Supporting Information

## MD-LAIs Software: Computing Whole-Sequence and Amino Acid-Level “Embeddings” for Peptides and Proteins

**Ernesto Contreras-Torres,1,2\* and Yovani Marrero-Ponce,2,3\***

*1Norewian Cruise Line Holdings Limited, Corporate Center Drive, Miami, FL, 33216, United States*

*2Facultad de Ingeniería, Universidad Panamericana, Augusto Rodin No. 498, Insurgentes Mixcoac, Benito Juárez, Ciudad de México 03920, México.*

*3Universidad San Francisco de Quito (USFQ), Grupo de Medicina Molecular y Traslacional (MeM&T), Colegio de Ciencias de la Salud (COCSA), Escuela de Medicina, Edificio de Especialidades Médicas, Quito, 170157, Pichincha, Ecuador.*

## Table of Contents

[MD-LAIs Software: Computing Whole-Sequence and Amino Acid-Level “Embeddings” for Peptides and Proteins 1](#_Toc179283462)

[Table of Contents 2](#_Toc179283463)

[1. MD-LAIs: Molecular Descriptors from Local Amino Acid Invariants. 4](#_Toc179283464)

[2. Macromolecular Vectors for Encoding Protein Sequences 4](#_Toc179283465)

[3. Group-based Macromolecular Vectors 10](#_Toc179283466)

[4. Fuzzy-Weighted Macromolecular Vectors 11](#_Toc179283467)

[5. Aggregation Operators for MDs Generalization. 13](#_Toc179283468)

[6. Window-based Generalized Amino Acid-Level MDs by Applying AOs. 20](#_Toc179283479)

[7. References 22](#_Toc179283480)

# MD-LAIs: Molecular Descriptors from Local Amino Acid Invariants.

Given a protein (or peptide) having *n* amino acids (*aa*), any **MD-LAI** descriptor is defined as:

|  |  |
| --- | --- |
|  | (1) |

where, is a macromolecular vector of size *n*, whose components *L1, L2,…,Ln*accounts for a physical-chemical (or structural) property of every *aa* within the protein sequence. Each component of is defined as follows:

|  |  |
| --- | --- |
|  | (2) |

where, *p*i is the value of the property *p* for the *ith* *aa* (see **Section 2**), takes the value of 1 if the *ith* *aa* belongs to the group *G* and zero otherwise (see **Section 3**), and is a weight (a value between 0 and 1) computed from a fuzzy membership function according to the position of the *ith* *aa* in the sequence (see **Section 4**). The vector is named as LAI (Local Amino Acidic Invariant) in analogy to the LOVI vector for small organic molecules (Local Vertex Invariant).1,2 From the LAI vector, total (whole-protein), group-based, unweighted, and fuzzy-weighted MDs can be obtained by applying aggregation operators (**AOs**) (see **Section 5**).3



**Scheme S1.** Major steps followed in the computation of the **MD-LAIs** descriptors.

# 2. Macromolecular Vectors for Encoding Protein Sequences

The encoding of small-to-medium sized organic molecules as molecular vectors has been described in previous reports.4–7 In the case of proteins, this concept can be adapted by considering every *aa* on a protein sequence as an “atomic unit” and the whole-sequence as a macromolecular vector. Each component of a macromolecular vector holds the value for a standard property of the corresponding *aa*.8,9 The properties used here are grouped into five sets. The first set includes **composition**, that is, the unitary property, thus, any descriptor computed with this property will be a count-based MD (e.g., *aa* or *k*-mer composition). The second set includes 16 **chemical-physical and structural** properties such as, volume (MV)10, Hopp-Woods hydropathy scale (HWS)11, electronic charge index (ECI)12, z-scales13, alpha-helix and beta-sheet relative frequencies (PAH and PBS)14 and so on. The third set includes 8 hydrophobic, steric, and topological (**VHSE**) scales.15 The fourth set contains 6 **MD-LOVIS**16 scales. The fifth set comprises 9 **QUBILS-MIDAS**17 scales. Figure **S1** shows seven macromolecular vectors computed with representative properties belonging to the above-mentioned sets. All numerical values of these properties for every *aa* are shown in Table **S1**.

**A screenshot of a computer

Description automatically generated**

**Figure S1.** Example of macromolecular vectors obtained from seven properties for the pentapeptide “VARGW” (**PDB ID**: 5WRX).

**Table S1.** Physicochemical properties for the 20 natural amino acids used for the computation of macromolecular vectors.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Amino acid** | **Code** | **z-scalea** | | | **ISAb** | **ECIc** | **PIEd** | **HWSe** | **KDSf** | **EPSg** | **GCP1h** |
| **z1** | **z2** | **z3** |
| Alanine | ALA | 0.07 | -1.73 | 0.09 | 62.9 | 0.05 | 6.01 | -0.5 | 1.8 | -433.66 | 19.2 |
| Arginine | ARG | 2.88 | 2.52 | -3.44 | 52.98 | 1.69 | 10.76 | 3 | -4.5 | -403.21 | 17.8 |
| Asparagine | ASN | 3.22 | 1.45 | 0.84 | 17.87 | 1.31 | 5.41 | 0.2 | -3.5 | -466.61 | 21.72 |
| Aspartate | ASP | 3.64 | 1.13 | 2.36 | 18.46 | 1.25 | 2.77 | 3 | -3.5 | -518.1 | 17.14 |
| Cysteine | CYS | 0.71 | -0.97 | 4.13 | 78.51 | 0.15 | 5.07 | -1 | 2.5 | -425.69 | 18.83 |
| Glutamate | GLU | 3.08 | 0.39 | -0.07 | 30.19 | 1.31 | 3.22 | 0.2 | -3.5 | -479.54 | 18.55 |
| Glutamine | GLN | 2.18 | 0.53 | -1.14 | 19.53 | 1.36 | 5.65 | 3 | -3.5 | -531.69 | 17.31 |
| Glycine | GLY | 2.23 | -5.36 | 0.3 | 19.93 | 0.02 | 5.97 | 0 | -0.4 | -420.86 | 19.48 |
| Histidine | HIS | 2.41 | 1.74 | 1.11 | 87.38 | 0.56 | 7.59 | -0.5 | -3.2 | -378.92 | 13.97 |
| Isoleucine | ILE | -4.44 | -1.68 | -1.03 | 149.77 | 0.09 | 6.02 | -1.8 | 4.5 | -449.27 | 20.76 |
| Leucine | LEU | -4.19 | -1.03 | -0.98 | 154.35 | 0.01 | 5.98 | -1.8 | 3.8 | -448.27 | 17.65 |
| Lysine | LYS | 2.84 | 1.41 | -3.14 | 102.78 | 0.53 | 9.74 | 3 | -3.9 | -446.97 | 17.05 |
| Methionine | MET | -2.49 | -0.27 | -0.41 | 132.22 | 0.34 | 5.74 | -1.3 | 1.9 | -435.34 | 17.88 |
| Phenylalanine | PHE | -4.92 | 1.3 | 0.45 | 189.42 | 0.14 | 5.48 | -2.5 | 2.8 | -376.77 | 16.81 |
| Proline | PRO | -1.22 | 0.88 | 2.23 | 122.35 | 0.16 | 6.48 | 0 | -1.6 | -422.17 | 18.55 |
| Serine | SER | 1.96 | -1.63 | 0.57 | 19.75 | 0.56 | 5.68 | 0.3 | -0.8 | -479.75 | 18.91 |
| Threonine | THR | 0.92 | -2.09 | -1.4 | 59.44 | 0.65 | 5.87 | -0.4 | -0.7 | -483.37 | 17.15 |
| Tryptophan | TRP | -4.75 | 3.65 | 0.85 | 179.16 | 1.08 | 5.89 | -3.4 | -0.9 | -365.49 | 20.94 |
| Tyrosine | TYR | -1.39 | 2.32 | 0.01 | 132.16 | 0.72 | 5.66 | -2.3 | -1.3 | -446.32 | 16.86 |
| Valine | VAL | -2.69 | -2.53 | -1.29 | 120.91 | 0.07 | 5.97 | -1.5 | 4.2 | -434.3 | 17.88 |
|  |  |  |  |  |  |  |  |  |  |  |  |

aZ-scales (Hellberg et al., 1987), bSide-chain isotropic surface area (Collantes and Dunn III, 1995), cAtomic charge (Collantes and Dunn III, 1995), dIsoelectric point (Hellberg et al., 1987), eHoop-Woods hydropathy index (Hopp and Woods, 1981), fKyte-Doolittle hydropathy index (Kyte and Doolittle, 1982)

Table S1. Physicochemical properties for the 20 natural amino acids used for the computation of macromolecular vectors (*continued*).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Amino acid** | **Code** | **GCP2i** | **MMj** | **MVk** | **PAHl** | **PBSm** | **VHSE1n** | **VHSE4**m | **VHSE6**m |
| Alanine | ALA | -77.85 | 89 | 88.6 | 1.29 | 0.9 | 0.15 | -0.92 | -0.91 |
| Arginine | ARG | 108.86 | 174 | 173.4 | 0.96 | 0.99 | -1.47 | 1.27 | 1.47 |
| Asparagine | ASN | -55.42 | 132 | 114.1 | 0.9 | 0.76 | -0.99 | 0.69 | 0.85 |
| Aspartate | ASP | 47.89 | 133 | 111.1 | 1.04 | 0.72 | -1.15 | -0.01 | 1.31 |
| Cysteine | CYS | 160.13 | 121 | 108.5 | 1.11 | 0.74 | 0.18 | -0.21 | 1.2 |
| Glutamate | GLU | 134.68 | 146 | 143.8 | 1.44 | 0.75 | -0.96 | 0.16 | 0.42 |
| Glutamine | GLN | 53.27 | 147 | 138.4 | 1.27 | 0.8 | -1.18 | 0.36 | -0.17 |
| Glycine | GLY | -148.03 | 75 | 60.1 | 0.56 | 0.92 | -0.2 | 2.28 | -1.18 |
| Histidine | HIS | -4.57 | 155 | 153.2 | 1.22 | 1.08 | -0.43 | 0.19 | 1.28 |
| Isoleucine | ILE | -104.8 | 131 | 166.7 | 0.97 | 1.45 | 1.27 | -1.8 | -1.61 |
| Leucine | LEU | -148.5 | 131 | 166.7 | 1.3 | 1.02 | 1.36 | -0.8 | -1.37 |
| Lysine | LYS | 47.61 | 146 | 168.6 | 1.23 | 0.77 | -1.17 | 0.8 | 0.67 |
| Methionine | MET | 46.37 | 149 | 162.9 | 1.47 | 0.97 | 1.01 | 0 | 0.1 |
| Phenylalanine | PHE | 47.67 | 165 | 189.9 | 1.07 | 1.32 | 1.52 | -0.16 | 0.28 |
| Proline | PRO | 169.73 | 115 | 112.7 | 0.52 | 0.64 | 0.22 | 0.05 | -1.34 |
| Serine | SER | 30.24 | 105 | 89 | 0.82 | 0.95 | -0.67 | -0.41 | 0.27 |
| Threonine | THR | 46.04 | 119 | 116.1 | 0.82 | 1.21 | -0.34 | -1.06 | -0.01 |
| Tryptophan | TRP | 178.69 | 204 | 227.8 | 0.99 | 1.14 | 1.5 | 0.75 | -0.13 |
| Tyrosine | TYR | 49.11 | 181 | 193.6 | 0.72 | 1.25 | 0.61 | 0.73 | 0.25 |
| Valine | VAL | -106.5 | 117 | 140 | 0.91 | 1.49 | 0.76 | -1.91 | -1.4 |

n,m,o,p Vector of Hydrophobic, Electronic and Steric Properties15.

Table S1. Physicochemical properties for the 20 natural amino acids used for the computation of macromolecular vectors (*continued*).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Amino acid** | **Code** | **vhse7o** | **VHSE8p** | **T2q** | **T3r** | **T4s** | **MDL1s** | **MDL2t** | **MDL3u** |
| Alanine | ALA | 0.36 | -0.48 | -1.63 | 0.63 | 1.04 | -1.18 | 0.45 | -1.14 |
| Arginine | ARG | 1.3 | 0.83 | 3.89 | -1.16 | -0.39 | 1.4 | 0.33 | 1.37 |
| Asparagine | ASN | 0.73 | -0.8 | 0.66 | 1.16 | -0.22 | -0.5 | -0.9 | 1.15 |
| Aspartate | ASP | 0.03 | 0.56 | 0.75 | 1.39 | -0.4 | -0.88 | -1.37 | 0.86 |
| Cysteine | CYS | -1.61 | -0.19 | -0.86 | -0.33 | 0.8 | -1.06 | 0.09 | 0.15 |
| Glutamate | GLU | -0.2 | -0.41 | 1.72 | 0.28 | -0.39 | 0.23 | -0.24 | 1.6 |
| Glutamine | GLN | 0.91 | 0.02 | 1.82 | 0.51 | -0.58 | -0.14 | -0.64 | 1.41 |
| Glycine | GLY | 2.01 | -1.34 | -1.21 | -0.12 | 0.75 | -2.23 | -0.23 | -1 |
| Histidine | HIS | 0.93 | 0.65 | -1.31 | 0.01 | -1.81 | 0.37 | -1.13 | -0.65 |
| Isoleucine | ILE | -0.16 | -0.13 | -0.28 | -0.15 | 1.4 | 0.63 | 1.45 | -1.21 |
| Leucine | LEU | 0.08 | -0.62 | 0.28 | -0.49 | 1.45 | 0.52 | 1.55 | 0.32 |
| Lysine | LYS | 1.63 | 0.13 | 2.34 | -1.69 | 0.41 | 0.92 | 1.38 | 0.87 |
| Methionine | MET | -0.86 | -0.68 | 0.98 | -2.34 | 1.64 | 0.22 | 1.31 | -0.63 |
| Phenylalanine | PHE | -1.33 | -0.2 | -0.94 | -0.63 | -1.27 | 1.01 | -0.66 | -0.81 |
| Proline | PRO | -0.19 | 3.56 | -3.54 | -0.53 | -0.36 | -0.66 | 0.57 | 0.98 |
| Serine | SER | -0.64 | 0.11 | -0.65 | 0.68 | -0.17 | -1.02 | -0.33 | -0.08 |
| Threonine | THR | -0.79 | 0.39 | -0.62 | 1.11 | 0.31 | -0.19 | -0.11 | -0.99 |
| Tryptophan | TRP | -1.01 | -0.85 | -2.67 | -0.07 | -1.96 | 1.73 | -1.56 | -1.49 |
| Tyrosine | TYR | -0.96 | -0.52 | -0.47 | 0.07 | -1.67 | 1.08 | -1.14 | -0.31 |
| Valine | VAL | -0.24 | -0.03 | -0.94 | 0.28 | 1.1 | -0.25 | 1.21 | -0.39 |
| q,r,s Topological-indices based scales; t,u,v,w,x **MD-LOVIS** scales | | | | | | | | | |

Table S1. Physicochemical properties for the 20 natural amino acids used for the computation of macromolecular vectors (*continued*).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Amino acid** | **Code** | **mdl4v** | **mdl5w** | **mdl6x** | **mid1y** | **mid2z** | **mid3z\*** |
| Alanine | ALA | 1.58 | -0.93 | 0.37 | 0.768 | -3.045 | 0.266 |
| Arginine | ARG | -0.06 | -0.94 | -0.28 | 0.583 | -0.596 | 1.816 |
| Asparagine | ASN | 0.57 | 0.62 | 0.41 | -0.731 | 0.147 | -0.242 |
| Aspartate | ASP | 0.22 | 1.15 | 1.49 | -1.394 | 0.214 | -0.418 |
| Cysteine | CYS | 0.69 | -0.25 | -1.11 | -1.765 | 0.819 | -0.26 |
| Glutamate | GLU | 0.15 | -0.57 | 0.56 | 0.911 | -0.365 | -0.773 |
| Glutamine | GLN | -0.17 | -0.81 | 1.29 | 0.899 | -0.442 | 0.174 |
| Glycine | GLY | -2.79 | -1.38 | -0.06 | -0.817 | 0.582 | -0.242 |
| Histidine | HIS | -0.26 | -0.53 | -0.14 | -1.082 | -0.114 | -0.203 |
| Isoleucine | ILE | 0.07 | 1.21 | 1.07 | 0.7 | 0.884 | 0.631 |
| Leucine | LEU | -0.37 | -0.91 | 0.89 | 0.56 | 0.759 | 0.553 |
| Lysine | LYS | -0.34 | -0.71 | -1.13 | -1.077 | -0.613 | -0.631 |
| Methionine | MET | 0.43 | -0.59 | -0.14 | 0.242 | 0.07 | 0.924 |
| Phenylalanine | PHE | -0.26 | 0.12 | -1.26 | -1.036 | -0.504 | 1.783 |
| Proline | PRO | -1.15 | 2.54 | -1.94 | 0.531 | -0.364 | 0.509 |
| Serine | SER | 1.55 | -0.44 | -1.42 | 0.986 | -0.069 | -0.861 |
| Threonine | THR | 1.61 | 0.71 | 0.08 | 1.161 | 1.764 | 0.174 |
| Tryptophan | TRP | -0.73 | 0.19 | 0.3 | 0.541 | -0.62 | -2.894 |
| Tyrosine | TYR | -0.17 | 0.12 | -0.48 | -1.218 | -0.033 | -0.065 |
| Valine | VAL | -0.59 | 1.39 | 1.49 | 1.236 | 1.526 | -0.239 |

y,z,z\* QuBiLS-MIDAS scales.

Table S1. Physicochemical properties for the 20 natural amino acids used for the computation of macromolecular vectors (*continued*).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Amino acid** | **Code** | **mid4aa** | **mid5bb** | **mid6cc** | **mid7dd** | **mid8ee** | **mid9ff** |
| Alanine | ALA | -0.57 | 0.949 | -1.335 | 1.998 | 0.345 | -1.63 |
| Arginine | ARG | 1.08 | 0.827 | 3.043 | 1.45 | 1.725 | -1.302 |
| Asparagine | ASN | 0.747 | 1.869 | 1.242 | 1.988 | 0.746 | 0.136 |
| Aspartate | ASP | 1.309 | 2.452 | -0.331 | 1.25 | -0.226 | 0.111 |
| Cysteine | CYS | -0.708 | 2.381 | 0.632 | 1.441 | -0.173 | -1.209 |
| Glutamate | GLU | -1.383 | 0.64 | 2.363 | 1.695 | -0.697 | -1.273 |
| Glutamine | GLN | 0.69 | 1.42 | -1.22 | 1.704 | 1.513 | -1.302 |
| Glycine | GLY | 0.949 | 0 | -0.559 | 0.042 | 0.454 | -0.578 |
| Histidine | HIS | -0.509 | 0.574 | -0.844 | 1.715 | -1.951 | -1.346 |
| Isoleucine | ILE | -1.161 | 1.218 | -1.363 | 1.806 | -1.7 | -0.853 |
| Leucine | LEU | 0.798 | 1.641 | 3.351 | 1.129 | 1.369 | 0.216 |
| Lysine | LYS | -0.606 | -0.413 | -0.988 | 2.835 | 1.626 | -0.707 |
| Methionine | MET | -0.196 | 2.604 | 1.324 | 0.819 | -1.548 | -0.681 |
| Phenylalanine | PHE | 0.334 | 2.256 | -0.772 | 1.087 | -1.424 | 0.802 |
| Proline | PRO | 0.956 | -1.484 | 1.319 | 1.389 | -2.205 | -1.126 |
| Serine | SER | -1.411 | 2.241 | 1.2 | 0.899 | -0.45 | -0.095 |
| Threonine | THR | 0.034 | 2.824 | 0.023 | 1.005 | -0.06 | -0.828 |
| Tryptophan | TRP | 1.627 | 0.504 | 0.004 | 1.8 | -2.473 | -0.303 |
| Tyrosine | TYR | -1.755 | 2.914 | 1.734 | 2.768 | 2.027 | -0.494 |
| Valine | VAL | -0.223 | 2.558 | -0.413 | 1.235 | 1.225 | 0.002 |

aa,bb,cc,dd,ee,ff QuBiLS-MIDAS scales.

# 3. Group-based Macromolecular Vectors

Given that the modeling endpoint for a set of proteins may depend not only on the whole sequence but also on certain subsets of *aas*,18 we present the group-based macromolecular vectors. In this approach, the *aas* are grouped in terms of their activity/properties on solution or their relative frequency of appearance in a certain secondary structure or motif. The groups included in **MD-LAIs** are **chemical-structural** (10 groups, see Table **S2**), **R-group** (20 groups, one per natural *aa*) and ***k*-mer** (400 *2*-mers, and 8000 *3*-mers). The entries of the Group-based Macromolecular Vectors are defined as (Figure **S1** andFigure **S2**):

|  |  |
| --- | --- |
|  | (3) |
| **Figure S2.** Numerical example of seven group-based macromolecular vectors computed from the following groups: beta-sheet favoring (FBS), aliphatic (ALG), apolar (RAP), polar uncharged (RPU), polar positively charged (RPC), alpha-helix favoring (FAH) and unfolding (UFG).  **Table S2.** Amino acid composition of the chemical-structural groups.   |  |  | | --- | --- | | Group | Amino acids | | FAHa | ALA, CYS, LEU, MET, GLU, GLN, HIS, LYS. | | FBSb | VAL, ILE, PHE, TYR, TRP, THR. | | UFGc | GLY, PRO. | | AFTd | GLY, SER, ASP, ASN, PRO. | | ALGe | GLY, ALA, PRO, VAL, LEU, ILE, MET. | | AROf | PHE, TYR, TRP. | | RPCg | LYS, HIS, ARG. | | RNCh | ASP, GLU. | | RAPi | PRO, ILE, ALA, VAL, LEU, PHE, TRP, MET. | | RPUj | ASN, CYS, GLY, SER, THR, TYR, GLN. | | aAlpha helix favoring amino acids; b Beta-sheets favoring amino acids; c Unfolding amino acids; dBeta-turn favoring amino acids; eAliphatic; fAromatic; gPolar positively charged; hPolar negatively charged; eApolar; jPolar uncharged. | | |  |

# 4. Fuzzy-Weighted Macromolecular Vectors

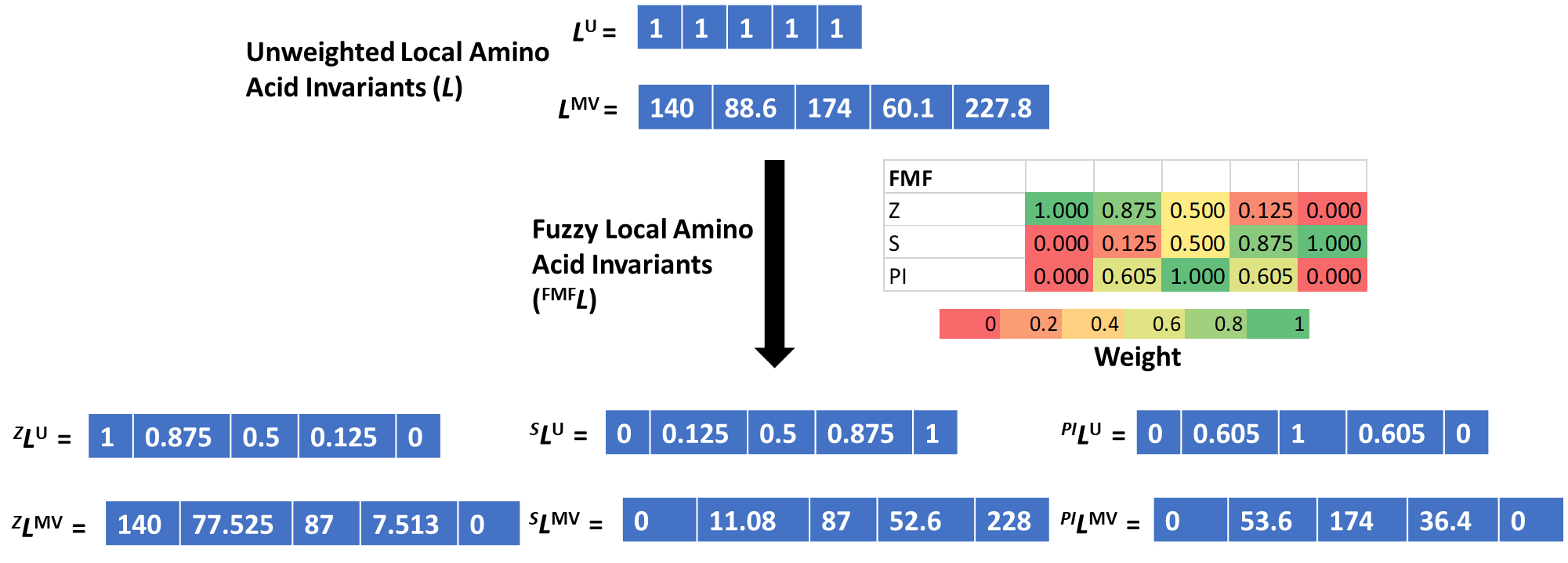
Let *A* be a *fuzzy set* on a universe of discourse *X* , then the *A* fuzzy set is characterized by a *fuzzy membership function* (FMF), ,which maps each element to a value in the interval [0,1]. This value represents the “membership degree” of *x* in *A*.[27](#_ENREF_27)

Let *p* be a reference point in a protein (or peptide) sequence (*i.e.,* the *N*-terminal, center, or *C*-terminal), then the fuzzy weight for *aa* *i* is obtained as follows:

(**4**)

where, is the topological distance of *aa* *i* (index *i* is always taken from *N*-terminal to *C*-terminal) to the point *p*. From now, *A* is a fuzzy set on the universe of discourse *X*, so that *X* is defined on the interval [0, *R*], where *R* is the maximum distance of any *aa* to the point *p*. The value of *R* depends on the reference point, that is, if *p* is the *N*-terminal or the *C*-terminal, then *R* is equal to *n-1*, whereas if *p* is the center, then the value of *R* is equal *n/2*, being *n* the number of *aa*s in the protein, and . Thus, is a FMF for the fuzzy set *A*. As the proteins will have different values of *R*, then the parameters and will be determined from percent values, from now indicated as and . So, and .

The FMFs listed on Table **S3** allow weighting aa contributions based on the proximity of *aas* with respect to a lower boundary , to an upper boundary , or to a middle region of the interval defined for the *A* fuzzy set. Thus, depending on the chosen FMF, each *aa* will have different weights (see Figure **S3**).



**Figure S3.** Numerical example of the calculation of the Macromolecular Vectors weighted with the FMFs: Z (gives more weight to *aas* close o *N*-terminal), S (gives more weight to *aas* near to the *C*-terminal) and PI (gives high weight to *aas* close to the center).

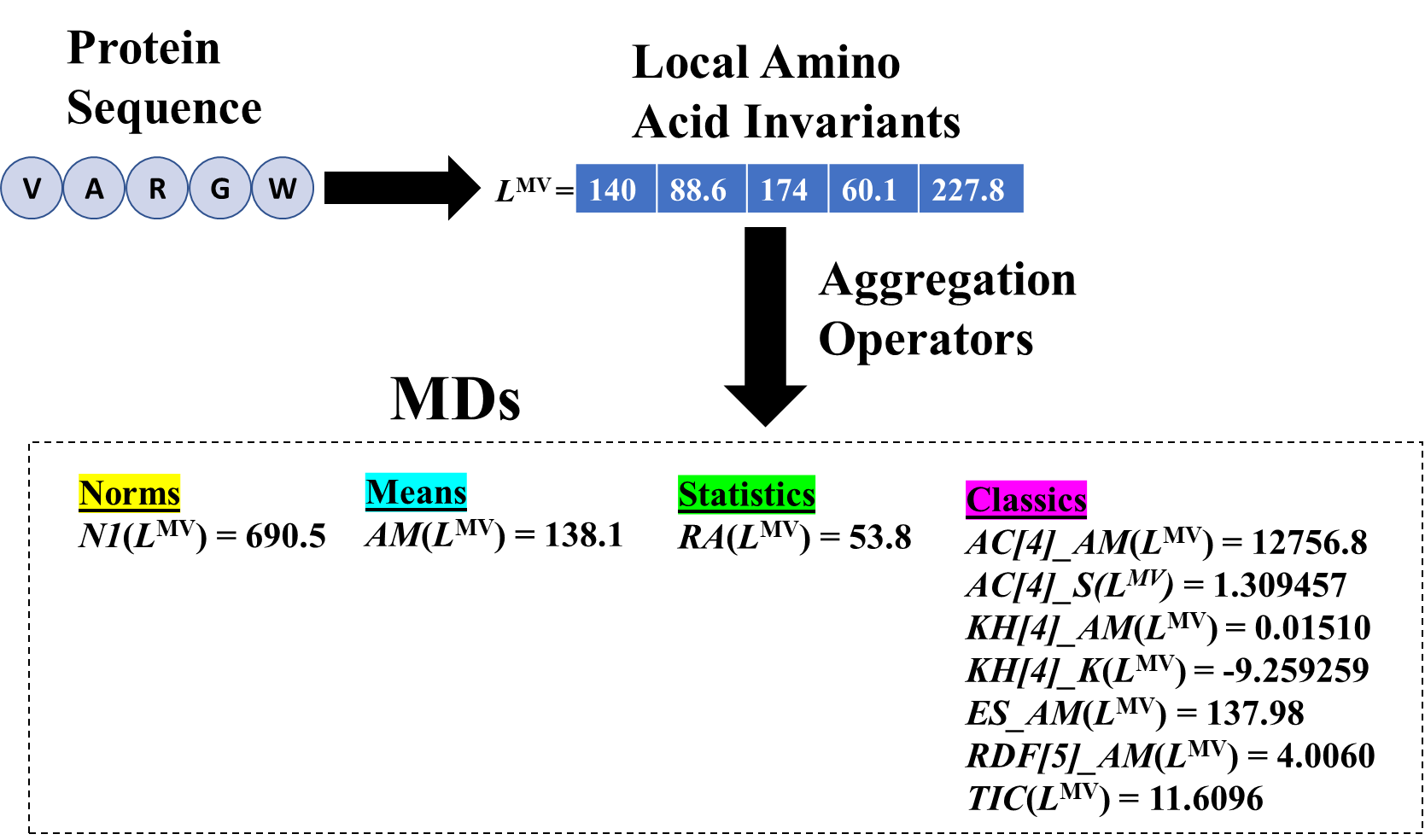
# 5. Aggregation Operators for MDs Generalization.

The additive (sum) approach for fusing chemical data has been widely used in the literature.19 However, it has been shown that other operators yielded better outcomes than those obtained from the sum.3,20 These operators also known as invariants, hereafter aggregation operators (**AOs**), constitute the basis of **MD-LAIs**. In our framework, the **AOs** are classified into four major groups: **a)** **Norms (or Metrics) Invariants** (Table **S4**)**:** Minkowski norms (**N1, N2, N3**).Note that **N1** is equal to the sum of the components of vector . **b)** **Mean Invariants (first statistical moment,** Table **S4**)**:** Geometric mean (**GM**), arithmetic mean (**AM**), quadratic mean (**P2**), power mean of third degree (**P3**) and harmonic mean (**HM**). **c)** **Statistical Invariants (highest statistical moments,** Table **S5):** 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 (**MX**) and minimum (**MN**). **d)** **Classical Invariants** (Table **S6**)**:** Autocorrelation (**AC**), Gravitational (**GV**), Total Sum (**TS**), Kier-Hall Connectivity (**KH**), Electro-topological State (**ES**), Ivanciuc-Balaban (**IB**), Radial Distribution Function (**RDF**), Morse (**MSE**), Inter-amino acid Interaction Spectrum (**IS**), Total Information Content (**TIC**), Mean Information Content (**MIC**), Standardized Information Content (**SIC**), N-binned Total Information Content (**TICN**), N-binned Mean Information Content (**MICN**), N-binned Standardized Information Content (**SICN**), Entropy (**H**), Geary Coefficient (**GC**), Potential of a Charge Distribution (**PCD**), Connective Eccentricity index (**CEI**), Beteringhe–Filip–Tarko (**BFT**) and Amphiplilic/Amphiphatic moments (**APM**).

It is important to remark that, in general, the classical **AOs** are applied coupled with non-classic **AOs**. In this way, information regarding the structural neighborhood of each *aa* is encoded before applying the non-classic **AO**. For instance, the Electro-topological **AO** considers the whole protein sequence in the calculation of the LAI of each *aa*, whereas the Kier-Hall **AO** considers subgraphs of type path for each *aa*. Finally, the application of **AOs** to the LAIs vector enables us to obtain sets of MDs that globally or locally characterize a protein or peptide (see Figures **S1** and **S4**).

**Table S3.** Fuzzy membership functions (FMFs) for weighting the amino acids contributions based on their positions in the sequence.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Function** | **Definition** | **Parameters** | |
| SHIFTING | Shifting1c |  | *x*=*dip* | |
| SWITCHING | Switchingc,m |  | *x*=*dip* | |
| S | S-shapeds  Parameter *a* defines the foot of the function, *b* defines its shoulder. |  | *x*=*dip*  *a*=*don*  *b*=*doff* | |
| Z | Z-shapedc  Parameter *a* defines the shoulder of the function, *b* defines its foot. |  | *x*=*dip*  *a*=*don*  *b*=*doff* | |
| PI | PI-shapedm  Parameters *a* and *d* define the feet of the function, *b* and *c* define its shoulders. |  | *x*=*dip*  *a*=*don*  *b*=(*don +doff*)×0.45  *c*=(*don +doff*)×0.55  *d*=*doff* | |
| GAUSSIAN | Gaussian-based  Parameter *a* is a measure of the width of the curve, and *c* defines the center of the curve. |  | *a*=(*doff -don*) ×0.5  *c*=*doff* | |
| **Note:** The superscripts *c*,*m* and *s* denote that the FMF gives more importance to amino acids close to the lower boundary (*e.g.,* *N*-terminal), middle region (*e.g.*, center) and upper boundary (e.g., *C*-terminal) of the analyzed interval, respectively. The parameter *d*ip is the distance of amino acid *i* to the point *p*. | | | |



**Figure S4.** Application of Aggregation Operators (**AOs**) to transform the LAI vector into a single parameter (MD). There are several classes of **AO**s, see Tables **S4**, **S5** and **S6**.

**Table S4.** Definition for Norms and Mean Aggregation Operators.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Group** | **Name** | **ID** | **Formula** |
| 1 | Norms (Metrics) | Minkowski norm (p = 1)  Manhattan norm | N1 | |  | | --- | |  | | |
| 2 | Minkowski norm (p = 2)  Euclidean norm | N2 | |  | | --- | |  | | |
| 3 | Minkowski norm (p = 3) | N3 | |  | | --- | |  | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 4 | Mean  (first  statistical moment) | Geometric Mean | GM |  |
| 5 |  | Arithmetic Mean  (Power mean of degree β = 1) | AM |  |
| 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) | HM |  |

**Table S5.** Definition for Statistical Aggregation Operators.

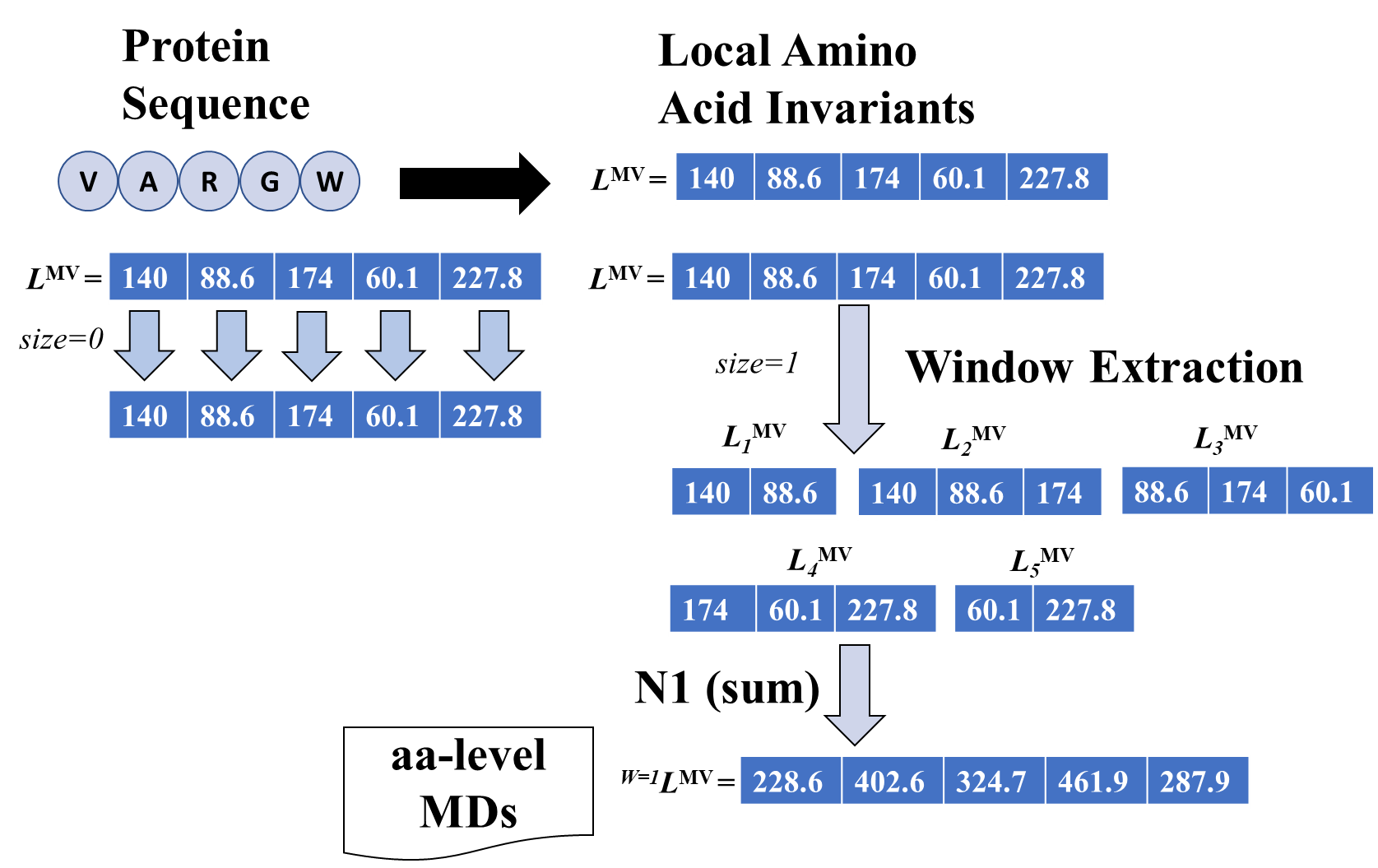
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Group** | **Name** | **ID** | **Formula** |
| 9 | Statistical  (*Highest statistical moments*) | Skewness | S | |  | | --- | |  | | |
| 10 |  | Variance | V |  |
| 11 |  | Kurtosis | |  | | --- | | K | | |  |
| 12 |  | Standard Deviation | SD |  |
| 13 |  | Variation Coefficient | VC |  |
| 14 |  | Range | R | - *L*min |
| 15 |  | Percentile 25 | Q1 |  |
| 16 |  | Percentile 50 | Q2 |  |
| 17 |  | Percentile 75 | Q3 |  |
| 18 |  | Inter-quartile Range | I50 |  |
| 19 |  | Maximum value | MX | MX = *L* max |
| 20 |  | Minimum value | MN | MN = *L*min |

**Table S6.** Definition for *classical* aggregation operators.

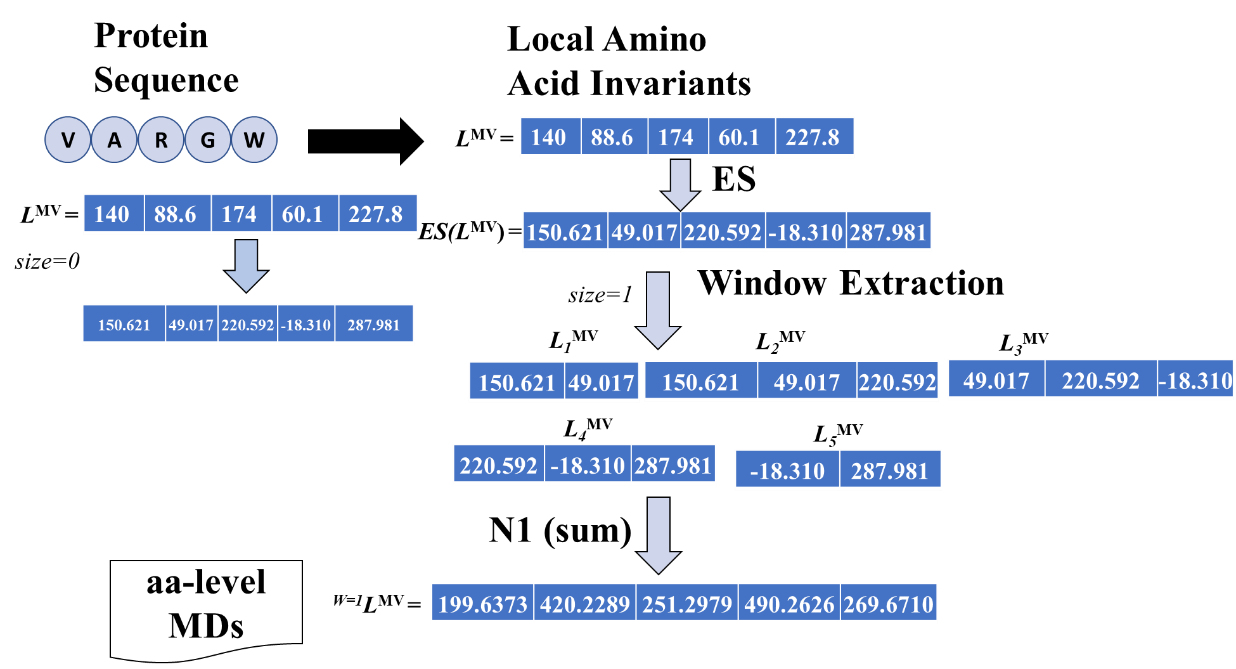
|  |  |  |
| --- | --- | --- |
| **Name** | **ID** | **Formula** |
| Autocorrelation | AC*k* | *where, dij is the topological distance between amino acids i and j and k is the cutoff distance.* |
| Gravitational | GV*k* |  |
| Total sum (lag k) | TS*k* |  |
| Kier-Hall Connectivity | KHk | *where, K is the number of path sub-graphs, nk is the number of amino acids in a fragment, λ is equal to ½, and k is the order of path subgraph (number of amino acids in the path considered). Here, k =1, 1,…,7* |
| Mean Information Content | MIC | *where, Ng is the number of amino acids with the same LAI value. No is the number of amino acids in a protein.* |
| Total Information Content | TIC |  |
| Standardized Information Content | SIC |  |
| Mean Information Content (N-binned) | MICN | *where, Ng is the number of amino acids that fall into different intervals (bins), No is the number of amino acids in a protein* |
| Total Information Content (N-binned) | TICN |  |
| Standardized Information Content (N-binned) | SICN |  |
| Entropy | H | , ,  *where, and are the arithmetic mean and standard deviation of LAIs vector (L), respectively.* |
| Electrotopological state (E-state index) | ES | *where, Ii is the intrinsic state of the ith amino acid, ΔIi is the field effect on the ith amino acid calculated as perturbation of the Ii of ith amino acid by all other amino acids in the protein, dij is the topological distance between the ith and the jth amino acids, and n is the number of amino acids.* |
| Ivanciuc-Balaban | IB | *where, the summation goes over all pairs of amino acids, but only pairs of adjacent amino acids are accounted for by means of the elements aij of the adjacency matrix. The n, B, and C are the number of amino acids, amide bonds, and rings (cyclomatic number), respectively.* |
| Geary Coefficient | GCk | *where is the LAI value for amino acid i, is its average value on the protein, n is the number of amino acids, k is the lag considered, dij is the topological distance between amino acid i and j, and is the Kronecker delta equal to 1 if , zero otherwise. is the number of vertex pairs at distance equal to k.* |
| Potential of a Charge Distribution | PCD | *where, is the LAI value for amino acid i, di is the distance from each amino acid to the center (half of sequence) of sequence.* |
| Connective Eccentricity index | CEI | *where, is the topological eccentricity of amino acid i, that is, the largest topological distance from amino acid i to n-1 amino acids.* |
| Radial Distribution Function | RDF (R) | *where, f is a scaling factor (here 1/n), Li characteristic amino acidic LAIs of the amino acids i and j, dij the topological distance between amino acid i and j, and n the number of amino acids. is a smoothing parameter (here 100 that defines the probability distribution of the individual interatomic distances; can be interpreted as a temperature factor that defines the movement of amino acids. Here, seven radiuses are employed R=(n/8,n/7,n/6,n/5,n/4,n/3 and n/2).* |
| Beteringhe–Filip–Tarko | BFT | *where, n is the number of amino acids and B the number of amide bonds.* |
| MoRSE | MSE | *where, s is the scattering parameter, dij is the topological distance between amino acid i and j.* |
| Inter-amino acid Interaction Spectrum | IS(R) | *dij is the topological distance between amino acid i and j. R is the radius. Here, seven radiuses are employed: R=(n/8,n/7,n/6,n/5,n/4,n/3 and n/2).* |
| Amphiplilic/Amphiphatic moments | APM | *where, is the topological distance of amino acid i to the farthest hydrophobic/hydrophilic/amphipatic amino acid.* |

# 6. Window-based Generalized Amino Acid-Level MDs by Applying AOs.

The availability of *aa*-level MDs is mandatory for several applications such as post-translational modification (PTM) prediction,21 catalytic residues recognition22 among others.23 However, this type of output is uncommon in current calculation software. To the best of our knowledge, only ProtDCal,24 BioSeq-analysis25 and Pfeature26 offer *aa*-level MDs. In this sense, we generalize the calculation of *aa*-level MDs by considering a sliding window centered in the target aa and running *k* steps to the left and to the right, where . When *k*=0, the *aa*-level MD is equal to the original LAI vector (Scheme **S2**). For *k*>0, each *aa*-level MD of the LAI vector is obtained by applying an **AO** to the sub-vector defined by the window, thus including information about the sequence-neighboring *aas*.

**A)**

**B)**



**Scheme S2.** Illustration of the calculation of *aa*-level MDs. **A**) based on a non-classic **AO** (N1), **B**) based on a classic operator (ES) coupled with the sum (N1).

# 7. References

(1) Todeschini, R.; Consonni, V. New Local Vertex Invariants and Molecular Descriptors Based on Functions of the Vertex Degrees. *MATCH - Communications in Mathematical and in Computer Chemistry* **2010**, *64*, 359–372. https://doi.org/10.1016/j.renene.2014.11.073.

(2) 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.

(3) 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.

(4) 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.

(5) 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.

(6) 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. *Bioorg Med Chem* **2006**, *14* (7), 2398–2408. https://doi.org/10.1016/j.bmc.2005.11.024.

(7) 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.

(8) 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.

(9) 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 Journal* **2010**, *277* (15), 3118–3146. https://doi.org/10.1111/j.1742-4658.2010.07711.x.

(10) 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.

(11) 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.

(12) 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.

(13) 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.

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

(15) Mei, H.; Liao, Z. H.; Zhou, Y.; Li, S. Z. A New Set of Amino Acid Descriptors and Its Application in Peptide QSARs. *Biopolymers - Peptide Science Section* **2005**, *80* (6). https://doi.org/10.1002/bip.20296.

(16) Martínez-López, Y.; Marrero-Ponce, Y.; Barigye, S. J.; Teran, E.; Martínez-Santiago, O.; Zambrano, C. H.; Torres, F. J. When Global and Local Molecular Descriptors Are More than the Sum of Its Parts: Simple, But Not Simpler? *Mol Divers* **2020**, *24* (4). https://doi.org/10.1007/s11030-019-10002-3.

(17) García-Jacas, C. R.; 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). https://doi.org/10.1002/jcc.23640.

(18) Liang, Y.; Yang, S.; Zheng, L.; Wang, H.; Zhou, J.; Huang, S.; Yang, L.; Zuo, Y. Research Progress of Reduced Amino Acid Alphabets in Protein Analysis and Prediction. *Computational and Structural Biotechnology Journal*. 2022. https://doi.org/10.1016/j.csbj.2022.07.001.

(19) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; 2010; Vol. 2. https://doi.org/10.1002/9783527628766.

(20) 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. *Current Computer Aided-Drug Design* **2013**, *9* (2), 164–183. https://doi.org/10.2174/1573409911309020003.

(21) Ruiz-Blanco, Y. B.; Marrero-Ponce, Y.; García-Hernández, E.; Green, J. Novel “Extended Sequons” of Human N-Glycosylation Sites Improve the Precision of Qualitative Predictions: An Alignment-Free Study of Pattern Recognition Using ProtDCal Protein Features. *Amino Acids* **2017**, *49* (2). https://doi.org/10.1007/s00726-016-2362-5.

(22) Zhang, T.; Zhang, H.; Chen, K.; Shen, S.; Ruan, J.; Kurgan, L. Accurate Sequence-Based Prediction of Catalytic Residues. *Bioinformatics* **2008**, *24* (20), 2329–2338. https://doi.org/10.1093/bioinformatics/btn433.

(23) Zhao, B.; Kurgan, L. Databases of Protein Structure and Function Predictions at the Amino Acid Level. In *Machine Learning in Bioinformatics of Protein Sequences*; 2023. https://doi.org/10.1142/9789811258589\_0013.

(24) Ruiz-Blanco, Y. B.; Paz, W.; Green, J.; Marrero-Ponce, Y. ProtDCal: A Program to Compute General-Purpose-Numerical Descriptors for Sequences and 3D-Structures of Proteins. *BMC Bioinformatics* **2015**, *16* (1). https://doi.org/10.1186/s12859-015-0586-0.

(25) Liu, B. BioSeq-Analysis: A Platform for DNA, RNA and Protein Sequence Analysis Based on Machine Learning Approaches. *Brief Bioinform* **2018**, *20* (4). https://doi.org/10.1093/bib/bbx165.

(26) Pande, A.; Patiyal, S.; Lathwal, A.; Arora, C.; Kaur, D.; Dhall, A.; Mishra, G.; Kaur, H.; Sharma, N.; Jain, S.; Usmani, S. S.; Agrawal, P.; Kumar, R.; Kumar, V.; Raghava, G. P. S. Pfeature: A Tool for Computing Wide Range of Protein Features and Building Prediction Models. *Journal of Computational Biology* **2023**, *30* (2). https://doi.org/10.1089/cmb.2022.0241.