A Safe-by-Design Tool for Functionalised Nanomaterials through the Enalos Nanoinformatics Cloud Platform

Dimitra-Danai Varsou, Antreas Afantitis, Andreas Tsoumanis, Georgia Melagraki*, Haralambos Sarimveis, Eugenia Valsami-Jones and Iseult Lynch*

Supplementary Information

Molecular descriptor details

- D133- Mean value of atomic composition index

  The atomic composition indices are molecular zero dimensional (0D) descriptors with high degeneracy, derived from the chemical formula of compounds and defined as information indices of the elemental composition of the molecule. They can be considered molecular complexity indices that take into account the molecular diversity in terms of different atom types.

  The mean information content on atomic composition is the mean value of the total information content, and is calculated as:

$$\bar{I}_{AC} = -\sum_g \frac{A_g}{A^h} \cdot \log_2 \frac{A_g}{A^h} = -\sum_g p_g \cdot \log_2 p_g \quad [1]$$

  where A^h is the total number of atoms (hydrogen included), A_g is the number of atoms of type g and p_g is the probability to randomly select a gth atom type.

- D173- Mohar order-2 index

  The first Mohar index ($TI1\_L$) and the second Mohar index ($TI2\_L$) are calculated from the eigenvalues of the Laplace matrix as follows:

$$TI1\_L = 2 \cdot \log\left(\frac{nBO}{nSK}\right) \cdot QW_L \quad [2]$$

$$TI2\_L = \frac{4}{nSK \cdot \lambda_{nSK-1}} \quad [3]$$

  where QW_L is the quasi-Wiener index, nBO and nSK are the number of non-H bonds and non-H atoms, respectively, and λ_{nSK-1} is the smallest nonzero eigenvalue of the Laplace matrix.

- D250- EXP5 of Path-distance/Walk-distance over all atoms

  The path-distance map matrix, denoted as PD, resembling the bond length-weighted distance matrix of a molecular graph, is defined as:

$$[PD]_{ij} = \min_{p_{ij}} \left([ED]_{kq}\right)_{ij} \quad [4]$$

  where $[ED]_{kq}$ denotes entries of the Euclidean-distance map matrix, p_{ij} is a path connecting vertices i and j, and the summation goes over all pairs of adjacent vertices along the considered path. Then, each entry of the path-distance map matrix is the shortest distance between two vertices measured along the path by summing the geometrical length of the edges connecting adjacent vertices along the path.

- D254- Radial centric index

Centric indices are molecular descriptors proposed to quantify the degree of compactness of molecules by distinguishing between molecular structures organized differently with respect to their centers. Based on the recognition of the graph center, these indices are mainly defined by information theory concepts applied to a partition of the graph vertices made according to their positions relative to the center.

The radial centric information index ($^{V}I_{C,R}$) is defined as

$$^{V}I_{C,R} = -\sum_{g=1}^{G}\frac{n_g}{A}\log_2\frac{n_g}{A} \quad [5]$$

where n_g is the number of graph vertices having the same atom eccentricity (the maximum distance from a vertex to any other vertex in the graph), G is the number of different vertex equivalence classes, and A is the number of graph vertices.

- D255- Vertex distance count equality index

The vertex distance counts indicate the frequencies of distances equal to 1, 2, 3 etc. from vertex v_i to any other vertex. The vertex distance count of first-order $^{1}f_i$ coincides with the vertex degree δ_i, that is, with the number of first neighbors, while $^{2}f_i$ and $^{3}f_i$ correspond to the connection number (i.e., number of second neighbors), and polarity number (i.e., number of third neighbors) for the i^{th} vertex, respectively.

- D269- Information content order-0 index

The information content based descriptors are calculated from the information content of a molecule (I_c). I_c is used to measure the degree of diversity of the atoms or bonds in a molecule (Eq. 6):

$$I_c = \sum_{c=1}^{C} n_c \cdot \log_2 n_c \quad [6]$$

where C is the number of different types of atoms or bonds and n_c is the number of atoms or bonds of the c^{th} type. Mean information content, mean information content on edge equality, and redundancy index are some examples of information content-based descriptors.

- D454- Geary topological structure autocorrelation length-8 weighted by atomic masses, D468- Geary topological structure autocorrelation length-6 weighted by atomic Sanderson electronegativities, D472- Geary topological structure autocorrelation length-2 weighted by atomic polarizabilities, and D473- Geary topological structure autocorrelation length-3 weighted by atomic polarizabilities

The Geary coefficient (c_k) is a general index of spatial autocorrelation that, if applied to a molecular graph and can be defined as:

$$c_k = \frac{\frac{1}{2\Delta_k}\sum_{i=1}^{A}\sum_{j=1}^{A}(w_i - w_j)^2 \cdot \delta(d_{ij};k)}{\frac{1}{A-1}\cdot\sum_{i=1}^{A}(w_i - \overline{w})^2} \quad [7]$$

where w_i is any atomic property as a weighting factor, $\overline{w}$ is its average value on the molecule, A is the number of atoms, k is the lag considered, d_{ij} is the topological distance between the ith and jth atoms, and $\delta(d_{ij};k)$ is the Kronecker delta equal to 1 if $d_{ij} = k$, zero otherwise. Δ_k is the number of vertex pairs at distance equal to k.

Geary coefficient is a distance-type function varying from zero to infinite. Strong autocorrelation produces low values of this index; moreover, positive autocorrelation translates into values between 0 and 1 whereas negative autocorrelation produces values larger than 1; therefore, the reference "no correlation" is $c_k = 1$.

- D522- Mean molecular topological order-2 charge index

  Descriptors related to the topological charge index are derived from the adjacency matrix and distance matrix of a molecule, which estimate the charge transfer between pairs of atoms, and therefore the global charge transfer in the molecule.

- D541 Lowest eigenvalue from Burden matrix weighted by van der Waals order 2

  Burden eigenvalues are demonstrated to reflect the topology of the whole molecule; the highest and the lowest eigenvalues reflect relevant aspects of molecular structure, and are useful for similarity searching, identification and ordering of molecular structures. In detail, Burden matrices (B) fall into the category of weighted adjacency matrices that encode information about proximity between vertices of molecular graphs that represent molecules containing heteroatoms and/or multiple bonds. The elements B_{ij} that represent two bonded atoms i and j are equal to $\pi^* 10^{-1}$, where π^* is the conventional bond order (0.1, 0.2, 0.3 and 0.15 for a single, double, triple and aromatic bond respectively; elements corresponding to terminal bonds are augmented by 0.01), the diagonal elements of the Burden matrix contain the atomic numbers Z_i of the atoms and all other matrix elements are set equal to 0.001. From Burden matrices, Burden eigenvalues are computed and are used in QSAR modeling. It is assumed that the smallest eigenvalues contain contributions from all atoms of the molecule and therefore reflect the topology of the whole molecule and have high discrimination power.

Bibliography

Todeschini, Roberto, and Viviana Consonni. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references*. Vol. 41. John Wiley & Sons, 2009.

Dehmer, Matthias, et al., eds. *Statistical modelling of molecular descriptors in QSAR/QSPR*. Weinheim, Germany: Wiley-VCH, 2012.

Singh, P., et al. "Topological descriptors in modeling malonyl coenzyme A decarboxylase inhibitory activity: N-Alkyl-N-(1, 1, 1, 3, 3, 3-hexafluoro-2-hydroxypropylphenyl) amide derivatives." *Journal of enzyme inhibition and medicinal chemistry* 24.1 (2009): 77-85.