^{ij}}$=1, [see Eq. (**7**)], only if the metric dependent (l_{ij}) and/or topological (p_{ij}) distances between amino acids *i* and *j* lie(s) within the pre-defined geometric (l_{min}-l_{max}) and/or topological (p_{min}-p_{max}) intervals or it is zero otherwise.

For instance, the use of the *length* criterion (together with exponent *k*) permits to take account only those non-covalent interactions among the functional groups of the

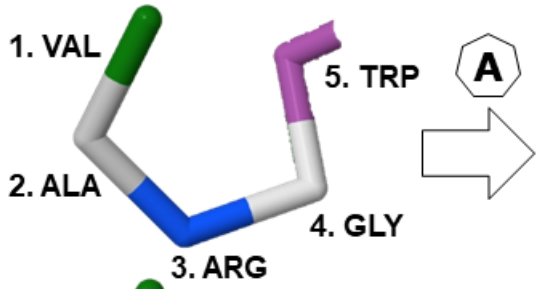
amino acids, which meaningfully contribute to the preservation of the 3D protein structure.

On the other hand, the *path* criterion allows to the selection of the non-covalent interactions for amino acids within a given topological distance. It should be noted that the topological distance between two amino acids *i* and *j* is determined by the *shortest path* between vertices *i* and *j* (e.g. C_{AB}^{i} , C_{AB}^{j}) of the graph.

Illustrations of the application of the length, path or both constrains to the computation of entries of the non-stochastic two or three tuple tensor considering k=1 at the lags l and/or p ($^{(\boldsymbol{p/l})}_{\boldsymbol{ns}}\mathbb{Z}^{\mathbf{1}}$) to characterize the 3D structure of a sample peptide could be found in Figure **9**.

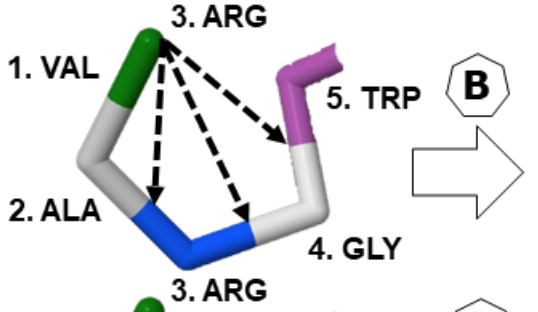
Lastly, the k^{th} simple- and mutual probability tensors at the lags l and/or p can be computed from the k^{th} non-stochastic tensor at the lags l and/or p, in the same way as described in Subsection **3.3**.

The constraints approach (both length and path thresholds) allows to unify geometric and topological information in the same tensor and they also permit to consider the most relevant interactions and at the same time excluding irrelevant chemical information due to long-range interactions. It is not mandatory to use any constraints for calculations, however, incorporating this approach may be beneficial as the “cut-offs” permits the discrimination of the interaction types.



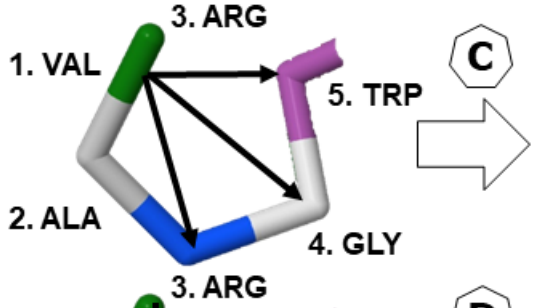
A

0.00	3.80	6.23	6.85	4.80
3.80	0.00	3.80	6.43	6.12
6.23	3.80	0.00	3.79	5.78
6.85	6.43	3.79	0.00	3.80
4.80	6.12	5.78	3.80	0.00



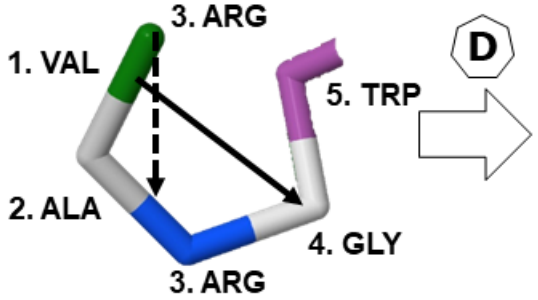
B

0.00	0.00	0.00	0.00	4.80
0.00	0.00	3.80	0.00	0.00
0.00	3.80	0.00	3.80	0.00
0.00	0.00	3.79	0.00	3.80
4.80	0.00	0.00	0.00	0.00



C

0.00	0.00	6.23	6.85	4.80
0.00	0.00	0.00	6.43	6.12
6.23	0.00	0.00	0.00	5.76
6.85	6.43	0.00	0.00	0.00
4.80	6.12	5.76	0.00	0.00



D

0.00	0.00	0.00	0.00	4.80
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00
4.80	0.00	0.00	0.00	0.00

Figure 9. Numerical calculation example considering 2 different types of cut-offs on a two-tuple spatial dis similarity tensor (**D-SDST**) employing the truncated peptide 5WRX. The two-tuple tensor was calculated considering an amide carbon protein representation (**AB**), Non stochastic tensor (**NS**), k=**1** and distance to the center was not considered. (**A**) **D-SDST** without cut-off. The cut-offs applied were: (**B**) Topological (lag p) constraint, cut-off interval [**3-5**], (**C**) Geometrical (lag l) constraint, cut-off interval [**4-7** Å], (**D**) Topological and geometrical constraints, cut-off interval [**3-5**] and [**4-7** Å], respectively.

4. N-linear algebraic forms as a strategy for MDs calculation

The definition for any k^{th} two or three-linear biomacro-molecular descriptors for a protein must consider a canonical basis set and the application of N-linear forms (two-linear or three-linear) in a $\mathbb{R}^n$ space; equations (**6**) and (**9**) indicate the mathematical expressions for these definitions:

$$ {}^{k}_{D}L = bl^{k}(\bar{x}, \bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} z_{ij}^{k} x^{i} y^{j} \quad (\mathbf{8}) $$

$$ {}^{k}_{T}L = tr^{k}(\bar{x}, \bar{y}, \bar{p}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} z_{ijl}^{k} x^{i} y^{j} p^{l} \quad (\mathbf{9}) $$

These two-linear and three-linear forms could be defined by using matrix notation as follows:

$$ {}^{k}_{D}L = [X]\,\mathbb{Z}^{k}[Y]^{T} = \mathrm{X}_{(1\times n)}\,{}^{D}\mathbb{Z}^{k}_{(n\times n)}\mathrm{Y}_{(n\times 1)} \quad (\mathbf{10}) $$

$$ {}^{k}_{T}L = [X]\,\mathbb{Z}^{k}[Y]^{T}[P]^{T} = \mathrm{X}_{(1\times n\times 1)}\,{}^{T}\mathbb{Z}^{k}_{(n\times n\times n)}\mathrm{Y}_{(n\times 1\times 1)}\mathrm{P}_{(1\times 1\times n)} \quad (\mathbf{11}) $$

where, ${}^{k}_{D}L$ *and* ${}^{k}_{T}L$ are the resulting two-linear and three-linear form MD, *n* is the number of amino acids (*aa*) present on the protein, $[X]$, $[Y]$, $[P]$ are the macro-molecular vectors containing $x^1,\ldots,x^n$, $y^1,\ldots,y^n$ and $p^1,\ldots,p^n$ elements, which are the physicochemical properties of every *aa* present in the protein structure (Section 2). The *k^{th} two and three-tuple- spatial (dis)similarity tensors* (D-SDST and T-SDST) (${}^{D}\mathbb{Z}^{k}$ *and* ${}^{T}\mathbb{Z}^{k}$) are a two and three-order tensors whose elements z_{ij}^{k} *and* z_{ijl}^{k} are calculated by using relationships (metrics and multi-metrics) between two and three *aa*, respectively (Section **3**) (See Figure 10).

Based on the physicochemical nature of the properties used for the macromolecular vectors conformation, the following algebraic forms could be defined: *Two-linear:* 1) Bilinear (**B**) (when all macromolecular vectors are configured differently, that is, using 2 different *aa* properties), 2) Quadratic (**Q**) (when both macromolecular vectors have the same configuration, that is, using the same *aa* property), 3) Linear (**F**) (when 1 macromolecular vector is the identity vector and the other one is an *aa* property).

Three-linear: 1) Trilinear Canonical (**Tr**) (when all macromolecular vectors are configured differently, that is, using 3 different *aa* properties), 2) Trilinear linear (**TrF**)

(when 2 of the macro-molecular vectors are the identity vector and the other one is an *aa* property), 3) Trilinear bilinear (**TrB**) (when 2 macromolecular vectors have the same configuration (that is to say, by using the same *aa* property) and the other one is the identity vector), 4) Trilinear quadratic bilinear (**TrQB**) (when 2 macromolecular vectors have the same configuration and the other one has a different aa property from the previous), and 5) Trilinear cubic (**TrC**) (when all the macromolecular vectors have the same configuration, *i.e.*, use the same aa property).

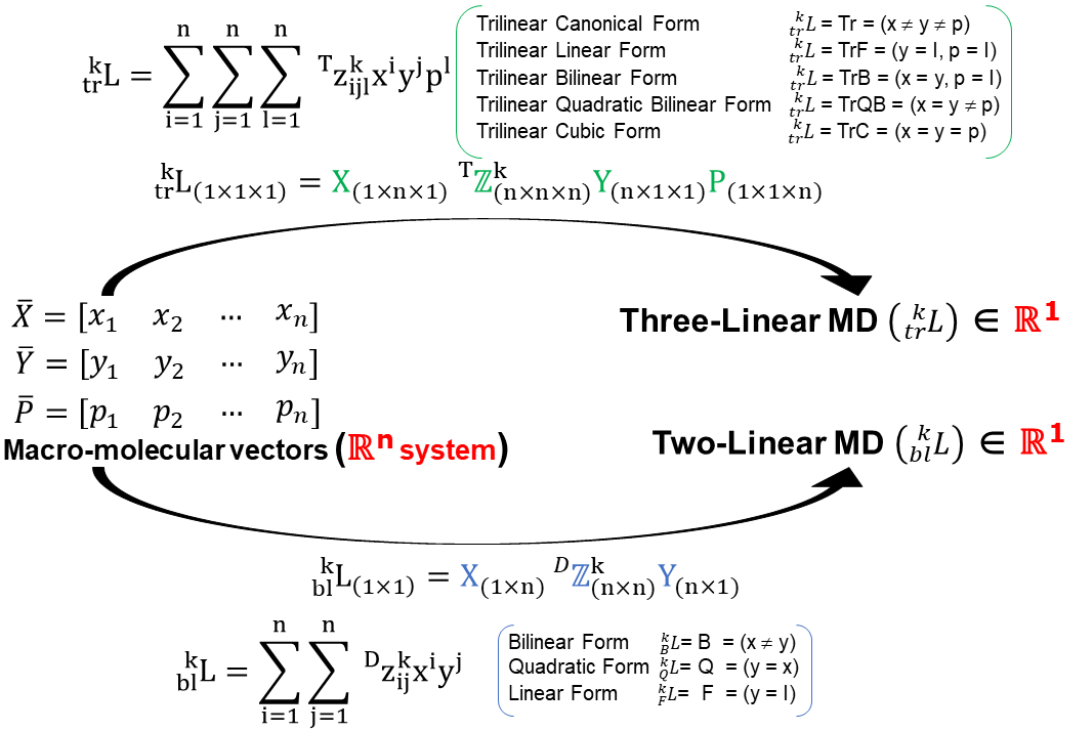


Figure 10. Schematic indication of the transformation of the information contained on macro-molecular vectors using spatial information of the protein (Two and Three-Tuple-Spatial Dis Similarity Tensors, **D-SDST** ($^{D}\mathbb{Z}^{k}$) and **T-SDST** ($^{T}\mathbb{Z}^{k}$), respectively) and algebraic forms. Here, **n** is the number of amino acids present on the protein, [***X***], [***Y***], [***P***] are macro-molecular vectors; $\mathbf{z^{k}_{ij}}$ and $\mathbf{z^{k}_{ijl}}$ are elements of the **D-SDST** and **T-SDST**, respectively, and $_{\mathbf{bl}}\mathbf{L}$ and $_{\mathbf{tr}}\mathbf{L}$ are the resulting two-linear and three-linear MDs. These algebraic forms are defined by the physicochemical nature of the macro-molecular vectors.

5. Amino acid-based MDs definition using N-linear algebraic forms

Considering that the structure of a protein comprises a defined number of amino acids (*aas*), then the k^{th} two or three-linear biomacro-molecular descriptors

for every amino acid are computed by applying two-linear forms (bilinear, quadratic and linear) and three linear forms (trilinear canonical, trilinear quadratic, trilinear quadratic bilinear, trilinear bilinear, trilinear cubic) in $\mathbb{R}^n$, using a canonical ('natural') basis set, and can be expressed by the following equations, respectively:

$$ {}^{k}_{bl}L_{aa} = bl^{aa,k}(\bar{x},\bar{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} z_{ij}^{aa,k} x^i y^i = [X]^T\, \mathbb{Z}^{aa,k}[Y]\,, \forall\ aa = 1,2,\dots,n \quad \textbf{(12)} $$

$$ {}^{k}_{tr}L_{aa} = tr^{aa,k}(\bar{x}\,,\bar{y},\bar{p}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} z_{ijl}^{aa,k} x^i y^j p^l = [X]\,\mathbb{Z}^{aa,k}[Y]^T[P]^T\ \forall\ aa = 1,2,\dots,r \quad \textbf{(13)} $$

where, *n* is the number of amino acids of the protein, $[X], [Y], [P]$ are the macro-molecular vectors containing $x^1,\dots,x^n$, $y^1,\dots,y^n$ and $p^1,\dots,p^n$ elements, respectively, which are physicochemical properties of every *aa* present on the structure.

The *k*th *amino acid-level two-tuple-spatial (dis)similarity tensors* (D-SDST) ($^{D}\mathbb{Z}^{aa,k}$) with elements $^{D}z_{ij}^{aa,k}$ are computed by considering the following rules:

$$ \begin{aligned} {}^{D}z_{ij}^{aa,k} &= {}^{D}z_{ij}^{k} && \text{if } \boldsymbol{i} \wedge \boldsymbol{j} = \mathbf{aa} \\ {}^{D}z_{ij}^{aa,k} &= \tfrac{1}{2}\,{}^{D}z_{ij}^{k} && \text{if } \boldsymbol{i} \vee \boldsymbol{j} = \mathbf{aa} \\ {}^{D}z_{ij}^{aa,k} &= 0 && \text{otherwise} \end{aligned} \quad \textbf{(14)} $$

The *k*th *amino acid-level three-tuple-spatial (dis)similarity tensors* (T-SDST) ($^{T}\mathbb{Z}^{aa,k}$) with elements $^{T}z_{ijl}^{aa,k}$ are computed by considering the following rules:

$$ \begin{aligned} {}^{T}z_{ijl}^{aa,k} &= {}^{T}z_{ijl}^{k} && \text{if } \boldsymbol{i} \wedge \boldsymbol{j} \wedge \boldsymbol{l} = \mathbf{aa} \\ {}^{T}z_{ijl}^{aa,k} &= \tfrac{2}{3}\,{}^{T}z_{ijl}^{k} && \text{if } \boldsymbol{i,j} \vee \boldsymbol{j,l} \vee \boldsymbol{i,l} = \mathbf{aa} \\ {}^{T}z_{ijl}^{aa,k} &= \tfrac{1}{3}\,{}^{T}z_{ijl}^{k} && \text{if } \boldsymbol{i} \vee \boldsymbol{j} \vee \boldsymbol{l} = \mathbf{aa} \\ {}^{T}z_{ijl}^{aa,k} &= 0 && \text{otherwise} \end{aligned} \quad \textbf{(15)} $$

Consequently, if a protein contains "B" *aa* in its structure, the D-SDST (${}^{D}\mathbb{Z}^{k}$) and the T-SDST (${}^{T}\mathbb{Z}^{k}$) can be expressed as the sum of "B" *aa*-level matrices (${}^{D}\mathbb{Z}^{aa,k}$ *and* ${}^{T}\mathbb{Z}^{aa,k}$) (see Figure **11**Figure **12**). From this concept, after the application of algebraic maps on every D-SDST and T-SDST, we will obtain "B" *aa*-level indices, denoted as ${}^{k}_{D}L_{aa}$ *and* ${}^{k}_{T}L_{aa}$ (see Eq. (**10** and **11**)), which will be stored on an array

This array will be designated as LAI (Local Amino Acidic Invariant) as a correspondence of the LOVI vector for organic molecules (Local Vertex Invariant).[37,38] From the LAI vector, the total (whole-protein) three-linear indices can be calculated by using aggregation operators (which is a generalization concept for merging components).[39] These aggregation operators will be discussed in **Section 7.** The general calculation scheme for these novel biomacro-molecular indices is shown in Scheme 1.

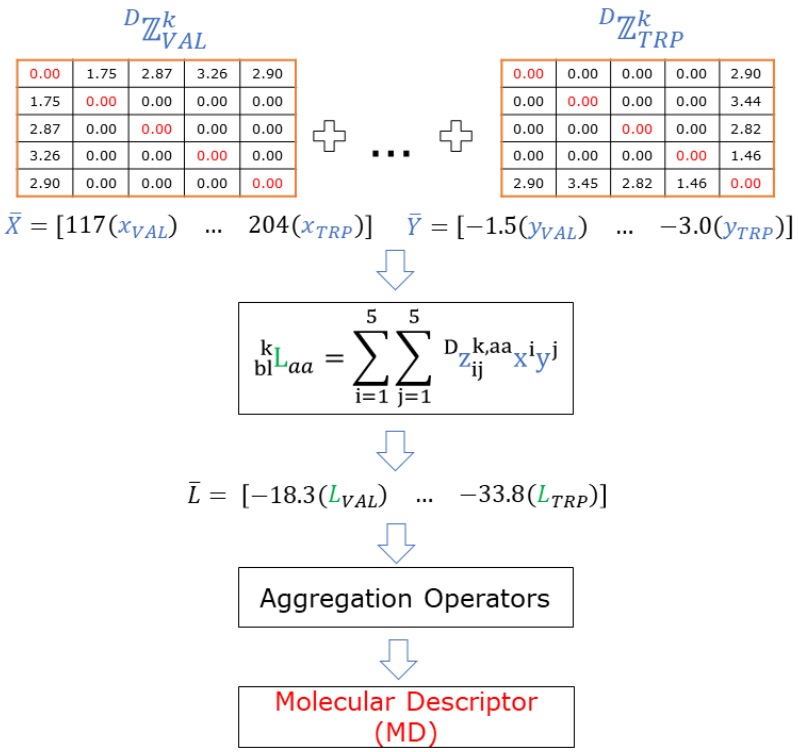


Figure 11. Generation of Local Amino Acid Invariant Vector (LAI vector, $\bar{\boldsymbol{L}}$), which contains every amino acid-based molecular descriptor. This operation considers the use of macro-molecular vectors along the two-tuple spatial dis similarity tensor (**D-SDST**) for every amino acid on the protein. $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{Y}}$ are the macromolecular vector generated considering Molecular Volume (**MV**) and Hopp and Woods Hydropathy index (**HWS**), respectively.

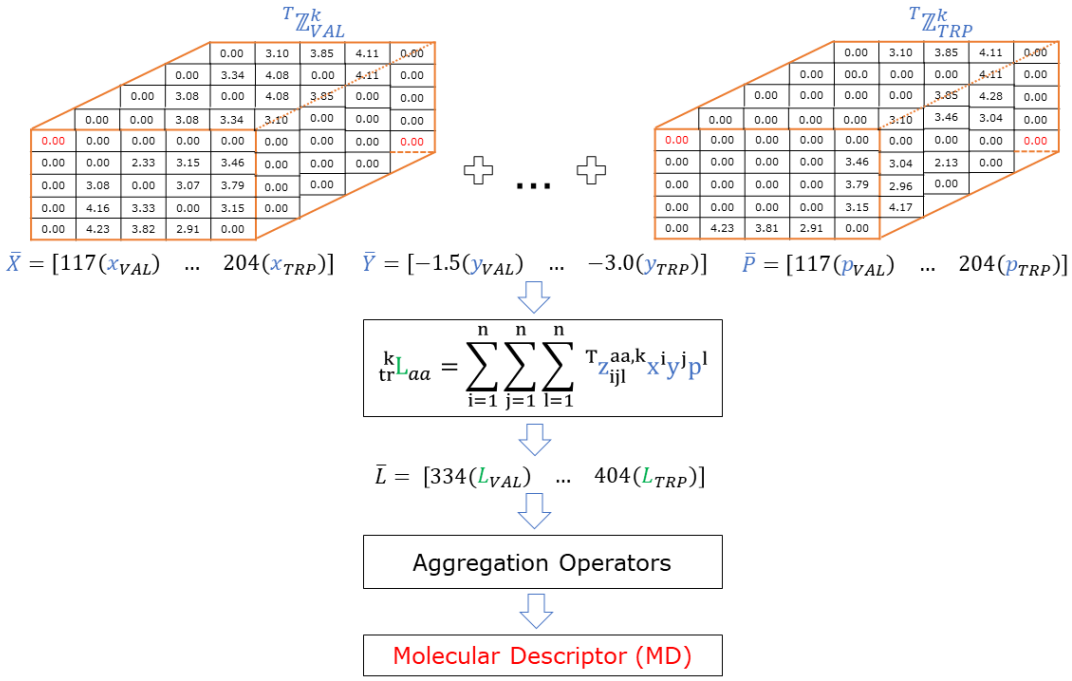


Figure 12. Generation of Local Amino Acid Invariant Vector (LAI vector, $\bar{L}$), which contains every amino acid-based molecular descriptor. This operation considers the use of macro-molecular vectors along the three-tuple spatial dis similarity tensor (**T-SDST**) for every amino acid on the protein. $\bar{X}$, $\bar{Y}$ and $\bar{P}$ are the macromolecular vector generated considering Molecular Volume (**MV**), Hopp and Woods Hydropathy index (**HWS**) and Electronic Charge Index (**ECI**), respectively.

6. Local (Group) based molecular descriptors

Group-based indices can be computed if groups of certain amino acids classified in terms of their activity/properties on solution or their probability to generate a certain secondary structure (see Table **7**) are considered. These indices can be generated by selecting amino acids from the chosen group on the LAI vector. As a consequence, a new vector will be generated (Local Group-based Amino Acidic Invariant (LAI_G)). This operation allows to evaluate the influence of certain *aa* in a variety of applications on protein science.

Table 7. Amino acids groups considered for the computation of the novel 3D algebraic biomacro-molecular descriptors for proteins.

Group	Amino acids
FAH[a]	ALA, CYS, LEU, MET, GLU, GLN, HIS, LYS.
FBS[b]	VAL, ILE, PHE, TYR, TRP, THR.
UFG[c]	GLY, PRO.
AFT[d]	GLY, SER, ASP, ASN, PRO.
ALG[e]	GLY, ALA, PRO, VAL, LEU, ILE, MET.
ARO[f]	PHE, TYR, TRP.
RPC[g]	LYS, HIS, ARG.
RNC[h]	ASP, GLU.
RAP[i]	PRO, ILE, ALA, VAL, LEU, PHE, TRP, MET.
RPU[j]	ASN, CYS, GLY, SER, THR, TYR, GLN.

[a]Alpha helix favoring amino acids; [b] Beta-sheets favoring amino acids; [c] Unfolding amino acids; [d]Beta-turn favoring amino acids; [e]Aliphatic; [f]Aromatic; [g]Polar positively charged; [h]Polar negatively charged; [e]Apolar; [j]Polar uncharged.

7. Aggregation operators as a tool for MDs generalization

The notion of using linear combination as a strategy for merging components has been widely used in the scientific field. However, in the articles of Barigye,[39,40] it was demonstrated that other mathematical merging operators yielded better results than the results obtained from the linear combination on chemical properties data. These invariants are classified in four major groups (see Table **8**, Table **9**,Table **10** for the mathematical expressions of all operators): **a) Norms (or Metrics) Invariants:** Minkowski norms (**N1, N2, N3**). Note that the **N1** in our case is equivalent to the summation of the components of vector $\bar{L}$. **b) Mean Invariants (first statistical moment):** Geometric mean (**GM**), arithmetic mean (**AM**), quadratic mean (**P2**), power mean of third degree (**P3**) and harmonic mean (**A**). **c) Statistical Invariants (highest statistical moments):** Variance (**V**), skewness (**S**), kurtosis (**K**), standard deviation (**SD**), variation coefficient (**CV**), range (**R**), percentile 25 (**Q1**), percentile 50 (**Q2**), percentile 75 (**Q3**), inter-quartile range (**I50**), maximum L_i (**MX**) and minimum L_i (MN). **d) Classical Invariants:** Autocorrelation (**AC**), Gravitational (**GV**), Total Information Content (**TIC**), Mean Information Content (**MIC**), Standardized Information Content (**SIC**), Total Sum (**TS**), Ivanciuc – Balaban (**IB**), Electrotopological State (**ES**) and Kier-Hall Connectivity (**KH**).

The use of these mathematical operators on the LAIs vector enables us to obtain a series of indices that globally or partially characterize a protein (See Figure **13**).

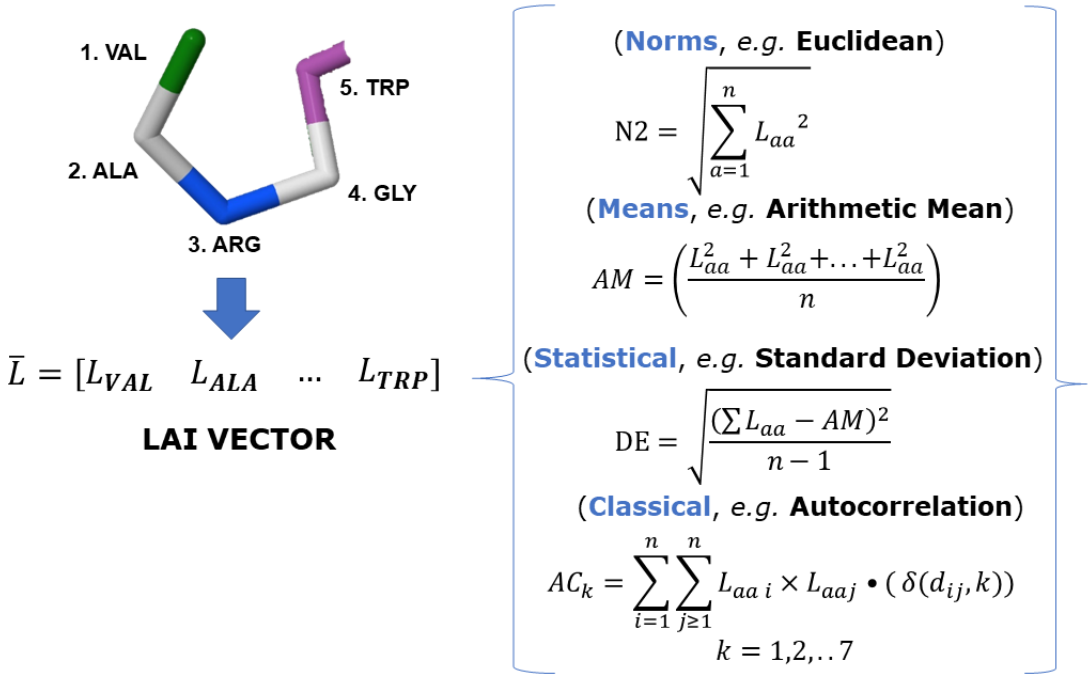


Figure 13. Use of aggregation operators (AOs) as a strategy to fuse amino acid-based molecular descriptors into a global molecular descriptor. There are several types of AOs proposed on this study, see Tables **8**, **9** and **10**.

8. Suggested theoretical configurations for in-software calculations

To reduce the number of MDs to evaluate after the calculation, several analyses (information redundancy and collinearity) were performed considering all theoretical consideration available. As a result, 15 suggested theoretical configurations for two-tuple descriptors (here designed as *projects*) and 10 suggested configurations for three tuple descriptors, were obtained. The project configuration for the two and three tuple descriptors are shown in Table **11** and Table **12**.

From these projects, a total of 13.648 were generated with the two-tuple approach and 20.263 MDs were generated with the three tuple approach on an HPC with the following computational characteristics: 16 cores Intel (R) Xeon (R) E5-2630 v3 @ 2.4 GHz and 64 GB of RAM using MuLiMs console version.

Table 8. Mathematical formulae for Norms and Mean Aggregation operators.

No.	Group	Name	ID	Formula
1	Norms (Metrics)	Minkowski norm (p = 1) Manhattan norm	N1	$\mathrm{N1} = \sum_{a=1}^{n} L_a$
2		Minkowski norm (p = 2) Euclidean norm	N2	$\mathrm{N2} = \sqrt{\sum_{a=1}^{n} {L_a}^2}$
3		Minkowski norm (p = 3)	N3	$\mathrm{N3} = \sqrt[3]{\sum_{a=1}^{n} {L_{aa}}^3}$
4	Mean (first statistical moment)	Geometric Mean	GM	$\mathrm{GM} = \sqrt[n]{\prod_{a=1}^{n} L_{aa}}$
5		Arithmetic Mean (Power mean of degree β = 1)	AM	$\mathrm{M}_\beta = \left(\frac{L_{aa}^\beta + L_{aa}^\beta + \ldots + L_{aa}^\beta}{n}\right)^{\frac{1}{\beta}}$
6		Quadratic Mean (Power mean of degree β = 2)	P2	
7		Power mean of degree β = 3	P3	
8		Harmonic Mean (Power mean of degree β = -1)	A	

Table 9. Mathematical formulae for Statistical Aggregation operators

No.	Group	Name	ID	Formula
9		Skewness	S	$S = \frac{n * (X_3)}{(n-1)(n-2)(SD)^3}$ $X_3 = \sum_{a=1}^{n} (L_{aa} - AM)^3$
10		Variance	V	$V = \frac{\sum_{a=1}^{n} (L_{aa} - AM)^2}{n-1}$
11		Kurtosis	K	$K = \frac{n(n+1)X_4 - 3(X_2)(X_2)(n-1)}{(n-1)(n-2)(n-3)(SD)^4}$ $X_j = \sum_{a=1}^{n} (L_{aa} - AM)^j$
12	Statistical (highest statistical moments)	Standard Deviation	DE	$\mathrm{SD} = \sqrt{\frac{(\sum L_{aa} - AM)^2}{n-1}}$
13		Variation Coefficient	CV	$\mathrm{CV} = {SD}/{AM}$
14		Range	R	$\mathrm{R} = L\min_{\max}$
15		Percentile 25	Q1	$\mathrm{Q1} = \left[\frac{N}{4} + \frac{1}{2}\right]$
16		Percentile 50	Q2	$\mathrm{Q2} = \left[\frac{N}{2} + \frac{1}{2}\right]$
17		Percentile 75	Q3	$\mathrm{Q3} = \left[\frac{3N}{4} + \frac{1}{2}\right]$
18		Inter-quartile Range	I50	$\mathrm{I50} = Q3 - Q1$
19		Maximum value	MX	MX = ${}_{\mathrm{tr}}L$ max
20		Minimum value	MN	MN = ${}_{\mathrm{tr}}L$ min

Table 10. Mathematical formulae for Classical Aggregation Operators.

No.	Group	Name	ID	Formula
21	Classical	Autocorrelation	AC^k	$AC_k = \sum_{i=1}^{n}\sum_{j\geq 1}^{n} L_i \times L_j \bullet (\delta(d_{ij},k))k = 1,2,..7$
22		Gravitational	GV^k	$GV_k = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{L_i L_j}{{}^k d_{ij}} \bullet \delta(d_{ij},k))k = 1,2,..7$
23		Total sum at lag k	TS^k	$TS_k = \sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij} \bullet \delta(d_{ij},k))\ \ k = 1,2,\ldots,7$
24		Kier-Hall connectivity	KH^m	$${}^m KH_t = \sum_{i=1}^{K}(\prod_{i=1}^{n_k} L_i, w)^{\lambda}_{k}$$ *where, K is the number of sub-graphs, n_k is the number of amino acids in a group, λ is equal to ½, m and t are the sub-graph order and type, respectively*
25		Mean Information Content	MIC	$$MIC = -\sum_{i=1}^{A} \frac{N_g}{N_o} \cdot \log_2 \frac{N_g}{N_o}$$ *where, Ng is the number of amino acids with the same LAI value. N_o is the number of amino acids in a molecule*
26		Total Information Content	TIC	$TIC = N_0 \cdot \log_2 N_0 - \sum_{g=1}^{G} N_g \cdot \log_2 \quad N_g$
27		Standardized Information Content	SIC	$SIC = \frac{IT}{N_0 \cdot \log_2 N_0}$

Table 11. Theoretical configurations for two-tuple MDs calculation

	Algebraic Form			*Matrix Form*			*Order*	*Metrics*	*Group*			*Properties*					*Cut-off*	
Proj	**B**	**Q**	**F**	**NS**	**SS**	**MP**	**k**	**Duple**	**Total**	**Group**	**Amino acid**	**Electr.**	**Hydr.**	**Steric**	**Aggregation Operators**	**Center**	**Topol.**	**Geom.**
proj1		x			x		(-12) to (0)	M8, M32	x	RPU	ALA, ARG	ISA, ECI		PBS	N1, GM, i50	x		
proj2		x				x	(0) to (12)	M11,M17	x	UFG	GLY, ASP	Z1		MV, PAH	N3, P2, K	x		
proj3		x		x			(-3) to (3)	M3,M24	x	RPC	PHE, TYR	PIE	HWS	Z3, PAH	N1, GM, P2, S, i50	x	>12	(4-11)
proj4		x			x		(-3) to (3)	M5, M26	x	RPU	ARG, ASP	ECI	KDS	MV, PBS	N3, AM, P3, S		(1-3)	(8.1-11)
proj5		x				x	(-12) to (-2)	M7, M11, M32	x	UFG	GLU, LYS, TYR	Z1		Z3	N2, S, i50		(1-3)	
proj6			x	x			(-2) to (2)	M8, M5, M16	x	RPC, FAH	GLU, LYS, TYR	ECI		MV, PAH	N1, GM, K			(6-8)
proj7			x		x	x	(0) to (6)	M7, M11, M26	x	RPU	ALA, TYR	ECI		MV, PAH	N2, P2	x		
proj8			x	x			(0) to (12)	M3, M15, M24	x	FAH, UFG	PHE	ISA		MV	N3, AM, K	x	>12	(4-5.9)
proj9	x					x	(0) to (10)	M16, M17, M24	x	RPC, RPU	ALA, ASP	ISA	HWS	PBS	N1, K	x		
proj10	x				x		(-8) to (0)	M5, M15, M32	x	RPC, UFG	ARG, ARG, LYS	ECI	HWS	MV	N1, K		>12	(4-11)
proj11	x	x	x	x			(-8) to (0)	M8, M11, M15	x	RPU	ALA	Z1	PBS		GM, i50	x		
proj12	x	x	x		x		(-6) to (0)	M8, M17, M24	x	UFG	ARG, LYS		HWS	PAH	N1, P2	x	(1-3)	(8.1-11)
proj13	x	x	x			x	(-5) to (5)	M7, M26	x	RPC	PHE, TYR	Z1		PBS	N3, S		>12	(4-5.9)
proj14	x	x	x	x	x	x	(-1) to (3)	M17, M32	x	FAH	GLU		KDS	Z3	VC, Q1	x	>12	(4-5.9)
proj15	x	x	x	x	x	x	(-1) to (3)	M3, M5	x	RPU	ALA	PIE		MM	P3, MX			

Table 12. Theoretical configurations for three-tuple MDs calculation

	Algebraic Forms					*Matrix Normalization*			*Order*	*Metrics*			*Group*		*Properties*				
Proj	**Tr**	**TrB**	**TrF**	**TrQB**	**TrC**	**NS**	**SS**	**MP**	**k**	**Duple**	**Ternary**	**Total**	**Groups**	**AA**	**Elect.**	**Hydro.**	**Steric**	**Aggregation Operators**	**Center**
proj1			x	x	x	x	x		(-3) to (2)	M5, M8	M41, M45, M59,	x			ISA		PAH, Z3	N1, K, MIC	x
proj2	x	x	x	x	x	x	x	x	(-12) to (-8)	M3, M17	M33, M37, M45, M57,M58	x	RPU, ARO, ALG, FAH, UFG		ECI	HWS	MV	N3, P2, S, I50, SIC	x
proj3		x	x					x	(4) to (8)	M5	M41, M50	x		ASP, TYR	PIE	KDS	PBS	N1, MX, GV	x
proj4				x	x		x	x	(-2) to (6)	M3	M38, M47, M50	x		ARG, PHE	ISA, ECI		MV	GM, P2, I50	x
proj5	x		x			x	x		(6) to (10)	M11, M16	M41,M46 , M57	x	RPC, RPU	ALA, TYR	Z1	KDS	Z3	N3, K, MIC	
proj6				x	x	x		x	(-6) to (2)	M12	M41,M48 , M59	x	ARO	ARG, ASP	ECI, PIE	HWS		N1, S, I50	
proj7			x	x	x		x		(-4) to (4)	M15	M45, M55, M58	x	FAH	GLU	ISA		Z3, PAH	N3, K, Q1	
proj8	x							x	(-12) to (-1)	M3	M33	x	RPU		ISA, ECI	HWS	Z3, PAH	N1, TS	x
proj9			x		x		x		(-2) to (2)	M13	M41, M46, M50	x	ALG, FAH	ALA, ARG	ECI, PIE	HWS		N3, GM, P2	x
proj10		x		x		x			(-12) to (0)	M5	M48, M57 M59	x	RPU, UFG	TYR, PHE			Z3, PBS, MV	N2, AM, MX	x

9. Bibliography

(1) Nelson, D. L.; Cox, M. M. *Lehninger Principles of Bichemistry,* Seventh Ed.; Macmillan Learning: New York, 2017.

(2) Lin, S.; Lapointe, J. Theoretical and Experimental Biology in One. *Biomed. Sci. Eng.* **2013**, *6* (April), 435–442.

(3) Di Paola, L.; De Ruvo, M.; Paci, P.; Santoni, D.; Giuliani, A. Protein Contact Networks: An Emerging Paradigm in Chemistry. *Chem. Rev.* **2013**, *113* (3), 1598–1613. https://doi.org/10.1021/cr3002356.

(4) Wilson, R. *Introduction to Graph Theory*, Fourth Edi.; Prentice Hall: Edinburgh, 1996.

(5) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Todeschini, R., Consonni, V., Eds.; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2009; Vol. 2. https://doi.org/10.1002/9783527628766.

(6) Gonzalez-Diaz, H.; Vilar, S.; Santana, L.; Uriarte, E. Medicinal Chemistry and Bioinformatics - Current Trends in Drugs Discovery with Networks Topological Indices. *Curr. Top. Med. Chem.* **2007**, *7* (10), 1015–1029. https://doi.org/10.2174/156802607780906771.

(7) Plaxco, K. W.; Simons, K. T.; Baker, D. Contact Order, Transition State Placement and the Refolding Rates of Single Domain Proteins. *J. Mol. Biol.* **1998**, *277* (4), 985–994. https://doi.org/10.1006/jmbi.1998.1645.

(8) Mishra, A.; Rana, P. S.; Mittal, A.; Jayaram, B. D2N: Distance to the Native. *Biochim. Biophys. Acta - Proteins Proteomics* **2014**, *1844* (10), 1798–1807. https://doi.org/10.1016/j.bbapap.2014.07.010.

(9) Castillo-Garit, J. A.; Martinez-Santiago, O.; Marrero Ponce, Y.; Casañola-

Martín, G. M.; Torrens, F. Atom-Based Non-Stochastic and Stochastic Bilinear Indices: Application to QSPR/QSAR Studies of Organic Compounds. *Chem. Phys. Lett.* **2008**, *464* (1–3), 107–112. https://doi.org/10.1016/j.cplett.2008.08.094.

(10) Castillo-Garit, J. A.; Marrero Ponce, Y.; Torrens, F.; Rotondo, R. Atom-Based Stochastic and Non-Stochastic 3D-Chiral Bilinear Indices and Their Applications to Central Chirality Codification. *J. Mol. Graph. Model.* **2007**, *26* (1), 32–47. https://doi.org/10.1016/j.jmgm.2006.09.007.

(11) Castillo-Garit, J. A.; Marrero Ponce, Y.; Torrens, F. Atom-Based 3D-Chiral Quadratic Indices. Part 2: Prediction of the Corticosteroid-Binding Globulin Binding Affinity of the 31 Benchmark Steroids Data Set. *Bioorganic Med. Chem.* **2006**, *14* (7), 2398–2408. https://doi.org/10.1016/j.bmc.2005.11.024.

(12) Marrero Ponce, Y.; Torrens, F.; García-Domenech, R.; Ortega-Broche, S. E.; Zaldivar, V. R. Novel 2D TOMOCOMD-CARDD Molecular Descriptors: Atom-Based Stochastic and Non-Stochastic Bilinear Indices and Their QSPR Applications. *J. Math. Chem.* **2008**, *44* (3), 650–673. https://doi.org/10.1007/s10910-008-9389-0.

(13) Marrero Ponce, Y.; Medina-Marrero, R.; Castillo-Garit, J. A.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. Protein Linear Indices of the 'Macromolecular Pseudograph α-Carbon Atom Adjacency Matrix' in Bioinformatics. Part 1: Prediction of Protein Stability Effects of a Complete Set of Alanine Substitutions in Arc Repressor. *Bioorg. Med. Chem.* **2005**, *13* (8), 3003–3015. https://doi.org/https://doi.org/10.1016/j.bmc.2005.01.062.

(14) Ortega-Broche, S. E.; Marrero Ponce, Y.; Díaz, Y. E.; Torrens, F.; Pérez-Giménez, F. Tomocomd-Camps and Protein Bilinear Indices - Novel Bio-

Macromolecular Descriptors for Protein Research: I. Predicting Protein Stability Effects of a Complete Set of Alanine Substitutions in the Arc Repressor. *FEBS J.* **2010**, *277* (15), 3118–3146. https://doi.org/10.1111/j.1742-4658.2010.07711.x.

(15) Collantes, E. R.; Dunn, W. J. Amino Acid Side Chain Descriptors for Quantitative Structure-Activity Relationship Studies of Peptide Analogues. *J. Med. Chem.* **1995**, *38* (14), 2705–2713. https://doi.org/10.1021/jm00014a022.

(16) Kyte, J.; Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157* (1), 105–132. https://doi.org/10.1016/0022-2836(82)90515-0.

(17) Hopp, T. P.; Woods, K. R. Prediction of Protein Antigenic Determinants from Amino Acid Sequences. *Proc. Natl. Acad. Sci. USA* **1981**, *78* (6), 3824–3828. https://doi.org/10.1073/pnas.78.6.3824.

(18) Sillero, A.; Ribeiro, J. M. Isoelectric Points of Proteins: Theoretical Determination. *Anal. Biochem.* **1989**, *179* (2), 319–325. https://doi.org/10.1016/0003-2697(89)90136-X.

(19) Hellberg, S.; Sjoestroem, M.; Skagerberg, B.; Wold, S. Peptide Quantitative Structure-Activity Relationships, a Multivariate Approach. *J. Med. Chem.* **1987**, *30* (7), 1126–1135. https://doi.org/10.1021/jm00390a003.

(20) Zamyatnin, A. A. Protein Volume in Solution. *Prog. Biophys. Mol. Biol.* **1972**, *24* (C), 107–123. https://doi.org/10.1016/0079-6107(72)90005-3.

(21) *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A., Eds.; Gordon and Breach Science Publishers, 1999.

(22) Nikolić, S.; Trinajstić, N.; Mihalić, Z.; Carter, S. On the Geometric-Distance Matrix and the Corresponding Structural Invariants of Molecular Systems. *Chem. Phys. Lett.* **1991**, *179* (1), 21–28. https://doi.org/https://doi.org/10.1016/0009-

2614(91)90285-H.

(23) Deza, E.; Deza, M.-M. Chapter 3 - Generalizations of Metric Spaces; Deza, E., Deza, M.-M. B. T.-D. of D., Eds.; Elsevier: Amsterdam, 2006; pp 36–43. https://doi.org/https://doi.org/10.1016/B978-044452087-6/50003-2.

(24) Warrens, M. *Similarity Coefficients for Binary Data: Properties of Coefficients, Coefficient Matrices, Multi-Way Metrics and Multivariate Coefficients*; 2008.

(25) Deza, E.; Deza, M.-M. Chapter 1 - General Definitions. In *Dictionary of Distances*; Deza, E., Deza, M.-M. B. T.-D. of D., Eds.; Elsevier: Amsterdam, 2006; pp 2–30. https://doi.org/https://doi.org/10.1016/B978-044452087-6/50001-9.

(26) Deza, E.; Deza, M.-M. Chapter 4 - Metric Transforms; Deza, E., Deza, M.-M. B. T.-D. of D., Eds.; Elsevier: Amsterdam, 2006; pp 44–49. https://doi.org/https://doi.org/10.1016/B978-044452087-6/50004-4.

(27) Garcia-Jacas, C.; Marrero-Ponce, Y.; Barigye, S. J.; Valdes-Martin, J. R.; Rivera-Borroto, O. M.; Olivero-Verbel, J. N-Linear Algebraic Maps for Chemical Structure Codification: A Suitable Generalization for Atom-Pair Approaches? *Curr. Drug Metab.* **2014**, *15*, 441–469. https://doi.org/10.2174/1389200215666140605124506.

(28) Marrero Ponce, Y.; González-Díaz, H.; Zaldivar, V. R.; Torrens, F.; Castro, E. A. 3D-Chiral Quadratic Indices of the 'Molecular Pseudograph's Atom Adjacency Matrix' and Their Application to Central Chirality Codification: Classification of ACE Inhibitors and Prediction of σ-Receptor Antagonist Activities. *Bioorg. Med. Chem.* **2004**, *12* (20), 5331–5342. https://doi.org/https://doi.org/10.1016/j.bmc.2004.07.051.

(29) Ramos de Armas, R.; González Díaz, H.; Molina, R.; Uriarte, E. Markovian

Backbone Negentropies: Molecular Descriptors for Protein Research. I. Predicting Protein Stability in Arc Repressor Mutants. *Proteins Struct. Funct. Bioinforma.* **2004**, *56* (4), 715–723. https://doi.org/10.1002/prot.20159.

(30) Gonzáles-Díaz, H.; Gia, O.; Uriarte, E.; Hernádez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; et al. Markovian Chemicals "in Silico" Design (MARCH-INSIDE), a Promising Approach for Computer-Aided Molecular Design I: Discovery of Anticancer Compounds. *J. Mol. Model.* **2003**, *9* (6), 395–407. https://doi.org/10.1007/s00894-003-0148-7.

(31) Klein, D. J.; Palacios, J. L.; Randić, M.; Trinajstić, N. Random Walks and Chemical Graph Theory. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1521–1525. https://doi.org/10.1021/ci040100e.

(32) Carbó-Dorca, R. Stochastic Transformation of Quantum Similarity Matrices and Their Use in Quantum QSAR (QQSAR) Models. *Int. J. Quantum Chem.* **2000**, *79* (3), 163–177. https://doi.org/10.1002/1097-461X(2000)79:3<163::AID-QUA2>3.0.CO;2-0.

(33) García-Jacas, C.; Marrero-Ponce, Y.; Acevedo-Martínez, L.; Barigye, S. J.; Valdés-Martiní, J. R.; Contreras-Torres, E. QuBiLS-MIDAS: A Parallel Free-Software for Molecular Descriptors Computation Based on Multilinear Algebraic Maps. *J. Comput. Chem.* **2014**, *35* (18), 1395–1409. https://doi.org/10.1002/jcc.23640.

(34) Gromiha, M.; Selvaraj, S. Comparison between Long-Range Interactions and Contact Order in Determining the Folding Rate of Two-State Proteins: Application of Long-Range Order to Folding Rate Prediction. *J. Mol. Biol.* **2001**, *310* (1), 27–32. https://doi.org/10.1007/s10853-005-0576-0.

(35) Gromiha, M.; Saraboji, K.; Ahmad, S.; Ponnuswamy, M. N.; Suwa, M. Role of

Non-Covalent Interactions for Determining the Folding Rate of Two-State Proteins. *Biophys. Chem.* **2004**, *107* (3), 263–272. https://doi.org/https://doi.org/10.1016/j.bpc.2003.09.008.

(36) Gromiha, M. Importance of Native-State Topology for Determining the Folding Rate of Two-State Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1481–1485. https://doi.org/10.1021/ci0340308.

(37) Todeschini, R.; Consonni, V. New Local Vertex Invariants and Molecular Descriptors Based on Functions of the Vertex Degrees. *MATCH - Commun. Math. Comput. Chem.* **2010**, *64*, 359–372. https://doi.org/10.1016/j.renene.2014.11.073.

(38) Balaban, A. Local versus Global (i.e. Atomic versus Molecular) Numerical Modeling of Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (2), 398–402. https://doi.org/10.1021/ci00018a028.

(39) Barigye, S. J.; Marrero Ponce, Y.; Martínez-López, Y.; Torrens, F.; Artiles-Martínez, L. M.; Pino-Urias, R. W.; Martínez-Santiago, O. Relations Frequency Hypermatrices in Mutual, Conditional, and Joint Entropy-Based Information Indices. *J. Comput. Chem.* **2012**, *34* (4), 259–274. https://doi.org/doi:10.1002/jcc.23123.

(40) Barigye, S.; Marrero-Ponce, Y.; Santiago, O.; Lopez, Y.; Perez-Gimenez, F.; Torrens, F. Shannon's, Mutual, Conditional and Joint Entropy Information Indices: Generalization of Global Indices Defined from Local Vertex Invariants. *Curr. Comput. Aided-Drug Des.* **2013**, *9* (2), 164–183. https://doi.org/10.2174/1573409911309020003.