Invariants or Aggregation Operators

The invariants are numerical quantities derived from the molecular structure and used to characterize local properties of a molecule; these numbers are calculated in such a way as to be independent of any arbitrary atom/bond numbering. Local invariants can be distinguished into LOcal Vertex Invariants (LOVIs) and LOcal Edge Invariants (LOEIs), depending on whether they refer to atoms or bonds.

LOVIs of a molecule are usually collected into an *N*-dimensional vector (*N* = number of atoms). LOVIs and LOEIs (also known as LOBIs) are used to calculate several molecular (namely, topological indices) by applying different operators. *L* is here adopted as the general symbol for local invariants.

Over the years, it has been generally accepted that the definition of global (or local) indices from LOVIs (L_i) implies the summation of the contributions of the elements that constitute a given molecular structure. However, summation (Minskowski's first norm (N1) in our specific case) is just one of the many invariants capable of globally characterizing given LOVIs.

In this program, are employed a series of *invariants* that generalize the traditional method of obtaining global (or local) invariants by summation of the LOVIs (and LOEIs used in bond-based indices). These are classified in four major groups:

1) **Norms (or Metrics)**:
    - **a)** Minkowski's norms (N1, N2, N3), and
    - **b)** Penrose's size (PN)
2) **Mean Invariants (first statistical moment)**:
    - **a)** Geometric Mean (G),
    - **b)** Arithmetic Mean (M),
    - **c)** Quadratic Mean (P2),
    - **d)** Potential Mean (P3) and
    - **e)** Harmonic Mean (A)
3) **Statistical Invariants (highest statistical moments):**
    - **a)** Variance (V),
    - **b)** Skewness (S),

c) Kurtosis (K),

d) Standard Deviation (DE),

e) Variation Coefficient (CV),

f) Range (R),

g) Percentile 25 (Q1),

h) Percentile 50 (Q2),

i) Percentile 75 (Q3),

j) Inter-quartile Range (I50),

k) X max (MX) and

l) X min (MN)

4) OWA Operators

a) S-OWA (SOWA)

b) Window-OWA (WOWA)

c) OWA with exponential initialization of type I (OWAE1)

d) OWA with exponential initialization of type II (OWAE2)

e) OWA depending on the aggregated objects of type I (OWAA1)

f) OWA depending on the aggregated objects of type II (OWAA2)

5) "Classical algorithms" Invariants (functions to derive MDs from LOVIs):

a) Autocorrelations AC(i),

b) Gravitational (GI(i)),

c) Total sum at *k* lags (TSk(i)),

d) Kier-hall connectivity (CN(i)),

e) Mean information content (MI(i)),

f) Total information content (TI),

g) Standardized information content (SI),

h) Ivanciuc-Balaban operator (IB),

i) Electro-topological state (ES(i)).

1. Norms (or Metrics): Mathematical definition

Name	ID	Formula
Minkowsky norm (p = 1) Manhattan norm	N1	$N1 = \sum_{a=1}^{n} \lvert L_a \rvert$
Minkowsky norm (p = 2) Euclidean norm	N2	$N2 = \sqrt{\sum_{a=1}^{n} \lvert L_a \rvert^2}$
Minkowsky norm (p = 3)	N3	$N3 = \sqrt[3]{\sum_{a=1}^{n} \lvert L_a \rvert^3}$
Penrose size	PN	$PN = \sqrt{\frac{1}{n^2}\left[\sum_{a=1}^{n} (L_a)\right]^2}$

Note 1. The general equation of Minkowsky's norms is, $\|\bar{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$.

Note 2. The formulae used in these invariants, are simplified forms of general equations given that the vector $\bar{y}$ is constituted of the coordinates of the origin. For example, in the case of the Euclidean norm (N2), the general formula is: $\|\bar{x}\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2 + (x_j - y_j)^2 + (x_z - y_z)^2}$.

But given that $\bar{y}$ = (0, 0, 0), this formula reduces to $\|\bar{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$.

2. Mean Invariants: Mathematical definition

Name	ID	Formula
Geometric Mean	GM	$G = \sqrt[n]{\prod_{a=1}^{n} L_a}$
Arithmetic Mean (Power mean of degree β = 1) Quadratic Mean (Power mean of degree β = 2) Power mean of degree β = 3 Harmonic Mean (Power mean of degree β = -1)	AM P2 P3 HA	$M_\beta = \left(\frac{L_1^\beta + L_2^\beta + \ldots + L_n^\beta}{n} \right)^{\frac{1}{\beta}}$

3. Statistical Invariants: Mathematical definition

Name	ID	Formula
Variance	V	$V = \frac{\sum_{a=1}^{n} (L_a - M)^2}{n - 1}$
Skewness	S	$S = \frac{n * (X_3)}{(n-1)(n-2)(DE)^3}$ $X_3 = \sum_{a=1}^{n} (L_a - M)^3$ M, arithmetic mean DE, standard deviation

Kurtosis	K	$k = \frac{n(n+1)X_4 - 3(X_2)(X_2)(n-1)}{(n-1)(n-2)(n-3)(DE)^4}$ M, $X_j = \sum_{a=1}^{n} (L_a - M)^j$ arithmetic mean DE, standard deviation
Standard Deviation	DE	$\mathrm{DE} = \sqrt{\frac{\left(\sum L_a - M\right)^2}{n-1}}$
Variation Coefficient	CV	$\mathrm{CV} = DE/M$
Range	R	$\mathrm{R} = L_{\max} - L_{\min}$
Percentile 25	Q1	$\mathrm{Q1} = \left[\frac{N}{4} + \frac{1}{2}\right]$ N, L_a number
Percentile 50	Q2	$\mathrm{Q2} = \left[\frac{N}{2} + \frac{1}{2}\right]$ N, L_a number
Percentile 75	Q3	$\mathrm{Q3} = \left[\frac{3N}{4} + \frac{1}{2}\right]$ N, L_a number
Inter-quartile Range	I50	$\mathrm{I50} = Q3 - Q1$
Maximum value	MX	MX = L_a max
Minimum value	MN	MN = L_a min

4. OWA Operators: Mathematical definition

The OWA operator is generally composed of the following three steps:

- Reorder the input arguments in descending order
- Determine the weights associated with the OWA operator by using a proper method
- Utilize the OWA weights to aggregate these reordered arguments

Formally, an OWA operator of dimension *n* is a mapping, *OWA*: $R^n \to R$, that has an associated *n* vector $w = (w_1, w_2, \ldots, w_n)^T$ such that $w_j \in [0,1]$ and $\sum_{j=1}^{n} w_i = 1$. Furthermore,

$$OWA_w(a_1, a_2, \ldots, a_n) = \sum_{j=1}^{n} w_j b_j$$

where b_j es el jth largest element of the collection of the aggregated objects $a_1, a_2, \ldots a_n$. Clearly, one key point in the OWA operator is to determine its associated weights, and thus 6 methods reported in the literature have been implemented in this works, which are explained to continue:

Name	ID	Formula
S-OWA	SOWA	$w_i \begin{cases} \frac{1}{n}(1-(\alpha+\beta))+\alpha, & i=1 \\ \frac{1}{n}(1-(\alpha+\beta)), & i=2,\ldots,n\text{-}1 \\ \frac{1}{n}(1-(\alpha+\beta))+\beta, & i=n \end{cases}$
Window-OWA	WOWA	$w_i \begin{cases} 0, & i<k \\ \frac{1}{m}, & k\le i<k+m \\ 0, & i\ge k+m \end{cases}$
Exponential Type I	OWAE1	$w_i \begin{cases} \alpha, & i=1 \\ w_{i-1}(1-w_1), & i=2,\ldots,n\text{-}1 \\ \frac{w_{i-1}(1-w_1)}{w_1}, & i=n \end{cases}$
Exponential Type II	OWAE2	$w_i \begin{cases} 1-\alpha, & i=n \\ w_i(1-w_n), & i=2,\ldots,n\text{-}1 \\ \frac{w_{i+1}(1-w_n)}{w_n}, & i=1 \end{cases}$
Aggregated Objects I	OWAA1	$OWA_w(a_1,a_2,\ldots,a_n)=\frac{\sum_{i=1}^{n} a_i^{\alpha+1}}{\sum_{i=1}^{n} a_i^{\alpha}}$
Aggregated Objects II	OWAA2	$OWA_w(a_1,a_2,\ldots,a_n)=\frac{\sum_{i=1}^{n}(1-a_i)^{\alpha} a_i}{\sum_{i=1}^{n}(1-a_i)^{\alpha}}$

5. Classical Invariants: Mathematical definition

Name	ID	Formula
Autocorrelation	AC^k	$AC_k=\sum_{i=1}^{n}\sum_{j\ge 1}^{n} L_i \times L_j \bullet (\delta(d_{ij},k))$ $k=1,2,..7$
Gravitational	GV^k	$GV_k=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{L_i L_j}{{}^{k}d_{ij}}\bullet\delta(d_{ij},k))$ $k=1,2,..7$

Total sum at lag k	TS^k	$TS_k = \sum_{i=1}^{n}\sum_{j=1}^{n} L_{ij} \bullet \delta(d_{ij}, k))$ $k = 1,2,\ldots,7$
Kier-Hall connectivity	KH^m	${}^{m}KH_t = \sum_{i=1}^{K} (\prod_{i=1}^{n_k} L_i, w)^{\lambda}_{k}$ *where, K is the number of sub-graphs, n_k is the number of atoms in a fragment, λ is equal to ½, m and t are the sub-graph order and type, respectively*
Mean Information Content	MIC	$MIC = -\sum_{i=1}^{A} \frac{N_g}{N_o} \cdot \log_2 \frac{N_g}{N_o}$ *where, Ng is the number of atoms with the same LOVI value. N_o is the number of atoms in a molecule*
Total Information Content	TIC	$TIC = N_0 \cdot \log_2 N_0 - \sum_{g=1}^{G} N_g \cdot \log_2 N_g$
Standardized Information Content	SIC	$SIC = \frac{IT}{N_0 \cdot \log_2 N_0}$
Electrotopological state (E-state index)	ES	$S_i = I_i + \Delta I_i = I_i + \sum_{j=1}^{n} \frac{I_i - I_j}{(d_{ij}+1)^2}$ *where, I_i is the intrinsic state of the i^{th} atom and ΔI_i is the field effect on the ith atom calculated as perturbation of the I_i of i^{th} atom by all other atoms in the molecule, d_{ij} is the topological distance between the i^{th} and the j^{th} atoms, and n is the number of atoms. The exponent k is 2.*
Ivanciuc-Balaban Type-Indices	IB	$J_k = \frac{n^2 \cdot B}{n + C + 1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} \left[L_i \times L_j\right]^{-\frac{1}{2}}$ *where, the summation goes over all pairs of atoms but only pairs of adjacent atoms are accounted for by means of the elements a_{ij} of the adjacency matrix. The n, B, and C are the number of atoms, bonds, and rings (cyclomatic number), respectively.*

Standardized option

In the standardization procedure, all values of *original* LOVIs are replaced by standardized LOVI values which are computed as follows: *Std. LOVIs* = (Original LOVI – mean of LOVIs)/Std. deviation of original LOVIs. With this re-scaling, each new LOVI has a mean of 0 and a standard deviation of 1.