

Description of features used for training

Glossary:

kDa - kilodaltons

Å² - Angstrom squared

Å³ - Angstrom cubed

DSSP – Dictionary of Secondary Structure of Proteins

‘O(I)>H-N(I±n)_H_bonds’ type descriptors explained:

Alpha helices are part of a protein’s secondary structure. They form when the backbone of the amino acid chain arranges into a helical structure with hydrogen bonds stabilising it.

The H-bond happens between the carbonyl oxygen from amino acid i and the N-H group of amino acid $i\pm n$. These bonds can happen between amino acids up to 5 positions away from each other. So ‘O(I)>H-N(I+2)_H_bonds’ for example represents the number of H-bonds that happen between the i -th amino acid oxygen and the $i+2$ amino acid N-H group.

Notes on some known mistakes and issues:

1. The ‘%polar’ descriptor in the training data is labelled wrong, it should be named ‘%apolar’, and it contains the percentage of apolar, not polar, amino acids in the protein.
2. For some reason the descriptors:
 - O(I)>H-N(I+0)_H_bonds
 - O(I)>H-N(I+0)_H_bonds%
 - O(I)>H-N(I+1)_H_bonds
 - O(I)>H-N(I+1)_H_bonds%are missing from the training data.
3. If a description is marked as ‘-’, it is because a description was not found for all the molecular descriptors calculated by Mordred and external information could not be found.

| Feature name | Description |
|--------------|---|
| prot_MW | Protein molecular weight (from sequence). In kDa |
| tot_aa_SASA | Total amino acid solvent accessible surface area (SASA), calculated by adding up individual |

| Feature name | Description |
|----------------------|---|
| | amino acid SASA. In Å ² |
| tot_VWvol | Total Van der Waals volume (VWV), calculated by adding up amino acid VWV. In Å ³ |
| %A | Percentage of alanine in the protein |
| %R | Percentage of arginine in the protein |
| %N | Percentage of asparagine in the protein |
| %D | Percentage of aspartic acid in the protein |
| %C | Percentage of cysteine in the protein |
| %E | Percentage of glutamic acid in the protein |
| %Q | Percentage of glutamine in the protein |
| %G | Percentage of glycine in the protein |
| %H | Percentage of histidine in the protein |
| %I | Percentage of isoleucine in the protein |
| %L | Percentage of leucine in the protein |
| %K | Percentage of lysine in the protein |
| %M | Percentage of methionine in the protein |
| %F | Percentage of phenylalanine in the protein |
| %P | Percentage of proline in the protein |
| %S | Percentage of serine in the protein |
| %T | Percentage of threonine in the protein |
| %W | Percentage of tryptophan in the protein |
| %Y | Percentage of tyrosine in the protein |
| %V | Percentage of valine in the protein |
| %polar | Percentage of <u>apolar</u> amino acids in the protein |
| %pol_unch | Percentage of polar uncharged amino acids in the protein |
| %pol_ch | Percentage of polar charged amino acids in the protein |
| %pol_ch_basic | Percentage of polar charged basic amino acids in the protein |
| %pol_ch_acid | Percentage of polar charged acidic amino acids in the protein |
| %aromatic | Percentage of aromatic amino acids in the protein |
| %aliphatic | Percentage of aliphatic amino acids in the protein |

| Feature name | Description |
|--------------------------|---|
| %hydrox | Percentage of hydroxylic amino acids in the protein |
| %amidic | Percentage of amidic amino acids in the protein |
| %w_S | Percentage of amino acids with a Sulfur atom in the protein |
| tot_aa_TPSA | Total topological polar surface area (TPSA) of all amino acids in the protein. In Å ² |
| apolar_SA | Apolar surface area of amino acids in the protein. In Å ² |
| pol_unch_SA | Polar uncharged surface area of amino acids in the protein. In Å ² |
| pol_ch_SA | Polar charged surface area of amino acids in the protein. In Å ² |
| pol_ch_basic_SA | Polar charged basic surface area of amino acids in the protein. In Å ² |
| pol_ch_acid_SA | Polar charged acidic surface area of amino acids in the protein. In Å ² |
| tot_SASA/actual_asa | SASA calculated from amino acids. In Å ² (tot_aa_SASA) divided by the actual SASA of the protein (from DSSP). In Å ² |
| length | Length of the protein sequence (all chains added) |
| n_chains | Number of individual chains in the protein |
| aa/chain | Average amino acids per chain |
| tot_SS_bri | Total S-S bridges |
| intrachain_SS_bri | Intrachain S-S bridges |
| interchain_SS_bri | Interchain S-S bridges |
| asa | *From here onwards, until stated otherwise, all data is from DSSP.* Accessible surface area of the protein . In Å ² |
| tot_O(I)>H-N(J)_H_bonds | Total H bonds of type O(I)>H-N(J) |
| O(I)>H-N(J)_H_bonds% | Percentage of H bonds of type O(I)>H-N(J) |
| parallel_bri_H_bonds | Total H bonds in parallel bridges |
| parallel_bri_H_bonds% | Percentage of H bonds in parallel bridges |
| antiparallel_bri_H_bonds | Total H bonds in antiparallel bridges |

| Feature name | Description |
|----------------------------------|--|
| antiparallel_bri_H_bonds% | Percentage of H bonds in antiparallel bridges |
| O(I)>H-N(I-5)_H_bonds | Total H bonds of type O(I)>H-N(I-5) |
| O(I)>H-N(I-5)_H_bonds% | Percentage of H bonds of type O(I)>H-N(I-5) |
| O(I)>H-N(I-4)_H_bonds | Total H bonds of type O(I)>H-N(I-4) |
| O(I)>H-N(I-4)_H_bonds% | Percentage of H bonds of type O(I)>H-N(I-4) |
| O(I)>H-N(I-3)_H_bonds | Total H bonds of type O(I)>H-N(I-3) |
| O(I)>H-N(I-3)_H_bonds% | Percentage of H bonds of type O(I)>H-N(I-3) |
| O(I)>H-N(I-2)_H_bonds | Total H bonds of type O(I)>H-N(I-2) |
| O(I)>H-N(I-2)_H_bonds% | Percentage of H bonds of type O(I)>H-N(I-2) |
| O(I)>H-N(I-1)_H_bonds | Total H bonds of type O(I)>H-N(I-1) |
| O(I)>H-N(I-1)_H_bonds% | Percentage of H bonds of type O(I)>H-N(I-1) |
| O(I)>H-N(I+2)_H_bonds | Total H bonds of type O(I)>H-N(I+2) |
| O(I)>H-N(I+2)_H_bonds% | Percentage of H bonds of type O(I)>H-N(I+2) |
| O(I)>H-N(I+3)_H_bonds | Total H bonds of type O(I)>H-N(I+3) |
| O(I)>H-N(I+3)_H_bonds% | Percentage of H bonds of type O(I)>H-N(I+3) |
| O(I)>H-N(I+4)_H_bonds | Total H bonds of type O(I)>H-N(I+4) |
| O(I)>H-N(I+4)_H_bonds% | Percentage of H bonds of type O(I)>H-N(I+4) |
| O(I)>H-N(I+5)_H_bonds | Total H bonds of type O(I)>H-N(I+5) |
| O(I)>H-N(I+5)_H_bonds% | Percentage of H bonds of type O(I)>H-N(I+5) |
| | End of DSSP data. |
| nAcid | *From here onwards, until stated otherwise, the descriptors pertain to the ligand.* Number of acidic groups in the ligand |
| nBase | Number of basic groups |
| nAromAtom | Number of atoms in an aromatic structure |
| nAromBond | Number of bonds in an aromatic structure |
| nAtom | Total number of atoms in the ligand |
| nHeavyAtom | Number of heavy atoms in the ligand |
| nSpiro | - |
| nBridgehead | - |
| nHetero | - |
| nH | Number of hydrogen atoms in the ligand |
| nB | Number of boron atoms in the ligand |
| nC | Number of carbon atoms in the ligand |

| Feature name | Description |
|-----------------|--|
| nN | Number of nitrogen atoms in the ligand |
| nO | Number of oxygen atoms in the ligand |
| nS | Number of sulphur atoms in the ligand |
| nP | Number of phosphorus atoms in the ligand |
| nF | Number of fluorine atoms in the ligand |
| nCl | Number of chlorine atoms in the ligand |
| nBr | Number of bromine atoms in the ligand |
| nI | Number of iodine atoms in the ligand |
| nX | Number of other atoms in the ligand (?) |
| BalabanJ | *topological indexes* Balaban J connectivity topological index |
| BertzCT | Bertz topological index meant to quantify "complexity" |
| nBonds | *bonds descriptors* Total number of bonds in the ligand |
| nBondsO | - |
| nBondsS | - |
| nBondsD | - |
| nBondsT | - |
| nBondsA | - |
| nBondsM | - |
| nBondsKS | - |
| nBondsKD | - |
| PNSA1 | *Charged Partial Surface Area (CPSA) descriptors* Partial negative surface areas |
| PNSA2 | |
| PNSA3 | |
| PNSA4 | |
| PNSA5 | |
| PPSA1 | Partial positive surface areas |
| PPSA2 | |
| PPSA3 | |
| PPSA4 | |

| Feature name | Description |
|--------------|---|
| PPSA5 | |
| DPSA1 | Differences in charged partial surface areas |
| DPSA2 | |
| DPSA3 | |
| DPSA4 | |
| DPSA5 | |
| FNSA1 | Fractional charged partial negative surface areas |
| FNSA2 | |
| FNSA3 | |
| FNSA4 | |
| FNSA5 | |
| FPSA1 | Fractional charged partial negative surface areas |
| FPSA2 | |
| FPSA3 | |
| FPSA4 | |
| FPSA5 | |
| WNSA1 | Surface weighted charged partial negative surface areas |
| WNSA2 | |
| WNSA3 | |
| WNSA4 | |
| WNSA5 | |
| WPSA1 | Surface weighted charged partial negative surface areas |
| WPSA2 | |
| WPSA3 | |
| WPSA4 | |
| WPSA5 | |
| RNCG | Relative negative charge ... |
| RPCG | Relative positive charge ... |
| RNCS | Relative negative charge surface area |
| RPCS | Relative positive charge surface area |
| TASA | Total hydrophobic surface area |
| TPSA | Total polar surface area |

| Feature name | Description |
|---------------------------|--|
| RASA | Relative hydrophobic surface area |
| RPSA | Relative polar surface area |
| | *end of CPSA descriptors* |
| ECIndex | - |
| fragCpx | Fragment complexity score |
| GeomDiameter | - |
| GeomRadius | - |
| GeomShapeIndex | - |
| GeomPetitjeanIndex | - |
| GRAV | - |
| GRAVH | - |
| GRAVp | - |
| GRAVHp | - |
| nHBAcc | Number of hydrogen bond acceptors |
| nHBDon | Number of hydrogen bond donors |
| Lipinski | Lipinski rule of 5. Checks whether ligand is drug-like or not |
| GhoseFilter | Ghose Filter. Checks whether ligand is drug-like |
| VMcGowan | McGowan volume |
| apol | Sum of the atomic polarizabilities (including implicit hydrogens) with polarizabilities |
| bpol | Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens) with polarizabilities |
| nRot | Number of rotatable bonds |
| RotRatio | - |
| SLogP | Log of the octanol/water partition coefficient (including implicit hydrogens) |
| SMR | Molecular refractivity. This property is an atomic contribution model [Crippen 1999] that assumes the correct protonation state (washed structures) |
| TopoPSA(NO) | - |
| TopoPSA | Topological polar surface area |

| Feature name | Description |
|----------------|--|
| Diameter | Largest value in the distance matrix of the ligand |
| Radius | Diameter/2 (?) |
| TopoShapeIndex | Topological shape index |
| PetitjeanIndex | Petit jean index |
| Vabc | - |
| MW | Molecular weight of the ligand |
| AMW | - |
| WPath | Wiener path number |
| WPol | Wiener polarity number |