MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **65** (2011) 347-364

ISSN 0340 - 6253

Two QSPR Methodologies, the Random, and the Super-Descriptors

Lionello Pogliani
Dipartimento di Chimica, Università della calabria,
via P. Bucci 14/C, 87036 Rende (CS), Italy, lionp@unical.it

(Received April 21, 2010)

ABSTRACT

The full combinatorial search methodology has been used with a set of molecular connectivity indices plus five experimental parameters and the molar mass to extract the best descriptors for twelve properties of a set of organic solvents. The performance of the full combinatorial methodology is compared with the performance of the greedy search methodology obtained in a previous paper. The molecular connectivity indices used with both methodologies belong to different configurations as they can encode different hydrogen and core electron contributions and can give rise to configuration-dependent descriptors. The indices of the best configuration-dependent descriptors for the twelve properties, either molecular connectivity indices and/or experimental indices have then been pooled together and used to derive super-descriptors. These super-descriptors achieve, a better description for four properties, among which, an impressive description for the melting points. A thorough investigation has also been performed on the model quality of random indices, which have been used to derive either 'zero-level' descriptors, or semi-random descriptors. This has not only allowed to have a concrete idea of the model quality of random indices but also do draw interesting considerations about the quality of semi-random descriptors that is, descriptors based both on random numbers and on molecular connectivity indices and/or experimental parameters. In fact, a few properties can advantageously be described with this type of descriptors. This last investigation has allowed to better focus the validity of the q^2 leave-one-out statistics and of the Topliss-Costello rule. On the other side, the full combinatorial technique, either with normal descriptors or with super-descriptors has shown the real limits of the greedy search algorithm, has confirmed previous conclusions about the contributions of the hydrogen atoms, and has underlined the importance of pseudoconnectivity indices, experimental indices, and of some *ad hoc* parameters.

INTRODUCTION

In a recent paper,[1] the satisfactory QSPR model of twelve properties of a highly heterogeneous class of organic solvents has been done with three different model strategies using the *greedy* algorithm, which is a forward combinatorial search that consists to add the next best index keeping constant the previous ones. Normally, this search algorithm that

drastically reduces the number of combinations to be searched, gives satisfactory results, but it is not guaranteed that it finds the best descriptor, a drawback that all inclusion stepwise methods share. Furthermore, it is not evident how far from optimal the found descriptor is. The *greedy* model strategy, which gave rather good results, was performed with a set of thirty configuration-dependent molecular connectivity indices shown in Table 1, where the Δ index encodes the number of electronegative atoms (n_{EA}), while the Σ index encodes the sum of the *S*-State index for the electronegative atoms: *N*, *O*, *F*, *Cl*, *Br* ($<S_{EA}>$ is the average value for a specific type of atom, sulphur has not been considered as an electronegative atom). To these indices the molar mass, *M*, was added, and for the boiling and melting points as well as for the dipole moment three *ad hoc* parameters were introduced: AH^b, AH^m, and ϕ (0, 1), respectively. These parameters helped to take care of the hydrogen bond problem and/or structural problems. It was also proposed and used a model *greedy* strategy that included five experimental parameters as indices: T_b, the boiling temperature; T_m, the melting temperature; ε, the dielectric constant; *d*, the density, and *RI*, the refractive index of the organic solvents. The model of the surface tension, for instance, could only be achieved by the aid of five experimental parameters, while the model of other six properties could be achieved with semiempirical descriptors made of molecular connectivity indices and experimental parameters.

Table 1. Definition of the *MCI* indices used in this study.

MCI	*pMCI*	*Dual MCI*	*Dual pMCI*
$D = \Sigma_i \delta_i$	${}^{S}\psi_1 = \Sigma_i I_i$	${}^{0}\chi_d = (-0.5)^N \Pi_i(\delta_i)$	${}^{0}\psi_{Id} = (-0.5)^N \Pi_i(I_i)$
${}^{0}\chi = \Sigma(\delta_i)^{-}$	${}^{0}\psi_1 = \Sigma(I_i)^{-}$	${}^{1}\chi_d = (-0.5)^{(N+\mu-1)} \Pi(\delta_i$	${}^{1}\psi_{Id} = (-0.5)^{(N+\mu-1)} \Pi(I_i$
${}^{1}\chi = \overset{\wedge}{\Sigma}$	${}^{1}\psi_1 = \overset{\wedge}{\Sigma}$	${}^{1}\chi_s = \Pi(\delta_i + \delta_j)^{-0.5}$	${}^{1}\psi_{Is} = \Pi(I_i + I_j)^{-0.5}$
$\chi_t = (\overset{\wedge}{\Pi}\delta_i)^{-}$	${}^{T}\psi_1 = (\overset{\wedge}{\Pi} I_i)$	$\Delta = \Sigma_{EA} n_{EA}$	$\Sigma = \Sigma_{EA} <S_{EA}>$

N is the number of atoms, *ij* means σ bond, μ is the cyclomatic number. [1, 4, 9] Replacing δ with δ^v and *I* with *S* the valence χ^v indices and ψ_E indices are obtained for a total of twenty-eight χ-ψ MC indices. Δ and Σ are special indices (see Introduction and method sections).

Present study uses the same indices and experimental parameters, but, here, the model strategy uses the full combinatorial algorithm, which in many cases encompasses a search over tens of millions of combinations. The purpose of the present study is (*i*) to obtain the overall best descriptor for each property, (*ii*) to compare it with the *greedy* descriptor, (*iii*) to check how random or 'zero-level' descriptors made of random indices work, (*iv*) to check how semi-random descriptors, a mixing of random indices plus MC indices and experimental parameters work, (*v*) to see if it is possible to derive from the best configuration-dependent

descriptors improved super-descriptors, (*v*) to check the importance of the hydrogen perturbation and how it depends on the type of search algorithm. Concerning point (*i*) let us remind that while the number of six-index combinations from a set of 36 indices (theoretical plus the empirical indices, plus *M*) entails millions of combinations, the number of greedy combinations for the same case entails only hundreds of combinations. Point (*ii*) will allow to measure the real utility of the *greedy* method and to check the importance of semiempirical graph-theoretical methods. Point (*iii*) and (*iv*) will allow not only to characterize a *zero-level* description for each property, but to check under which conditions sets of descriptors made of random indices plus experimental and/or graph-theoretical indices may describe a property. There are some 'urban legends' that try to put on an equal foot the model quality of random and of molecular connectivity indices, it is then highly interesting to know, if, how and when random indices work. As many subjects in physics (quantum gravity, quantum chromodynamics, disordered physical systems, amorphous materials), use randomized matrices,[2] it is a no minor task to check the role of random numbers in computational chemistry.

METHOD

The valence delta number, δ^v, is the basic parameter for the valence MCI (χ^v), and for the *I*- and *E*-State pesudoconnectivity indices. In fact, *I* and *S* are δ^v-dependent:[3]

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

Here, 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, and, as it can be negative, *S* can also be negative. To avoid imaginary ψ_E values, every *S* value (as some atoms have $S < 0$) has to be rescaled.[4, 5] Throughout the present model and as already done in Ref. 1, the rescaling value is 6.611. This rescaling procedure brings about that ${}^S\psi_I \neq {}^S\psi_E$, while, normally, the electrotopological state concept implies $\Sigma_i S_i = \Sigma_i I_i$.

Our δ^v number shown in eq. (2) takes care of the core electrons, by the aid of complete graphs, and of the depleted hydrogen atoms, by the aid of a perturbation parameter,[1, 5-9]

$$\delta^v = \frac{(q + f^n_\delta)\delta^v(ps)}{(p \cdot r + 1)} \qquad (2)$$

Here, $\delta^v(ps)$ is the valence delta for a chemical pseudograph (or general graph) only. Parameters p is the order of a complete graph, K_p, which is a graph where every pair of its vertices is adjacent, and r r ($= p - 1$) is its regularity.[10] A K_p is always r-regular, i.e., all its vertices have the same r. The first order complete graph, a K_1 graph (a vertex), has normally been used to encode second row atoms, and especially the carbon atom. Usually, $q = 1$ or p, where p can be either odd-valued (p = 1, 3, 5, 7,..) or sequential-valued (p = 1, 2, 3, 4,...). Four representations for δ^v are possible: K_p-(p-*odd*) for $q = 1$, and p = *odd*; K_p-(p-*seq*) for q = 1, and p = *seq* (sequential); K_p-(*pp-odd*) for $q = p$, and p = *odd*, and K_p-(*pp-seq*) for $q = p$, and p = *seq*. To keep the number of MCI indices under control only two representations have been chosen as in ref. 1: the K_p-(p-*odd*) representation with the smallest δ^v values (p = 1, 3, and 5 and $q = 1$), and the K_p-(*pp-odd*) representation with the second largest δ^v values, that is, the (p = 1, 3, and 5 and $q = p$). The f_δ fractional perturbation parameter encodes the depleted hydrogen atoms (or K_0 null complete graphs), and is defined in the following way,

$$f_\delta = [\delta^v{}_m(ps) - \delta^v(ps)] / \delta^v{}_m(ps) = 1 - \delta^v(ps)/\delta^v{}_m(ps) = n_H / \delta^v{}_m(ps) \qquad (3)$$

Here, $\delta^v{}_m(ps)$ 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 δ and δ^v are no more equal, as $\delta^v(ps) = \delta$, but, $\delta^v = (1 + f_\delta^n)\delta$ (with, $p = 1$). For quaternary carbons $f_\delta = 0$ and $\delta^v = \delta$. Exponent n quantifies the importance of the perturbation, i.e., the higher the n values the lower the importance of the perturbation. It is no optimization parameter, as different values of n give rise to different sets of indices, where in each set n is constant, and, consequently, the corresponding δ^v is constant. In this study, n = - 1, - 0.5, - 0.1, 0.1, 0.5, 1, 2, 5, 8, 50. Relatively to our greedy study two new n values are here considered, n = -0.1 and 0.1.

Table 2. The f_δ^n values for n = 8, 5, 2, 1, 0.5, 0.1, - 0.5, and -1 for hydrogenated carbon, nitrogen and oxygen atoms. ≡*CH* means: →*CH,* ≥*CH*, and ≡*CH*, while =CH_2 means: >CH_2 and =CH_2.

Groups	*8**	*5*	*2*	*1*	*0.5*	*0.1*	*-0.1*	*-0.5*	*-1*
≡ *CH*	$1.5 \cdot 10^{-5}$	0.001	0.06	0.25	0.5	0.87	1.15	2.00	4
= CH_2	0.004	0.031	0.25	0.50	0.71	0.93	1.07	1.41	2
- CH_3	0.100	0.237	0.56	0.75	0.87	0.97	1.03	1.15	1.333
- *OH*	0	0.0001	0.03	0.17	0.41	0.84	1.20	2.45	6
= *NH*	0	0.0003	0.04	0.20	0.45	0.85	1.17	2.24	5
- NH_2	0.001	0.010	0.16	0.4	0.63	0.91	1.10	1.58	2.5

*The n = 50 values correspond to zero perturbation, and they are similar to the values in this column.

Table 2 shows the f_δ^n values for n < 50 (n = 50 means no perturbation) for hydrogenated carbon, nitrogen and oxygen atoms. Here, ≡CH means the three ways a carbon atom can be thrice-bonded: →CH (three σ-bonds), ≥CH (a σ-bond and a π-bond), and ≡CH (a triple π-bond); =CH_2 means the two ways a carbon atom can be twice-bonded: >CH_2 (two σ-bonds), and =CH_2 (a π-bond). From now, indices which have been derived either with a K_p(*p-odd*) or with a K_p(*pp-odd*) representation for the core electrons and with a particular type of f_δ^n hydrogen perturbation will belong to either to the K_p(*p-odd*) / f_δ^n or K_p(*pp-odd*) / f_δ^n configuration.

Fig. 1 top shows the (hydrogen depleted) simple graph of Br-CH=CH-Br from which it is possible to derive the δ values (here: 1, 2, 2, 1). Fig. 1 middle shows the pseudograph (or general graph) of the same molecule from which it is possible to derive δ^v(ps) values (here: 7, 3, 3, 7), and Fig. 1 bottom shows the chemical pseudograph plus complete graph of the same molecule whose valence δ^v are: 7/21, 3.1875, 3.1875, 7/21, where, the core electrons of Br have been encoded with K_5 odd complete graphs (p = 5, q = 1, and n_H = 0), while the two carbon atoms throughout the three graphs have been encoded with K_1 vertices and their valence delta values have been calculated with a f_δ^2 hydrogen perturbation (n_H = 1, $\delta^v{}_m(ps)$ = 4).

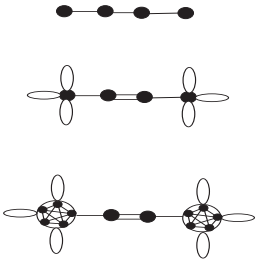

Fig. 1. The simple (top) graph, the pseudograph, and the pseudograph plus K_5 complete graph (bottom) of Br-CH=CH-Br.

All indices in Table 1 can be considered molecular connectivity indices, as, formally, all of them have their origin in Randić's index.[11-13] MCI of subset $\{D, {}^0\chi, {}^1\chi, \chi_t, {}^0\chi_d, {}^1\chi_d, {}^1\chi_s, \Delta\}$ are independent of the hydrogen content of the solvent molecules as well as from the complete graph representation for the core electrons. To the thirty indices of Table 1, the molar mass, M, and the experimental parameters $\{T_b, T_m, \varepsilon, d, RI\}$ are be added for a total of thirty-six indices (graph-theoretical plus empirical indices). The huge number of possible combinations these indices give rise should be multiplied by twenty as two are the representations for the core electrons: K_p-(*p-odd*) and K_p-(*pp-odd*). For each representation, there are ten different values for f_δ^n. The full combinatorial space for the best descriptor has been searched with the help of Statistica 6.0 of Statsoft Inc.

A particular attention in this study will be given to the Topliss-Costello rule, i.e., at least $r > 0.84$ ($r^2 > 0.71$) for a ratio $N°$ data/$N°$ indices ≥ 5.[14] Usually, r^2 is given as it can be compared with the q^2 leave-one-out statistics. The correlation parameters (c_i) of the regressions together with their errors s_i (see ref. 15 for their relevance) are collected into the vector, $\boldsymbol{C} = [c_1(s_1), c_2(s_2), \ldots, c_n(s_n), c_0(s_0)]$. Vector $\boldsymbol{C}$ can be read as the ordered list of regression values of a property P with respect to the vector of the MC indices, $\boldsymbol{\chi} = (\chi_1, \chi_2, \ldots, \chi_n, \chi_0 \equiv 1)$, and the property can be computed as the scalar product of the row vector $\boldsymbol{C}$ (the c_i only) with the column $\boldsymbol{\chi}$ vector: $P = \boldsymbol{C}\cdot\boldsymbol{\chi}$. Observed vs. calculated plots for all models can be obtained from the author (as well as the index values), as sometimes 'good' statistics hide rather poor plots.[5, 16-19] The prediction q^2 coefficient, used to check the validity of the leave-one-out method will also be given. Each observation is removed one at a time and during the removal, it is assumed that the descriptor does not change. The prediction coefficient q^2 equals (SD- $PRESS$)/SD, where $SD = \Sigma(y_i - \langle y \rangle)^2$ is the squared deviation of the observed value from their mean, and $PRESS = (y_i - y_{iloo})^2$, where y_{iloo} is a predicted value of the studied property where the prediction has been made by the leave-one-out method.[13, 20] It has been suggested that $q^2 > 0.6$. The present study will show that this is a too optimistic choice. The leave-one-out method has some drawbacks with small data sets and with strong clusterization, which is here rarely the case.[21, 22]

The model quality of the optimal descriptor of the given properties has been previously tested on a smaller training set of compounds by leaving out those compounds with '°' in Table 3, and the chosen descriptor is both the best descriptors for both the training and the full set of compounds. Strong outliers, as few as possible, whenever advantageous, were excluded from the model. The super-descriptors have been obtained with a search over the set of

indices of the optimal configuration-dependent descriptors for the twelve properties. The best 'zero-level' descriptors have been searched among two thirty-eight sets, *r1-r38*, and *rd1-rd38* of 0-1 random numbers. A combinatorial space made of the *r1-r38* and then (due to the limits of our PC) of the *rd1-rd38* sets of random indices plus the given MC indices, the five experimental parameters, and *M* is searched for the best semi-random or *mc-exp-rn*-descriptor (*mc* means molecular connectivity, *exp* experimental and *rn* random). Whenever it is not computationally feasible, only the indices of the best descriptors will be added to the *r1-r38* and/or *rd1-rd38* random sets. Random numbers have been obtained with the algorithm of the 2003 Microsoft Excel electronic sheet. The strongest correlations ($r > 0.98$ [23]) among the indices of the descriptor will be given to check, about the possibility of a loss of meaning of the structure-property relationship. In this case, it might be advantageous to work with orthogonalized regressions, which can easily be obtained with the Randić's stepwise orthogonalization method, as performed in Ref. 1.[24, 25] It has been suggested that even strongly correlated indices are not that damaging.[2]

Table 3. The properties of organic solvents plus their molar mass M (g·mol^{-1}): T_b, boiling points (K, in parenthesis AH^b values); T_m, melting points (K, in parenthesis AH^m values); *RI*, refractive index (20°C); *d*, density (at 20°C±5°C relative to water at 4°C, g/cc); ε, dielectric constant; *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* ($-\chi \cdot 10^6$), magnetic susceptibility (emu mol^{-1}, 1 emu = 1 cm^3, temperatures cover a range from 15°C to 32°C); and *EV*, Elutropic value (silica).

Solvents	*M*	T_b	T_m	*RI*	*d*	ε	*FP*	η	γ	*UV*	μ	*MS*	*EV*
(°)Acetone	58.1	329	179	1.359	0.791	20.7	256	0.32	23.46	330	2.88	0.46	0.43
(°)Acetonitrile	41.05	355 *(1)*	225 *(1)*	1.344	0.786	37.5	278	0.37	28.66	190	3.92	0.534	0.50
Benzene	78.1	353	278 *(1)*	1.501	0.84	2.3	262	0.65	28.22	280	0	0.699	0.27
Benzonitrile	103.1	461	260	1.528	1.010	25.2	344	1.24[1]	38.79				
1-Butanol	74.1	391 *(1)*	183 *(1)*	1.399	0.810	17.1	308	2.95	24.93	215			
(°)2-Butanone	72.1	353	186	1.379	0.805	18.5	270	0.40	23.97	330			0.39
Butyl Acetate	116.2	398	195 *(-1)*	1.394	0.882	5.0	295	0.73	24.88	254			
CS_2	76.1	319	161	1.627	1.266	2.6	240	0.37	31.58	380	0	0.532	
CCl_4	153.8	350	250 *(1)*	1.460	1.594	2.2		0.97	26.43	263	0	0.691	0.14
Cl-Benzene	112.6	405	228	1.524	1.107	5.6	296	0.80	32.99	287			
1Cl-Butane	92.6	351	150	1.4024	0.886	7.4	267	0.35	23.18	225			
$CHCl_3$	119.4	334	210	1.446	1.492	4.8		0.57	26.67	245	1.01	0.740	0.31
Cyclohexane	84.2	354	280 *(1)*	1.426	0.779	2.0	255	1.00	24.65	200	0	0.627	0.03
(°)Cyclopentane	70.1	323	179 *(1)*	1.400	0.751	2.0	236	0.47	21.88	200		0.629	
1,2-diCl-Benzene	147.0	453	257	1.551	1.306	9.9	338	1.32		295	2.50	0.748	
1,2-diCl-Ethane	98.95	356	238	1.444	1.256	10.4	288	0.79	31.86	225	1.75		
diCl-Methane	84.9	313	176	1.424	1.325	9.1		0.44	27.20	235	1.60	0.733	0.32
N,N-diM-Acetamide	87.1	438 *(1)*	253 *(1)*	1.438	0.937	37.8	343			268	3.8		
N,N-diM-Formamide	73.1	426 *(1)*	212 *(1)*	1.431	0.944	36.7	330	0.92		268	3.86		
1,4-Dioxane	88.1	374	285	1.422	1.034	2.2	285	1.54	32.75	215	0.45	0.606	
Ether	74.1	308	157	1.353	0.708	4.3	233	0.24	16.95	215	1.15		0.29
Ethyl acetate	88.1	350	189 *(-1)*	1.372	0.902	6.0	270	0.45	23.39	260	1.8	0.554	0.45
(°)Ethyl alcohol	46.1	351 *(1)*	143	1.360	0.785	24.3	281	1.20	21.97	210	1.69	0.575	
Heptane	100.2	371	182	1.387	0.684	1.9	272		19.65	200			0.00
Hexane	86.2	342	178	1.375	0.659	1.9	250	0.33	17.89	200			0.00
2-Methoxyethanol	76.1	398 *(1)*	188	1.402	0.965	16.0	319	1.72	30.84	220			
(°)Methyl alcohol	32.0	338 *(1)*	175 *(1)*	1.329	0.791	32.7	284	0.60	22.07	205	1.70	0.530	0.73
(°)2-Methylbutane	72.15	303	-	1.354	0.620	1.8	217						
4-Me-2-Pentanone	100.2	391	193	1.396	0.800	13.1	286			334			
2-Me-1-Propanol	74.1	381 *(1)*	165 *(1)*	1.396	0.803	17.7	310					0.534	
2-Me-2-Propanol	74.1	356 *(1)*	298 *(2)*	1.387	0.786	10.9	277		19.96		1.66		
DMSO	78.1	462 *(2)*	292 *(2)*	1.479	1.101	46.7	368	2.24	42.92	268	3.96		
(°)Nitromethane	61.0	374	244	1.382	1.127	35.9	308	0.67	36.53	380	3.46	0.391	

1-Octanol	130.2	469 *(1)*	258 *(1)*	1.429	0.827	10.3	354	10.6	27.10				
(°)Pentane	72.15	309	143	1.358	0.626	1.8	224	0.23	15.49	200			0.00
3-Pentanone	86.1	375	233	1.392	0.853	17.0	279		24.74				
(°)1-Propanol	60.1	370 *(1)*	146 *(1)*	1.384	0.804	20.1	288	2.26	23.32	210			
(°)2-Propanol	60.1	356 *(1)*	184 *(1)*	1.377	0.785	18.3	295	2.30	20.93	210			0.63
Pyridine	79.1	388	231	1.510	0.978	12.3	293	0.94	36.56	305	2.2	0.611	0.55
tetraCl-Ethylene	165.8	394	251	1.506	1.623	2.3		0.90				0.802	
(°)tetra-Hydrofuran	72.1	340	165	1.407	0.886	7.6	256	0.55		215	1.75		0.35
Toluene	92.1	384	180	1.496	0.867	2.4	277	0.59	27.93	285	0.36	0.618	0.22
1,1,2triCl,triFEthane	187.4	321	238	1.358	1.575	2.4		0.69		230			0.02
2,2,4-triMe-Pentane	114.2	372	166 *(-1)*	1.391	0.692	1.9	266	0.50		215			0.01
o-Xylene	106.2	417	249	1.505	0.870	2.6	305	0.81	29.76				
p-Xylene	106.2	411	286	1.495	0.866	2.3	300	0.65	28.01				
(°)Acetic acid	60.05	391 *(1)*	290 *(1)*	1.372	1.049	6.15			27.10		1.2	0.551	
Decaline	138.2	465	230	1.476	0.879	2.2						0.681	
diBr-Methane	173.8	370	221 *(1)*	2.497	1.542	7.8			39.05		1.43	0.935	
1,2-diCl-Ethylen(Z)	96.9	334	193	1.449	1.284	9.2					1.90	0.679	
(°)1,2-diCl-Ethylen(E)	96.9	321	223	1.446	1.255	2.1					0	0.638	
1,1-diCl-Ethylen	96.9	305	151 *(-1)*	1.425	1.213	4.7					1.34	0.635	
Dimethoxymethane	76.1	315	168 *(-1)*	1.356	0.866	2.7						0.611	
(°)Dimethylether	46.1	249 *(-0.5)*	134 *(-0.5)*			5.0							
Ethylen Carbonate	88.1	511 *(2)*	310 *(2)*	1.425	1.321	89.6					4.91		
(°)Formamide	45.0	484 *(2)*	276 *(2)*	1.448	1.133	109			57.03		3.73	0.551	
(°)Methylchloride	50.5	249 *(-0.5)*	175 *(-0.5)*	1.339	0.916	12.6					1.87		
Morpholine	87.1	402 *(1)*	270 *(1)*	1.457	1.005	7.3						0.631	
Quinoline	129.2	510	258	1.629	1.098	9.0			42.59		2.2	0.729	
(°)SO_2	64.1	263 *(-1)*	200		1.434	17.6					1.6		
2,2-tetraCl-Ethane	167.8	419	229	1.487	1.578	8.2			35.58		1.3	0.856	
tetraMe-Urea	116.2	450	272	1.449	0.969	23.1					3.47	0.634	
triCl-Ethylen	131.4	360	200	1.480	1.476	3.4						0.734	

(°) compounds excluded for a checking of the model.

RESULTS AND DISCUSSION

Tables 3 and 4 show the values of the twelve properties (for the origin of the values see 1). Table 5 shows the best *greedy* descriptions obtained in Ref. 1. Superscripts on the left and right side of a combination means the type of configuration, for instance, superscript *f1* means that the (δ^v-based) valence MC indices have been obtained with the f_δ^1 hydrogen perturbation, while superscript *pp-odd* on the right means that these indices have been obtained with the K_p(*pp-odd*) representation for the core electrons where p = odd and $q = p$. Notice that even with the full combinatorial algorithm, T_b and T_m and μ properties need, for an optimal model, their corresponding *ad hoc* parameters, AH^b, AH^m, and ϕ(0, 1), respectively. The best descriptors for the twelve properties are collected in Table 6, the zero-level descriptors in Table 7, and some good and also very good semi-random descriptors are shown in Table 8.

Table 5. The *greedy* descriptors for the twelve properties (P) of the organic solvents, together with their statistics. Superscripts specify the configuration of the δ^v-based MC indices (N = nº of points).

P	*Greedy Descriptors*	N	r^2	s	q^2
T_b	$^{f1}\{D^v, {}^0\psi_{Id}, \varepsilon, M, AH^b\}^{pp\text{-}odd}$	63	0.912	17.8	0.826
T_m	$^{f8}\{D^v, {}^0\chi_d, {}^0\psi_{Ed}, {}^1\chi, {}^0\psi_{Id}, AH^m\}^{pp\text{-}odd}$	62	0.741	25	0.654
ε	$^{f8}\{\Sigma^3/M^{1.7}, T_b, {}^1\chi_s\}^{p\text{-}odd}$	62	0.903	5.3	0.867
d	$^{f8}\{{}^0\chi^v/M, {}^T\psi_I/M, \Delta, T_b\}^{pp\text{-}odd}$	62	0.972	0.04	0.965
RI	$^{f5}\{\chi_t^v, D^v, {}^1\psi_{Is}, {}^S\psi_I, {}^1\chi_s\}^{p\text{-}odd}$	61	0.922	0.04	0.861

FP	$^{f\text{-}0.5}\{T_b, {}^0\psi_E, \Delta, {}^S\psi_I, {}^1\chi^v\}^{p\text{-}odd}$	41	0.984	4.7	0.979
η	$^{f\text{-}0.5}\{{}^0\psi_{Ed}, {}^1\psi_{Ed}, \Sigma, {}^0\psi_I\}^{p\text{-}odd}$	39	0.929	0.5	0.839
γ	$\{T_b, d, \varepsilon, RI, M\}$	40	0.916	2.5	0.847
UV	$^{f50}\{1/RI, ({}^S\psi_I/M)^2, (1/\eta)^{0.5}, ({}^T\psi_I/M)^2\}^{pp\text{-}odd}$	33	0.891	17.8	0.854
μ	$^{f50}\{\phi(\Sigma\cdot T_b)/M^{0.5}, \phi(\Sigma/M)^{2.3}, \phi(\varepsilon/(\Sigma+5))^{0.6}, \phi(\Sigma\cdot{}^0\psi_{Id})^{0.4}\}^{pp\text{-}odd}$	34	0.903	0.4	0.866
$-\chi\cdot10^6$	$^{f50}\{(M+\varepsilon^{0.5}+T_b{}^{0.7}), ({}^S\psi_I+5RI), ({}^0\chi_d{}^v/M\varepsilon), ({}^1\psi_{Ed}/\varepsilon\}^{pp\text{-}odd}$	32	0.885	0.04	0.854
EV	$^{f0.5}\{{}^1\chi^v, T_b, \Sigma, RI\}^{p\text{-}odd}$	20	0.920	0.07	0.820

***Boiling Points*, T_b.**

The ***C*** vector of descriptor in Table 6 is: [0.88 (0.1), 44.1 (4.2), 0.68 (0.07), 26.1 (1.9), -295 (31), 2.27 (0.6), 195.2 (7)]. The quality of the training set obtained excluding items (°) in Table 3 is: N = 45, r^2= 0.939, s = 14, q^2 = 0.802, F = 97. The strongest correlated indices instead are: $r(D, {}^1\psi_E)$ = 0.97, while the strongest correlation of T_b is: $r(T_b, D)$ = 0.71. The search for a six-index descriptor is outside the range of our PC. If the search for a six-index descriptor is restricted to *r1-r38* plus *rd1-rd38* and the indices of descriptor, $^{f50}\{\varepsilon, AH^b, M, D, {}^1\psi_E, \Sigma\}^{pp\text{-}od}$, we

land on the same six-index descriptor. The search for a six-index semi-random descriptor with *r1-r38* and *rd1-rd38* plus indices of the set, $\{\varepsilon, AH^b, M, D\}$, whose quality is: r^2 = 0.870, s = 21, q^2 = 0.826, lands on the descriptor shown in Table 8.

Table 6. The full combinatorial descriptors for the twelve properties (*P*) and their statistics.

P	*Full Descriptors*	*N*	r^2	*s*	q^2
T_b	$^{f50}\{\varepsilon, AH^b, M, D, {}^1\psi_E, \Sigma\}^{pp\text{-}odd}$	63	0.951	13.4	0.933
T_m	$^{f\text{-}0.5}\{T_b, AH^m, \chi_t, {}^0\chi_d, {}^1\chi_d, {}^1\psi_E, \Delta, \Sigma\}^{p\text{-}odd}$	62	0.799	22.0	0.698
ε	$^{f8}\{T_b, {}^0\psi_E, T_{\Sigma/M}\}^{p\text{-}odd}$	62	0.916	4.9	0.902
d	$^{f50}\{D/M, {}^S\psi_I/M, {}^0\psi_E/M, \Delta\}^{pp\text{-}odd}$	62	0.975	0.4	0.969
RI	$^{f5}\{M, D, {}^1\chi_s, \chi_t{}^v, {}^0\psi_I, {}^1\psi_{Is}, \Delta\}^{p\text{-}odd}$	61	0.944	0.04	0.915
FP	$^{f\text{-}0.5}\{T_b, d, RI, {}^S\psi_I, {}^S\psi_E\}^{p\text{-}odd}$	41	0.984	4.7	0.979
η	$^{f\text{-}0.5}\{T_b, \varepsilon, {}^1\psi_{Id}, {}^0\psi_{Ed}, \Sigma\}^{p\text{-}odd}$	39	0.963	0.3	0.863
γ	$^{f1}\{\varepsilon, d, RI, \chi_t{}^v, {}^T\psi_I\}^{p\text{-}odd}$	41	0.959	1.8	0.915
UV	$^{f1}\{RI, AH^b, {}^0\psi_I, {}^S\psi_E, {}^0\psi_E\}^{p\text{-}odd}$	33	0.876	19.4	0.802
μ	$^{f0.5}\{\phi\varepsilon, \phi^I\chi, \phi D^v, \phi\Sigma, \phi T_{\Sigma/M}\}^{pp\text{-}odd}$	34	0.926	0.4	0.818
$-\chi\cdot10^6$	$^{f50}\{M, \chi_t, {}^0\chi^v, {}^1\chi_s{}^v\}^{p\text{-}odd}$	32	0.848	0.5	0.789
EV	$^{f0.5}\{{}^1\chi^v, {}^0\psi_I, {}^0\psi_E, \Sigma\}^{pp\text{-}odd}$	20	0.949	0.06	0.906

Table 7. The full combinatorial zero-level descriptors on the *r1-r38* plus *rd1-rd38* space.

P	*Zero-level Descriptor*	*N*	r^2	*s*	q^2
T_b	{*rd2*, *rd6*, *rd9*, *rd24*, *rd34*, *rd37*}	63	0.410	45	0.255
T_m	{*r1*, *r23*, *r27*, *r29*, *r30*, *r31*, *r35*, *r38*}	62	0.427	37	0.252
ε	{*rd9*, *rd14*, *rd16*, *rd20*, *rd27*, *rd32*, *rd34*, *rd37*}	62	0.479	13	0.265
d	{*rd21*, *rd36*, *r1*, *r348*}	62	0.336	0.2	0.213
RI	{*r1*, *r5*, *r13*, *r20*, *r30*, *r35*, *r36*}	61	0.328	0.1	0.14
FP	{*rd6*, *rd11*, *rd16*, *rd19*, *rd20*}	41	0.702	20	0.605
η	{*rd8*, *rd28*, *rd34*, *r9*, *r20*}	39	0.567	1.2	0.01
γ	{*rd27*, *r10*, *r27*, *r30*, *r35*}	41	0.677	4.9	0.545
UV	{*rd11*, *r13*, *r21*, *r27, r38*}	33	0.649	33	0.480
μ	{*rd9*, *rd14*, *rd16*, *rd24*, *rd28*}	34	0.516	1.0	0.326
$-\chi \cdot 10^6$	{*rd7*, *rd16*, *rd21*, *r25*}	32	0.568	0.1	0.379
EV	{*rd14*, *rd22*, *rd25*, *r30*}	20	0.864	0.1	0.751

***Melting Points*, T_m.**

The ***C*** vector of descriptor in Table 6 is: = [0.21 (0.08), 34.1 (5.2), 63.9 (26), 32.5 (6.5), 0.014 (0.003), 109 (25), 7.15 (2.6), - 1. 72 (0.4), 43.2 (35)]. The training set without items (°) has: N = 45, r^2=0.833, s = 20, q^2= 0.684, F = 22. The restricted search for a six-index descriptor using the *r1-r38* and then *rd1-rd38* sets plus the six the indices of set ${}^{f50}\{AH^m, D, \chi_t, {}^0\chi_d, {}^1\chi_d, {}^v\chi_t\}^{pp\text{-odd}}$ confirm these six indices.

***Dielectric Constant*, ε.**

Ethylencarbonate is a strong outlier and it was left out of the model. The ***C*** vector of descriptor in Table 6 is: = [0.11 (0.02), - 8.22 (1.5), 25.4 (1.8), - 19.9 (4.8)]. Leaving out compounds with (°) we have (no ethylencarbonate): N = 44, r^2 = 0.883, s = 3.7, q^2 = 0.846, F = 100. An interesting correlations is: $r(\varepsilon, T_{\Sigma/M})$ = 0.93. Notice that the $T_{\Sigma/M}$ term encodes the information about the molar mass but also about the overall *E*-State index of the electronegative atoms, i.e., it reflects the charge distribution due to these atoms, normalized to the molar mass. The search for a three-index *mc-exp-rn*-description (with no Ethylencarbonate) with *r1-r38* plus *rd1-rd38* plus the *MCI*- K_p(*p-odd*) / f_δ [8] indices ($T_{\Sigma/M}$ inclusive) finds the already found three-index descriptor. If the search for a five-index descriptor is done with the previous three-index descriptor plus all random indices, the interesting descriptor of Table 8 is found. Leaving out compounds with (°, no ethylencarbonate) the semi-random descriptor shows: N = 44, r^2 = 0.897, s = 3.5, q^2 = 0.854, F = 66. For the final model, we prefer to stick to the unique three-index descriptor. These

results are illuminating about the potential uses of random numbers, when they can count on good indices.

***Refractive Index*, *RI*.**

The ***C*** vector of descriptor in Table 6 is: [0.0019(0.0006), 0.054 (0.006), 0.88 (0.1_5), 0.82 (0.05), - 0.17 (0.02), - 2.73 (0.3), - 0.043 (0.01), 1.41 (0.03)]. The training test that excludes items (°) shows: $N = 45$, r^2=0.957, $s = 0.04$, $q^2 = 0.878$, $F = 117$. The search for a five-index *mc-exp-rn*-description throughout *r1-r38* and then *rd1-rd38* plus the MC- K_p*(p-odd)/*f_δ^5 indices and *M* finds the same five-index descriptor already found. If the search for a seven-index semirandom descriptor is done with the five indices of descriptor ${}^{f5}\{M, D, {}^1\chi_s, \chi_t^v, {}^0\psi_I, {}^1\psi_{Is}\}^{\text{p-odd}}$ ($N = 61$, r^2= 0.887, $s = 0.05$, $q^2 = 0.792$, $F = 86$) plus *r1-r38* and then *rd1-rd38* we obtain the best q^2 descriptor of Table 8, which confirms nonetheless the five MC indices. Excluding items (°) in Table 3 we have: $N = 45$, r^2=0.954, $s = 0.04$, $q^2 = 0.911$, $F = 108$.

***Flash Point*, *FP*.**

FP is the lowest temperature at which there is enough fuel vapor to ignite. It is here possible to use as indices parameters T_b, ε, *d*, and *RI*, but no T_m as it has no data for 2-methylbutane. The vector ***C*** of descriptor in Table 6 is: [0.85 (0.03), 34.6 (7.9),- 95.6 (18), 1.71 (0.2), 1.06 (0.1), 95.2 (22)]. Notice that $r(FP, T_b) = 0.93$. The training test has: $N = 29$, r^2=0.985, $s = 4.3$, $q^2 = 0.981$, $F = 362$. The search for a *mc-exp-rn*-description with *r1-r38* and then *rd1-rd38* plus T_b, *d*, *RI* and the set of MC indices finds the optimal descriptor of Table 8, where only ${}^S\psi_I$ is configuration-dependent. Notice that ${}^{f-0.5}\{T_b, D, {}^S\psi_I, \Delta\}^{p\text{-}odd}$ is a quite good descriptor with: $r^2 = 0.982$, $s = 5.0$, $q^2 = 0.976$. Excluding items (°) in Table 3 the semi-random descriptor shows: $N = 29$, $r^2 = 0.989$, $s = 4.0$, $q^2 = 0.982$, $F = 405$.

Table 8. The best semi-random descriptors obtained with the full combinatorial method applied to the *r1-r38* plus *rd1-rd38* indices plus some of the indices of the best descriptors.

P	*Semi-random Descriptor*	*N*	r^2	*s*	q^2
T_b	$\{\varepsilon, AH^b, M, D, rd14, rd25\}$	63	0.901	19	0.858
T_m	None found (due to PC limits)				
ε	${}^{f8}\{T_b, {}^0\psi_E, T_{\Sigma/M}, r21, r32\}^{\text{p-odd}}$	62	0.939	4.3	0.923
d	${}^{f50}\{D/M, {}^S\psi_I/M, {}^0\psi_E/M, rd31\}^{pp\text{-}odd}$	62	0.960	0.06	0.951
RI	${}^{f5}\{D, {}^1\chi_s, \chi_t^v, {}^0\psi_I, {}^1\psi_{Is}, rd6, rd25\}^{p\text{-}odd}$	61	0.941	0.04	0.924
FP	${}^{f-0.5}\{T_b, D, {}^S\psi_I, \Delta, rd38\}^{p\text{-}odd}$	41	0.987	4.2	0.983
η	${}^{f-0.5}\{T_b, \varepsilon, {}^0\psi_{Ed}, \Sigma, r32\}^{p\text{-}odd}$	39	0.956	0.4	0.791

γ	${}^{f1}\{\varepsilon, d, RI, \chi_t^v, rd33\}^{p\text{-}odd}$	41	0.948	2.0	0.822
UV	$\{RI, rd11, rd34, r21, r38\}$	33	0.767	27	0.645
μ	${}^{f0.5}\{\phi\varepsilon, \phi\Sigma, \phi T_{\Sigma/M}, \phi rd10, \phi rd34\}^{pp\text{-}odd}$	34	0.953	0.3	0.908
$-\chi\cdot 10^6$	${}^{f50}\{M, \chi_t, {}^T\psi_I, r30\}^{p\text{-}odd}$	32	0.887	0.04	0.842
EV	${}^{f0.5}\{{}^1\chi^v, rd1, rd10, rd25\}^{\text{p-odd}}$	20	0.895	0.1	0.812

Viscosity, η.

The Vector ***C*** of descriptor in Table 6 is: = [0.014 (0.002), - 0.029 (0.006), $0.25\cdot10^{-4}$ ($6\cdot10^{-6}$), - $0.37\cdot10^{-5}$ (10^{-7}), 0.07 (0.01), - 4.23 (0.6)]. The training test has: $N = 28$, $r^2 = 0.972$, $s = 0.3_5$, $q^2 = 0.786$, $F = 155$. The search for a *mc-exp-rn*-description with *r1-r38* and then *rd1-rd38* plus T_b, ε, and the set of MC-$K_p(p\text{-}odd)/f_\delta^{-0.5}$ indices finds exactly the already found five-index descriptor. The descriptor searched among *r1- r38* plus *rd1-rd38* plus the five indices of the optimal descriptor finds again the same descriptor. The descriptor is now chosen among *r1- r38* plus *rd1-rd38* plus ${}^0\psi_{Ed}$, Σ, T_b and ε. The search lands on the descriptor of Table 8, which, relatively to the optimal descriptor, has a worse q^2 statistics.

Surface Tension, γ.

Vector ***C*** of descriptor in Table 6 is: [0.27 (0.01), 21.0 (1.4), 48.8 (3.5), -42.2 (4.5), 39.4 (9.1), - 64.7 (5.4)]. Here there is no need to exclude CH_2Br_2 as required by the greedy algorithm. Notice that solvents with high dielectric constants and high density have usually high surface tension. The training set (no items (°) in Table 3) has: $N = 29$, r^2=0.948, $s = 1.7$, $q^2 = 0.916$, $F = 84$. The search for a *mc-exp-rn*-description with *r1-r38* and then *rd1-rd38* plus ε, d, RI, and the set of MC- $K_p(p\text{-}odd)/f_\delta^1$ indices finds the already found five-index descriptor. Instead, the *mc-exp-r*-descriptor searched among *r1- r38* plus *rd1-rd38* plus ε, d, RI, and χ_t^v, finds the semi-random descriptor of Table 8 with a worse q^2 quality [${}^{f1}\{\varepsilon, d, RI, \chi_t^v\}^{p\text{-}odd}$: $r^2 = 0.937$, $s = 2.2$, $q^2 = 0.721$].

Cutoff UV Values, *UV*.

The full combinatorial search, like the greedy search, needs to exclude the four strong outliers: heptane, 4-Me-2-pentanone, N,N-diMe-acetamide, and acetonitrile, but here, there is no need of any strange composite indices. Vector ***C*** of descriptor in Table 6 is: [736.820 (62),- 106.109 (13),500.009 (79), 39.7165 (4.7),- 2065.97 (277), - 785.973 (88)]. There is a strong correlations with $r({}^0\psi_I, {}^0\psi_E) = 0.98$. The training set has been modeled with a four-index descriptor (excluding the last index) to obey the Topliss-Costello rule, and has: $N = 23$,

r^2=0.879, s = 16, q^2 = 0.783, F = 33. The search for a *mc-exp-rn*-descriptor that uses *r1-r38* and then *rd1-rd38* plus *RI*, AH^b and the MC- K_p(*p-odd*)/ f_δ^1 indices (excluding the previous strong outliers) finds the already found five-index descriptor. Limiting the search to *r1-r38* plus *rd1-rd38* plus *RI*, AH^b, ${}^0\psi_I$, and ${}^S\psi_E$, the mediocre descriptor of Table 8 is found.

***Dipole Moment*, μ.**

Acetonitrile is even here a strong outlier and even here the model takes advantage of the two-valued $\phi(0, 1)$ symmetry parameter that zeroes all indices of those properties with $\mu = 0$ and leaves them unchanged if $\mu \neq 0$, i.e., for $\mu = 0$, $\phi\chi = 0$, while for $\mu \neq 0$, $\phi\chi = \chi$. [1] The ***C*** vector of descriptor in Table 6 is: [0.048 (0.004), 0.64 (0.3),- 0.05 (0.03), 0.19 (0.03),- 0.31 (0.05), - 0.13 (0.1_5)]. A rather strong correlations is, $r(\phi^l\chi, \phi D^v) = 0.97$. The positive role of the dielectric constant and of Σ, the overall *E*-State index for the electronegative atoms, and of $T_{\Sigma/M}$ (= $\Sigma^3/M^{1.7}$) is not unexpected. The training set (no items (°) in Table 3) with a four-index descriptor (no ϕD^v), due to the Topliss-Costello rule, has: N = 24, r^2=0.928, s = 0.4, q^2 = 0.742, F = 63. The search encompassing a combinatorial space made of $\phi r1$-$\phi r38$ and then $\phi rd1$-$\phi rd38$ plus ϕMCI and $\phi\varepsilon$ (no acetonitrile) finds the *mc-exp-rn*-descriptor of Table 8, which is even better (and easier) than the *greedy* convoluted descriptor (only $\phi\Sigma$ and $\phi T_{\Sigma/M}$ are configuration-dependent). Notice that the quality of ${}^{f0.5}\{\phi\varepsilon, \phi\Sigma, \phi T_{\Sigma/M}\}^{pp\text{-}odd}$ is: r^2= 0.903, s = 0.4, q^2 = 0.771, F = 93. This means that the two random indices bring a striking contribution to q^2. Without items (°) in Table 3 and with a four-index descriptor without $\phi rd10$ (to hold the Topliss-Costello rule), we have: N = 24, r^2= 0.960, s = 0.3, q^2 = 0.915, F = 114.

***Magnetic Susceptibility*, $-\chi\cdot10^6$.**

The found greedy descriptor (Table 5) is superior but quite convoluted. The training set shows: N = 23, r^2=0.850, s = 0.04, q^2 = 0.758, F = 26. A four-index descriptor searched among *MCI* plus *M*, plus *r1-r38* and then *rd1-rd38* finds the optimal *mc-exp-rn*-descriptor of Table 8 with a single random index. This descriptor is statistically similar to the highly convoluted greedy descriptor but more 'down-to-earth'. The vector ***C*** of this descriptor is [0.0019 (0.0002), - 0.13 (0.04), 0.36 (0.07), 0.14 (0.03), 0.037 (0.03)]. Without *r30* we have: r^2 = 0.788, s = 0.05, q^2 = 0.729, F = 34. Here only ${}^T\psi_I$ is configuration-dependent. Leaving-out items (°) in Table 3 we have: N = 23, r^2=0.858, s = 0.04, q^2 = 0.777, F = 27.

***Elutropic Values*, *EV*.**

Vector ***C*** of descriptor in Table 6 is: [-2.24 (0.2), 1.99 (0.2), -2.02 (0.2), 0.05 (0.01), 0.08 (0.07)]. It should here be noticed that we are bordering the Topliss-Costello rule ($N°$ data/$N°$ indices ≥ 5). The strongest correlations are: $r({}^{1}\chi^{v}, {}^{0}\psi_{I}) = 0.99$ and $r({}^{0}\psi_{I}, {}^{0}\psi_{E}) = 0.97$. Notice that the quality of the single descriptor ${}^{f\,0.5}\{{}^{1}\chi^{v}\}^{\text{p-odd}}$ is: $r^2 = 0.707$, $s = 0.1$, $q^2 = 0.661$. The training test (no (°) items in Table 3) is useless as the number of data become quite few for an interesting model. The random descriptor of Table 7 let us guess that giving up the Topliss-Costello rule, i.e., with a descriptor encompassing seven random indices it should be possible to reach a very satisfactory model, even at the q^2 level, as it is the case with the following random descriptor, which, nevertheless, continue to have a lower q^2 quality than the previous descriptor with four indices,

{*rd3*, *rd10*, *rd15*, *rd22*, *rd25*, *rd32*, *r37*}: $N = 20$, $r^2 = 0.954$, $s = 0.06$, $q^2 = 0.856$, $F = 36$

The attempt to model this property with a four-index descriptor searched among the sets of MC indices plus the *r1-r38* and then *rd1-rd38* random numbers gave no new results (the previous descriptor was confirmed). If the search is restricted to *r1-r38* plus *rd1-rd38* plus ${}^{1}\chi^{v}$, and ${}^{0}\psi_{I}$, then the rather good descriptor of Table 8 is obtained. Notice that its q^2 is far from the q^2 of the optimal descriptor (only ${}^{1}\chi^{v}$ is configuration-dependent).

Super-Descriptors.

All MC or empirical indices of the best descriptors of Table 6 are joined together to form a space of super-indices of differing configurations, which will be used for a full combinatorial search of the best super-descriptors in a kind of configuration interaction of the best indices. This super-descriptor space gave no remarkable results with the *greedy* algorithm, but it does find improved descriptors for the following four properties, which are shown in Table 9.

***Melting Points*, T_m.**

The very good super-descriptor for this property is too large to enter in Table 9 and it is:

$\{AH^{m}, {}^{0}\chi_{d}, D^{v}(f^{0.5}\text{-}ppo), {}^{1}\psi_{E}(f^{-0.5}\text{-}po), {}^{0}\psi_{Ed}(f^{-0.5}\text{-}po), \Sigma(f^{-0.5}\text{-}po), \Sigma(f^{0.5}\text{-}ppo), {}^{1}\psi_{E}(f^{50}\text{-}ppo)\}$

Its statistics are shown in Table 9. It is probably the best descriptor ever obtained of a set of sixty-two melting points with graph-theoretical methods, and we would dare to say, with any theoretical method. The ***C*** vector of this descriptor is: [40.3 (3.9), 36.0 (5.9), 3.75 (0.9), 286 (45), $-1\cdot10^{-5}(10^{-6})$, -3.90 (0.5), 4.14 (1.1), -288 (49), 142.7 (8.0)]. The largest correlations is: $r[{}^{1}\psi_{E}(f^{-0.5}\text{-}po), {}^{1}\psi_{E}(f^{50}\text{-}ppo)] = 0.97$. The training test has: $N = 45$, $r^2 = 0.853$, $s = 18.7$ $q^2 = 0.760$, $F = 26$. A feasible search for a seven-index semi-random super-descriptor with the

eight indices of the super-descriptor plus the *r1-r38* and then the *rd1-rd38* excluded index $\Sigma(f^{0.5}\text{-}ppo)$ and confirmed the other indices ($N = 62$, $r^2 = 0.81$, $s = 21$, $q^2 = 0.750$).

Table 9. The super-descriptors for four properties (P) of the organic solvents and their statistics.

P	*Super-descriptors*	N	r^2	s	q^2
T_m	*see correponding T_m paragraph*	62	0.852	19.0	0.804
FP	$\{T_b, \Delta, \Sigma(f^{-0.5}\text{-}po), {}^{S}\psi_E(f^{1}\text{-}po), rd8\}$	41	0.989	4.0	0.984
UV	$\{RI, AH^b, {}^{1}\chi_s, {}^{S}\psi_I/M(f^{50}\text{-}ppo), D^v(f^{0.5}\text{-}ppo)\}$	33	0.904	17.0	0.807
μ	$^{f0.5}\{\phi\varepsilon, \phi\Delta, \phi\Sigma(f^{50}\text{-}ppo), \phi T_{\Sigma/M}(f^{0.5}\text{-}ppo), \phi rd5\}^{pp\text{-}odd}$	34	0.960	0.3	0.914

***Flash Points*, *FP*.**

A search throughout a combinatorial space made of T_b, the thirty-two super-indices plus the *r1-r38* and then *rd1-rd38* indices finds the optimal *mc-exp-rn*-super-descriptor of Table 9, which improves slightly in s and q^2 over the previous optimal semi-random descriptor. Notice that excluding *rd8* the remaining four-index descriptor is quite good: $r^2 = 0.982$, $s = 5.0$, $q^2 = 0.975$, $F = 485$, i.e., *rd8* brings about a small but noticeable improvement. The raining test has: $N = 29$, $r^2 = 0.988$, $s = 4.2$, $q^2 = 0.980$, $F = 374$.

***Cutoff UV Values*, *UV*.**

Without the four strong outliers, heptane, 4-Me-2-pentanone, N,N-diMe-acetamide, and acetonitrile, the optimal super-descriptor shown in Table 9 can be found. The training test, leaving out the last index of the previous descriptor (due to the Toplis-Costello rule), has: $N = 23$, $r^2 = 0.902$, $s = 15$, $q^2 = 0.798$, $F = 41$. The combinatorial space made of *RI*, AH^b, the thirty-two super-indices plus the *r1-r38* and then *rd1-rd38* random sets has been searched and no interesting results have been obtained. The given super-descriptor is by far the best descriptor. A search with the five indices of the previous descriptor plus the two sets of random indices finds no different descriptor. A search with the first four indices of the previous descriptor [$r^2 = 0.733$, $s = 28$, $q^2 = 0.425$] plus the random indices finds a poorer semi-random descriptor with an even poorer q^2 statistics (0.667).

***Dipole Moment*, μ.**

The optimal semi-random super-descriptor of Table 9, which improves over the previous semi-random descriptor, is found by the aid of a combinatorial space made of the ϕ-super-indices (inclusive $\phi\varepsilon$ and ϕM) plus $\phi r1$-$\phi r38$ and then $\phi rd1$-$\phi rd38$ random sets. Without *rd5* the statistics of the remaining super-descriptor is: $r^2 = 0.929$, $s = 0.3$, $q^2 = 0.883$, $F = 94$. The training test has: $N = 24$, $r^2 = 0.967$, $s = 0.3$, $q^2 = 0.909$, $F = 106$.

CONCLUSIONS

With the full combinatorial method, an optimal model for the twelve properties has been achieved and for some of them the model is even outstanding. Normally the zero-level descriptors show their weakness especially at the *s* and q^2 level. The random tests of the *EV* property confirm in a clear way the importance of the Topliss-Costello rule. The very good model of *EV* by a semi-random descriptor with three random indices borders the limit of five for the ratio *N°data*/*N°indices*. The importance of this rule and of the q^2 statistics is confirmed by the 'zero-level' descriptors for *EV* and *FP*. The zero-level model suggests that the lowest value for q^2 should be as high as possible and we would suggest that $q^2 \geq 0.8$. They also suggest that the Topliss-Costello rule should always hold with no exceptions, a fact that had already been underlined by other authors.[27-29] The semi-random tests (Table 8) tell us if $q^2 \geq 0.8$ should hold, the semi-random model for *UV* should be rejected, while the semi-random model for the viscosity, η, should be considered with a critical eye. The other good semi-random models confirm the importance of having at hand several good MC indices or/and experimental parameters that by themselves could give rise to a satisfactory description. If the positive side of a semi-random description is that it allows an economy of the number of MC indices to be used for a model, the negative side is that it is not possible to know from scratch which MC indices or experimental parameters are the best ones for the random indices. Four properties are optimally modeled with the help of random indices: the refractive index, *RI*, the flash point, *FP*, the dipole moment, μ, and the magnetic susceptibility, $-\chi \cdot 10^6$. Random indices seem to tell us that in every property there are minor contributions that can be seen as random on a macro-level, due to a lack of detailed information about structure, interaction, and dynamics at the micro-level. The very good model of the twelve properties with normal (MCI), semiempircal, semi-random, and super-descriptors underlines the power of the full combinatorial algorithm. This algorithm confirms the utility of the experimental parameters and *M* as well as of the *ad hoc* parameters AH^b, AH^m, and ϕ. Due to the huge number of combinations of the full combinatorial space, we should expect that, normally, the configuration of the full combinatorial descriptors be different from the configuration of the *greedy* descriptors. Looking at our best descriptors, throughout Tables 5 - 9, we notice that things are not exactly that way. Properties, ε, *d*, *RI*, and η, share the same (nearly the same for *d*) configuration with both search algorithms. This is good news for the *greedy* algorithm. Assuming the q^2 statistics as the critical statistics the full combinatorial algorithm does an optimal job with T_b (the *greedy* algorithm could not find a better descriptor with higher number of indices), T_m (excellent with a super-descriptor), ε, *RI*, η, γ, μ (this last with a super-descriptor), *FP* (a super-descriptor), and *EV*. Concerning *UV* the full combinatorial descriptor has not the same q^2 of the greedy descriptor but it is much simpler. The tendency to choose semiempirical descriptors is nearly the same with both search algorithms (nine properties), whiles the *greedy* algorithm for the surface tension, γ, chooses a pure empirical descriptor. Normally, the full combinatorial algorithm does not need highly convoluted descriptor for an

optimal model. Concerning the normal descriptors the full combinatorial algorithm prefers a stronger hydrogen perturbation, i.e., $f_\delta^{\,n}$ with n = - 0.5, 0.5, and 1 (in 7/12 cases see text; the *greedy* algorithm in 4/12 cases, see Table 5), and also a K_p(*p-odd*) encoding for the core electrons (in 8/12 cases see text; 6/12 cases with the *greedy* method). Practically, the full combinatorial algorithm prefers to compensate a strong hydrogen contribution with a weaker electron core contribution. The model of properties with the highest and similar number N of compounds (T_b, T_m, ε, d, and RI) confirms that the hydrogen perturbation is mainly property-dependent.

The advantage to work with configuration-dependent indices is emphasized by possibility to use super-descriptors, i.e., to work with a combinatorial space made of indices of the best descriptors of all properties, and thus performs a kind of configuration interaction of the best indices. Last but not least, throughout the model of the twelve properties the pseudoconnectivity indices together with the electrotopological index, Σ, which encode specific information on the electronic environment, bear the main responsibility for the success of the model.

REFERENCES AND NOTES

(1) L. Pogliani, Model of twelve properties of a set of organic solvents with graph–theoretical and/or experimental parameters, *J. Comput. Chem*. **31** (2010) 295–307.

(2) E. Brézin, V. Kazakov, D. Serban, P. Wiegman, A. Zabrodin, *Applications of Random Matrices in Physics*, NATO Science Series, Springer, Dordrecht, 2006.

(3) K. B. Kier, L. H. Hall, *Molecular Structure Description. The Electrotopological State*, Academic press, New York, 1999.

(4) L. Pogliani, Novel molecular connectivity indices: Pseudoconnectivity, dual, *cis-trans* indices and indices based on a new valence delta, in: I. Gutman, B. Furtula (Eds.) *Novel Molecular Structure Descriptors – Theory and Applications I*, Univ. Kragujevac, Kragujevac, 2010, pp. 39–72.

(5) R. Garcia–Domenech, J. Galvez, J. V. de Julian-Ortiz, L. Pogliani, Some new trends in chemical graph theory, *Chem. Rev.* **108** (2008) 1127–1169.

(6) L. Pogliani, Model of the physical properties of halides with complete graph–based indices, *Int. J. Quant. Chem.* **102** (2005) 38–52.

(7) L. Pogliani, A natural graph–theory model for partition and kinetic coefficients, *New J. Chem* **29** (2005)1082–1088.

(8) L. Pogliani, The hydrogen perturbation in molecular connectivity computations, *J. Comput. Chem.* **27** (2006) 868–872.

(9) L. Pogliani, Implementing molecular connectivity theory, a basic tool in modeling drugs, *J. Pharm. Sci.* **96** (2007) 1856–1871.

(10) N. Trinajstić*, Chemical Graph Theory*, CRC Press, Boca Raton, 1992.

(11) M. Randić, On characterization of molecular branching, *J. Am. Chem. Soc*. **97 (**1975) 6609–6615.

(12) L. B. Kier, L. H. Hall, *Molecular Connectivity in Structure–Activity Analysis*, Wiley, New York, 1986.

(13) R. Todeschini, V. Consonni, *The Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000. A new edition is on the way, and it will include an interesting version of many of the indices here used.

(14) J. G. Topliss, R. J. Costello, Chance correlation in structure–activity studies using multiple regression analysis, *J. Med. Chem*. **15** (1972) 1066–1069.

(15) M. Randić, Curve–fitting paradox, *Int. J. Quant. Chem.: Quant. Biol. Symp.* **21** (1994) 215–225.

(16) L. Pogliani, J. V. de Julian–Ortiz, Plot methods in quantitative structure–property studies, *Chem. Phys. Lett.* **393** (2004) 327–330.

(17) E. Besalu, J. V. de Julian–Ortiz, M. Iglesias, L. Pogliani, An overlooked property of plot methods, *J. Math. Chem.* **39** (2006) 475–484.

(18) E. Besalu, J. V. de Julian–Ortiz, L. Pogliani, Some plots are not that equivalent, *MATCH Commun. Math. Comput. Chem*. **55** (2006) 281–286.

(19) E. Besalu, J. V. de Julian–Ortiz, L. Pogliani, Trends and plot methods in MLR studies, *J. Chem. Inf. Model.* **47** (2007) 751–760.

(20) R. Carbó–Dorca, D. Robert, L. Amat, X. Girones, E. Besalu, *Molecular Quantum Similarity in QSAR and Drug Design*, Springer, Berlin, 2000.

(21) L. Eriksson, E. Johansson, M. Muller, S. J. Wold, On the selection of training set in environmental QSAR when compounds are clustered, *Chemometrics* **14** (2000) 599–616.

(22) H. A. Martens, P. Dardenne, Rapid fuel quality surveillance through chemometric modeling of near–infrared spectra, *Chemom. Intell. Lab. Syst.* **14** (1998) 99–121.

(23) Z. Mihalić, S. Nikolić, N. Trinajstić, Comparative study of molecular descriptors derived from the distance matrix, *J. Chem. Inf. Comput. Sci.* **32** (1992) 28–37.

(24) M. Randić, On computation of optimal parameters for multivariate analysis of structure–property relationship, *J. Comput. Chem.* **12** (1991) 970–980.

(25) M. Randić, Orthogonal molecular descriptors, *New J. Chem.* **15** (1991) 517–525.

(26) S. C. Peterangelo, P. G. Seybold, Synergistic interactions among QSAR descriptors, *Int. J. Quant. Chem.* **96** (2004) 1–9.

(27) A. Golbraikh, A Tropsha, Beware of q^2! *J. Mol. Graph. Modell.* **20** (2002) 269–276.

(28) D. M. Hawkins, S. C. Basak, D. Mills, Assessing model fit by cross–validation, *J. Chem. Inf. Comput. Sci.* **43** (2003) 579–586.

(29) D. M. Hawkins, The problem of overfitting, *J. Chem. Inf. Comput. Sci*. **44** (2004) 1–12.