**Finding needles in a haystack: determining key molecular descriptors associated with the blood-brain barrier entry of chemical compounds using machine learning**

Subhabrata Majumdara\*, Subhash C. Basakb, Claudiu N. Lunguc, Mircea V. Diudeac and Gregory D. Grunwaldd

aUniversity of Florida Informatics Institute, 432 Newell Dr, CISE Bldg E251, Gainesville FL 32611, USA;   
bDepartment of Chemistry and Biochemistry, University of Minnesota, 246 Chemistry Building, 1039 University Drive, Duluth MN 55812, USA;  
cDepartment of Chemistry, Babes-Bolyai University, Strada Arany János 11, Cluj-Napoca 400028, Romania;  
dNatural Resources Research Institute, University of Minnesota, 5013 Miller Trunk Highway, Duluth MN 55811, USA.

\*Currently at AT&T Labs Research. Correspondence e-mail: [subho@research.att.com](mailto:subho@research.att.com)

**Supplementary Table 4.** Symbols, definitions and classification of structural molecular descriptors

|  |  |
| --- | --- |
| *Potential energy descriptors (PS)* | |
| Enb | Value of the potential energy with all bonded terms disabled |
| Estb | Bond stretch-bend cross-term potential energy |
| *Topochemical (TC)* | |
| GCUTPEOE2 | Eigenvalues of a modified graph distance adjacency matrix. Each *ij* entry of the adjacency matrix takes the value 1/sqr(*dij*) where *dij* is the (modified) graph distance between atoms *i* and *j*. The diagonal takes the value of the PEOE partial charges. |
| oprnring | The number of ring bonds |
| oprnrot | The number of rotatable bonds |
| oprviolation | The number of violations of Oprea's lead-like test |
| radius | ri is the largest matrix entry in row i of the distance matrix D, radius is defined as the smallest of the ri |
| *Physicochemical combined descriptors (PS)* | |
| SlogPVSA9 | Sum of *vi* such that *Li* > 0.40 |
| vsurfHL1 | Hydrophilic-Lipophilic surface |
| vsurfIW6 | Hydrophilic integy moment |
| logS | Log of the aqueous solubility (mol/L). This property is calculated from an atom contribution linear atom type model , with r2 = 0.90, ~1,200 molecules. |
| GCUTSLOGP0 | The GCUT descriptors using atomic contribution to logP (using the Wildman and Crippen SlogP method) instead of partial charge. |
| *Geometrical combined (3D)* | |
| vsurfCW5 | Capacity factor |
| vsurfDD13 | Contact distances of vsurf\_DDmin |
| SMRVSA6 | Sum of vi such that Ri is in (0.485,0.56] |