**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 2: Brief description of the formulae and software used for the calculation of molecular descriptors.**

1. **Cluj Topological Descriptors**

Given below are the formulae and description of Cluj topological descriptors

**1*.* Cluj matrices** CJ(*G*) have been proposed by Diudea [1,2]; they are defined on Cluj fragmentswhich collect vertices *v* lying closer to *i* than to *j*, the endpoints of a path *p*(*i*,*j*). These fragments represent the vertex proximity (see also [3,4]) of *i* vs. any vertex *j*, joined by the path *p*, with the distances measured in the subgraph *G*\*p*:



The entries in CJ matrices are defined as the maximum cardinality among all such fragments: , in graphs containing rings, with more than one path joining the pair (*i, j*). Cluj matrices are defined by using both *distance* (UCJDI) and *detour* (UCJDE) concepts: they are non-symmetric matrices, except for some symmetric graphs.

A symmetric Cluj matrix SCJ can be obtained by the Hadamard multiplication of the non-symmetric matrix UCJ with its transpose:.The edge-based symmetric matrix is calculable from the path-based one as: . In trees, the path joining any two vertices is unique, then  is also unique while the distances superpose to detours. The entries in the Cluj matrix of a tree represent the number of paths passing through i to j. Indices derived from topological matrices (see above) are calculated as half sum of matrix entries:



Fragmental property indices [5,6], encoding fragmental topo-chemical information and based on Cluj and related matrices, can be computed by means of layer matrices. The fragmental properties include mass, partial charges, covalent atomic radius, van-der-Waals radius, etc.

2. Layer and Shell matrices [7-9] are built on layers *G*(*v*)*k* of vertices surrounding the vertex *v* at distance *k*: . A *Layer Matrix* LM is a collection of vertex property shells, up to *k* =*d*(*G*), the diameter of graph (*i.e*., the largest distance in *G*):

****

The zero column is just the column of vertex properties, *pv*;when *pv*=1 (i.e., the counting property), LMmatrix is named LC (Layer matrix of Counting). Any vertex/atomic property can be considered as *pv* and any square matrix M can be taken as *info matrix* (*i.e.*, the matrix supplying local/vertex properties, as row-sum *RS*, column-sum *CS* or diagonal entries, as implemented in TopoCluj software [10].

*Shell* *Matrices* ShM, collecting a vertex-pair property, [9,10], are defined as:

****

The “shell-operator” transforms a square matrix, according to the vertex-distance partitions in the graph, in a layer-type matrix. The zero column entries ****can be the diagonal entries of the info matrix. Column sums in LM-ShM represent the coefficients of Hosoya-Diudea property polynomials [10]; the 1st derivative (in x=1) of such polynomials provides informative topological descriptors.

**3.** **Indices of Centrality**

Indices of centrality [7-10] are calculated by formula:



This is a highly discriminating local index enabling to order the vertices of a molecular graph, in non-increasing order of centrality; the vertex centrality classes superpose over the equivalence classes found by permuting the entries in the adjacency matrix. A global centrality (eventually normalized to the number of vertices/atoms) can also be calculated.

*Centrocomplexity* index *X*(*LM*) [10] is calculated by formula:



The above index expresses the location with respect to a vertex of high complexity (*e.g*., a vertex of high degree or a heteroatom). A global index can be eventually weighted and/or normalized.

The above Centrality indices can be calculated on the layer/shell matrices built by the matrix operator [11] **W**(**M**1,**M**2,**M**3), defined by the principle of the *single endpoint characterization of a path*; its entries are defined by relation:



which combines the information of three square topological matrices; RS stands for the row sum in the mixed matrix.

The presence of a *heteroatom in molecules* is accounted by the “property fragmental indices” [5,6] and/or by means of partial charges [9], calculated, by TopoCluj software, as the interaction between the atom *i* and the remaining *j* atoms in the molecule:

****

## In the above, Si, Sj represent the Sanderson group electronegativities, calculated for the hydride groups (i.e., the heavy atoms with their surrounding hydrogen atoms) in the molecule. The log function provides the sign for the partial charge chij, viewed as a distance decreasing perturbation of the ith atom produced by the atom j (see the exponent, where dij is the Euclidean distance separating atoms i and j).

The *N*x*N* array collecting the entries *chij* is the charge matrix **CH**, whose row sums *chi* represent the total partial charge on hydride group/atom *i* in the molecule. It can be rearranged in a LM-format by applying the shell operator.

Supplementary **Table 4** gives a list of all Cluj descriptors used in this Study.

**B. Descriptors calculated by Basak’s team at the University of Minnesota Duluth**

Basak’s team at the University of Minnesota Duluth has used descriptors calculated by the following software:

1. S. C. Basak, D. K. Harriss and V. R. Magnuson, "POLLY v2.3," Copyright of the University of Minnesota, 1988.
2. MolconnZ v4.05, Quincy, MA: Hall Ass. Consult., 2003.
3. S. C. Basak, G. Grunwald and A. Balaban, "TRIPLET," Copyright of the Regents of the University of Minnesota, 1993.
4. J. Stewart, MOPAC Version 6.00, QCPE #455, Frank J. Seiler Research Laboratory: US Air Force Academy, CO, 1990.

Supplementary **Table 2** gives a list of all descriptors calculated by Basak group in this study.

**References**

1. M. V. Diudea; Cluj matrix invariants; J. Chem. Inf. Comput. Sci., 37 (2), 1997, 300-305

2. M. V. Diudea, B. Parv and I. Gutman; Detour-Cluj matrix and derived invariants; J. Chem. Inf. Comput. Sci., 37 (6), 1997, 1101-1108

3. M.V. Diudea, O.M. Minailiuc, G. Katona, I. Gutman, Szeged matrices and related numbers, MATCH Commun. Math. Comput. Chem. 1997, 35, 129-143. |41|;

4. P. V. Khadikar, S. Karmarkar, V. K. Agrawal, J. Singh, A. Shrivastava, I. Lukovits, M. V. Diudea, Szeged index - Applications for drug modeling, Letters in Drug Design & Discovery, 2005, 2 (8) 606-624. |46|

5. L. Jäntschi, G. Katona, M.V. Diudea, Modeling molecular properties by Cluj indices, MATCH Commun. Math. Comput. Chem. 2000, 41, 151-188. |46|

6. O.M. Minailiuc, G. Katona, M.V. Diudea, M. Strunje, A. Graovac, I. Gutman, Szeged fragmental indices, Croat. Chem. Acta, 1998, 71 (3), 473-488. |40|

7. M. V. Diudea; Molecular Topology. 16. Layer Matrices in Molecular Graphs; J. Chem. Inf. Comput. Sci., 34 (5), 1994, 1064-1071.

8. M.V. Diudea, M. Topan, A. Graovac, Molecular topology. 17. Layer matrixes of walk degrees, J. Chem. Inf. Comput. Sci. 1994, 34 (5), 1072-1078. |47|

9. M.V. Diudea, O. Ursu, Layer matrices and distance property descriptors, Indian J. Chem. A, 2003, 42 (6), 1283-1294. |52|

10. M. V. Diudea, Nanomolecules and Nanostructures - Polynomials and Indices, MCM, No. 10, Univ. Kragujevac, Serbia, 2010.

11. M. V. Diudea, Valences of property; Croat. Chem. Acta, 72 (4), 1999, 835-851.