RSC Advances

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances

QSPR with Descriptors based on Averages of Vertex Invariants. An Artificial Neural Network Study

Lionello Pogliani, Jesus Vicente de Julián-Ortiz
MOLware SL, c/Burriana 36-3, 46005, Valencia, Spain, and Unidad de Investigación de Diseño de Farmacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, 46100 Burjassot (València), Spain; liopo@uv.es, jejuor@uv.es.

Abstract

A new type of indices, the mean molecular connectivity indices (MMCI), based on nine different concepts of mean are proposed to model, together with molecular connectivity indices (MCI), experimental parameters and random variables, eleven properties of organic solvents. Two model methodologies are used to test the different descriptors: the multilinear least-squares (MLS) methodology and the Artificial Neural Network (ANN) methodology. The top three quantitative structure-property relationships (QSPR) for each property are chosen with the MLS method. The indices of these three QSPRs were used to train the ANNs that selected the best training sets of indices apt to estimate the evaluation sets of compounds. The best ANN relationships for most properties are of the semiempirical types that include mean molecular connectivity indices (MMCI), molecular connectivity indices (MCI) and experimental parameters. Refractive index, *RI*, viscosity, η, and surface tension, γ, prefer a semiempirical relationship made of MCI and an experimental parameter only. In our previous study with no MMCI, random variables contributed to semiempirical relationships for two properties at the ANN level (*MS*, and *El*), here the use of MMCIs undo the contribution of such variables. Most of the MMCIs that contribute to improve the model of the properties are valence-delta-dependent (δ^{v}), that is, they encode both the hydrogen atom contribution and the core electrons of higher-row atoms.

1. Introduction

Molecular connectivity became a full grown-up branch of chemical graph theory with Randić [1] and Kier and Hall [2] and nearly quarter of a century later Todeschini and Consonni [3] were able to write an *opus magnum* on descriptors. In it they elegantly stated that "a descriptor is the final result of a logico-mathematical procedure, which transforms an information, encoded within a symbolic representation of an event, into *useful* numbers". Descriptors are critical in QSAR/QSPR modeling studies, thus, finding new and useful ones is an important task for those working in the field. [4]

In the literature, nine definitions of mean between two numbers (see Appendix) can be found, and these definitions are used to define new type of indices, the mean molecular connectivity indices (MMCI). These new indices are here used together with molecular connectivity indices, experimental parameters, and random variables to build optimal semiempirical quantitative structure-property relationships (QSPR) for eleven properties of a set of organic solvents. These properties were recently modeled [5] with a semiempirical set of descriptors that encompassed only the molecular connectivity indices (MCI), empirical parameters, and random variables. The cited work [5] emphasized the advantage in using ANN for model purposes. The two main aims of the present study are: (*i*) test the usefulness of the new MMCI indices and (*ii*) the related usefulness of the ANN methodology.

2. Computational Tools

In the following are the definitions of the mean molecular connectivity indices (MMCI),

$$^{A}M = \Sigma_i \delta_i / N \quad (1)$$

$$^{G}M = \Sigma_{ij}(\delta_i\delta_j)^{1/2} \quad (2)$$

$$^{H}M = 2\Sigma_{ij} (\delta_i^{-1}+\delta_j^{-1})^{-1} \quad (3)$$

$$^{R}M = \Sigma_{ij}[(\delta_i^{2}+\delta_j^{2})/2]^{1/2} \quad (4)$$

$$^{S}M = \Sigma_{ij} (\delta_i^{2} + \delta_j^{2}) / (\delta_i + \delta_j) \quad (5)$$

$$^{U}M = \Sigma_{ij}[\delta_j - \delta_i + (\delta_i^{2} - 2\,\delta_i\delta_j + 5\delta_j^{2})^{0.5}]/2 \quad (6)$$

$$^{Ho}M = \Sigma_{ij}(\delta_i^{p}+ \delta_j^{p})^{1/p} / 2 \quad (7)$$

$$^{L}M = \Sigma_{ij}(\delta_i^{p}+ \delta_j^{p}) / (\delta_i^{p-1} + \delta_j^{p-1}) \quad (8)$$

$$^{St}M = \Sigma_{ij}[(\delta_i^{p} - \delta_j^{p}) / (p\delta_i + p\delta_j)]^{1/(p-1)} \quad (9)$$

Here, *i* (and *j*) assigns the *N* atoms of a hydrogen-depleted molecule, *ij* means two atoms directly bonded through a σ bond, and p = *N*, even if other values are possible. It should be underlined that *N* for the studied molecules (see Table 1) is not that large. The reader may notice, among other similarities, that the Lehmer mean, ^{L}M, for $p = 2$ equals the symmetrical mean, ^{S}M.

Replacing throughout these definitions δ, with: (*i*) the valence delta, δ^{v}, (*ii*) the Intrinsic-*I*-State indices, and (*iii*) the Electrotopological-*S*-State indices (see Refs. 5-7 and Appendix), it is possible to obtain the three subsets: the valence MMCI,{$^{A}M^{V}$, $^{G}M^{V}$, $^{H}M^{V}$, $^{R}M^{V}$, $^{S}M^{V}$, $^{U}M^{V}$, $^{Ho}M^{V}$, $^{L}M^{V}$, $^{St}M^{V}$}, the *I*-State MMCI: {$^{A}M_{I}$, $^{G}M_{I}$, $^{H}M_{I}$, $^{R}M_{I}$, $^{S}M_{I}$, $^{U}M_{I}$, $^{Ho}M_{I}$, $^{L}M_{I}$, $^{St}M_{I}$}, and *E*-State MMCI: {$^{A}M_{E}$, $^{G}M_{E}$, $^{H}M_{E}$, $^{R}M_{E}$, $^{S}M_{E}$, $^{U}M_{E}$, $^{Ho}M_{E}$, $^{L}M_{E}$, $^{St}M_{E}$}. The basic notions of delta, valence delta, *I*- and *S*-indices belong to the origins of the molecular connectivity theory, and are based on graph concepts.[1-3, 8-10] To avoid imaginary *S*-State MMCIs, as some *S* values for highly electropositive atoms can be negative, a rescaling of the *S* value is undertaken (see Ref. 5). Summing up we have thirty-six MMCIs. Other MMCI can be derived following different types of bonding and branching as suggested by Kier and Hall [9] but for our present purpose these are enough. To model our eleven properties we will also use thirty MCI (see Table 2 in Ref. 5), fifty random variable *rn1* - *rn50* (where $0 < rn < 1$). The five experimental variables, {*M*, T_b, ε, *d*, *RI*}, of Table 1 will also be used as indices throughout the present calculations, i.e., in some cases they will show up on the right-side of the modeling relationships, and then the relationship will be labeled semiempirical. The final number of independent variables sums up to 121. The best relationship for each property might then encompass these four different type of indices: MMCIs, MCIs, experimental variables, and random variables.

The MMCI have been obtained with a visual basic home-made program that uses both adjacency and distance matrices [6] and that runs on a PC. The number of indices of the present multilinear relationships equals the number of indices of the corresponding relationship of ref. 5 that obeyed the *Topliss-Costello* rule: [11] the ratio of data points to the number of variables should be higher or equal to five and should provide a correlation coefficient $r > 0.84$ ($r^2 > 0.70$).

The multilinear least-squares procedure of Statistica 8 is used to find out the best relationship for the training compounds of Table 1, which is then used to evaluate the left-out compounds [those with (°) in Table 1]. It should be underlined that in principle the experimental values of the evaluated points are unknown, and they have to be guessed from the predictive relationship obtained with the training points. This equation will check how much the guessed evaluated points will deviate from the true values, and how symmetrically are the residuals (deviations) placed around the zero line in a residual plot. [12] The overall quality of the model for each property, that is, r^2, *s,* and *N* (here number of compounds), is obtained with the EXCEL spreadsheet plotting the observed property (*P*) vs. the calculated one, P_{clc}. The quality of the training regression equation is given also by the q^2 leave-one-out statistics[5] (Table 2).

Our previous ANN study [5] has shown that, as a rule, ANN models fit the data better than the MLS ones and this is the reason that three best sets of MLS descriptors, with similar quality, for the training set of compounds have been passed over to the ANN method. Additionally, the ANN program chooses a small set (20%) of test compounds (underlined and bold compounds in Table 1) belonging to the training compounds to achieve a rapid convergence and to avoid overtraining.

ANN methods, which are capable of performing regression and data validation, carry out both tasks in a non-parametric way that makes no assumption regarding the relationship between *y* and *x*, where $y = f(x)$. This means that the function $Property = f(indices)$ is not known *a priori*. In short, a non-parametric model is a kind of *black box* that tries to discover the mathematical function that can approximate the relationship between the *indices* and the *property* well enough. It uses highly flexible transfer functions with adaptable parameters that can model a wide spectrum of functional relationships.[13] ANN results were obtained with the built-in utility of Statistica 8, the multilayer perceptron neural network (MLP). The ANN-MLP network used here has three-layered feedforward architecture with unidirectional full connections between successive layers and with error backpropagation (or backprop). The three layers are: *input units - hidden units - output units*, that correspond to: *indices - hidden units – 1* (one), where the only output unit or neuron is the targeted property. The connections between the units (here two sets of links: input-hidden, and hidden-output) are the weights that determine the values assigned to the nodes. There exist additional weights assigned to the bias values that act as node value offsets. The weights that are adjusted by

the training process are initially random and are passed to all nodes of the following layer. The training process is iterative and each iteration is called an epoch. The weights are slightly varied in each epoch to minimize the sum-of-squares error function: $SOS = \sum_{i=1\text{-}N}(P_{iclc} - P_i)^2$, where P_{iclc} (*clc* = calculated) is the i^{th} predicted value (network outputs) of the property, P_i (target value), to be predicted. This function is the sum of differences between the prediction outputs and the target defined over the entire training set of points (compounds) *N*. The number of hidden nodes in Statistica 8 is set, by default, between 3 and 11. For UV, *MS*, and *El* this number is set between 3 and 10. This means that the final weight values for a single property of, for instance, a [5-7-1] network could fill an entire page. In Table 3 are given, as in our previous work, only the sensitivity values, which are the values that are due to the sensitivity analysis that rates the importance of the models' input variables. The activation functions for both hidden and output nodes in Statistica 8 are: identity (*i*), logistic sigmoid (*l*), hyperbolic tangent (*t*), sine (*s*), and exponential (*e*). The detailed activation function together with the neuronal architecture will be given in Table 3 together with the statistics, r^2 and *s*, for each property, that were obtained with the EXCEL spreadsheet plotting the observed *P* vs. the calculated P_{clc} ANN-MLP values.

Statistica 8 allows one to set only the number of networks to train and retain (100 / 40), without taking into account the number of training cycles/epochs. The type of algorithm that optimizes the network is the BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm that ensures a fast convergence rate. [14] In Table 3 are given the *number of epochs* for which it runs even if the actual number of cycles used to train the model might be greater. As the *number of epochs* is not definitive it cannot be held as an unfailing parameter (it can exceed the given number).

It is not rare the case that the model becomes exceedingly good giving rise to *overfitting* with exceedingly poor externally evaluated values. The choice of training (here 80%) and test (here 20%) sets normally avoids overfitting because the network is repeatedly trained for a number of cycles so long as the test error is on the decrease, as soon as it increases again the training is halted.

3. Studied Properties

The eleven properties of organic solvents are listed in Table 1. The source of the experimental values is given in ref. 7. Compounds with (°) in Table 1 build the evaluation set of compounds while the remaining compounds build the training set used to find out, with a full combinatorial least-squares regression, the best descriptors.

Table 1 Eleven properties of organic solvents plus their molar mass M (g·mol^{-1}): T_b, boiling points (K); ε, dielectric constant; d, density (at 20°C±5°C relative to water at 4°C, g/cc); RI, refractive index (20°C); FP, FlashPoint (K); η, viscosity (Cpoise, 20°C; [1]at 25°C, [2] at 15°C); γ, surface tension (mN/m at 25°C); UV, Cutoff UV values (nm); μ, dipole moments in Debye (1D = 10^{-18} esu cm = 3.3356 10^{-3} C m); MS ($-\chi\cdot10^6$), magnetic susceptibility (also, - $\chi\cdot10^6$, in emu mol^{-1}, 1 emu = 1 cm^3, temperatures cover a range from 15°C to 32°C); and El, Elutropic value (silica).

Solvents	*M*	T_b	ε	*d*	*RI*	*FP*	η	γ	*UV*	μ	*MS*	*El*
(°)Acetone	58.1	329	20.7	0.791	1.359	256	0.32	23.46	330	2.88	0.46	0.43
(°)Acetonitrile	41.05	355	37.5	0.786	1.344	278	0.37	28.66	190	3.92	0.534	0.50
Benzene	78.1	353	2.3	0.84	1.501	262	**0.65**	28.22	280	0	0.699	0.27
Benzonitrile	103.1	461	25.2	1.010	1.528	344	1.24[1]	38.79				
1-Butanol	74.1	391	17.1	0.810	1.399	308	2.95	24.93	215			
(°)2-Butanone	72.1	353	18.5	0.805	1.379	270	0.40	23.97	330			0.39
Butyl Acetate	116.2	398	5.0	0.882	1.394	**295**	**0.73**	**24.88**	254			
CS_2	76.1	319	2.6	1.266	1.627	**240**	0.37	31.58	380	0	0.532	
CCl_4	153.8	350	2.2	1.594	1.460		**0.97**	**26.43**	263	**0**	0.691	0.14
Cl-Benzene	112.6	405	5.6	1.107	1.524	296	0.80	32.99	**287**			
1Cl-Butane	92.6	351	7.4	0.886	1.4024	267	0.35	23.18	225			
$CHCl_3$	119.4	334	4.8	1.492	1.446		0.57	26.67	245	1.01	0.740	**0.31**
Cyclohexane	84.2	354	2.0	0.779	1.426	255	1.00	24.65	**200**	**0**	0.627	0.03
(°)Cyclopentane	70.1	323	2.0	0.751	1.400	236	0.47	21.88	200		0.629	
1,2-diCl-Benzene	147.0	453	9.9	1.306	1.551	**338**	1.32		**295**	2.50	0.748	
1,2-diCl-Ethane	98.95	356	10.4	1.256	1.444	288	0.79	31.86	225	1.75		
diCl-Methane	84.9	313	9.1	1.325	1.424		0.44	27.20	235	1.60	**0.733**	0.32
N,N-diM-Acetamide	87.1	**438**	**37.8**	**0.937**	**1.438**	343			268	3.8		
N,N-diM-Formamide	73.1	**426**	**36.7**	**0.944**	**1.431**	**330**	0.92		268	3.86		
1,4-Dioxane	88.1	374	2.2	1.034	1.422	285	**1.54**	**32.75**	215	0.45	0.606	
Ether	74.1	308	4.3	0.708	1.353	**233**	0.24	16.95	**215**	1.15		0.29
Ethyl acetate	88.1	350	6.0	0.902	1.372	270	0.45	23.39	260	1.8	0.554	**0.45**
(°)Ethyl alcohol	46.1	351	24.3	0.785	1.360	281	1.20	21.97	210	1.69	0.575	
Heptane	100.2	371	1.9	0.684	1.387	272		19.65	200			0.00
Hexane	86.2	**342**	**1.9**	**0.659**	**1.375**	**250**	0.33	17.89	200			0.00
2-Methoxyethanol	76.1	398	16.0	0.965	1.402	319	**1.72**	30.84	220			
(°)Methyl alcohol	32.0	338	32.7	0.791	1.329	284	0.60	22.07	205	1.70	0.530	0.73
(°)2-Methylbutane	72.15	303	1.8	0.620	1.354	217						
4-Me-2-Pentanone	100.2	391	13.1	0.800	1.396	286			334			
2-Me-1-Propanol	74.1	381	17.7	0.803	1.396	310						
2-Me-2-Propanol	74.1	356	10.9	0.786	1.387	277		19.96		**1.66**	**0.534**	
DMSO	78.1	462	46.7	1.101	1.479	368	2.24	**42.92**	268	3.96		
(°)Nitromethane	61.0	374	35.9	1.127	1.382	308	0.67	36.53	380	3.46	0.391	
1-Octanol	130.2	469	10.3	0.827	1.429	354	10.6[2]	27.10				
(°)Pentane	72.15	309	1.8	0.626	1.358	224	0.23	15.49	200			0.00
3-Pentanone	86.1	375	17.0	0.853	1.392	279		24.74				
(°)1-Propanol	60.1	370	20.1	0.804	1.384	288	2.26	23.32	210			
(°)2-Propanol	60.1	356	18.3	0.785	1.377	295	2.30	20.93	210			0.63
Pyridine	79.1	388	12.3	0.978	1.510	293	0.94	36.56	305	2.2	0.611	0.55
tetraCl-Ethylene	165.8	394	2.3	1.623	1.506		0.90				**0.802**	
(°)tetra-Hydrofuran	72.1	340	7.6	0.886	1.407	256	0.55		215	1.75		0.35
Toluene	92.1	384	2.4	**0.867**	1.496	277	0.59	27.93	285	0.36	0.618	0.22
1,1,2triCl,triFEthane	187.4	321	2.4	**1.575**	1.358		**0.69**		**230**			**0.02**
2,2,4-triMe-Pentane	114.2	372	1.9	0.692	1.391	266	0.50		215			0.01
o-Xylene	106.2	**417**	**2.6**	**0.870**	**1.505**	**305**	0.81	29.76				
p-Xylene	106.2	411	2.3	0.866	1.495	300	0.65	28.01				
(°)Acetic acid	60.05	391	6.15	1.049	1.372			27.10		1.2	0.551	
Decaline	138.2	**465**	**2.2**	**0.879**	**1.476**						0.681	
diBr-Methane	173.8	370	7.8	1.542	2.497			**39.05**		**1.43**	0.935	
1,2-diCl-Ethylen(Z)	96.9	334	9.2	1.284	1.449					1.90	0.679	
(°)1,2-diCl-Ethylen(E)	96.9	321	2.1	1.255	1.446					0	0.638	
1,1-diCl-Ethylen	96.9	**305**	**4.7**	**1.213**	**1.425**					1.34	0.635	
Dimethoxymethane	76.1	315	2.7	0.866	1.356						0.611	
(°)Dimethylether	46.1	249	5.0									
Ethylen Carbonate	88.1	511	89.6	1.321	1.425					4.91		
(°)Formamide	45.0	484	109	1.133	1.448			57.03		3.73	0.551	
(°)Methylchloride	50.5	249	12.6	0.916	1.339					1.87		
Morpholine	87.1	402	7.3	1.005	1.457						0.631	
Quinoline	129.2	510	9.0	1.098	1.629			42.59		2.2	0.729	
(°)SO_2	64.1	263	17.6	1.434						1.6		
2,2-tetraCl-Ethane	167.8	**419**	**8.2**	**1.578**	**1.487**			35.58		1.3	0.856	
tetraMe-Urea	116.2	450	23.1	0.969	1.449					**3.47**	0.634	
triCl-Ethylen	131.4	360	3.4	1.476	1.480						**0.734**	

(°) externally validated compounds. Underlined **bold** values: test compounds used in ANN-MLP calculations.

As we already explained in the previous section the ANN-MLP methodology further subdivides the training compounds into training (80%) and test compounds (20%, underlined and bold in Table 1). Concerning the model for dipole moments, all indices were multiplied by a two-valued symmetry indicator variable which is zero for symmetric molecules (with $\mu = 0$ in Table 1) and 1 otherwise. Due to PC limitations the entire space of {MMCI, MCI, Rn, ExpPar} could not be searched for the best descriptor. The search was done in two different ways: (*i*) search for the best descriptor within the set {MMCI, MCI, Exp.Par.}, i.e., best (MMCI, MCI, Exp.Par.), and, finally, (*ii*) search for the best descriptor within the set {best(MMCI, MCI, Exp.Par.), Rn}.

4. Results

Table 2 shows the best relationships and their statistical parameters obtained with the stepwise multilinear least-squares (MLS) search procedure. The quality of each training equation is also accounted by the errors ($\pm \Delta c_i$) of the regression parameters c_i (in vector form in parenthesis with no ± signs). Each training equation (obtained without (°) compound in Table 1) has then been applied to model the evaluated points of Table 1 (those with (°)).

Table 3 shows the ANN-MLP results: the 1st column describes the MLP architecture with the abbreviation for the activation functions for the hidden and output layers, the number of epochs, and the training and test errors. The second column shows the best set of indices of the ANN-MLP method, the values of the sensitivity analysis for the indices (2nd line), the statistical parameter for the training [*N*(Tr)], and for all compounds [*N*(Tr+n°Te+n°EV)]] (3rd line). Notice that the training (TR) set of Table 2 throughout the ANN-MLP calculations of Table 3 is subdivided into training (Tr) and test (Te) sets.

The reader can notice that in Table 2 viscosity, η, is silent as MMCI contribute no model equation with improved quality relatively to the one given in ref. 5, where there was no talk about MMCI. For the same reason in Table 3 refractive index, *RI*, viscosity, η, and surface tension, γ, are silent.

5. Discussion

Tables 2 and 3 show that the best regression equations (relationships) are always of semiempirical type, i.e., composed of MCI, MMCI, experimental parameters, and, in two MLS cases (*FP*, and *El* in Table 2), of a random variable also. ANN-MLP calculations, instead, show no preference for random variables.

Table 2. The best MLS results for ten out of the eleven properties. 1st column: δ^v type [5] for the valence-dependent indices. 2nd column: relationships and statistical results for the training, *N*(TR), and training plus evaluation compounds, *N*(T). In the last line, the excluded strong outliers (those with residuals > 3*s*).

δ^v - *type*†	*Regression equations*	
$\delta^v_{ppo}(0.5)$	$\boldsymbol{T_b} = 183.6 + 1.807\boldsymbol{\varepsilon} + 5.556{}^{R}\boldsymbol{M}^{v} + 7.02{}^{G}\boldsymbol{M_E} - 41.59\boldsymbol{D} + 102.02{}^{I}\boldsymbol{\chi}^{v} + 8.241\boldsymbol{T_{\Sigma/M}}$ (8.3, 0.1, 1.0, 0.6, 4.4, 8.4, 1.01) *N*(TR) = 45, q^2 = 0.948, r^2 = 0.963, *s* = 10; *N*(T) = 62, r^2 = 0.874, *s* = 20 Excluded strong outliers in EV: Formamide.	(10)
$\delta^v_{po}(50)$	$\boldsymbol{\varepsilon} = -8.668 + 0.145\boldsymbol{T_b} - 0.069\boldsymbol{M} - 5.934{}^{Ho}\boldsymbol{M_E} + 4.802\boldsymbol{D}^{v} + 17.14{}^{I}\boldsymbol{\psi_I} + 13.07\boldsymbol{T_{\Sigma/M}}$ (5.3, 0.01, 0.01, 1.0, 0.8, 3.4, 3.0) *N*(TR) = 44, q^2 = 0.905, r^2 = 0.941, *s* = 2.5; *N*(T) = 60, r^2 = 0.937, *s* = 2.8 Excluded strong outliers: Ethylencarbonate (TR), HAc & Formamide (EV)	(11)
$\delta^v_{ppo}(2)$	$\boldsymbol{d} = -1.840 + 0.001\boldsymbol{T_b} + 0.592{}^{A}\boldsymbol{M}^{v} - 0.031\,\boldsymbol{D}^{v} + 0.239\,{}^{0}\boldsymbol{\chi}^{v} + 1.737\,{}^{T}\boldsymbol{\psi_I}$ (0.1, 0.0001, 0.01, 0.002, 0.01, 0.1) *N*(TR) = 45, q^2 = 0.981, r^2 = 0.986, *s* = 0.03; *N*(T) = 60, r^2 = 0.953, *s* = 0.06 Excluded strong outliers in EV: formamide, MeOH; (*N* = 62, r^2 = 0.906, *s* = 0.08**)	(12)
$\delta^v_{ppo}(1)$	$\boldsymbol{RI} = 1.287 + 0.0007\boldsymbol{T_b} - 0.131{}^{Ho}\boldsymbol{M_I} + 0.011\boldsymbol{M} - 0.479{}^{I}\boldsymbol{\chi} + 0.071\boldsymbol{D}^{v} - 0.080\boldsymbol{\Delta}$ (0.03, 0.0001, 0.005, 0.0003, 0.02, 0.002, 0.006) *N*(TR) = 45, q^2 = 0.970, r^2 = 0.983, *s* = 0.02; *N*(T) = 61, r^2 = 0.979, *s* = 0.02	(13)
$\delta^v_{po}(-0.5)$	$\boldsymbol{FP} = -75.22 + 0.873\boldsymbol{T_b} + 21.26\boldsymbol{d} + 7.018{}^{A}\boldsymbol{M}^{v} - 1.112{}^{G}\boldsymbol{M}^{v} + 13.72\boldsymbol{Rn41}$ (8.2, 0.02, 4.1, 0.6, 0.07, 2.8) *N*(TR) = 29, q^2 = 0.986, r^2 = 0.992, *s* = 3.1; *N*(T)= 41, r^2 = 0.967, *s* = 6.4	(14)
$\delta^v_{ppo}(2)$	$\boldsymbol{\gamma} = -14.25 + 0.153\,\boldsymbol{T_b} + 3.467\,\boldsymbol{RI} + 2.345\,{}^{G}\boldsymbol{M_I} + 0.475\,{}^{0}\boldsymbol{\psi_{Id}} - 0.902\,{}^{S}\boldsymbol{\psi_E}$ (2.3, 0.01, 1.2, 0.2, 0.09, 0.05) *N*(TR) = 29, q^2 = 0.953, r^2 = 0.977, *s* = 1.1; *N*(T) = 40, r^2 = 0.865, *s* = 3.0 Excluded strong outlier in EV: methanol.	(15)
$\delta^v_{po}(5)$	$\boldsymbol{UV} = -776.0 + 682.0\,\boldsymbol{RI} - 35.44{}^{H}\boldsymbol{M} + 7.259{}^{H}\boldsymbol{M}^{v} + 27.69\boldsymbol{D}$ (68, 40, 5.6, 0.8, 6.2) *N*(TR) = 25, q^2 = 0.928, r^2 = 0.955, *s* = 9.1; *N*(T) = 33, r^2 = 0.919, *s* = 12 Excluded strong outlier: 4-Me-2-Pentanone (TR); 2-butanone, MeCl, nitromethane (EV)	(16)
$\delta^v_{ppo}(50)$ φ = 0, 1	$\boldsymbol{\mu} = 0.0311 + 0.043\boldsymbol{\varepsilon} + 0.327{}^{H}\boldsymbol{M}^{v} - 0.293{}^{S}\boldsymbol{M_E} + 3.317{}^{I}\boldsymbol{\chi} + 0.188\boldsymbol{\Sigma}$ (0.1, 0.003, 0.04, 0.03, 0.3, 0.02) *N*(TR) = 24, q^2 = 0.939, r^2 = 0.984, *s* = 0.2; *N*(T) = 34, r^2 = 0.897, *s* = 0.4 Excluded strong outlier in EV: formamide.	(17)
$\delta^v_{po}(50)$	$-\boldsymbol{\chi 10^6} = 0.231 + 0.004\boldsymbol{M} - 0.008{}^{L}\boldsymbol{M_E} - 0.004{}^{St}\boldsymbol{M_E} - 0.137\,{}^{1}\boldsymbol{\psi_I} + 0.252{}^{I}\boldsymbol{\psi_{Es}}$ (0.03, 0.0003, 0.001, 0.0006, 0.02, 0.05) *N*(TR) = 23, q^2 = 0.842, r^2 = 0.945, *s* = 0.02; *N*(T) = 31, r^2 = 0.911, *s* = 0.03 Excluded strong outlier in EV: nitromethane.	(18)
$\delta^v_{po}(5)$	$\boldsymbol{El} = -1.479 + 0.006\boldsymbol{T_b} + 0.332{}^{A}\boldsymbol{M}^{v} - 0.021{}^{S}\boldsymbol{M_E} - 0.166\boldsymbol{Rn12}$ (0.1, 0.0003, 0.01, 0.001, 0.03) *N*(TR) = 15, q^2 = 0.945, r^2 = 0.986, *s* = 0.02; *N*(T) = 20, r^2 = 0.831, *s* = 0.1 pentane and tetrahydrofurane ∈{TR}	(19)

*TR = Training compounds; *N*(T) = Training plus evaluation compounds.
** With no outliers to allow comparison with previous results [5] (present are better).
† For the meaning of *po* and *ppo* see Appendix 2.

Table 3. ANN - MLP results for eight out of eleven properties. 1st column: the MLP architecture, the abbreviation for the activation functions for the hidden and output layers, the number of epochs, and the training and test errors; 2nd column: indices of the ANN relations, sensitivity values for the indices, and statistical parameters for the training (Tr), and training plus test (Te) and evaluation (EV) compounds.

MLP	δ^{v} (*type*) – (***Descriptors***) → ***Property***	
6 - 3 - 1 (*e*, *t*)* 30 0.001/0.001	$\delta^{v}_{ppo}(0.5)$ - $(\varepsilon, {}^{R}M^{v}, {}^{G}M_{E}, D, {}^{1}\chi^{v}, T_{\Sigma/M}) \rightarrow T_b$ (12.9, 17.63, 149.0, 86.34, 24.26, 2.443) N(Tr) = 38, r^2 = 0.968, s = 9.4, N(Tr + 7Te + 16EV) = 61 r^2 = 0.909, s = 17 Excluded strong outlier in EV: SO_2.	(20)
6 - 3 - 1 (*l*, *i*) 54 0.0002/0.0003	$\delta^{v}_{po}(50)$ - $(T_b, M, {}^{Ho}M_{E}, D^{v}, {}^{1}\psi_{I}, T_{\Sigma/M}) \rightarrow \varepsilon$ (53.26, 2.150, 240.0, 108.1, 24.37, 2.646) N(Tr) = 38, r^2 = 0.985, s = 2.0, N(Tr+ 7Te + 16EV) = 61, r^2 = 0.984, s = 2.5 Excluded strong outliers in EV: Acetonitrile, and HAc.	(21)
5 - 6 - 1 (*e*, *t*) 37 0.0004/0.0001	$\delta^{v}_{ppo}(1)$ - $(T_b, {}^{A}M^{v}, D^{v}, {}^{0}\chi^{v}, {}^{T}\psi_{I}) \rightarrow d$ (5.100, 138.9, 111.7, 74.96, 61.29) N(Tr) = 36, r^2 = 0.992, s = 0.03, N(Tr + 9Te + 15EV) = 60, r^2 = 0.992, s = 0.03 Excluded strong outliers in EV: formamide, and Me-Cl.	(22)
5 - 5 - 1 (*t*, *i*) 13 0.0004/0.001	$\delta^{v}_{ppo}(5)$ - $(T_b, RI, {}^{G}M_{I}, {}^{1}\chi^{v}, T_{\Sigma/M}) \rightarrow FP$ (133.3, 3.009, 4.263, 5.809, 2.443) N(Tr) = 22, r^2 = 0.990, s = 3.3, N(Tr + 7Te + 12EV) = 41, r^2 = 0.979, s = 5.1	(23)
4 - 5 - 1 (*e*, *t*) 29 0.0009/0.0004	$\delta^{v}_{ppo}(2)$ - $(RI, {}^{H}M, {}^{G}M^{v}, \Delta) \rightarrow UV$ (20.90, 138.9, 186.1, 19.98) N(Tr) = 20, r^2 = 0.969, s = 7.7, N(Tr + 5Te + 8EV) = 33, r^2 = 0.936, s = 11 Excluded strong outliers: 4M2-Pentanone in Tr, nitromethane, 2-butanone, and acetone in EV	(25)
5 - 5 - 1 (*l*, *s*) 53 0.0004/0.00006	$\delta^{v}_{ppo}(5)$ [φ = 0, 1] - $(\varepsilon, {}^{L}M_{E}, {}^{1}\chi_{d}, {}^{S}\psi_{E}, {}^{0}\psi_{Ed}) \rightarrow \mu$ (89.93, 98.12, 69.91, 127.5, 212.0) N(Tr) = 19, r^2 = 0.989, s = 0.1, N(Tr + 5Te + 10EV) = 34, r^2 = 0.937, s = 0.3 Excluded strong outliers in EV: MeOH.	(26)
5 - 5 - 1 (*t*, *e*) 33 0.0009/0.0006	$\delta^{v}_{po}(0.5)$ - $(M, {}^{L}M_{E}, {}^{St}M_{E}, {}^{1}\psi_{I}, {}^{1}\psi_{Es}) \rightarrow -\chi 10^{6}$ (79.09, 57.12, 15.52, 38.21, 8.276) N(Tr) = 19, r^2 = 0.968, s = 0.02, N(Tr + 4Te + 8EV) = 31, r^2 = 0.932, s = 0.03 Excluded strong outliers in EV: nitromethane.	(27)
4 - 9 - 1 (*e*, *l*) 33 0.0002/0.00004	$\delta^{v}_{po}(0.5)$ - $(\varepsilon, {}^{L}M^{v}, {}^{1}\psi_{E}, \Delta) \rightarrow El$ (37.46, 247.3, 489.7, 18.30) N(Tr) = 12, r^2 = 0.995, s = 0.01, N(Tr + 3Te + 5EV) = 20, r^2 = 0.932, s = 0.06 pentane and tetrahydrofurane belong here to {TR}	(28)

* Activation functions (in parenthesis): *e* = exponential, *i* = identity, *l* = logistic, *t* = tanh, *s* = sin

For six properties out of eleven (*RI*, *FP*, γ, *UV*, μ, *El*) ANN-MLP methodology do not choose the best descriptive equation of the MLS method. In general, ANN-MLP overall model quality (training + test + evaluation) improve (Table 3) over MLS (training + evaluation) model ability (Table 2). This confirms our previous findings [5]. Let us now compare MLS results of the present Table 2 with results of the corresponding results of Tables 3 (MLS for - $\chi \cdot 10^6$, and *El*), and 4 (MLS all other properties) of ref. 5. On the whole, semiempirical equations with one or more MMCI fare better than the corresponding semiempirical equations with no MMCI. Only exception being, as already told in the previous paragraph, the viscosity, η, whose semiempirical equation has no use for MMCI. The training semiempirical relationship for the magnetic susceptibility (- $\chi \cdot 10^6$) with no MMCI (ref. 5) show a better quality, but it is made of three random variables also, while the corresponding present training semiempirical regression with MMCI+MCI has no use for random variables.

Going over to the overall model ability (training plus evaluated properties) of the MLS regressions things change a bit: T_b, *FP*, γ, and μ, show preference for relationships with MCI only (Table 4 of ref. 5), *UV*, and - $\chi \cdot 10^6$ show no interesting improvement with MMCI, while for the ε (dielectric constant) r^2 prefers MCI only (ref. 5) but *s* improves when MMCI are added. Only for the remaining three properties, *d*, *RI*, and *El* there is a clear preference for model equations that include MMCI.

Comparison of ANN-MLP results of Table 3 with the corresponding results of Table 5 in ref. 5, show that more often than not training semiempirical equations with MMCI+MCI fare better. There are some exceptions: the already cited case for η, *RI*, γ, and for - $\chi \cdot 10^6$. Comparison, instead, of the overall description (i.e., training plus evaluated points) shows improvements with nearly the same exceptions: η, *RI*, and γ. The overall description of - $\chi \cdot 10^6$ improves consistently, while for *UV*, and *El* the model quality is rather similar. All in all mean molecular connectivity indices (MMCI) are useful both to improve a model quality and also to get rid of the random variables.

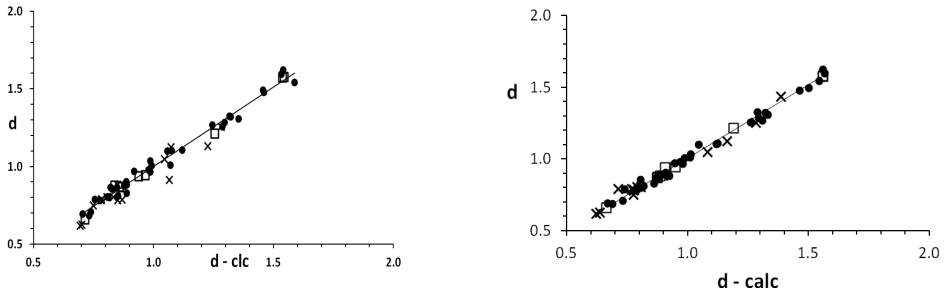


Fig.1. Plot for density (*d*) obtained in our previous ANN study [5] (left) and in the present study (right). Full circles: training compounds; crosses: evaluation compounds; empty squares: test compounds.

The model plots for the different properties improve relatively to the previous ones [5]. To give the reader an idea of the improvement two model plots for the density, *d*, are here shown. Fig. 1 on the left side shows the plot obtained in Ref. 5 with ANN-MLP-{T_b, M, ${}^0\chi^v$, χ_t^v, ${}^1\psi_{Is}$}, and on the right side the plot obtained with the present ANN-MLP -{T_b, M, ${}^AM^v$, D^v, ${}^T\psi_1$}, where ${}^AM^v$ is a MMCI.

The detected asymmetry in the residual plots for the evaluated points (more deviations are located on one side of the zero line than on the other side), is not as drastic as in ref. 5 but it continues to show up. This detail that is probably due to the fact that higher order regressions are needed to model the present properties does not thwart the predictive character of the present relationships.

Before closing this section let us add some words about the descriptive quality of relationships made either with MMCI or with MCI alone [15]: *RI*, *FP*, *UV*, and *El* are advantageously described with equations made of pure MMCI indices, μ is indifferent to the type of indices, while, T_b, d, ϵ, η, γ, and *MS* (- $\chi \cdot 10^6$) are advantageously described with pure MCI descriptors. Needless to say the present semiempirical equations perform much better.

6. Conclusion

Once E. Bright Wilson remarked (cited in ref. 20): "it is always worthwhile to explore a region which is really new. Unexpected results can generally be relied upon under these circumstances". Now, of the two main aims of the present work one is unexpected while the other is partially unexpected: (*i*) the new indices here proposed are really useful, and (*ii*) the ANN methodology gives rise to better estimations than the normal least-squares methods. Even if this was already known from ref. 5, the unexpected finding is that the quality of ANN calculations can be improved if they were allowed to choose, by the aid of a combinatorial search algorithm, the best subset of indices. Present ANN computations rely on prior least-squares-combinatorial calculations that choose the first, second, and third best subset of indices. These three subsets are then passed over to ANN that chooses the optimal subset of indices that is usually better and different from the very best one chosen with least-squares method. The message is that coupling ANN with a combinatorial search algorithm could surely help to improve the modeling relationships. Present paper even if it is not a study into the details and complexity of ANN nonetheless suggests how to improve it.

The implementation brought about by the new mean indices is a good hint that strategies to find new descriptors, even if not always successful, are always worth to trying [16-18]. Thirteen out of thirty-six MMCI show up in our semiempirical equations side-by-side with MCI and experimental parameters. MCI nearly double in number the MMCI, and the five experimental parameters are always a good help. Concerning the MMCIs ${}^AM^v$ shows up four times, ${}^LM_E$, thrice, ${}^HM$, ${}^HM^v$, ${}^GM_I$,

$^{G}M_E$, $^{G}M^{v}$, $^{S}M_E$, $^{R}M^{v}$, $^{Ho}M_E$, and $^{St}M_E$, twice, and $^{Ho}M_I$, and $^{L}M^{v}$, only once. A brief look at these indices shows that they are mainly δ^{v}-dependent (only exception being ^{H}M), either directly or through the intrinsic *I*-State and *electrotopological S*-State indices. This means that, not only they depend on properties of general graphs, but that both the hydrogen contribution and the complete graph contribution for the core electrons (see appendix) play an important role in the descriptive quality of the MMCI. Notice that in eqs. (13) and (17) the simple and seminal Randić $^{1}\chi$ index [1] shows up underlining how the corresponding more complex $^{1}\chi^{v}$ index [2, 9] does not cover all the properties.

The index $^{Ho}M_E$ brings about the greatest improvement in the model of a property, the dielectric constant, ε: with no MMCIs we had, $N(\mathrm{T}) = 61$, $r^2 = 0.933$, $s = 3.9$ [5], while with $^{Ho}M_E$ we have, $N(\mathrm{T}) = 61$, $r^2 = 0.984$, $s = 2.5$. MMCI help to reduce the importance of random variables while MMCI together with the ANN-MLP calculations have no use of them.

Concerning the different types of configurations of the MMCI & MCI indices due to the different types of valence delta, $\delta^{v}{}_{\mathrm{po}\,/\,\mathrm{ppo}}(n)$, it is possible to notice that in MLS calculations half of the properties prefer the *ppo* configuration, while in ANN-MLP calculations five out of eight properties prefer this configuration. Concerning the *n* values of $\delta^{v}{}_{\mathrm{po}/\,\mathrm{ppo}}(n)$ practically all values show up. Thus, general and complete graph characteristics like multiple bonds and core electrons of heteroatoms (here chlorine and bromine), as well as the hydrogen contribution, are an important factor in modeling studies.

Appendix

1. *The Original Means*

In literature [19] the following nine definitions of means between numbers *a* and *b* can be found,

$$\text{Arithmetic mean: } AM = (a+b)/2 \tag{A1}$$

$$\text{Geometric mean: } GM = (ab)^{1/2} \tag{A2}$$

$$\text{Harmonic mean: } HM = 2/(a^{-1} + b^{-1}) \tag{A3}$$

$$\text{Root mean square: } RM = [(a^2 + b^2)/2]^{1/2} \tag{A4}$$

$$\text{Symmetric mean: } SM = (a^2 + b^2)/(a + b) \tag{A5}$$

$$\text{Unsymmetrical mean: } UM = [b - a + (a^2 - 2ab + 5b^2)^{0.5}]/2 \tag{A6}$$

$$\text{Hölder mean : } HoM(p) = (a^p + b^p)^{1/p} / 2 \tag{A7}$$

$$\text{Lehmer's mean: } LM(p) = (a^p + b^p)/(a^{p-1} + b^{p-1}) \tag{A8}$$

$$\text{Stolarsky's mean : } StM(p) = [(a^p - b^p)/(pa - pb)]^{1/(p-1)} \tag{A9}$$

The reader can notice that the Stolarsky's mean has a minus instead of a plus sign in the denominator. The plus sign in eq. 9 was introduced to avoid an undefined value, zero/zero, whenever $\delta_i = \delta_j$ ($a = b$ in A9) had we hold on to the minus sign.

2. *The Valence Delta*

All χ, ψ, Δ, Σ, and $T_{\Sigma/M}$ indices employed in the present study are defined in Table 2 of Ref. 5. Here we will only define some concepts that will help to understand Tables 2 and 3. The δ^v number used throughout present and previous works [5] is defined in the following way,

$$\delta^v = \frac{(q + f_\delta^n)\delta^v(ps)}{(p \cdot r + 1)} \qquad \text{(A10)}$$

$\delta^v(ps)$ is the valence of a vertex in a chemical pseudograph (or general graph) that allows multiple bonds and self-connections (or loops). Normally, in chemical graph theory simple graphs (with no multiple bonds and loops) and general graphs (or pseudographs) are hydrogen-depleted. Parameters p (= 1, 2, 3, 4, ...) is the order of a complete graph, K_p, and r is its regularity ($r = p - 1$). A complete graph is a graph where every pair of its vertices is adjacent. The first order complete graph, K_1, is just a vertex and it is usually used to encode second row atoms. Parameter q in Eq. (A1) is two-valued: $q = 1$ or p. Generally, two representations (or configurations) for δ^v are useful (see Tables 2, and 3): $\delta^v{}_{po}(n)$ where $q = 1$, and $p = odd$, and $\delta^v{}_{ppo}(n)$ where q = p and $p = odd$. Number n that appears in the two deltas is the value of exponent n in f_δ (eq. A10). It quantifies the importance of the hydrogen perturbation: the higher the n values the lower the importance of the perturbation. The values for n here used that generate different sets of indices are: n = - 0.5, 0.5, 1, 2, 5, 50. This parameter could be used as a fine-tuning optimization variable, something like (but not quite) the Randić's variable chi index, [21, 22] that was proposed as an alternative way of characterizing heteroatoms in molecules. The f_δ fractional hydrogen perturbation parameter that encodes the depleted hydrogen atoms is defined in the following way,

$$f_\delta = 1 - \delta^v(ps)/\delta^v{}_m(ps) = n_H / \delta^v{}_m(ps) \qquad \text{(A11)}$$

$\delta^v{}_m(ps)$ is the maximal $\delta^v(ps)$ value a heteroatom (a vertex) can have in a hydrogen depleted chemical pseudograph when all bonded hydrogen atoms are substituted by heteroatoms, and n_H equals the number of hydrogen atoms bonded to a heteroatom. For completely substituted heteroatoms, $f_\delta = 0$ as $\delta^v{}_m(ps) = \delta^v(ps)$ (i.e., $n_H = 0$). In hydrocarbons $\delta^v(ps) = \delta$, which is the delta number in simple chemical graphs with no multiple bonds and loops. In this case: $\delta^v = (1 + f_\delta^n)\delta$ (for $p = 1$). For quaternary carbons $f_\delta = 0$ and $\delta^v = \delta$.

3. ***The Intrinsic I-State and Electrotopological S-State indices***

The *I*- and *E*-State indices (in $\psi_{E,I}$: *E* means Electrotopological, and *I* intrinsic) are related to δ^v in the following way, [10]

$$I = (\delta^v + 1) / \delta, \; S = I + \Sigma\Delta I, \text{ with } \Delta I = (I_i - I_j) / r^2_{ij} \qquad \text{(A12)}$$

r_{ij} counts the atoms in the minimum path length separating atoms *i* and *j*, which equals the graph distance, $d_{ij} + 1$; $\Sigma\Delta I$ incorporates the information about the influence of the remainder of the molecular environment.

References

1 M. Randić, *J. Am. Chem. Soc*. 1975, ***97***, 6609-6615.

2 L.B. Kier, L.H. Hall, W.J. Murray and M. Randic, *J.Pharm.Sci*. 1975, 64, 1971-1974.

3 R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, 2nd edn, Wiley-VCH, Weinheim, 2000.

4 J. Devillers and A.T. Balaban, (Eds.) *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach, UK, 1999.

5 L. Pogliani and J.V. de Julián-Ortiz, *RSC Advances* 2013, **3**, 14710-14721.

6 R. García-Domenech, J. Gálvez, J.V. de Julián-Ortiz and L. Pogliani, *Chem. Rev.* 2008, **108**, 1127-1169.

7 L. Pogliani, *J. Comput. Chem.* 2010, **31**, 295-307.

8 L. B. Kier and L.H. Hall, *J. Pharm. Sci* 1981, **70**, 583-589.

9 L.B. Kier and L.H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Wiley, NY, 1986.

10 L.B. Kier and L.H. Hall, *Molecular Structure Description. The Electrotopological State*, New York: Academic Press, 1999.

11 J.G. Topliss and R.J. Costello, *J. Med. Chem*. 1972, **15**, 1066-1069.

12 E. Besalu, J.V. de Julian-Ortiz and L. Pogliani, *MATCH - Commun.Math.Comput.Chem*. 2006, 55, 281-286.

13 J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design*: *An Introduction*, 2nd Edition, Wiley-VCH: Weinheim, 1999.

14 E Castillo, B. Guijarro-Berdiñas, O. Fontenla-Romero and A Alonso-Betanzos, *J. of Machine Learning Research* 2006, **7**, 1159–1182.

15 L. Pogliani and J. V. de Julián-Ortiz, *Int. J. Chem. Mod.* 2014, in press.

16 R. García-Domenech, J.V. de Julián-Ortiz, M.J. Duart, J.M. García-Torrecillas, G.M. Antón-Fos, I. Ríos-Santamarina, C. de Gregorio Alapont and J. Gálvez, J., *SAR and QSAR Environ. Res*. 2001, 12, 237-254.

17 M.J. Duart, G.M. Antón-Fos, J.V. de Julián-Ortiz, R. Gozalbes, J. Gálvez, and R. García-Domenech, *Int. J. Pharm*. 2002, 246, 111-119.

18 L. Pogliani, J.V. Julian-Ortiz and E. Besalu, *Int. J. Chem. Mod.* 2013, 5, 295-302.

19 Wolfram MathWorld: http://mathworld.wolfram.com/

20 M. Randić, *Chem.Rev.* 2003, 103, 3449-3605 (p. 3470).

21 M. Randić, *Chemom. Intell. Lab. Sys*t. 1991, *10*, 213-227

22 M. Randić, *J. Mol. Graphics Modelling* 2001*, 20*, 19-35.

QSPR with Descriptors based on Averages of Vertex Invariants. An ANN Study

Lionello Pogliani, Jesus Vicente de Julián-Ortiz

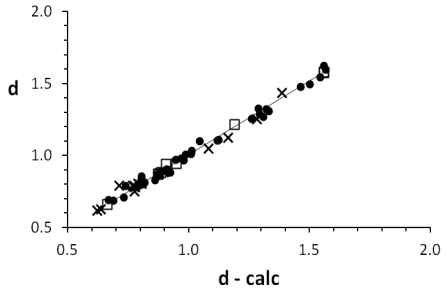
[a] Unidad de Investigación de Diseño de Farmacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, 46100 Burjassot (València), [b]MOLware SL, c/Burriana 36-3, 46005, València, Spain; liopo@uv.es, jejuor@uv.es.

Table of Contents

input units → hidden units → output units

$(T_b, {}^{A}M^{v}, D^{v}, {}^{0}\chi^{v}, {}^{T}\psi_I)$ → ***6 hidden units*** → **d - calc**

(**d** = density experimental values)



Mean molecular connectivity indices (MMCI) defined as averages of vertex invariants together with molecular connectivity indices (MCI) and experimental parameters build optimal semiempirical relationships for eight out of eleven properties of organic solvents studied with artificial neural network.