# LIMITS WITH MODELING DATA AND MODELING DATA WITH LIMITS

***Lionello Pogliani***

*Dipartimento di Chimica, Università della Calabria, 87030 Rende (CS), Italy.*
*Email : lionp@unical.it*

## *ABSTRACT*

*Modeling of the solubility of amino acids and purine and pyrimidine bases with a set of sixteen molecular descriptors has been thoroughly analyzed to detect and understand the reasons for anomalies in the description of this property for these two classes of compounds. Unsatisfactory modeling can be ascribed to incomplete collateral data, i.e, to the fact that there is insufficient data known about the behavior of these compounds in solution. This is usually because intermolecular forces cannot be modeled. The anomalous modeling can be detected from the rather large values of the standard deviation of the estimates of the whole set of compounds, and from the unsatisfactory modeling of some of the subsets of these compounds. Thus the detected abnormalities can be used (i) to get an idea about weak intermolecular interactions such as hydration, self-association, the hydrogen-bond phenomena in solution, and (ii) to reshape the molecular descriptors with the introduction of parameters that allow better modeling. This last procedure should be used with care, bearing in mind that the solubility phenomena is rather complex.*

**Keywords**: Modeling, Solubility, Amino acids, Bases, Incomplete Data, Molecular Connectivity Indices, Configuration Interaction of Graph-Type Basis Indices.

## 1 INTRODUCTION

Science is being exposed to a rapidly increasing flood of data and the possibility of modeling the properties or activities of the rising number of compounds with the aid of mathematical functions could be the only way to keep us from drowning in this rising sea of data. Recently, a review on modeling with higher-order molecular connectivity descriptors (Pogliani, 2000a,b), some work on pseudoconnectivity descriptors (Pogliani, 2000c, 2001), as well as works by other authors on different topological descriptors have demonstrated that it is not at all illusory to achieve an optimal description of compounds' properties and activities using graph-theoretical descriptors, especially molecular connectivity descriptors (Basak, Balaban, Grunwald, & Gute, 2000; Diudea, 2001; Estrada & Rodriguez, 1999; Galvez, Garcia-Domenech, Gomez-Lechon & Castell, 2000; Gutman & Tomović, 2000; Kier & Hall, 1986, 1999; Klein, Randić, Basić, Lucić, Nikolić & Trinajstić, 1997; Kuanar & Mishra, 1998; Nikolić & Raos, 2001; Randić, Mills & Basak, 2000; Randić & Basak, 2000; Reinhard & Drefahl, 1999; Rouvray, 1989; Seybold, 1999).

Throughout the present paper we will be concerned with the fact that one of the main problems in modeling can be phrased as the 'absent col-data' problem, i.e., the failure to model in a satisfactory way properties or activities of a class of compounds whose collateral data are either missing or incomplete, while the main body of data actually seems complete. The fact that the modeling is unsatisfactory can be detected at many statistical levels, but in some cases only a critical analysis of the standard deviation of the estimate, *s*, like in the present case, reveals that something is 'faulty' with the modeling. This is, in fact, what happens with the modeling of the solubility of amino acids and of purine and pyrimidine bases. The basis descriptors used throughout this study belong to a medium-sized set made up of two subsets of eight molecular connectivity indices and eight molecular pseudoconnectivity I/E-State indices, recently defined, and which will be elaborately discussed in the next section. These graph-theoretical molecular indices (Kier & Hall, 1986; Kier & Hall 1999; Pogliani, 2000, 2001) like many other indices of the same type, are, nevertheless, rather insensitive to weak intermolecular interactions. Nevertheless, the modeling of the solubility for these two classes of compounds when examined in detail with these molecular descriptors can help to detect at which level the modeling fails, how consistent the failure is, and what can be done to prevent it.

The solubility of solids is a rather complex process, which is influenced by the magnitude of the enthalpy change on the fusion of the pure solute, $\Delta H_{fus}$, and the melting point of solute, $T_{fus}$, i.e., $-\ln x = (\Delta H_{fus}/R)(1/T - 1/T_{fus})$ (Atkins, 1990), where x is the mole fraction solubility at T. But, other factors, such as the association or self-association phenomena in solution, which can gives rise to supramolecular species, can influence solubility. The importance of such phenomena can be seen with the hydration numbers, n, of some cations in aqueous solvent: $n(Cs^{+}) = 6$; $n(K^{+}) = 7$, $n(Na^{+}) = 13$, $n(Li^{+}) = 22$, $n(Cd^{+2}) = 39$, and $n(Zn^{+2}) = 44$ (Van der Sluys, 2001). Association and self-association, are, surely the main, even if not the only, phenomena that influences the solubility of amino acids and of purine and pyrimidine bases. Actually, self-association in solution has been clearly detected for only four purine and pyrimidine bases (Pogliani, 2000a; Pogliani, 1993; Agostini, Bonacchi, Dapporto, Paoli, Fedi & Manzini, 1990; Agostini, Bonacchi, Dapporto,

Paoli, Pogliani & Toja, 1994; Nagashima & Suzuki, 1984; Guttman & Higuchi, 1957; Bolton, Guttman & Higuchi, 1957). For all other compounds, similar phenomena can only be indirectly inferred from the irregular characteristics of the modeling, which are useful if one remains aware of the pitfalls of a circular reasoning. In practice, modeling the solubility of these two classes of compounds is influenced by missing data about concerning association, self-association, and even by missing thermodynamic data. If this information were at hand a full set of supramolecular or semiempirical descriptors could be introduced therefore for the whole set of compounds, which could be used to refine the modeling.

It cannot be excluded that a wider set of molecular descriptors could achieve better modeling, but the reader is reminded that graph-theoretical molecular descriptors are rather insensitive to weak non-covalent intermolecular interactions at long range and to van der Waals forces at close range. These type of interactions constitute a tremendous challenge not only for chemical graph theory but, also, for the whole of modern chemistry (Dykstra & Lisy, 2000).

## 2 METHOD

The Structure (S)-Property (P) relation is usually approximated by Linear equation (1),

$$P = c_1 S + c_0 U_0 \tag{1}$$

where P is the modeled property, $c_1$, and $c_0$ are the regression coefficients, $U_0 \equiv 1$ is the unitary index and S is any structural descriptor, which can either be a molecular connectivity (MC) term, $X = f(\chi)$, a molecular pseudoconnectivity term, $Y = f(\psi)$, or a mixed molecular connectivity-pseudo- connectivity higher-order term, $Z = f(X,Y)$ (Pogliani 2000c, 2001, 2002). This last term can also have the form $Z = f(X, Y, \beta)$, where $\beta$ is a basis MC index. The linear relation can also be written as a dot vectorial product: $P = \mathbf{C}\cdot\mathbf{S}$, where $\mathbf{C} = (c_1, c_0)$, and $\mathbf{S} = (S, U_0)$. To avoid negative calculated P values, with no biological or physical meaning, which can further reduce the quality of the modeling it is better to use the modulus modeling equation: $P = |c_1 S + c_0 U_0|$. Here bars stand for absolute value. This modeling equation normally enhances the description, provided that the experimental activities or properties are all positive. If some experimental activity, A, or property, P, values are negative then the modulus bars should be omitted and the normal modeling equation should be used. Clearly, any molecular descriptor can be introduced and used for S, such as graph-theoretical descriptors, geometrical descriptors, quantum mechanical descriptors, thermodynamic descriptors, and even for more 'ad hoc' descriptors (Kier & Hall, 1986). The basis descriptors of this study we will be a set, $\{\beta\} = \{\{\chi\},\{\psi\}\}$, of basis indices known as the molecular connectivity and pseudoconnectivity indices. With these basis indices more complex S descriptors will be derived. To avoid huge calculation problems the following medium-sized set of molecular connectivity and pseudoconnectivity indices will be used.

$$\{\chi\} = \{D, {}^0\chi, {}^1\chi, \chi_t, D^v, {}^0\chi^v, {}^1\chi^v, \chi^v_t\} \tag{2}$$
$$\{\psi\} = \{{}^S\psi_I, {}^0\psi_I, {}^1\psi_I, {}^T\psi_I, {}^S\psi_E, {}^0\psi_E, {}^1\psi_E, {}^T\psi_E\} \tag{3}$$

Basis $\chi$ indices are directly based on the $\delta$ and $\delta^v$ connectivity numbers of a graph and a pseudo-graph respectively (Kier & Hall, 1986, Pogliani, 2000). Basis $\psi$ indices are, on the other hand indirectly based on $\delta$ and $\delta^v$ numbers through the I-State ($\psi_I$ subset) and S-State ($\psi_E$ subset) indices (Kier & Hall, 1999; Pogliani 2000c, 2001), which are defined in Eqs. (4) and (5)

$$I_i = [(2/N)^2\delta_i^v + 1] / \delta_i \tag{4}$$
$$S_i = I_i + \Sigma_j \Delta I_{ij} \tag{5}$$

Here, N = principal quantum number, $\Delta I_{ij} = (I_i - I_j) / r^2_{ij}$, and $r_{ij}$ = counts of atoms in the minimum path length separating two atoms *i* and *j*, which is equal to the usual graph distance $d_{ij} + 1$. From the factor $\Sigma_j \Delta I_{ij}$ it is evident that S incorporates, at the atomic level, information about the influence of the remainder of the molecular environment, and that it can also be negative. These two atom-level indices encode simultaneously the graph and pseudograph representation of a molecule, as they are directly (I) and indirectly (S) based on $\delta$ and $\delta^v$ numbers of a graph and a pseudograph, respectively. Indices of subsets (2) and (3) and their subsets are formally similar as can be seen from the following definitions

$$D = \Sigma_i \delta_i \tag{6}$$
$${}^S\psi_I = \Sigma_i I_i \tag{7}$$
$${}^0\chi = \Sigma_i (\delta_i)^{-0.5} \tag{8}$$
$${}^0\psi_I = \Sigma_i (I_i)^{-0.5} \tag{9}$$
$${}^1\chi = \Sigma (\delta_i \delta_j)^{-0.5} \tag{10}$$
$${}^1\psi_I = \Sigma (I_i I_j)^{-0.5} \tag{11}$$

$$\chi_t = (\delta_1 \cdot \delta_2 \cdot \delta_3 \cdot \ldots\ldots \cdot \delta_N)^{-0.5} \quad (12)$$

$$^{T}\psi_I = (I_1 \cdot I_2 \cdot I_3 \cdot \ldots\ldots \cdot I_N)^{-0.5} \quad (13)$$

Index $\chi_t$ (and $\chi^v_t$) is the total molecular connectivity index, and has as its $\psi$ counterpart the total molecular pseudoconnectivity index, $^{T}\psi_I$ (and $^{T}\psi_E$). Sums in Eqs. (6-9), as well as products in Eqs. (12) and (13), are taken over all the N atoms (vertices in graph terminology) of a molecule. Sums in eq. 10, and 11 are over all edges ($\sigma$ bonds in a molecule) of the chemical graph. By replacing $\delta$ with $\delta^v$ in Eqs. (6, 8, 10, and 12) the subset of valence $\chi$ indices $\{D^v, {}^{0}\chi^v, {}^{1}\chi^v, \chi^v_t\}$ is obtained. By replacing $I_i$ with $S_i$ in Eqs. (7, 9, 11, and 13) the pseudoconnectivity $\psi_E$ subset $\{{}^{S}\psi_E, {}^{0}\psi_E, {}^{1}\psi_E, {}^{T}\psi_E\}$ is obtained. Peaks S and T in $\psi$ indices stand for sum and total, the other peaks follow the established denomination for $\chi$ indices (Ker & Hall, 1986).

One of the results of the $I_S$ concept (Kier & Hall, 1999) states that $\Sigma_i S_i = \Sigma_i I_i$, with the consequence that $^{S}\psi_I = {}^{S}\psi_E$. In this case set 3 will consist of seven $\psi$ indices only. Now, to avoid negative $S_i$ values for carbon atoms bonded to highly electronegative atoms, which could give rise to imaginary $\psi_E$ values, every $S_i$ value of a class of compounds whose carbon atoms show negative $S_i$ values has been rescaled to the S value of the carbon atom in $CF_4$ (S = - 5.5). This is the lowest S values a carbon atom can assume. Inevitably, this rescaling invalidates the cited result of the $I_S$ concept, with the consequence that $^{S}\psi_I \neq {}^{S}\psi_E$. This rescaling procedure is mandatory for amino acids, and purine and pyrimidine bases. For further information About the influence of the rescaling procedure on the quality of the modeling see Pogliani (2001).

The procedure used to construct the molecular connectivity, X = f($\chi$) and the molecular pseudo-connectivity terms, Y = f($\psi$) is a trial-and-error procedure (Pogliani, 2000-2001). This procedure, which optimizes not only the basis indices but also the optimization parameters, normally converges quite rapidly or does not work at all. The general form of these terms looks like a rational function,

$$S = [a(\beta_1)^m + b(\beta_2)^n]^q / [c(\beta_3)^o + d(\beta_1)^p]^r \quad (14)$$

Here $\beta$ is a basis index, S = X or Y for $\beta = \chi$ or $\beta = \psi$, respectively, and *a* - *d*, *m* - *q*, and *r* are optimization parameters that can be either negative, or zero or one. In these last two cases the rational function can be condensed into a much simpler form. As can be seen from Eq. (14) the power of each basis index is again optimized, which means that the original power (- 1/2, see Eqs. (8-13)), looses its restrictive meaning. The method of constructing terms could loosely be called for Configuration Interaction of Graph-Type Basis Indices (CI-GTBI) because of its vague resemblance with the quantum method, Configuration Interaction of Molecular Orbitals made up of Gaussian type basis functions. Throughout the present study mixed connectivity-pseudoconnectivity terms, Z = f(X, Y) will be derived and used, whenever possible. The construction of the higher-level mixed Z terms is performed with the aid of a search procedure, which consists of trying the different mathematical operations that can be used to combine X and Y together. For the sake of briefness this search procedure will also be called a trial-and-error search.

The statistical performance of the graph-structural MC invariant, S, is controlled by a quality factor, Q = r / s, and by the Fischer ratio $F = fr^2/[(1-r^2)\nu]$, where *r* and *s* are the correlation coefficient and the standard deviation of the estimates, respectively, *f* is the number of freedom degrees = n-($\nu$+1), $\nu$ is the number of variables, and *n* is the number of data. Parameter Q has no absolute meaning as it is an 'intra' statistical parameter used to compare the descriptive power of different descriptors for the same property, however this property should always be given in the same scale. The F ratio, which has the character of an 'inter'-statistical parameter, tells us, even if Q improves, which additional descriptor endangers the statistical quality of the combination. For every invariant S, $\beta$, and $U_0$, the fractional utility, $u_k = |c_k/s_k|$, where $s_k$ is the confidence interval of $c_k$, as well as the average fractional utility $\langle u \rangle = \Sigma u_k/(\nu+1)$, will be given. If the modeling relation is linear, with only one structural descriptor, S, and with $U_0$, then : $\langle u \rangle = (u_1+u_0)/2$. The utility statistics allows descriptors that give rise to unreliable coefficient values ($c_k$), whenever they have a high deviation interval ($s_k$) to be detected. Thus, this statistics gives an indirect information about the importance of a descriptor in the modeling equation. The reader should be aware that specific modeling is always under the control of all of these statistical parameters, and an improved Q is not a good recipe for a good modeling. To avoid citing the dimensions of the modeled properties every time each property P should be read as P/P° where P° is the unitary value of the property. This allows the property P to be read as a pure numerical value (Berberan-Santos & Pogliani, 1999).

## 3 RESULTS AND DISCUSSION

Table 1 shows the experimental values of the modeled properties for the amino acids, and the purine and pyrimidine bases. Tables 2 through 5 show the connectivity and pseudoconnectivity values of amino acids, and purine and pyrimidine bases, respectively. Notice that the solubility values are given with the corresponding temperatures, which for the amino acids is 25°C. The temperatures for the purine and pyrimidine bases is given in parenthesis

beside each solubility value. The original source for the experimental values are Weast (1984-1985), Lide (1991-1992), Guttman & Higuchi (1957), and Bolton et al., (1957). Throughout these sources there is no direct mention about experimental errors, but from a comparison done on different results for Leucine, mentioned in Weast (1984-1985), a 7%-10% error for the found solubility values can be assumed.

**Table 1.** Solubility of amino acids, Sol, in grams per kg of water (T=25°C); Solubility, Sol, of purines and pyrimidines bases in grams per 1000 ml of water at the given temperature (in parenthesis)

| AA | Sol | AA | Sol | PP* | Sol (T°C) | PP* | Sol (T°C) |
|---|---|---|---|---|---|---|---|
| Gly | 251 | Asp | 5 | 7I8MTp | 6.3 (20) | UA | 0.02 (20) |
| Ala | 167 | Lys | 6 | 7B8MTp | 4.5 (20) | OA | 1.8 (18) |
| Cys | | Hyp | 361 | 7Itp | 27 (20) | X | 0.5 (20) |
| Ser | 422 | Gln | 42 | 7BTp | 3.7 (30) | IsoG | 0.06 (25) |
| Val | 58 | Glu | 8.6 | 1BTb | 5.6 (30) | G | 0.04 (40) |
| Thr | 97 | His | 43 | 7PTp | 231.1 (30) | HypoX | 0.7 (19) |
| Met | 56 | Arg | 181 | 1PTb | 13.8 (30) | A | 0.9 (25) |
| Pro | 1622 | Phe | 29 | 7ETp | 36.6 (30) | T | 4.0 (25) |
| Leu | 23 | Tyr | 0.5 | 1ETb | 39.8 (30) | 5MC | 4.5 (25) |
| Ile | 34 | Trp | 12 | Cf | 25.8 (30) | U | 3.6 (25) |
| Asn | 25 | | | Tp | 08.1 (30) | C | 7.7 (25) |
| | | | | Tb | 0.54 (30) | | |

* A = Adenine, G = Guanine, U = Uracil, T = Thymine, C = Cytosine, OA = orotic acid, UA = uric acid, X = Xanthine, M = methyl, P= propyl, B = butyl, I = isobutyl, Cf = Caffein =137MMMX = 7MTp, Tb = Theobromine = 37MMX, Tp = theophylline = 13MMX.

The modeling power of a linear relation with connectivity or pseudoconnectivity terms is very dependent on the quality of the data used to derive the modeling equation. Now, there are cases where the data are not complete, in the sense that solubility values are not enough to give a full picture of the solubility problem. To solve the solubility problem of the amino acids and bases, information about their association in solution should be at hand. Now, for some of these compounds (some bases) the information exists and this uncovers and underlines the importance of this kind of information for all the remaining compounds. For most compounds the association phenomena in solution can only be guessed at by the unsatisfactory modeling that can be detected at the level of the standard deviation of the estimates, *s*, which is degraded by the presence of strong outliers. Whilst it makes things easier throwing away outliers is scientifically unsatisfying, especially if they do not represent any form of experimental error. The solubility of amino acids and purine and pyrimidine bases has another interesting aspect in that it is a classic example of how it is possible to derive a modeling equation that works pretty well for the training set of compounds, but that it does not work on the chosen subsets of compounds. Thus indicating a case of overfitting for the training set.

**Table 2.** Molecular connectivity indices, $\chi$, for 21 amino acids

| AA | D | $D^v$ | $^0\chi$ | $^0\chi^v$ | $^1\chi$ | $^1\chi^v$ | $\chi_t$ | $\chi_t^v$ |
|---|---|---|---|---|---|---|---|---|
| Gly | 8 | 20 | 4.28446 | 2.63992 | 2.27006 | 1.18953 | 0.40825 | 0.03727 |
| Ala | 10 | 22 | 5.15470 | 3.51016 | 2.64273 | 1.62709 | 0.33333 | 0.03043 |
| Ser | 12 | 28 | 5.86181 | 3.66448 | 3.18074 | 1.77422 | 0.23570 | 0.00962 |
| Val | 14 | 26 | 6.73205 | 5.08751 | 3.55342 | 2.53777 | 0.19245 | 0.01757 |
| Thr | 14 | 30 | 6.73205 | 4.53473 | 3.55342 | 2.21862 | 0.19245 | 0.00786 |
| Met | 16 | 26.7 | 7.27602 | 6.14607 | 4.18074 | 4.04355 | 0.11785 | 0.01859 |
| Pro | 16 | 28 | 5.98313 | 4.55413 | 3.80453 | 2.76688 | 0.08333 | 0.00932 |
| Leu | 16 | 28 | 7.43916 | 5.79462 | 4.03658 | 3.02094 | 0.13608 | 0.01242 |
| Ile | 16 | 28 | 7.43916 | 5.79462 | 4.09142 | 3.07578 | 0.13608 | 0.01242 |
| Asn | 16 | 36 | 7.43916 | 4.70278 | 4.03658 | 2.30434 | 0.13608 | 0.00254 |
| Asp | 16 | 38 | 7.43916 | 4.57273 | 4.03658 | 2.23927 | 0.13608 | 0.00196 |
| Lys | 18 | 32 | 7.98313 | 5.91594 | 4.68074 | 3.36624 | 0.08333 | 0.00439 |
| Hyp | 18 | 34 | 6.85337 | 4.87159 | 4.19838 | 2.84158 | 0.06804 | 0.00340 |
| Gln | 18 | 38 | 8.14627 | 5.40997 | 4.53658 | 2.80434 | 0.09623 | 0.00179 |
| Glu | 18 | 40 | 8.14627 | 5.27984 | 4.53658 | 2.73927 | 0.09623 | 0.00139 |
| His | 22 | 42 | 8.26758 | 5.81918 | 5.19838 | 3.15529 | 0.03402 | 0.00080 |
| Arg | 22 | 42 | 9.56048 | 6.70883 | 5.53658 | 3.60022 | 0.04811 | 0.00078 |
| Phe | 24 | 42 | 8.97469 | 6.60402 | 5.69838 | 3.72222 | 0.02406 | 0.00069 |
| Tyr | 26 | 48 | 9.84493 | 6.97388 | 6.09222 | 3.85651 | 0.01964 | 0.00027 |
| Trp | 32 | 54 | 10.83650 | 8.10402 | 7.18154 | 4.71624 | 0.00567 | 0.00009 |

### 3.1 Solubility of Amino Acids

A detailed analysis of the modeling of the solubility, Sol, of 20 amino acids (no Cys, but with Hyp), at once uncovers four strong outliers: Arg, Ser, Hyp, and Pro. To take care of these outliers a weighting parameter *a* has to be introduced, which weights the reciprocal basis indices (Pogliani, 2000) that have to be used to model this property of amino acids ($R = 1/\beta$)

$$\{R_S(\chi)\} = \{a/D, a/^0\chi, a/^1\chi, 1/a\chi_t, a/D^v, a/^0\chi^v, a/^1\chi^v, 1/a\chi_t^v\} \quad (15)$$

$$\{R_S(\psi)\} = \{a/^S\psi_I, a/^0\psi_I, a/^1\psi_I, 1/a^T\psi_I, a/^S\psi_E, a/^0\psi_E, a/^1\psi_E, 1/a^T\psi_E\} \quad (16)$$

Here, *a*(Pro) = 8, *a*(Ser, Hyp, Arg) = 2, and *a*(others) = 1. The rationale for such a choice will be elaborated further throughout this section. The resulting two subsets of suprareciprocal basis indices, $R_S(\chi)$ and $R_S(\psi)$ of Eqs. (14) and (15), represent the best basis descriptors up to now detected for this property. Note that the suprareciprocal descriptors of these two sets can be read as very simple forms of the molecular connectivity and pseudoconnectivity terms.

Let us look closer at the character of parameter *a*. As already underlined in an another paper (Pogliani, 2000a) the concept of outliers has a meaning only in the context of a model and the reasons that give rise to them should be determined. Alas, in many cases these reasons are unclear as there is a lack of experimental data, and then these can only be guessed at from a faulty modeling. Thus, parameter *a* could be seen as a weighting factor, loosely representing an association parameter. Improving the modeling will tell us if it is an appropriate choice. Practically, this parallels the method that subjectively gives outliers different weights, which asserts that the model is correct but the data needs to be adjusted. The fact that the total connectivity, $\chi_t$ and $\chi_t^v$, and the pseudoconnectivity, $^T\psi_I$ and $^T\psi_E$, indices have to be divided by *a*, instead of multiplied, resides in their definition: in fact, Eqs. (12) and (13) show that their values decrease with the increasing complexity of the chemical graph.

The trial-and-error search for the best mixed higher-order connectivity pseudoconnectivity term for the solubility of amino acids gives the following term and statistical parameters (**C** is the correlation vector, **u** is the utility vector of each parameter of the regression, and $Z = [a/^0\chi^v + 0.06(a/^0\psi_I)]^{0.9}$)

$$Z'_{sol} = [Z_{sol} + 40\cdot(a/^0\chi)^{0.7}] \quad (17)$$

$Q = 0.040$, $F = 3980$, $r = 0.998$, $s = 25$, $n = 20$, $\langle u \rangle = 49$, $\mathbf{u} = (63, 34)$, $\mathbf{C} = (38.6589, -337.592)$

**Table 3.** Molecular pseudoconnectivity indices, $\psi$ for 21 amino acids*

| AA | $^S\psi_I$ | $^0\psi_I$ | $^1\psi_I$ | $^T\psi_I$ | $^S\psi_E$ | $^0\psi_E$ | $^1\psi_E$ | $^T\psi_E$ |
|---|---|---|---|---|---|---|---|---|
| Gly | 20.17 | 2.87653 | 1.64846 | 0.04876 | 55.59 | 1.79888 | 0.61430 | 0.00533 |
| Ala | 22.00 | 3.63425 | 2.32607 | 0.03661 | 55.01 | 2.14940 | 0.78058 | 0.00179 |
| Cys | 24.72 | 4.3009 | 2.87594 | 0.02356 | 63.23 | 2.52101 | 0.94227 | 0.00065 |
| Ser | 27.50 | 4.15189 | 2.75426 | 0.01726 | 66.02 | 2.52046 | 0.96944 | 0.00061 |
| Val | 25.33 | 5.20847 | 3.69109 | 0.02244 | 69.35 | 2.92156 | 1.10863 | 0.00026 |
| Thr | 29.33 | 4.90961 | 3.43195 | 0.01296 | 73.34 | 2.91805 | 1.16460 | 0.00024 |
| Met | 26.83 | 6.00647 | 4.23467 | 0.01804 | 76.37 | 3.29648 | 1.22896 | 0.00001 |
| Pro | 23.00 | 5.50909 | 4.38551 | 0.03564 | 76.01 | 2.94957 | 1.21321 | 0.00029 |
| Leu | 26.83 | 6.02497 | 4.35520 | 0.01833 | 76.36 | 3.30240 | 1.25155 | 0.00010 |
| Ile | 26.83 | 6.02497 | 4.36330 | 0.01833 | 73.33 | 3.41587 | 1.26010 | 0.00013 |
| Asn | 34.17 | 5.39543 | 3.73214 | 0.00618 | 83.68 | 3.26195 | 1.30864 | 0.00008 |
| Asp | 36.17 | 5.30368 | 3.66114 | 0.00505 | 85.68 | 3.27170 | 1.33558 | 0.00008 |
| Lys | 30.00 | 6.69313 | 4.82917 | 0.01150 | 85.01 | 3.66577 | 1.38908 | 0.000035 |
| Hyp | 28.83 | 5.96795 | 4.82216 | 0.01545 | 78.34 | 3.32198 | 1.42521 | 0.00010 |
| Gln | 35.67 | 6.21193 | 4.39881 | 0.00505 | 90.66 | 3.64139 | 1.44242 | 0.00003 |
| Glu | 37.67 | 6.12018 | 4.32781 | 0.00412 | 92.67 | 3.64889 | 1.46916 | 0.00003 |
| His | 32.17 | 7.19659 | 5.43098 | 0.00654 | 92.66 | 4.01363 | 1.63775 | 0.000012 |
| Arg | 36.67 | 7.78291 | 5.53389 | 0.00345 | 102.66 | 4.36498 | 1.67151 | 0.000004 |
| Phe | 33.17 | 8.05300 | 6.14710 | 0.00578 | 99.17 | 4.40001 | 1.78495 | 0.000005 |
| Tyr | 38.84 | 8.52796 | 6.55737 | 0.00258 | 110.32 | 4.76156 | 1.97446 | 0.000002 |
| Trp | 39.01 | 10.2331 | 8.32851 | 0.00219 | 121.50 | 5.51765 | 2.35735 | 0.0000002 |

* $\psi_E$ values have been rescaled (see method section)

From the solubility values of amino acids in Table 1, let us note that the value of the *s* statistics of our optimal term is not that small. To further enhance this modeling it is better to use the modulus Eq. (18), otherwise amino acids Tyr and Trp will show negative calculated solubility values,

$$\mathrm{Sol(AA)} = \left| 38.66 \cdot Z'_{sol} - 337.8 \right| \qquad (18)$$

Use of the modulus equation enhances the description, $Q(S_{calc}/S_{exp}) = 0.04$, $F(S_{calc}/S_{exp}) = 3980$ becomes $Q(|S_{calc}|)/S_{exp}) = 0.05$, and $F(|S_{calc})|/S_{exp}) = 5308$. With the modulus Eq. (18) the calculated, $Sol_{clc}$, of Table 6 have been obtained. Table 6 also shows the leave-one-out values, $Sol_{loo}$. Between calculated ($Sol_{calc}$) and leave-one-out values ($Sol_{loo}$), no consistent disagreement can be detected, while a noticeable disagreement can be detected between calculated and experimental solubility values. The column of ratio values, $Ra=Sol_{clc}/Sol$, shows that only twelve solubility values are modeled in a satisfactory way, if an interval for this ratio ranging between 0.5 and 1.5 for this range is allowed.

Now, let us examine the influence of the association parameter *a* on the modeling. This parameter has been inferred to avoid throwing away the strong outliers Arg, Hyp, Pro, and Ser. If we choose $a = 1$ for Arg, Hyp, Pro, and Ser, both $Z'_{Sol}$, and ${}^0R^v$ become very por descriptor of the solubility of the amino acids. In the following lines the utility values have not been reported as the description is bad enough at the level of the remaining statistics

$\{Z'_{Sol}\}$: Q = 0.001, F = 3, r = 0.380, s = 344, n = 20
$\{{}^0R^v\}$ : Q = 0.0008, F = 1.4, r = 0.269, s = 398, n = 20

Let us now try to model the solubility of different subclasses of amino acids. Excluding Arg, Hyp, Pro, and Ser from the modeling for the remaining n = 16 amino acids, for which $a = 1$ we obtain,

$\{Z'_{Sol}\}$: Q = 0.035, F = 82, r = 0.924, s = 27, n = 16, <u> = 8.3, **u** = (9.1, 7.6)
$\{{}^0R^v\}$ : Q = 0.029, F = 59, r = 0.899, s = 31, n = 16, <u> = 6.5, **u** = (7.7, 5.3)

Here ${}^0R^v = 1/{}^0\chi^v$, as $a = 1$. The best descriptors for these sixteen amino acids is (here $a = 1$)

$\{{}^0R = 1/{}^0\chi\}$ : Q = 0.038, F = 94, r = 0.935, s = 25, n = 16, <u> = 9.7, **u** = (9.9, 7.6)

An interesting description of these last compounds is given by the reciprocal of the molar mass,

{1/M}: Q = 0.033, F = 76, r = 0.918, s = 28, n = 16, <u> = 7.5, **u** = (8.7, 6.4)

From now on let us keep an eye on the modeling quality of the reciprocal of the molar mass, notice that for the entire class of amino acids we have: {a/M}; Q = 0.024, F = 1494, r = 0.994, s = 41, n = 20, <u> = 25, **u** = (39, 11).
If we exclude also Asp, Gln, Lys, Met, Thr, and Tyr from the modeling, leaving a total of n = 10 amino acids (with $a = 1$), we obtain the following results, where even here as for the n = 16 case, ${}^0R$ is the best descriptor

$\{Z'_{Sol}\}$: Q = 0.034, F = 66, r = 0.944, s = 28, n = 10, <u> = 7.3, **u** = (8.1, 6.6)
$\{{}^0R\}$ : Q = 0.038, F = 82, r = 0.955, s = 25, n = 10, <u> = 7.8, **u** = (9.1, 6.6)
{1/M}: Q = 0.033, F = 63, r = 0.942, s = 28, n = 10, <u> = 6.7, **u** = (7.9, 5.5)

Even here the mixed term $Z'_{Sol}$ is (as for n = 16 case) a discrete descriptor, and this underlines the reliability of this term. Note that, up to now, *s* has always been rather 'unhealthy'. Let us see further how the modeling of the ten amino acids left-out behave, i.e., Tyr, Thr, Met, Lys, Gln, Asp, Ser, Pro, Hyp, and Arg, for which $a = 1$:

$\{Z'_{Sol}\}$: Q = 0.041, F = 3721, r = 0.9989, s = 24, n = 10, <u> = 44, **u** = (61, 27)
{1/M}: Q = 0.020, F = 850, r = 0.995, s = 51, n = 10, <u> = 17, **u** = (29, 5.7)

Term $Z'_{Sol}$ is here the best descriptor while 1/M is not a good descriptor, as its *s* value is unsatisfactory. If we enlarge the search to the modeling of a subclass (**I**) made up of six amino acids with very different solubility values from among the sixteen amino acids with $a = 1$, i.e., Gly, Ala, Thr, Asp, Lys, and Tyr, and to a subclass (**II**) made up of six amino acids with very similar solubility values, i.e., Leu, Asn, Phe, Ile, Gln, and His, we note (i) the good quality of the Z' term (ii) the good quality of ${}^0R$, and (iii) the poor results of a/M in modeling subclass **II**.
For subclass **I** we have (Gly, Ala, Thr, Asp, Lys, and Tyr),

$\{Z'_{Sol}\}$: Q = 0.033, F = 59, r = 0.968, s = 29, n = 6, <u> = 6.9, **u** = (7.7, 6.1)

{1/M}: Q = 0.036, F = 69, r = 0.972, s = 27, n = 6, <u> = 7.0, **u** = (8.3, 5.7)
{$^{S}R_{E}$}: Q = 0.038, F = 80, r = 0.976, s = 26, n = 6, <u> = 7.5, **u** = (8.9, 6.0)

Where we have added the best descriptors for this subclass and only for this subclass, $^{S}R_{E} = 1/ {}^{S}\psi_{E}$
For subclass **II** (Leu, Asn, Phe, Ile, Gln, His) we finally have expected a much better *s* value:

{$Z'_{Sol}$}: Q = 0.188, F = 13, r = 0.872, s = 4.7, n = 6, <u> = 3.8, **u** = (3.6, 4.0)
{$^{1}R_{E}$}: Q = 0.253, F = 23, r = 0.923, s = 3.6, n = 6, <u> = 5.4, **u** = (4.8, 5.9)
{1/M}: Q = 0.054, F = 1.1, r = 0.459, s = 8.4, n = 6

Where we have added the best descriptors of this subclass and only for this subclass, $^{1}R_{E} = 1/ {}^{1}\psi_{E}$. All in all the $Z'_{Sol}$ term seems the most effective descriptor of every chosen subclass of amino acids.

The negative point in the simulation of the solubility of these different subclasses of amino acids arises from the standard deviation of the estimates, *s*. Not only this value is too large for some amino acids with low solubility, but looking at the different subclasses we notice that it is only satisfactory, i.e., *s* = 3-5, only with subclass (**II**), where the differences in solubility are not dramatic, i.e., Sol(Leu)=23, Sol(Asn)=25, Sol(Phe)=29, Sol(Ile)=34, Sol(Gln)=42, Sol(His)= 43. The behavior of *s* could be explained assuming the existence in solution of associative phenomena not taken into due consideration by the weighting parameter *a*, which is used here for only four amino acids. Thus, our inferred *a* values are only partially useful. We could always infer a more precise set of *a* values valid for other amino acids, but lacking experimental evidence renders such a choice highly questionable.

Before closing this section on the solubility of amino acids let us note that an attempt to develop semiempirical terms with the $T_{fus}$ of amino acids (see introduction), following the method outlined by Pogliani (2000a), gives only poor results. Instead semiempirical terms which, instead, include $\Delta H_{fus}$ and $\Delta H_{fus}$ plus $T_{fus}$ cannot be derived as $\Delta H_{fus}$ for the whole set of amino acids is missing. Here we face here a second case of incomplete information. The solubility of amino acids is mainly influenced by rapid association (with the solvent) or self-association phenomena in solution and has suggested the next section on the solubility of twenty-three bases.

### 3.2 Solubility of Purine and Pyrimidine Bases

Before getting into the details of this description it should be noted that some of the original experimental solubility

**Table 4.** Molecular connectivity indices, $\chi$, for 23 Purine and Pyrimidine bases*

| PP | D | $D^{v}$ | $^{0}\chi$ | $^{0}\chi^{v}$ | $^{1}\chi$ | $^{1}\chi^{v}$ | $\chi_{t}$ | $\chi_{t}^{v}$ |
|---|---|---|---|---|---|---|---|---|
| 7I8MTp | 38 | 62 | 13.61036 | 11.38981 | 8.34111 | 5.97071 | 0.003564 | 8.51E-05 |
| 7B8MTp | 38 | 62 | 13.44723 | 11.22667 | 8.48527 | 6.11486 | 0.003086 | 7.37E-05 |
| 7ITp | 36 | 60 | 12.74012 | 10.46716 | 7.93043 | 5.53989 | 0.004365 | 9.82E-05 |
| 7BTp | 36 | 60 | 12.57699 | 10.30402 | 8.07459 | 5.68405 | 0.00378 | 8.51E-05 |
| 1BTb | 36 | 60 | 12.57699 | 10.30402 | 8.07459 | 5.68405 | 0.00378 | 8.51E-05 |
| 7PTp | 34 | 58 | 11.86988 | 9.59691 | 7.57459 | 5.18405 | 0.005346 | 0.00012 |
| 1PTb | 34 | 58 | 11.86988 | 9.59692 | 7.57459 | 5.18405 | 0.005346 | 0.00012 |
| 7ETp | 32 | 56 | 11.16277 | 8.88981 | 7.07459 | 4.68405 | 0.00756 | 0.00017 |
| 1ETb | 32 | 56 | 11.16277 | 8.88981 | 7.07459 | 4.68405 | 0.00756 | 0.00017 |
| Cf | 30 | 54 | 10.45567 | 8.1827 | 6.53658 | 4.10793 | 0.01069 | 0.00024 |
| Tp | 28 | 52 | 9.58542 | 7.23549 | 6.1259 | 3.71758 | 0.013095 | 0.000269 |
| Tb | 28 | 52 | 9.58542 | 7.23549 | 6.10906 | 3.7135 | 0.013095 | 0.000269 |
| UA | 26 | 54 | 8.71518 | 5.72474 | 5.6647 | 3.11237 | 0.01604 | 0.00013 |
| OA | 22 | 50 | 8.43072 | 5.24931 | 5.09222 | 2.66333 | 0.03928 | 0.00027 |
| X | 24 | 48 | 7.84493 | 5.34106 | 5.27086 | 2.92873 | 0.01964 | 0.00034 |
| IsoG | 24 | 46 | 7.84493 | 5.45738 | 5.27086 | 2.96049 | 0.01964 | 0.00043 |
| G | 24 | 46 | 7.84493 | 5.45738 | 5.27086 | 2.96049 | 0.01964 | 0.00043 |
| HypoX | 22 | 42 | 6.97469 | 4.95738 | 4.87701 | 2.74509 | 0.02406 | 0.00085 |
| A | 22 | 40 | 6.97469 | 5.07369 | 4.87701 | 2.77277 | 0.02406 | 0.00108 |
| T | 18 | 36 | 6.85337 | 4.89385 | 4.19838 | 2.4856 | 0.06804 | 0.00301 |
| 5MC | 18 | 34 | 6.85337 | 5.01016 | 4.19838 | 2.51736 | 0.06804 | 0.0038 |
| U | 16 | 34 | 5.98313 | 3.9712 | 3.78769 | 2.06893 | 0.08333 | 0.00347 |
| C | 16 | 32 | 5.98313 | 4.08751 | 3.78769 | 2.1007 | 0.08333 | 0.00439 |

* For an explanation of the names see footnote of table 1

values of these purine and pyrimidine bases are scattered throughout four different publications (Guttman & Higuchi, 1957; Bolton et al., 1957; Agostini et. al., 1990, 1994), and in Pogliani (1995).

For this modeling the following suprasquared basis indices have to be introduced, where: *a*(7PTp) = 4, *a*(1ETb, 7ETp, Cf) = 2, *a*(7ITp) = 1.5, and *a*(others) = 1 (Pogliani, 2000), the rationale for this choice is explained in the following lines

$$\{S_S(\chi)\} = \{(aD)^2, (a^0\chi)^2, (a^1\chi)^2, (\chi_t/a)^2, (aD^v)^2, (a^0\chi^v)^2, (a^1\chi^v)^2, (\chi_t^v/a)^2\} \quad (19)$$

$$\{S_S(\psi)\} = \{(a^S\psi_I)^2, (a^0\psi_I)^2, (a^1\psi_I)^2, ({}^T\psi_I/a)^2, (a^S\psi_E)^2, (a^0\psi_E)^2, (a^1\psi_E)^2, ({}^T\psi_E/a)^2\} \quad (20)$$

The fact that the total, $\chi_t$ and $\chi_t^v$, connectivity indices and the total ${}^T\psi_I$ and ${}^T\psi_E$ pseudoindices have to be divided, instead of multiplied, by the association parameter *a* is again because of their definition: their values decrease with increasing complexity of the chemical graph. The presence of such strong outliers as, 7PTp, 1Etb, 7ETp, Cf, and 7Itp, oblige us to introduce the weighting parameter, *a*, which has been already introduced for the solubility of amino acids. But things with purine and pyrimidine bases are a little different. Actually, the weighting parameter for the cited outliers really represents of an experimental association parameter (Pogliani, 1995; Guttman & Higuchi, 1957; Bolton at al., 1957). Remarkable (i) as for the amino acids, the type of descriptor found for the solubility of purine and pyrimidine bases (i.e. suprasquared indices) is similar for both $\chi$ and $\psi$ indices, and that (ii) the optimal basis descriptors for the solubility of amino acids and for the solubility of purine and pyrimidine bases are completely different completely from each other (i.e., suprareciprocal and suprasquared indices and pseudoindices).

Even for these bases the following molar mass descriptor is a very good simulator for the solubility, nearly as good as the best suprasquared index, ${}^1S = (a^1\chi)^2$

$\{(aM)^2\}$ : Q = 0.170, F = 1455, r = 0.993, s = 5.8, n = 23, <u> = 21, **u** = (38, 4.8)
$\{{}^1S\}$: Q = 0.176, F = 1553, r = 0.993, s = 5.7, n = 23, <u> = 22, **u** = (39, 4.9)

The statistics of the best molecular pseudoconnectivity suprasquared index, ${}^0S_I = (a^0\psi_I)^2$, is

$\{{}^0S_I\}$: Q = 0.170, F = 1457, r= 0.993, s = 5.8, n = 23, <u>= 21, **u** = (38, 4.4)

**Table 5.** Molecular pseudoconnectivity indices, $\psi$, for 23 Purine and Pyrimidine bases. *

| PP | ${}^S\psi_I$ | ${}^0\psi_I$ | ${}^1\psi_I$ | ${}^T\psi_I$ | ${}^S\psi_E$ | ${}^0\psi_E$ | ${}^1\psi_E$ | ${}^T\psi_E$ |
|---|---|---|---|---|---|---|---|---|
| 7I8MTp | 44.17 | 12.5454 | 10.0227 | 0.001016 | 143.17 | 6.68910 | 2.82330 | 0* |
| 7B8MTp | 43.83 | 12.6051 | 10.0013 | 0.001106 | 143.03 | 6.68882 | 2.80642 | 0 |
| 7ITp | 42.50 | 11.7708 | 9.38844 | 0.001312 | 135.99 | 6.30651 | 2.66817 | 0 |
| 7BTp | 42.17 | 11.8306 | 9.36703 | 0.001428 | 135.67 | 6.31100 | 2.65343 | 0 |
| 1BTb | 42.17 | 11.8306 | 9.36703 | 0.001428 | 135.67 | 6.32036 | 2.66325 | 0 |
| 7PTp | 40.67 | 11.0141 | 8.70037 | 0.001749 | 128.67 | 5.92957 | 2.50774 | $10^{-7}$ |
| 1PTb | 40.67 | 11.0141 | 8.70037 | 0.001749 | 128.67 | 5.93724 | 2.51493 | $10^{-7}$ |
| 7ETp | 39.17 | 10.1976 | 8.03370 | 0.002142 | 121.74 | 5.54520 | 2.36020 | $3\cdot10^{-7}$ |
| 1ETb | 39.17 | 10.1976 | 8.03370 | 0.002142 | 121.66 | 5.55188 | 2.36612 | $3\cdot10^{-7}$ |
| Cf | 37.67 | 9.38110 | 7.37900 | 0.002623 | 114.66 | 5.16141 | 2.21297 | $7\cdot10^{-7}$ |
| Tp | 36.17 | 8.59935 | 6.76839 | 0.003318 | 107.66 | 4.78055 | 2.06777 | $1.8\cdot10^{-6}$ |
| Tb | 36.17 | 8.59935 | 6.76336 | 0.003318 | 107.66 | 4.77243 | 2.05551 | $1.8\cdot10^{-6}$ |
| UA | 39.33 | 7.53631 | 5.99677 | 0.002408 | 105.34 | 4.39022 | 1.98122 | $4.3\cdot10^{-6}$ |
| OA | 40.67 | 6.61224 | 4.84903 | 0.002244 | 101.17 | 4.01094 | 1.73494 | 0.000011 |
| X | 33.17 | 7.03583 | 5.53712 | 0.005309 | 93.66 | 4.00635 | 1.75822 | 0.000012 |
| IsoG | 30.67 | 7.10276 | 5.54628 | 0.006412 | 91.17 | 3.97952 | 1.70240 | 0.000012 |
| G | 30.67 | 7.10276 | 5.54628 | 0.006412 | 91.16 | 3.98575 | 1.71225 | 0.000012 |
| HypoX | 27.00 | 6.53535 | 5.07746 | 0.011707 | 82.00 | 3.62601 | 1.54067 | 0.000035 |
| A | 24.50 | 6.60228 | 5.09033 | 0.014138 | 79.50 | 3.60499 | 1.49359 | 0.000035 |
| T | 28.00 | 5.75861 | 4.19746 | 0.013275 | 77.49 | 3.27150 | 1.32728 | 0.000093 |
| 5MC | 25.50 | 5.82554 | 4.20662 | 0.016031 | 75.00 | 3.25085 | 1.28068 | 0.000093 |
| U | 26.33 | 4.98409 | 3.54991 | 0.017139 | 70.33 | 2.88585 | 1.16430 | 0.000243 |
| C | 23.83 | 5.05102 | 3.55907 | 0.020698 | 67.83 | 2.87011 | 1.12403 | 0.000246 |

* a value < $10^{-7}$ was assumed equal to zero; $\psi_E$ values have been rescaled (see method section)

While the best two-pseudoindex combination has the following statistical level, where, $^{T}S_I = (^{T}\psi_I/a)^2$,

$$\{^{0}S_I, {}^{T}S_I\} : Q= 0.232, F = 1352, r = 0.996, s = 4.3, n = 23, \langle u\rangle = 21, \mathbf{u} = (51, 4.3, 7.4)$$

While no improved combination is obtained with the empirical descriptor $(aM)^2$, the following homogeneous combination seems to be an optimal descriptor even at the level of the F statistics

$$\{^{1}S, S_t\} : Q = 0.240, F= 1446, r = 0.997, s = 4.2, n = 23, \langle u\rangle = 22, \mathbf{u} = (53, 4.4, 8)$$

For sheer curiosity let us now see if we can improve the F value of the $\{^{1}S, S_t\}$ combination just by algebraically adding its two descriptors,

$$\{(^{1}S + S_t)\}: Q= 0.176, F = 1553, r = 0.993, s = 5.7, n = 23, \langle u\rangle = 22, \mathbf{u} = (39, 4.9)$$

The artifice of merging two descriptors into one to enhance F statistics has in fact worsened both *r* and *s* (i.e., Q) and brought no improvement in utility. This fact tells us that a CI-GTBI cannot be based on the simple algebraic sum of the best indices and/or pseudoindices, but that it must cover a (i) basis index optimization, (ii) an exponent optimization, and (iii) an optimization of the coefficient of the basis index. The best overall descriptor for the solubility of bases is the following CI-GTBI or term,

$$Z'_{Sol} = [Z_{Sol} + 7\cdot 10^{6}\cdot(\chi_t^{v}/a)^2]^{0.95} \tag{21}$$

$$Q = 0.336, F = 5662, r = 0.998, s = 3.0, n = 23, \langle u\rangle = 40, \mathbf{u} = (75, 5.1), \mathbf{C} = (0.08858, -3.49115)$$

Here, $Z_{Sol} = [X + 0.01\cdot(Y)^{1.2}]^{1.2}$, $X = (a^{1}\chi)^2$, and $Y = (a^{0}\psi_I)^2$. The final modeling equation can then be written in the following concise form,

$$\text{Sol (PP)} = 0.09\cdot Z'_{Sol} - 3.49 \tag{22}$$

**Table 6.** The experimental (Sol) and calculated ($Sol_{calc}$) solubility values of amino acids, and their calculated solubility with leave-one-out method ($Sol_{loo}$). The experimental (Sol) and calculated ($Sol_{clc}$) solubility values of bases, and their calculated solubility with the leave-one-out method ($Sol_{loo}$). In parenthesis are the corresponding molar mass (*M*) values. Ra stands for the ratio $Sol_{clc}$ / Sol. Also shown are the assumed association *a* values (see text).

| AA(M) | Sol | $Sol_{clc}$ | *a* | $Sol_{loo}$ | Ra | PP (M) | Sol | $Sol_{clc}$ | *a* | $Sol_{loo}$ | Ra |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gly (75) | 251 | 238 | 1 | 237 | 0.9 | 7I8MTp (250) | 6.3 | 8.5 | 1 | 8.6 | 1.4 |
| Ala (89) | 167 | 166 | 1 | 166 | 1.0 | 7B8MTp (250) | 4.5 | 8.9 | 1 | 9.1 | 2.0 |
| Ser (105) | 422 | 414 | 2 | 414 | 1.0 | 7ITp (236) | 27 | 24 | 1.5 | 23 | 0.9 |
| Val (117) | 58 | 79 | 1 | 80 | 1.4 | 7BTp (236) | 3.7 | 7.6 | 1 | 7.7 | 2.1 |
| Thr (119) | 97 | 80 | 1 | 79 | 0.8 | 1BTb (236) | 5.6 | 7.6 | 1 | 7.7 | 1.4 |
| Met (149) | 56 | 56 | 1 | 56 | 1.0 | 7PTp (222) | 231.1 | 231 | 4 | 232 | 1.0 |
| Pro (115) | 1622 | 1625 | 8 | 1647 | 1.0 | 1PTb (222) | 13.8 | 6.0 | 1 | 5.7 | 0.4 |
| Leu (131) | 23 | 50 | 1 | 52 | 2.2 | 7ETp (208) | 36.6 | 36.7 | 2 | 36.7 | 1.0 |
| Ile (131) | 34 | 50 | 1 | 51 | 1.5 | 1ETb (208) | 39.8 | 36.7 | 2 | 36.5 | 0.9 |
| Asn (132) | 25 | 52 | 1 | 54 | 2.1 | Cf (194) | 25.8 | 30.0 | 2 | 30.2 | 1.2 |
| Asp (133) | 5 | 52 | 1 | 55 | 10 | Tp (180) | 8.1 | 2.4 | 1 | 2.1 | 0.3 |
| Lys (146) | 6 | 32 | 1 | 33 | 5.3 | Tb (180) | 0.54 | 2.3 | 1 | 2.4 | 4.5 |
| Hyp (132) | 361 | 334 | 2 | 332 | 0.9 | UA (168) | 0.02 | 1.3 | 1 | 1.4 | 71 |
| Gln (146) | 42 | 27 | 1 | $26_{.5}$ | 0.7 | OA (156) | 1.8 | 0.3 | 1 | 0.2 | $0.1_3$ |
| Glu (147) | 8.6 | 28 | 1 | 29 | 3.2 | X (152) | 0.5 | 0.7 | 1 | 0.7 | 1.3 |
| His (155) | 43 | 23 | 1 | 22 | 0.5 | IsoG (151) | 0.06 | 0.7 | 1 | 0.7 | 12 |
| Arg (174) | 181 | 193 | 2 | 194 | 1.1 | G (151) | 0.04 | 0.7 | 1 | 0.8 | 19 |
| Phe (165) | 29 | 2.6 | 1 | 0.9 | 0.09 | HypoX (136) | 0.7 | 0.3 | 1 | 0.3 | 0.4 |
| Tyr (181) | 0.5 | 19 | 1 | 20 | 37 | A (135) | 0.9 | 0.5 | 1 | 0.5 | $0.5_4$ |
| Trp (204) | 12 | 40 | 1 | 44 | 3.3 | T (126) | 4.0 | 3.3 | 1 | 3.3 | 0.8 |
| | | | | | | 5MC (125) | 4.5 | 5.8 | 1 | 5.9 | 1.3 |
| | | | | | | U (112) | 3.6 | 4.2 | 1 | 4.2 | 1.2 |
| | | | | | | C (111) | 7.7 | 7.5 | 1 | 7.5 | 1.0 |

This relation, Eq. (22), has no absolute value bars as every calculated solubility value of bases is positive. Considering that some solubility values are very low, i.e., down to 0.02 for Sol(UA) (see Table 1), this seems to underline the good quality of found mixed higher-order term, $Z'_{Sol}$, even if its *s* value (*s* = 3) seems effectively too large in relation to the lowest solubility values. Table 6 shows the calculated solubility values with Eq. (22) and the calculated solubility values with the leave-one-out method. The similarity between these two sets of values underlines the low sensitivity of the leave-one-out method in detecting irregular behavior in the simulation of a property. Only a comparison between experimental and calculated values, as we did for amino acids, tells us that the modeling is anomalous. From the ratio of calculated to experimental solubility values, $Ra = Sol_{clc}/ Sol$, we note that, if a $\Delta Ra = \pm 0.5$ is accepted as a limit for a good simulation, then only the solubility of thirteen purine and pyrimidine bases are fairly described. The standard deviation of estimates, *s*, for the purine and pyrimidine bases is much lower than the *s* for the amino acids, as s(PP) = 3.0 and s(AA) = 25. Nevertheless, it should not be forgotten that the scale of the solubility values of amino acids goes up to 1600 for Pro (remember that we are dealing with adimensional $P/P_0$ values), and that some solubility values of our bases are as low as 0.02. Thus, things are not at all rosy even for our bases, and even here we will need more data to tell us about what is going on in solution for each compound. Due to the low *s* value the simulation of purines and pyrimidines solubility seems more homogeneous than the simulation of the solubility of amino acids. In fact, a too high solubility is predicted for seven amino acids while a too low solubility is predicted for only one. For purines and pyrimidines the spectrum of solubility values is more symmetrical as six solubility values are too high and four are too low.

Now, let us model some subclasses of these bases, and first of all let us see how the optimal term, Z′, models the entire class of bases when $a = 1$ for every compound. Let us also look for the best descriptor, and the quality of the molar mass descriptor, $(aM)^2$

$\{Z'_{Sol}\}$: Q = 0.004, F = 1.0, r = 0.214, s = 48, n = 23
$\{{}^1S = ({}^1\chi)^2\}$: Q = 0.006, F = 1.8, r = 0.284, s = 47, n = 23
$\{(M)^2\}$: Q = 0.006, F = 1.8, r = 0.279, s = 47, n = 23

The very poor quality of these descriptors with $a = 1$, means that there is no description without supraindices. Let us now eliminate those compounds with $a \neq 1$ from the description, i.e., 7PTp, 1Etb, 7ETp, 7Itp, and Cf , and model only those compounds with $a = 1$. The result is,

$\{Z'_{Sol}\}$: Q = 0.221, F = 11, r = 0.643, s = 2.9, n = 18
$\{(M)^2\}$: Q = 0.11, F = 2.9, r = 0.393, s = 3.5, n = 18

The description has improved compared to the preceding case, especially for Z′ term, which is now the best descriptor. Even if the improvement is noteworthy nevertheless it remains unsatisfactory. Interestingly, note the low *s* value of these new descriptions compared with the preceding case. Let us determine which compounds endanger this last description when the five compounds with $a \neq 1$ have been excluded. For the following nine compounds, 1PTb, 1BTb, OA, A, Hypo-X, X, Iso-G, G, and UA, the description improves and begins to be decent compared with the previous cases

$\{Z'_{Sol}\}$: Q = 0.273, F = 12, r = 0.799, s = 2.9, n = 9, <u> = 2.9, **u** = (3.5, 2.3)
$\{{}^0S^v\}$: Q = 0.295, F = 14, r = 0.820, s = 2.8, n = 9, <u> = 2.7, **u** = (3.8, 1.6)
$\{(M)^2\}$: Q = 0.272, F = 12, r = 0.797, s = 2.9, n = 9, <u> = 2.8, **u** = (3.5, 2.1)

If we delete UA, OA, and 1PTb from this description it improves showing that we have detected a further three 'bad' compounds

$\{Z'_{Sol}\}$: Q = 2.015, F = 92, r = 0.979, s = 0.5, n = 6, <u> = 7.6, **u** = (9.6, 5.7)
$\{(M)^2\}$: Q = 1.333, F = 41, r = 0.954, s = 0.7, n = 6, <u> = 5.1, **u** = (6.4, 3.8)

The Z′ term is the best descriptor here, while the squared molar mass enhances its quality but also its gap from $Z'_{Sol}$. Let us see how much the description improves if to these six optimal compounds we add, now, the five compounds with $a \neq 1$: $a$(7PTp) = 4, $a$(1ETb, 7ETp, Cf) = 2, $a$(7ITp) = 1.5,

$\{Z'_{Sol}\}$: Q = 0.454, F = 9355, r = 0.9995, s = 2.2, n = 11, <u> = 51, **u** = (97, 4.7)
$\{(aM)^2\}$: Q = 0.198, F = 1786, r = 0.997, s = 5.0, n = 11, <u> = 24, **u** = (42, 5.2)

There is a very interesting improvement in *r*, F and the utility, while *s* and consequently Q worsen. Nevertheless the modeling of these eleven compounds can be considered good, especially the one achieved using the Z′ term, which

exceeds by far the modeling quality of the suprasquared molar mass. Now let us check if the nine excluded compounds with $a = 1$, are really the poor ones together with UA, OA, and 1PTb. The description for the following excluded compounds, Tp, C, 7I8MTp, 7B8MTp, 5MeC, T, 7BTp, U, Tb, is, in fact, deceptive

$$\{Z'_{Sol}\}: Q = 0.115, F = 0.58, r = 0.28, s = 2.4, n = 9$$
$$\{(M)^2\}: Q = 0.005, F = 0.001, r = 0.013, s = 2.5, n = 9$$

Even here Z′ is the most interesting descriptor while the squared molar mass is a very bad descriptor. From all these models we can infer that we need further experimental data to achieve a satisfactory modeling for twelve compounds, and we especially need data that explain their behavior in solution. The 'poor behavior' of these twelve compounds disappears when they are combined with the other compounds to give a class of twenty-three compounds. In this case their 'poor behavior' is averaged out by the others 'good behavior'. Notice that for most of these descriptions the Z′ term is the optimal or the nearly optimal descriptor performing better than the squared or suprasquared molar mass. It is not even possible to develop a semiempirical term with $T_{fus}$ and $\Delta H_{fus}$ for purine and pyrimidine bases as the complete set of these values for these bases is also missing.

## 4 CONCLUSION

The '*incomplete data*' issue in modeling of the solubility of amino acids and purine and pyrimidine bases uncovers one of the main problems in QSAR/QSPR studies: the need for additional collateral data on 'nearby' properties to achieve an optimal modeling. An anomalous modeling can normally be uncovered by the large value of the standard deviation of the estimates, *s*, of the description, which can even be larger than many experimental values. Sometimes the underestimated statistic, *s*, is much more efficient than any other kind of statistic (inclusive of the leave-one-out method) for detecting 'anomalous' situations. Incomplete information can be of two types, information totally missing, that is as in the case of amino acids and information partially missing that is the case of the purine and pyrimidine bases. The only way is to introduce an undifferentiated weighting parameter. This parameter, in the case of amino acids solubility, to make up for that missing information, can be freely interpreted as an association constant based on the experimental results taken from a series of solubility values of purine and pyrimidine bases that were also studied. Even after the introduction of the weighting parameter in the case of some amino acids, and after the introduction of the association constant in the case of some bases a poor value of the standard deviation of the estimates, *s*, is detected. underlining the fact that a complete set of data about the given compounds behavior in solution of is missing. Incomplete information includes not only data on the association phenomena in solution, but also data on $\Delta H_{fus}$, which deprive us of the possibility of building semiempirical terms. Nevertheless, modeling of the properties of compounds whose collateral experimental data are either totally or partially missing is always worthwile. In fact, it offers interesting hints not only about the quality and quantity of the incomplete information, but also suggests the practical possibility of defining supramolecular basis descriptors that can take care of some non-covalent interactions. Clearly, there is here the risk of ending up with a circular reasoning of the kind: the model does not work, a new parameter is introduced to make it work, and finally it works. To avoid this, the new parameter (i) should have a clear physical meaning, (ii) should at least have been detected in some cases at least, and (iii) should be used parsimoniously, until further evidence, i.e., new experimental data are at hand.

This study on 'imperfect' information has also shown that the CI-GTBI method is not able to model everything, as has been suggested because it claims that they mimic or can be mimicked by random numbers. Apart from the fact that it is not possible to mimic any property whatsoever (Kier & Hall, 1986) with random numbers, such a possibility would deprive the random numbers of their random character, as they are either random or they show trends and therefore are no more random. Let us end this paper with the wise words of E.T.Bell (Taine, 1964), "*Things in the real universe don't all fit together like the pieces of a puzzle*".

## 5 ACKNOWLEDGEMENTS

The author would like to the anonymous reviewers for their valuable hints to improve the paper.

## 6 REFERENCES

Agostini, O., Bonacchi, G., Dapporto, P., Paoli, P., Fedi, M., & Manzini, S. (1990) Physico-chemical properties of new the antibronchospastic agent isbufylline. *Arzneim.-Forsch./Drug Res*. *40* (10), 1089-1092.

Agostini, O., Bonacchi, G., Dapporto, P., Paoli, P., Pogliani, L., & Toja, E. (1994) Structure and Dynamics of theophylline derivatives by X-ray, NMR and molecular mechanics studies. *J.Chem.Soc.,Perkin Trans. 2* (5), 1061-1066.

Atkins, P.W. (1990) *Physical Chemistry*, (p. 170) Oxford, UK: Oxford Univ.Press.

Basak, S.C., Balaban, A.T., Grunwald, G.D., & Gute, B.D. (2000) Topological indices: their nature and mutual relatedness. *J. Chem. Inf. Comput. Sci. 40* (4), 891-898.

Berberan-Santos, M.N., & Pogliani, L. (1999) Two alternative dwerivations of Bridgam's theorem. *J. Math. Chem*. *26*, 255-261.

Bolton, S., Guttman, D., & Higuchi, T. (1957) Complexes formed in solution by homologs of caffeine. *J. Am. Pharm. Assoc. 46* (1), 38-41.

Diudea, M.V. (Ed.), (2001) *QSPR/QSAR Studies by Molecular Descriptors*, New York: Nova Science.

Dykstra, C.E., & Lisy, J.M. (2000) Experimental and theoretical challenges in the chemistry of noncovalent intermolecular interaction and clustering. *J. Mol. Struct.(THEOCHEM) 500*, 375-390.

Estrada, E., & Rodriguez, L. (1999) Edge-connectivity indices in QSPR/QSAR studies. 1. Comparison to other topological indices in QSPR studies. *J. Chem. Inf. Comput. Sci*. *39* (6) , 1037-1041.

Galvez, J., Garcia-Domenech, R., Gomez-Lechon, M.J., & Castell, J.V. (2000) Use of molecular topology in the selection of new cytostatic drugs. *J. Mol. Struct. (THEOCHEM) 504*, 241-248.

Gutman, I., & Tomović, Z. (2000) More on the line ggraph model for predicting physico-chemical properties of alkanes. *Models in Chemistry 137* (4), 439-445.

Guttman, D., & Higuchi, T. (1957) Reversible association of caffeine and some caffeine homologs in aqueous solution. *J. Am. Pharm. Assoc*. *46* (1), 4-10.

Kier, L.B. & Hall, L.H. (1986) *Molecular Connectivity in Structure-Activity Analysis*, New York: Wiley (and references therein).

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

Klein, D.J., Randić, M., Basić, D., Lucić, B., Nikolić, S., & Trinajstić, N. (1997) Hierarchical orthogonalization of descriptors. *Int. J. Quant. Chem*. *63* (1), 215-222.

Kuanar, M., & Mishra, B.K. (1998) Optimization of regression model for a quantitative structure. Mutagenicity relationship of some natural amino acids. *Bull. Chem. Soc. Jpn 71* (1), 191-198.

Lide, D.R. (Editor-in-Chief) (1991-1992) *CRC Handbook of Chemistry and Physics*, (pp. 7-1--7-3) 72nd, Boca Raton, FL: CRC Press.

Nagashima, N., & Suzuki, E.-I. (1984) Studies of hydration by broas-line pulsed nmr. *Appl. Spectr. Rev*. *20* (1), 1-53.

Nikolić, S., & Raos, N. (2001) Estimation od stability constants of mixed amino acid complexes with Copper(II) from topological indices. *Croat. Chim. Acta 74* (3), 621-631.

Pogliani, L. (1993) Molecular connectivity model for determination of T1 relaxation times of α-carbons of amino acids and cyclic dipeptides. *Computers Chem*. *17* (3), 283-286.

Pogliani, L. (1995) Molecular modeling by linear combinations of connectivity indices. *J. Phys. Chem. 99* (3), 925-937.

Pogliani, L. (2000a) From molecular connectivity indices to semiempirical connectivity terms: recent trends in graph theoretical descriptors. *Chem.Revs*. *100* (10), 3827-3858.

Pogliani, L. (2000b) Graph-Theoretical modeling of experimental data: a data quality problem. *Proc. of the 17th International CODATA Conference* (pp 138). Baveno, Italy.

Pogliani, L. (2000c) Modeling with molecular pseudoconnectivity descriptors. A useful extension of the intrinsic I-State concept. *J. Phys. Chem.A 104* (39), 9029-9045.

Pogliani, L. (2001) How far are molecular connectivity descriptors from $I_S$ molecular pseudoconnectivity descriptors. *J. Chem. Inf. Comput. Sci*. *41* (3), 836-847.

Pogliani, L. (2002) Mixed Higher-Order Connectivity-Pseudoconnectivity Terms (ch 8, pp 208-242). In Bruce King, R. & Rouvray, D. H. (Eds.), *Topology in Chemistry*, Chichester, England: Horwood Pub.Lim.

Randić, M., & Basak, S.C. (2000) Construction of high-Quality Structure-Property-Activity Regressions: the boiling points of sulfides. *J. Chem. Inf. Comput. Sci*. *40* (4), 899-905.

Randić, M., Mills, D., & Basak, S.C. (2000) On characterization of physical properties of amino acids. *Int. J. Quant. Chem. 80*, 1199-1209.

Reinhard, M. & Drefahl, A. (1999) *Handbook for Estimating Physicochemical Properties of Organic Compounds,* New York: Wiley.

Rouvray, D.H. (1989) The limits of applicability of topological indices. *J. Mol. Struct. (THEOCHEM) 185*, 187-201.

Seybold, P.G. (1999) Exploration of molecular structure-property relationships. *SAR and QSAR in Environ.Res. 10*, 101-115.

Taine, J. (Eric Temple Bell) (1964) *The Time Stream, Three Science Fiction Novels*, New York: Dover.

Van der Sluys, W.G. (2001) The solubility rules: why are all acetates soluble? *J. Chem. Ed*. *78* (1), 111-115.

Weast, R.C. (1984-1985) (Editor-in-chief), *CRC Handbook of Chemistry and Physics*, (pp C-7237) 65th, Boca Raton, Fl: CRC Press.