Internet Electronic Journal of Molecular Design **2003**, *2*, 137–159 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

March 2003, Volume 2, Number 3, Pages 137–159

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 7

Guest Editor: Jun–ichi Aihara

Compound Similarity Used in Solvent–Solute Interaction Modeling

Guido Sello

Dipartimento di Chimica Organica e Industriale, Università degli Studi di Milano, via Venezian 21, 20133 Milano, Italy

Received: October 28, 2002; Revised: November 29, 2002; Accepted: December 15, 2002; Published: March 31, 2003

Citation of the article:
G. Sello, Compound Similarity Used in Solvent–Solute Interaction Modeling, *Internet Electron. J. Mol. Des.* **2003**, *2*, 137–159, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Compound Similarity Used in Solvent–Solute Interaction Modeling[#]

Guido Sello*

Dipartimento di Chimica Organica e Industriale, Università degli Studi di Milano, via Venezian 21, 20133 Milano, Italy

Received: October 28, 2002; Revised: November 29, 2002; Accepted: December 15, 2002; Published: March 31, 2003

Internet Electron. J. Mol. Des. **2003**, *2* (3), 137–159

Abstract

Motivation. Solvent–solute interactions greatly influence the behavior of compounds. In chemical reactivity the solvent should favor the desired reaction disfavoring competitive reactions and enhancing reaction rate. In biological activity the ubiquitous presence of water, as the solvent, in contrast to the lipidic composition of many tissues, affects many macroscopic results, such as compound delivery or excretion. The possibility of modeling such interactions can reduce experiments and permit better understanding of compound activity.

Method. In recent years, we introduced a novel approach for the prediction of the best solvent in a synthetic reaction. We generally applied the principle of similarity in solvation, *i.e*. we calculated an approximate similarity between reactants, transition states and solvent molecules. We are now proposing an extension of this concept to the general evaluation of the interactions between organic compounds and solvents.

Results. The model has been applied to three sets of compounds in order to predict or understand the solvent role in their biological behavior.

Conclusions. The results show that modeling solvation using structure similarity can represent an alternative to classical descriptors, also giving an insight into solvation at molecular level.

Keywords. Solvation; similarity; log *P*; dermal penetration; aromatic hydrocarbons; aromatic amines.

Abbreviations and notations

ADME, adsorption, distribution, metabolism, excretion	PAH, polycyclic aromatic hydrocarbon
DNA, deoxyribonucleic acid	QSAR, quantitative structure–activity relationship
log *P*, water–octanol partition coefficient	SSS, solute–solvent similarity
MW, molecular weight	

1 INTRODUCTION

The capability of predicting molecule–molecule interaction is the basis of the modeling of many chemical and biological interactions. Chemical reactivity, enzymatic recognition, drug efficacy, compound toxicity, and many other effects, are all influenced by the interaction of two or more molecules, that often determines the positive or negative result of each process. In this view, it is

[#] Dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday.

* Correspondence author; E–mail: sello@mailserver.unimi.it.

necessary to consider all the molecules participating to a complex group of interactions and determining one particular outcome. Solvent effects are an important part of the intermolecular interactions and they have been consequently studied to develop models capable of predicting the experimental behavior of compounds in solvents.

There are basically two different active areas in solvent modeling. The first studies the molecules soaked in the solvent using diverse calculation methods (*e.g*. molecular dynamics, continuum solvent simulation), while the second deduces the solvent effect calculating some macroscopic variables by the use of molecular descriptors. These two approaches are evidently different in both aim and scope. In particular, the first precisely models the solute–solvent interaction, whilst the second predicts the influence of the solvent on the solute behavior from the calculated variables. It is clear that the applications are very different. In the second area there are many models available that can be divided on the basis of the level of the theory used in the descriptor calculation, from complete experimental to complete theoretical. In addition, the number and the use of the descriptors can greatly vary.

Recently, many papers [1–8] have discussed the calculation of compound solubility, mainly in water, because there is increasing interest in the modeling of the ADME properties of molecules. These approaches can be divided into three groups: (*a*) the first group calculates aqueous solubility using experimental data, thus requiring the availability of such data [1,2]; (*b*) the second group calculates different molecular properties that are then combined to estimate water solubility [3–7]; (*c*) the third group uses atom group contribution to calculate the desired values [8]. All the three groups have their own positive and negative features that have been already discussed by the authors. However, two rather common characteristics are the need of a specific choice of descriptors for each case and the use of a training set in order to deal with such choice.

It is our aim to introduce a different way to model solute–solvent interaction; in particular, our approach will always use a single calculated descriptor and will only modify its mathematical manipulation to adapt the descriptor to the current problem.

Recently [9], we have proposed a scheme for the evaluation of the best reaction conditions in organic synthesis. The scheme calculates, among other characteristics, an ordered list of the solvents that should enhance the reaction rate of the main reaction with respect to the rates of the competing reactions. The procedure to compile the list includes the calculation of the similarity between the solvents and both the reactants and the hypothetical transition states. The basic principle is that a good solvent for a reaction will destabilize the reactants and stabilize the transition state, in order to decrease the activation energy of the reaction. We apply the principle of "similarity in solvation", *i.e*. we calculate an approximate similarity between reactants, transition states and solvent molecules. The compounds more similar to the solvent molecules are better solvated.

The calculation of the solute–solvent similarity is based on the comparison, atom by atom, of the corresponding structures, following the procedure reported in the next section. The obtained values were used in the procedure to order the solvent list and to choose the best solvent for the reaction. In this paper, this descriptor will be used to evaluate the solute–solvent interactions in three data sets.

2 MATERIALS AND METHODS

2.1 Model Description

The first operation applied to the molecules participating to the interaction consists of the calculation of a coefficient (γ) correlated to the electronic nature of each atom. The coefficient is calculated for the atoms of both the solute and the solvent. γ derives from the combination of two parts: one calculated by the Pauling [10] electronegativity (χ) of the atoms and the other dependent on the ionic or covalent radii of the atoms, obtained from literature data [10]:

$$\gamma_i = f(\chi_i) + f(\chi, r)_i \tag{1}$$

Given an atom *i*, γ_i is the value of γ; $f(\chi_i)$ is the part calculated using the electronegativity of the atom and of its neighbors, and $f(\chi, r)_i$ is the part derived from its ionic or covalent radius. The function $f(\chi_i)$ is the mean of the electronegativity differences between the atom and its neighbors:

$$f(\chi_i) = \frac{1}{Val_i} \times \sum_{k=1}^{Val_i} (\chi_k - \chi_i) \tag{2}$$

where Val_i is the number of the atoms bonded to atom *i*, χ_i is its electronegativity, and χ_k the electronegativity of its neighbors.

The function $f(\chi, r)_i$ is the ratio of $f(\chi_i)$ and the ionic or covalent radius multiplied by a constant:

$$f(\chi, r)_i = \frac{1}{100} \times \frac{f(\chi)_i}{r_i} \tag{3}$$

where $f(\chi, r)_i$ represents the sensitivity of the atom to the electron distribution and can be associated to its polarizability.

Using the γ of all the atoms we can calculate the solvation value of each solvent with respect to each molecule. The system is based on the similarity between the γ's of the solute atoms and those of the solvent atoms. Having calculated the γ values we locate, for each atom of the molecule, excluding the hydrogen atoms bonded to carbon atoms, the value of γ most similar to the γ of a corresponding atom of the solvent, *i.e*. for each atom of the molecule:

$$\Delta\gamma_{\min} = \min_{k=1}^{M} \left| \gamma - \gamma_{Sk} \right| \tag{4}$$

where $\Delta\gamma_{min}$ is the absolute difference between the two most similar γ's, M is the number of the solvent atoms, γ the value of the molecule atom, and γ_{Sk} the value of the solvent atom k. Then, the $\Delta\gamma_{min}$ are summed obtaining a first similarity value, that only considers the most similar atoms:

$$S_{max} = \sum_{i=1}^{N} \Delta\gamma_{min_i} \tag{5}$$

where S_{max} is the value of the highest similarity between the molecule and the solvent, *i.e.* the value that we should have if the solvent is made only by the atoms most similar to those of the molecule. N is the number of the molecule atoms.

The next step is the calculation of the mean similarity between the molecule and the solvent; this value is obtained summing all the Δγ resulting from the comparison of all the γ's of the solute atoms and of the solvent atoms, and dividing the result by the number of the solvent atoms:

$$S_{mean} = \frac{1}{M} \times \sum_{j=1}^{N} \sum_{k=1}^{M} \left| \gamma_j - \gamma_{Sk} \right| \tag{6}$$

Finally, we can calculate the similarity of each molecule atom to each solvent atom. It is obtained summing S_{max} and S_{mean} and dividing by the number of the molecule atoms:

$$SSS = \frac{1}{N} \times \left(S_{max} + S_{mean} \right) \tag{7}$$

The scheme reported above was used to order the solvents in the reaction simulation model. However, even keeping the essential characteristic (*i.e.* the similarity hypothesis) of the approach we decided to make the calculation more exact in the perspective of extending its use. In Table 1, we report all the variables used in the previous equations with their definitions.

Table 1. Definitions of the variables used in equations 1–7

Variable name	Definition	Variable name	Definition
γ	Atomic similarity coefficient	S_{max}	Solute–solvent maximum similarity
χ	Atom electronegativity	S_{mean}	Solute–solvent medium similarity
Val	Atomic valence	M	Number of solvent atoms
r	Atom covalent radius	N	Number of solute atoms
$\Delta\gamma_{min}$	Smallest difference in similarity coefficients	*SSS*	Solute–solvent similarity

The principal source of uncertainty in the previous calculation is represented by the used electronegativities. In fact, the direct use of Pauling's electronegativities is not sufficient to take care of all the intramolecular perturbations in the molecule. For example, in this approach all the sp^3 carbon atoms are equal if they are bonded to similar atoms, a situation that does not consider the influence of atoms on the successive spheres. However, we have available a calculation scheme of electronegativity that is more precise [11–13]. It recursively considers all the intramolecular effects, even the distant ones.

Consequently, we modified the calculation scheme for what concerns the calculation of the atomic electronegativities, maintaining all the rest. This permits an improved correspondence between the model and the real interaction.

There are still other improvements that can be imagined, *e.g*. the use of the three–dimensional calculation scheme of electronegativities, which we have also available [14], but we presently prefer to test the model in a simpler general scheme to have an impression of its potential.

The second significant difference is the need to compare different compounds with respect to the same or different solvents. This is not a problem in the ordering of the solvents in the reaction modeling, because, in that case, the comparison is performed only between the competing reactions of a single molecular set; consequently, there is no need of calculating an absolute value. Now, the situation is different; we wish to compare compounds of different sets and, in addition, we wish to get quantitative prediction and not only solvent orders. The greater precision introduced by the new calculation guarantees a better quantitative estimation of the values.

More complex is the problem of what kind of similarities must be compared. It is in fact clear that speaking about similarity we cannot assign absolute similarity values, but only comparative similarities (a measure of the similarity of an object with respect to another object and considering a precise attribute) [15]. In our view, the model should be capable of estimating the similarity between a solute and a solvent (*i.e*. inter–compound similarity). This means that we should be able to tune the model to the specific problem (*i.e*. we can choose the model as a function of the attribute). For example, we can choose the solvent when modeling the aqueous solubility (water) or when modeling the passage through cell walls (a lipophilic solvent simulating a membrane). The choice of the specific model will depend on the current problem.

2.2 Chemical Data

To test the solute–solvent similarity (*SSS*) as a measure of solute–solvent interaction we developed three models using three different data sets. The first data set contains the compounds used by Gute *et al*. [16] to study the dermal penetration of polycyclic aromatic hydrocarbons (PAH). The second data set contains the aromatic amines used by Franke *et al*. to study their carcinogenicity [17]. The third data set contains 95 aromatic primary amines used by Basak *et al*. in a study on mutagenicity prediction [18].

This choice has been made because the three sets show a good variability of compound structures. In addition, the solubility of the compounds in the first set is directly applied to model a biological activity, thus the literature results can be directly compared to ours; differently, the solute–solvent interaction of compounds in the other two sets is only one descriptor in a group used to rationalize a complex biological effect, thus our results cannot be directly compared to the biological data.

We are going to develop one model for each problem and to discuss the agreement with the literature analyses; then, we will try to better understand the potential applicability of our approach; finally, we will briefly discuss the reasons that can support our idea of developing many models based on only one descriptor to represent solute–solvent interactions.

3 RESULTS AND DISCUSSION

3.1 Dermal Penetration of PAHs

In their paper Gute *et al*. [16] study the possibility of predicting the dermal penetration of PAHs using a hierarchical QSAR approach. They use some calculated topological molecular descriptors, together with the calculated log *P* [19] (water–octanol partition coefficient) and MW (molecular weight), to calibrate a function to predict the PAH dermal penetration. All their best models use shape and/or size descriptors, thus they conclude that this kind of descriptors are the most important for this biological effect. Dermal penetration is defined as the percentage of the applied dose (40 nmoles per cm^2 skin surface) which penetrate the skin [20].

From a slightly different viewpoint we would like to check if our calculation of *SSS* can also predict this effect, taking in due consideration that ours is not a shape and size descriptor. In particular, we will compare our results with both the reported log *P*'s and the dermal penetration data.

The first step, in our approach, consists in the preparation of the hypothesis concerning the model structure. From the Gute *et al*. [16] discussion it appears that PAH dermal penetration is governed by the direct interaction between the derma and the compound; in other words, the more lipophilic is the compound the better it passes through the dermal barrier. In this view, a classical descriptor of lipophilicity is log *P*, where the compounds that preferentially partition in octanol are the most lipophilic. As a consequence, we can assume that the compounds more similar to octanol should penetrate better through the derma. We are going to calculate the *SSS* of the PAHs to octanol. One more consideration is worth. Because the data set is highly homogeneous the crude similarity measures are sufficient to get an acceptable description of the biological results.

The calculated *SSS* are reported in Table 2 together with all the other necessary data and the compound structures are shown in Figure 1. The first important point concerns the calculated log *P*. In homogeneous data set this values are correlated to the number of non–hydrogen atoms in each compound. In fact, the regression analysis of the reported log *P* with the number of non–hydrogen atoms gives: log $P = 0.272\ N + 1.029$, $n = 60$, $r^2 = 0.85$, $sd = 0.34$, F = 314.3. The regression of log *P* with our *SSS* values gives: $SSS = 1.013 \log P + 0.104$, $n = 60$, $r^2 = 0.88$, $sd = 0.3$, F = 428.7. It is clear that the *SSS*s are in agreement with the calculated log *P*'s.

We calculated two more regression lines, correlating either the reported log *P*'s or our values to the experimental dermal penetration data. The results are, respectively: %DP = –11.375 log *P* + 91.710, *n* = 60, r^2 = 0.60, *sd* = 8.2, F = 86.6, and %DP = –11.122 *SSS* + 92.184, *n* = 60, r^2 = 0.667, *sd* = 7.45, F = 117.2, r^2_{LOO} = 0.662. Overall, our model is comparable to those reported by Gute *et al*. The regression lines are presented in Figure 2.

Table 2. Literature and calculated data of PAH dermal penetration

No.	log *P*	*SSS*	Dermal penetration	Atom number	No.	log *P*	*SSS*	Dermal penetration	Atom number
1	7.044	7.772	0.7	24	**31**	6.128	6.037	20	18
2	7.298	7.860	2	24	**32**	5.664	5.958	20	18
3	8.266	8.666	6	26	**33**	5.942	5.598	20	16
4	6.124	6.569	7	20	**34**	6.838	7.260	20	22
5	7.298	7.868	8	24	**35**	5.664	5.973	20	18
6	7.067	7.121	8	21	**36**	5.599	5.691	22	17
7	5.858	6.667	8	20	**37**	6.962	6.806	24	20
8	7.422	7.415	8.3	22	**38**	5.399	5.774	25	17
9	6.584	7.163	9	22	**39**	5.399	5.770	25	17
10	6.838	7.263	9.4	22	**40**	6.312	6.859	26	20
11	6.124	6.560	10	20	**41**	6.432	6.908	29	20
12	6.584	7.168	10	22	**42**	4.95	5.270	30	14
13	6.432	6.893	10	20	**43**	5.668	5.454	30	16
14	6.842	6.739	10	20	**44**	6.773	6.993	32	21
15	6.916	7.166	11	21	**45**	5.139	5.099	33	15
16	6.313	6.384	14	19	**46**	5.599	5.701	33	17
17	6.124	6.573	14	20	**47**	5.788	5.518	33	16
18	6.124	6.572	15	20	**48**	5.664	5.972	35	18
19	6.128	6.034	18	18	**49**	4.225	4.471	36	13
20	6.716	6.281	20	18	**50**	5.139	5.096	38	15
21	6.466	6.186	20	18	**51**	5.273	5.155	38	15
22	6.378	6.679	20	20	**52**	5.139	5.094	40	15
23	7.067	6.394	20	19	**53**	4.784	4.800	40	14
24	6.313	6.393	20	19	**54**	5.214	5.505	40	16
25	6.313	6.384	20	19	**55**	4.49	4.690	42	14
26	4.674	4.965	20	14	**56**	4.95	5.272	42	16
27	5.783	6.477	20	19	**57**	4.874	4.891	49	14
28	5.942	5.541	20	16	**58**	5.139	5.100	50	15
29	7.186	6.731	20	20	**59**	4.685	5.078	50	15
30	6.977	7.051	20	21	**60**	4.49	4.673	50	14

In literature there are other analyses of this data set in addition to that by Gute *et al*. For the whole set of 60 PAHs, Roy *et al*. [19] obtained the following QSAR equation: %DP = 111.9 – 14.7 log *P* – 22.0 SHDW6, r^2 = 0.640, *sd* = 7.7, F = 54, where log *P* is the calculated octanol–water partition coefficient, and SHDW6 is the normalized area of the two–dimensional projection of the molecule onto the Y–Z plane.

An improved model is reported by Ivanciuc *et al*. [21] and it is obtained with SD1, the average electrophilic reactivity: %DP = – 50.10 + 12542 SD1, r^2 = 0.711, *sd* = 6.9, F = 142.8; a second

correlation was obtained by these authors using two descriptors, further enhancing the result: %DP = 8897 + 12025 SD1 – 2269 SD11, r^2 = 0.748, *sd* = 6.5, F = 84.4, where SD11 is the average valence of a carbon atom. These models do not use a size and shape descriptor, nor a classic solvation index; consequently, their comparison to our result is less straightforward.

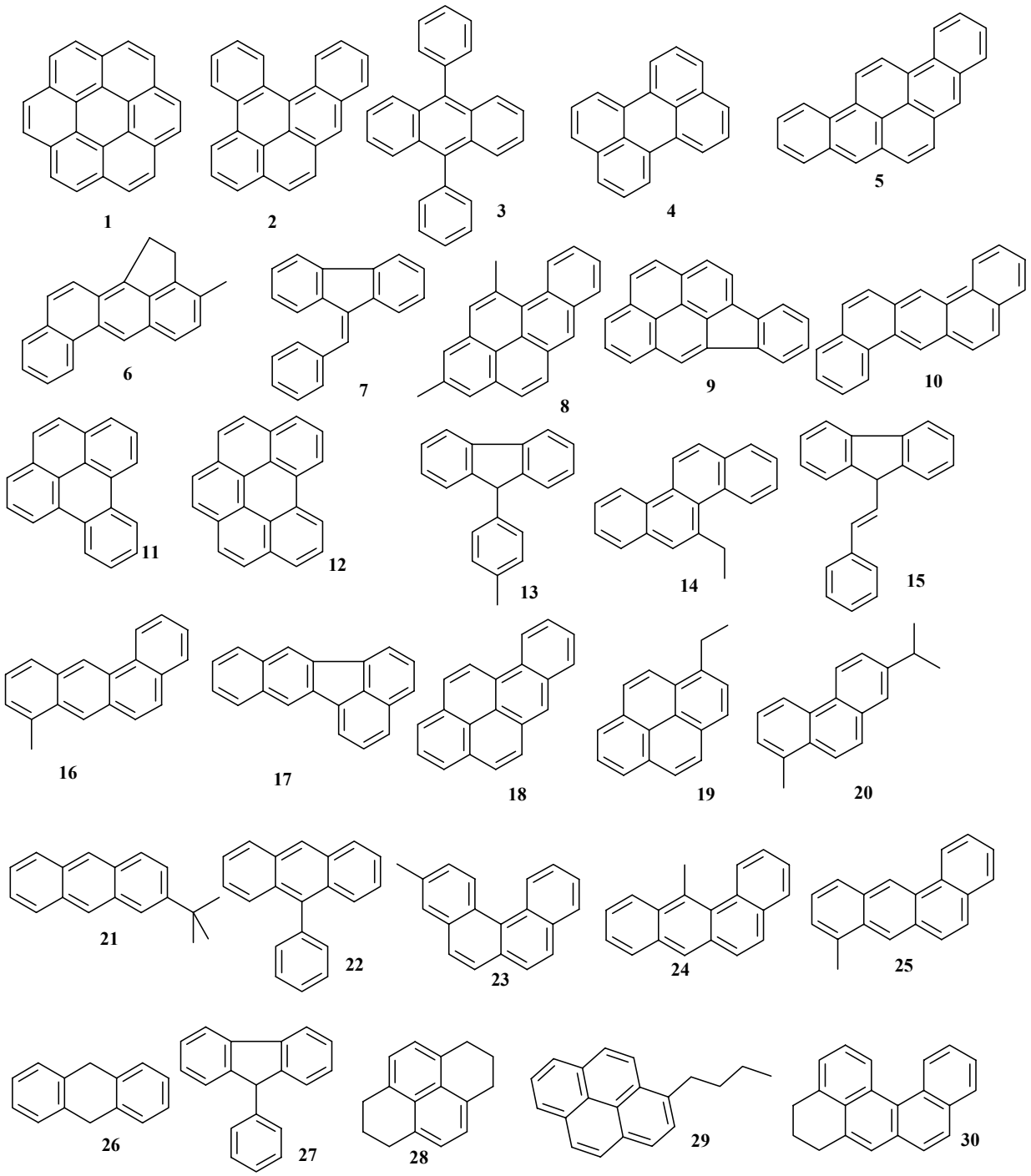

Figure 1. Structure of polycyclic aromatic hydrocarbons.

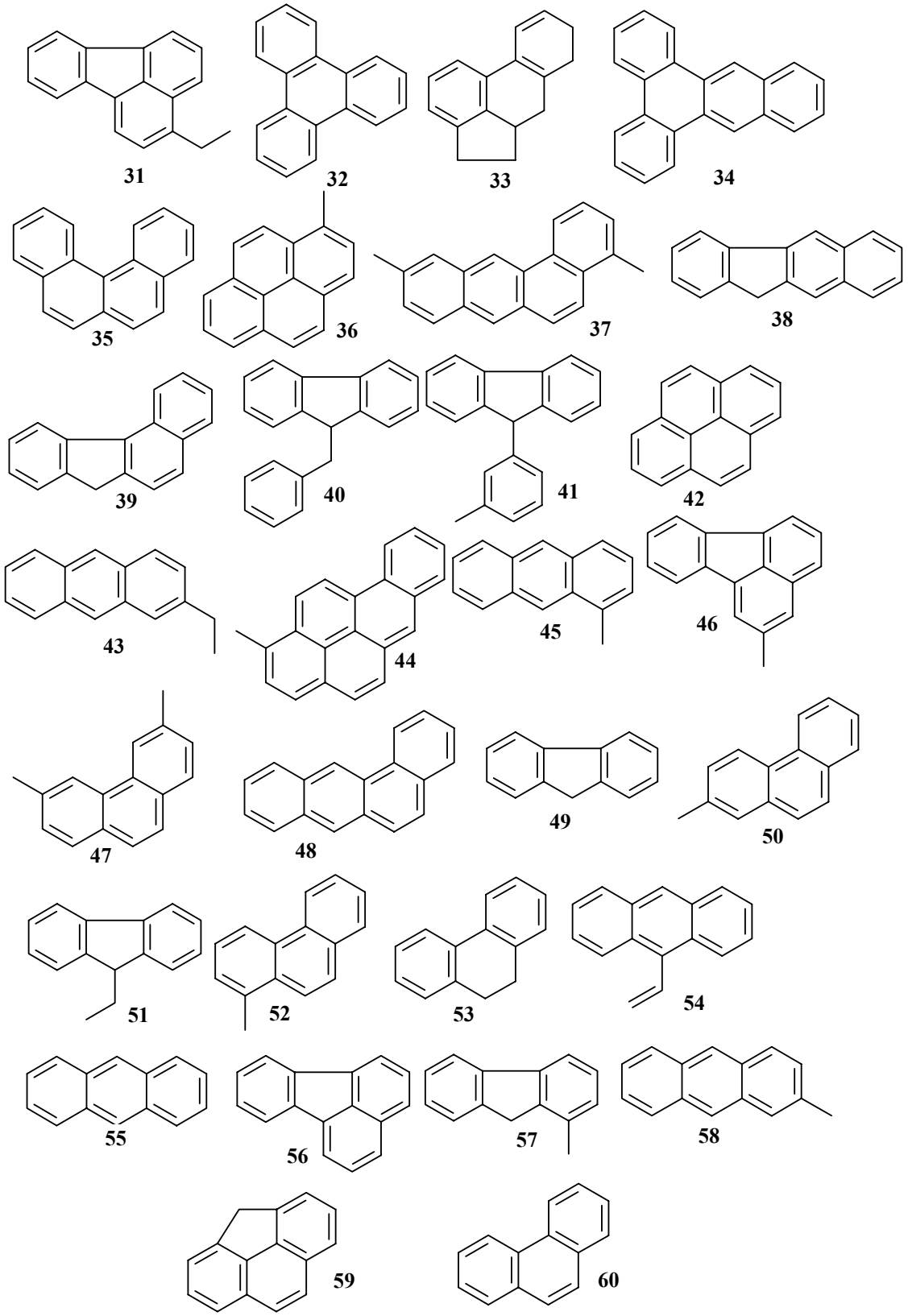

Figure 1. (Continued).

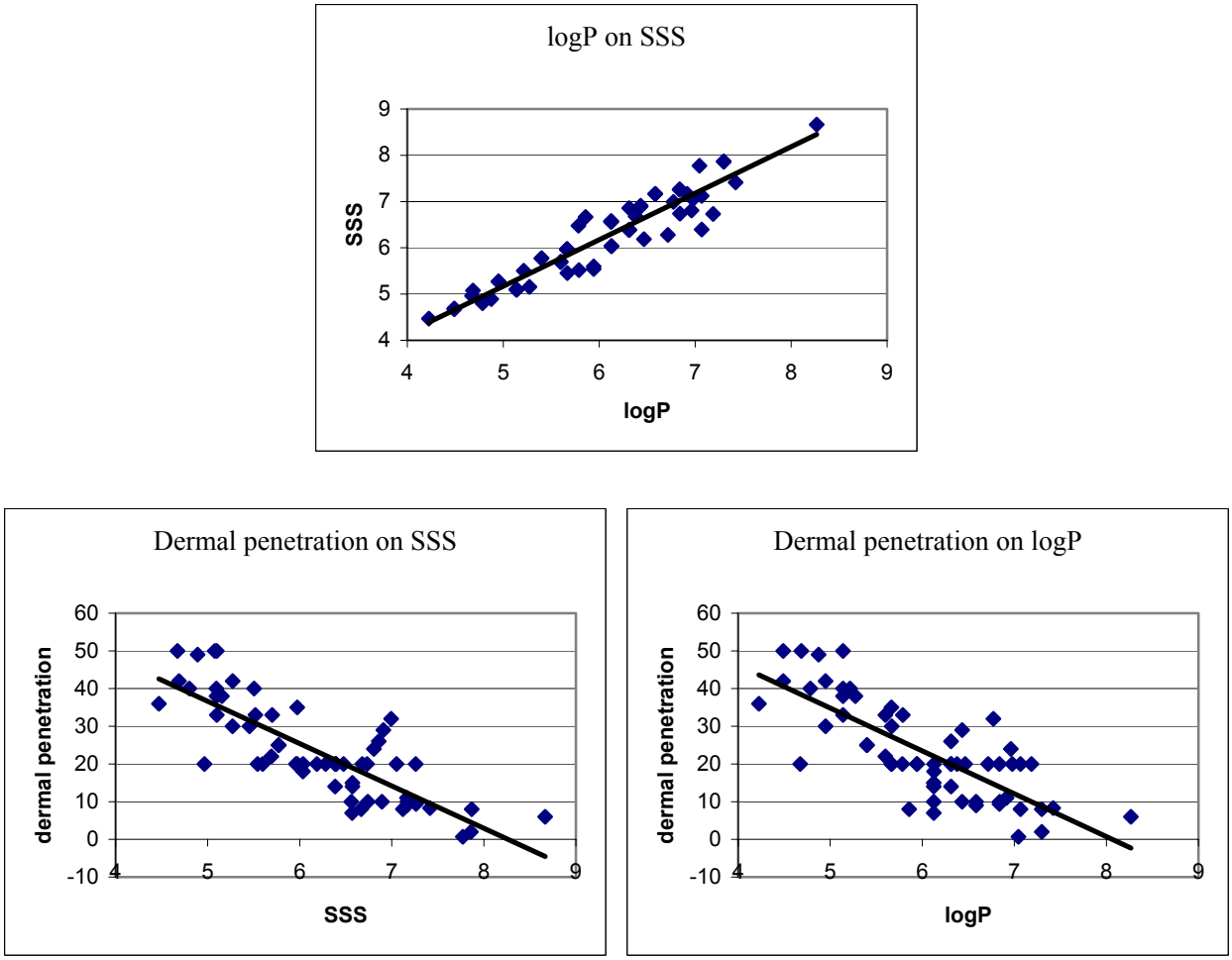

Figure 2. Regression lines relative to PAH dermal penetration analysis.

3.2 Aromatic Amine Carcinogenicity

The second set we are going to study includes 82 aromatic amines studied by Franke *et al*. [17] in order to predict their carcinogenicity against mice and rats. The set includes very different compounds, containing different number of aromatic rings and diverse functional groups (see Figure 3). Therefore, it represents a challenge for the method.

In their paper Franke *et al*. [17] show that in a correlation equation between molecular descriptors and carcinogenicity it is often present one term that can be attributed to the transport/ penetration capability of compounds. This part of their model requires the use of either calculated log *P* [22] or MR (molar refractivity, *i.e*. an index of steric properties). Consequently, we are going to concentrate on the possibility of correlation between their log *P* and our *SSS*. Here, again the first step is the choice of the model (solvent or solvents) to use. In this case, we have not a barrier to cross, but we must consider the interaction between a compound and its environment, that is mainly

water. Therefore, even if water and octanol would have been the natural choice to correlate to log *P*, we will use water alone. We will see that also the use of the water/octanol pair gives a similar result. The relevant data are reported in Table 3.

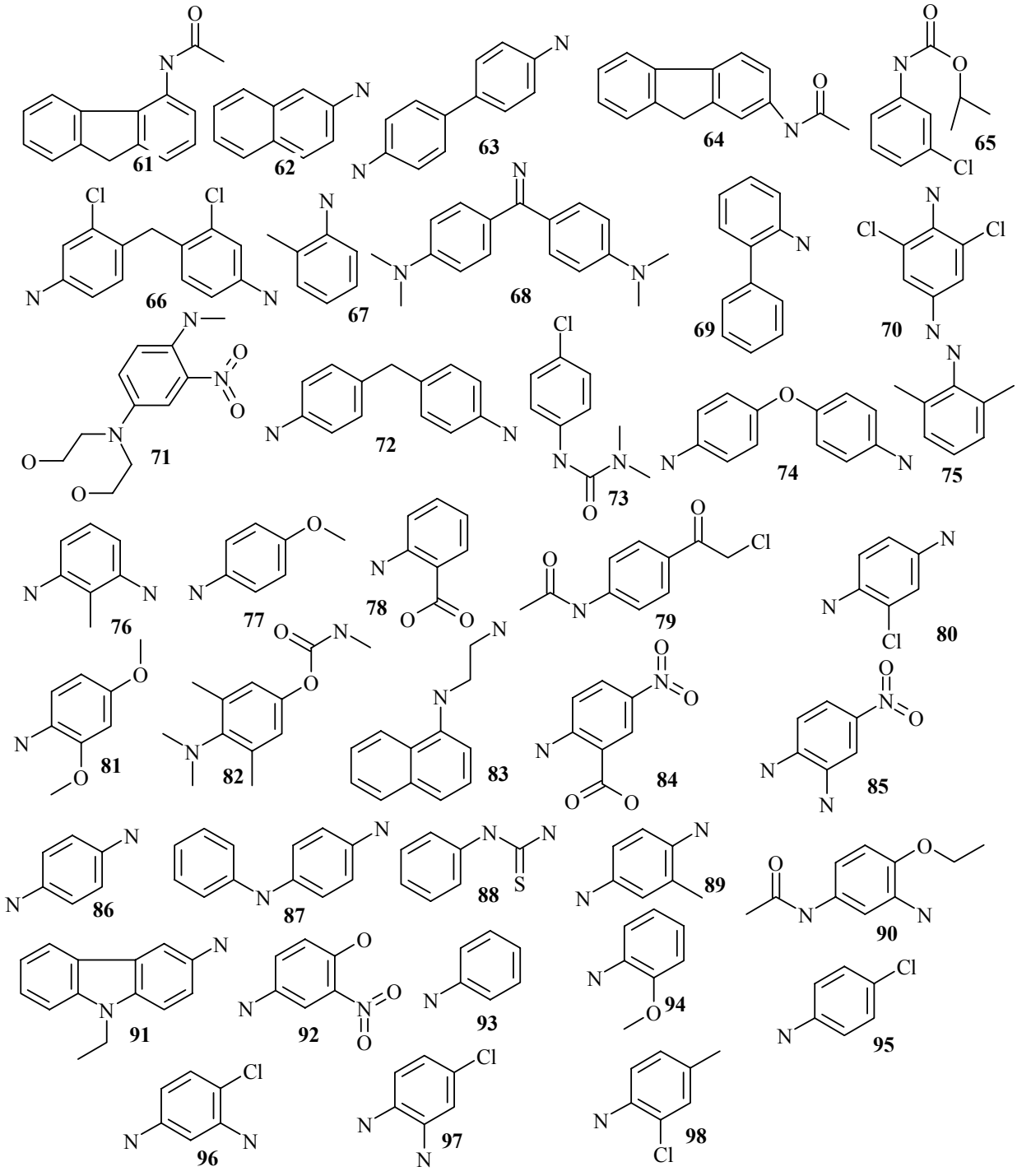

Figure 3. Compounds in Franke *et al*. study.

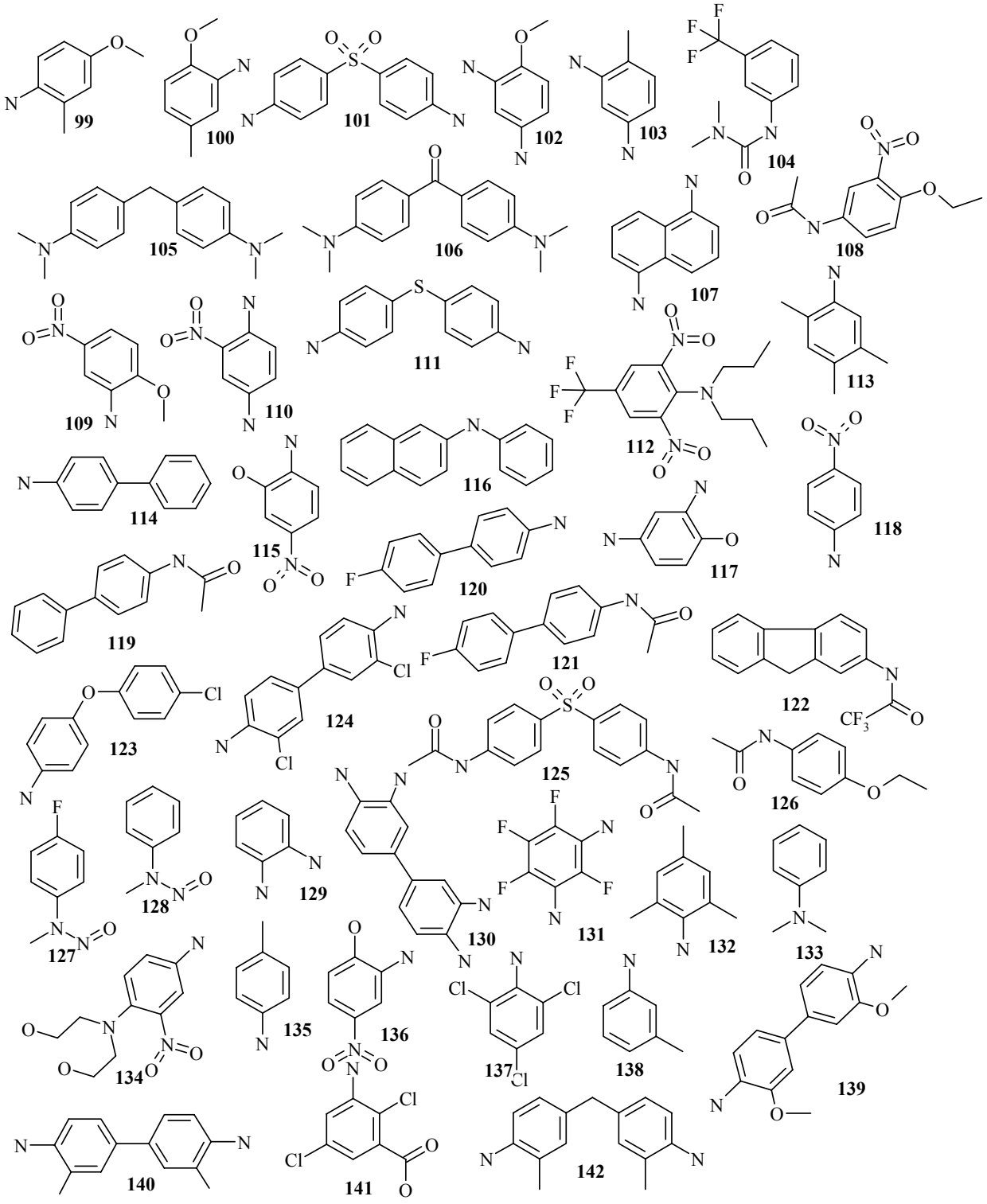

Figure 3. (Continued).

Table 3. Literature and calculated data of compounds in the carcinogenicity study

Compound	log *P*	water/ self [a]	water/ octanol [b]	Compound	log *P*	water/ self [a]	water/ octanol [b]
61	2.61	5.191	4.106	**102**	0.23	3.190	2.509
62	2.27	4.860	3.741	**103**	0.95	3.560	2.653
63	2.16	4.179	3.213	**104**	2	2.692	2.452
64	2.61	5.195	4.112	**105**	3.71	7.801	4.671
65	2.79	3.687	3.058	**106**	2.85	6.206	4.205
66	3.6	4.135	3.065	**107**	1.48	3.954	3.047
67	1.73	4.276	3.258	**108**	0.94	3.805	3.155
68	3.02	5.985	4.056	**109**	0.96	3.535	2.795
69	2.95	5.066	3.895	**110**	0.43	3.308	2.545
70	1.52	3.302	2.442	**111**	2.25	4.322	3.312
71	0.34	3.153	2.859	**112**	4.25	3.324	2.780
72	2.56	4.293	3.234	**113**	2.67	4.690	3.419
73	1.64	4.039	3.345	**114**	2.95	5.084	3.899
74	1.91	3.750	2.975	**115**	0.93	2.805	2.562
75	2.2	4.488	3.348	**116**	4.16	6.059	4.550
76	0.95	3.553	2.653	**117**	0.2	2.615	2.255
77	1.01	3.637	2.970	**118**	1.22	3.813	2.969
78	0.96	2.628	2.483	**119**	2.58	4.884	4.056
79	0.8	3.686	3.154	**120**	3.09	4.025	3.408
80	1	3.321	2.494	**121**	2.72	4.081	3.582
81	0.76	3.441	2.834	**122**	3.73	3.177	2.890
82	2.25	4.085	3.269	**123**	3.21	4.265	3.354
83	1.69	4.446	3.400	**124**	3.2	4.011	3.019
84	0.92	2.661	2.440	**125**	0.57	4.477	3.507
85	0.43	3.278	2.535	**126**	0.99	3.864	3.351
86	0.48	3.391	2.558	**127**	1.83	4.281	3.505
87	2.38	4.435	3.426	**128**	1.69	5.887	4.277
88	1.86	3.865	3.007	**129**	0.48	3.374	2.545
89	0.95	3.560	2.652	**130**	0.6	3.426	2.570
90	0.2	3.387	2.844	**131**	1.04	1.972	1.820
91	2.39	5.072	3.837	**132**	2.67	4.684	3.412
92	0.93	2.854	2.522	**133**	1.84	6.000	4.657
93	1.26	4.071	3.162	**134**	0.2	2.901	2.678
94	1.01	3.604	2.957	**135**	1.73	4.296	3.267
95	1.78	3.946	3.003	**136**	0.93	2.819	2.495
96	1	3.321	2.493	**137**	2.82	3.901	2.755
97	1	3.326	2.495	**138**	1.73	4.293	3.259
98	2.25	4.150	3.080	**139**	1.66	3.727	2.989
99	1.48	3.838	3.053	**140**	3.1	4.404	3.298
100	1.48	3.838	3.053	**141**	2	2.669	2.416
101	1.31	4.125	3.139	**142**	3.5	4.631	3.324

[a] water/self is the ratio between the *SSS* of the compound to water and to itself
[b] water/octanol is the ratio between the *SSS* of the compound to water and to octanol

The use of the *SSS* calculated for the similarity to water is not sufficient to explain all the variance. But, if we modulate the water *SSS* using the compound self–similarity, we get an acceptable correlation (log $P = 0.775\ SSS_{wat}/SSS_{self} - 1.496$, $n = 73$, $r^2 = 0.575$, $sd = 0.63$, F = 96.0, $r^2_{LOO} = 0.569$). This result, shown in Figure 4, has been obtained excluding 9 compounds that contain more than one halogen atom (**66**, **70**, **104**, **112**, **122**, **124**, **131**, **137**, **141**).

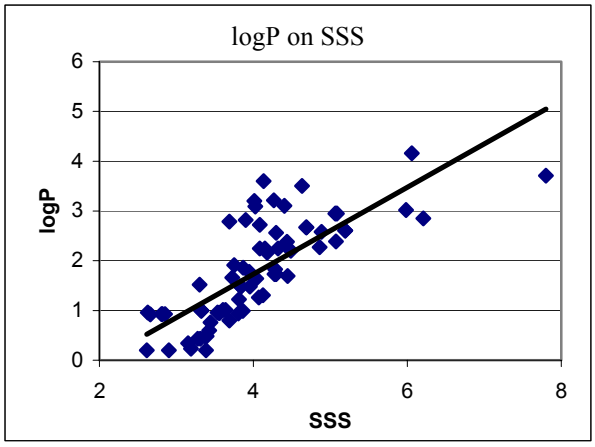

Figure 4. Correlation line between calculated and literature log *P* of Franke *et al*. data set.

If we use the *SSS* of water and octanol the result is comparable to the previuos one (log *P* = 1.277 SSS_{wat}/SSS_{oct} – 2.396, r^2 = 0.57, *sd* = 0.63, F = 96.0); however, we prefer our first choice because it compares the interaction of a compound with water and with itself, similarly to a solvation index.

It is worth to comment on the meaning of the similarity ratio. In the present context, *i.e.* comparing atom γ's, two compounds are more similar if they have many atoms that are electronically similar and few atoms that are electronically dissimilar. If we assume that this similarity can measure the intermolecular interactions, we can affirm that two similar compounds have many interactions and, as a consequence, they can mutually exchange without needing much energy. When we consider the interactions between hydrophobic compounds and water we can expect a limited similarity, thus the compounds are not sufficiently characterized. On the contrary, the self similarity represents an ideal situation because it must be the best possible similarity; the ratio is a method to balance the similarity to water using the best reference similarity. In the log *P* the comparison is between the solubility in water and in octanol, where this last is the hydrophobic solvent; thus, the similarity to octanol assumes the role of the self similarity, because the compounds are hydrophobic. The consequence is that the two ratios are in agreement.

3.3 Primary Aromatic Amine Mutagenicity

The last data set contains the compounds used by Basak *et al*. [18] to predict their mutagenicity, an experimental measure of the interaction of each compound with DNA in the well–known Ames test [23]. The set used to develop the model contains 95 highly varied compounds. The structures of the compounds are presented in Figures 5. This set contains 16 compounds that are also present in the previous set, but we will maintain the set as it is in order to make possible a comparison between the two.

In their paper Basak *et al*. [18] analyze the possibility to predict mutagenicity using some diverse equations that include from 4 to 9 descriptors, both topological and geometric and electronic. The correlation equations show different levels of predictive power. One of the descriptors that the authors have selected is the log *P* of the compounds; some log *P*'s are experimental, others are calculated [24]. Thus, we considered the possibility to correlate also in this case the log *P*'s to our *SSS*'s. Here, the choice of the model (solvent or solvents) to use is constrained; in fact, if we would like to compare these results to those of Franke *et al*. we need to use the same model, i.e. the ratio between the similarity of each compound to water and the self similarity.

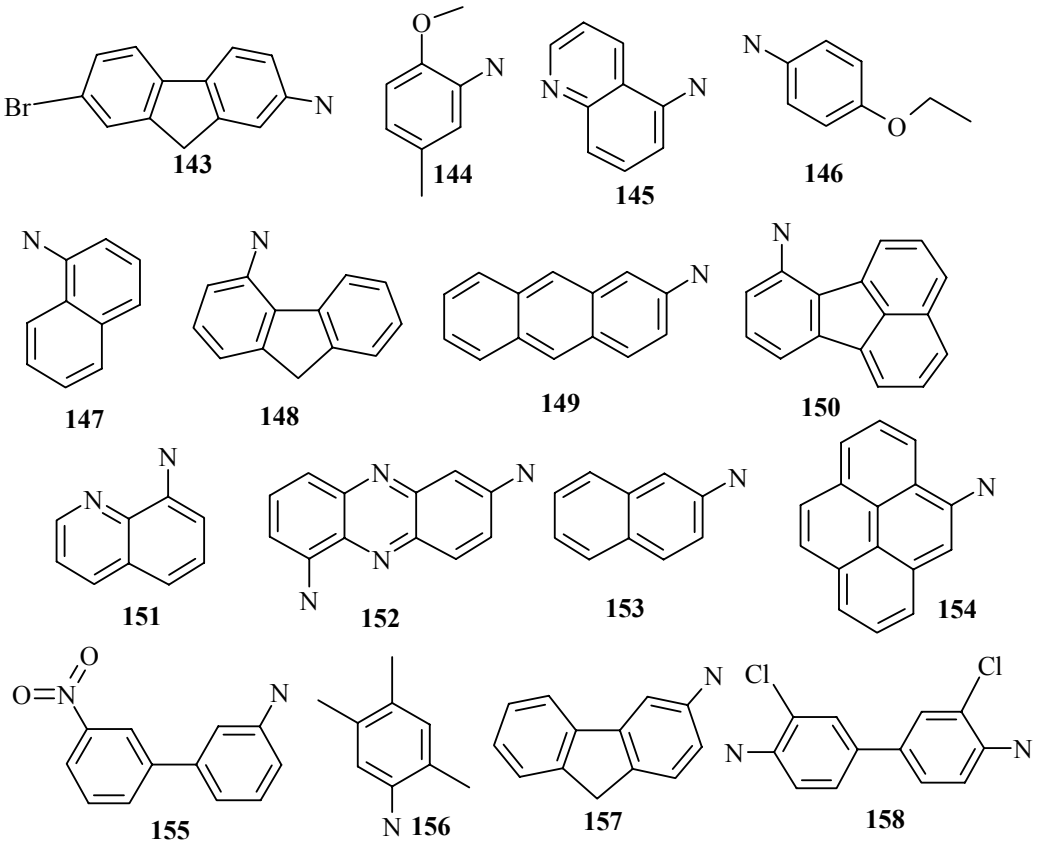

Figure 5. Compounds in Basak *et al*. study

The procedure applied to the Basak *et al*. [18] set gives a better correlation than in the Franke case (log $P = 1.029\ SSS_{wat}/SSS_{self} - 2.266$, $n = 90$, $r^2 = 0.741$, $sd = 0.46$, F = 251.66, $r^2_{LOO} = 0.738$); also here we excluded the 5 compounds containing more than one atom of chlorine or fluorine (**158**, **178**, **191**, **208**, **223**). The data are reported in Table 4 and the result in Figure 6. It is interesting to note that the correlation between the experimental log *P*'s and our *SSS* gives $r^2 = 0.68$, that is not as good as we could have preferred.

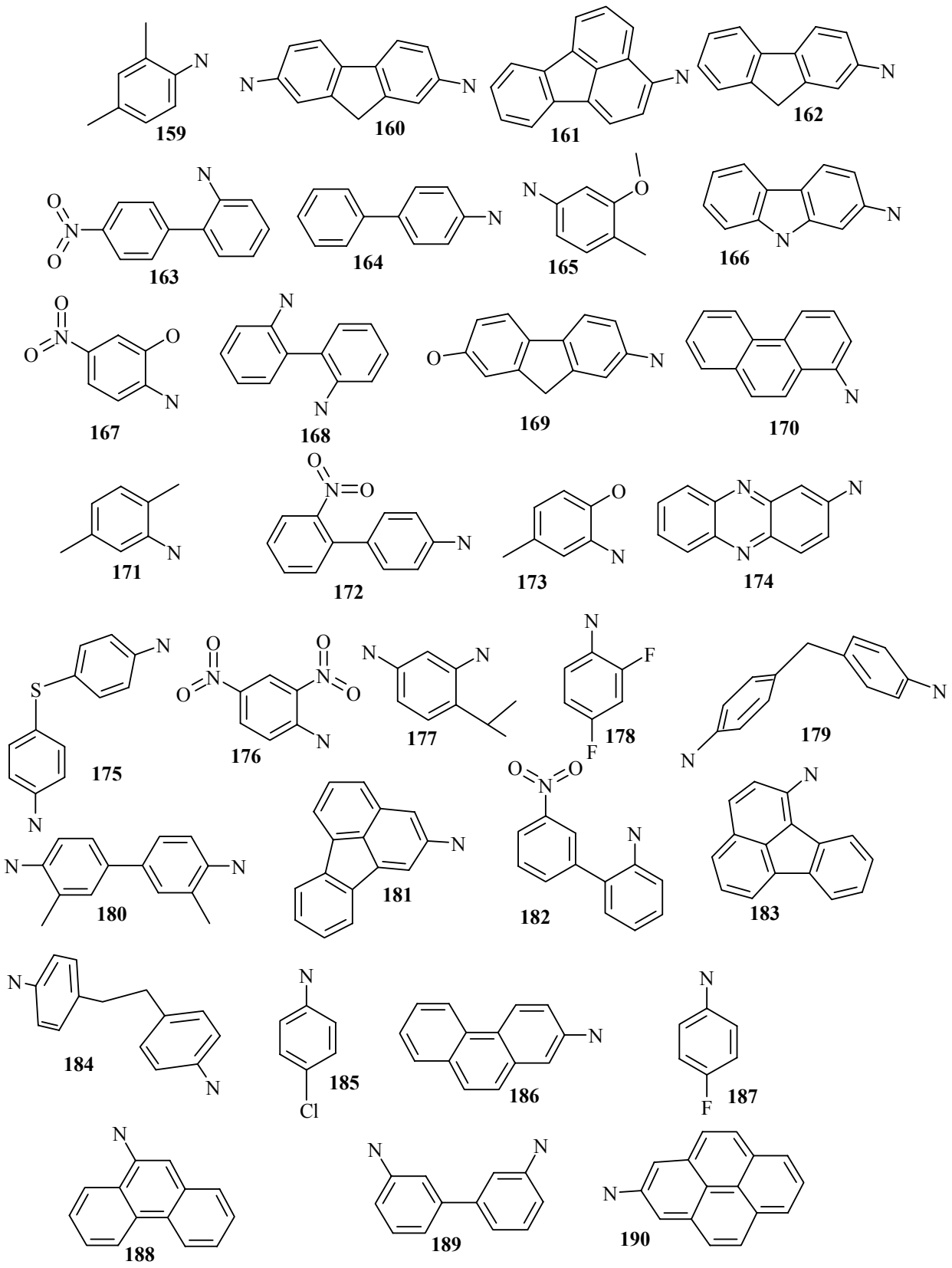

Figure 5. (Continued).

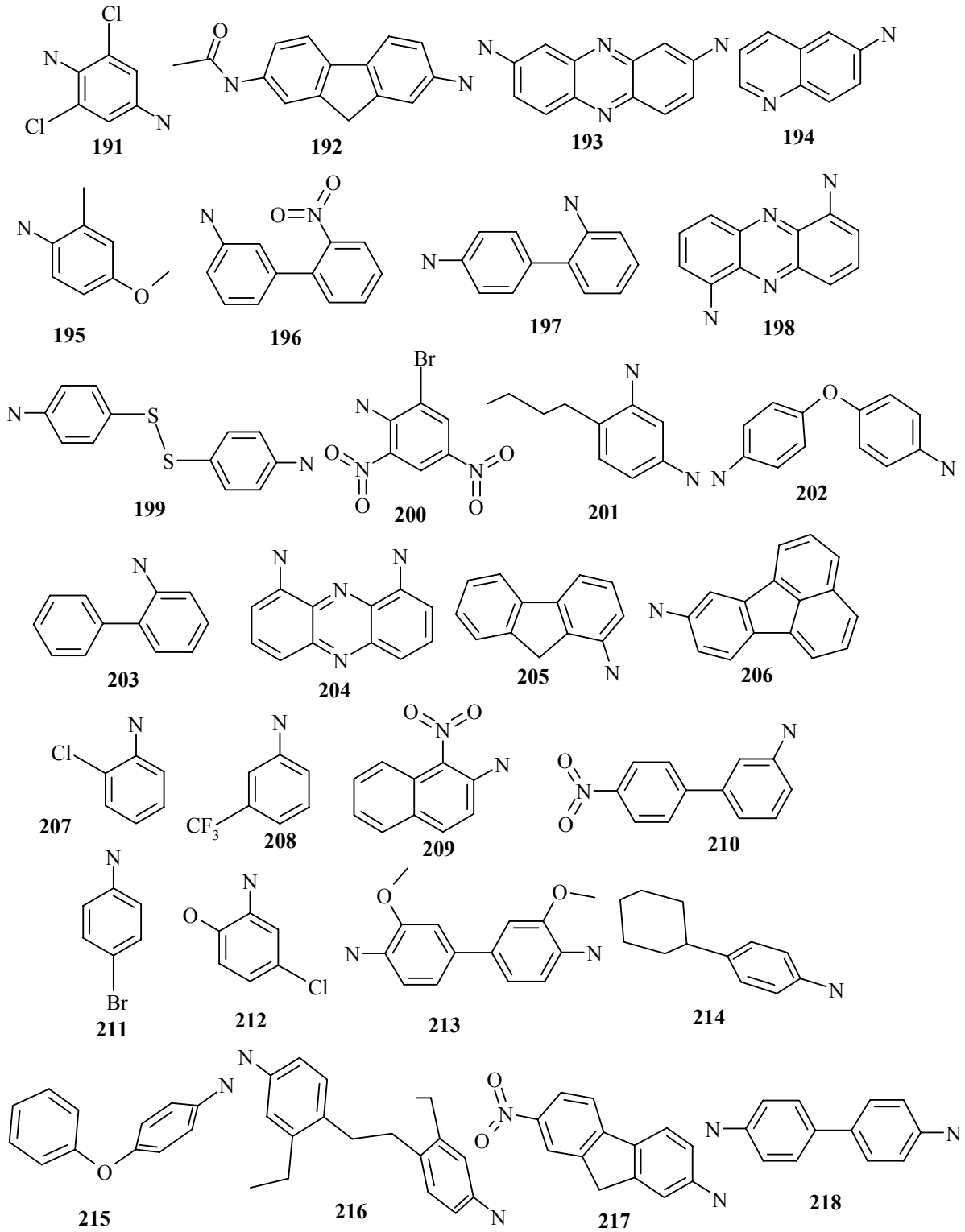

Figure 5. (Continued).

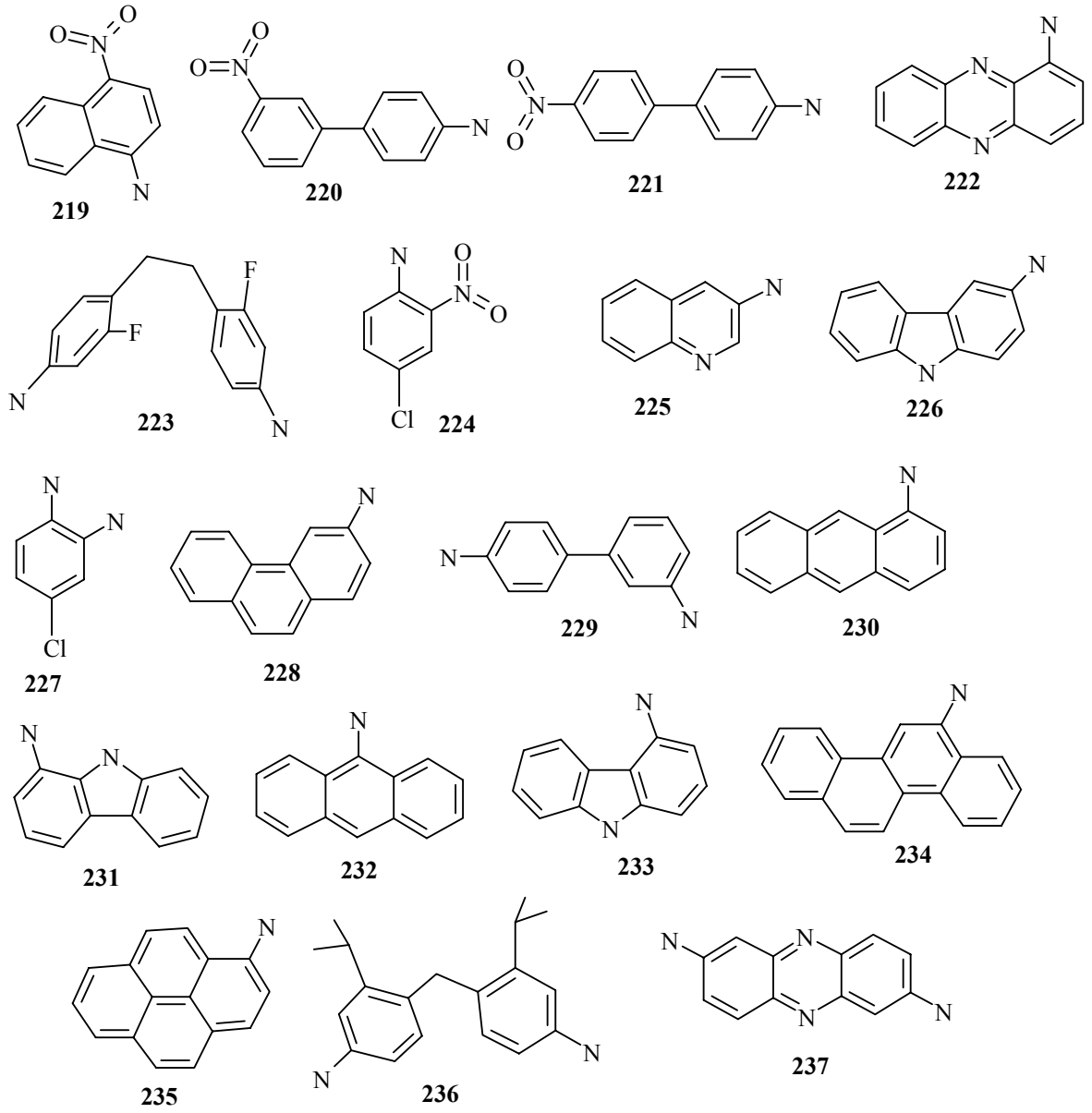

Figure 5. (Continued).

On the other hand, Basak *et al*. sometimes calculated the same log *P*'s for different isomeric structures; this is not the case with our calculations that are sensitive to regioisomer effects. Last, the correlation of the log *P*'s of the 16 compounds that are common to Franke *et al*. [17] (**62–153**, **63–218**, **69–203**, **70–191**, **72–179**, **74–202**, **95–185**, **97–227**, **99–195**, **100–144**, **113–156**, **114–164**, **115–167**, **124–158**, **139–213**, **140–180**) gives $r^2 = 0.65$, a demonstration of the difficulty that is still present in the calculation of this property.

Table 4. Literature and calculated data of compounds in the mutagenicity study.

No	log *P*	water/self [a]	water/octanol [b]	No	log *P*	water/self [a]	water/octanol [b]
143	3.92	5.644	4.005	**191**	1.79	3.302	2.442
144	1.74	3.819	3.048	**192**	1.72	4.357	3.458
145	1.16	4.392	3.388	**193**	1.64	3.916	2.943
146	1.24	3.779	3.100	**194**	1.28	4.391	3.390
147	2.25	4.846	3.741	**195**	1.23	3.838	3.053
148	2.7	5.412	3.952	**196**	2.68	4.570	3.532
149	3.26	5.441	4.099	**197**	1.58	4.166	3.211
150	3.72	5.817	4.268	**198**	1.64	3.878	2.923
151	1.79	4.363	3.373	**199**	1.99	4.470	3.406
152	1.64	3.895	2.933	**200**	2.78	3.809	2.857
153	2.28	4.860	3.741	**201**	1.77	3.917	2.964
154	3.72	5.820	4.287	**202**	1.36	3.750	2.975
155	2.68	4.491	3.530	**203**	2.84	5.066	3.895
156	2.41	4.690	3.419	**204**	1.64	3.878	2.923
157	2.7	5.435	3.952	**205**	3.18	5.416	3.954
158	3.51	4.011	3.019	**206**	3.72	5.838	4.268
159	1.68	4.503	3.352	**207**	1.9	3.914	2.982
160	1.47	4.422	3.278	**208**	2.29	2.360	2.133
161	4.2	5.829	4.277	**209**	2.95	4.539	3.448
162	3.14	5.434	3.952	**210**	2.68	4.555	3.539
163	2.68	4.618	3.564	**211**	2.26	4.312	3.305
164	2.86	5.084	3.899	**212**	1.81	2.788	2.503
165	1.52	3.833	3.072	**213**	1.81	3.727	2.989
166	2.3	4.499	3.443	**214**	3.65	5.030	3.817
167	1.36	2.805	2.562	**215**	2.96	4.337	3.486
168	1.58	4.157	3.208	**216**	3.66	5.458	3.512
169	2.03	3.912	3.329	**217**	3.06	4.939	3.671
170	3.26	5.423	4.102	**218**	1.34	4.179	3.213
171	1.83	4.503	3.354	**219**	2.48	4.471	3.428
172	2.68	4.652	3.570	**220**	2.68	4.492	3.535
173	1.16	2.975	2.653	**221**	2.68	4.619	3.561
174	2.18	4.544	3.391	**222**	2.18	4.517	3.376
175	2.18	4.322	3.312	**223**	2.5	3.320	2.852
176	1.84	3.701	2.802	**224**	2.72	3.823	2.888
177	1.12	3.790	2.846	**225**	1.63	4.372	3.387
178	1.54	2.565	2.362	**226**	2.3	4.496	3.443
179	1.59	4.293	3.234	**227**	1.28	3.326	2.495
180	2.34	4.404	3.298	**228**	3.26	5.439	4.101
181	3.72	5.844	4.269	**229**	1.58	4.176	3.209
182	2.68	4.483	3.533	**230**	3.69	5.425	4.100
183	3.72	5.827	4.273	**231**	2.3	4.481	3.433
184	2.13	4.391	3.309	**232**	3.26	5.401	4.090
185	1.88	3.946	3.003	**233**	2.3	4.486	3.443
186	3.26	5.439	4.100	**234**	4.98	5.885	4.360
187	1.15	3.068	2.682	**235**	4.31	5.791	4.273
188	3.56	5.445	4.114	**236**	4.46	5.538	3.516
189	1.58	4.175	3.204	**237**	1.64	3.916	2.943
190	3.72	5.809	4.265				

[a] water/self is the ratio between the *SSS* of the compound to water and to itself
[b] water/octanol is the ratio between the *SSS* of the compound to water and to octanol

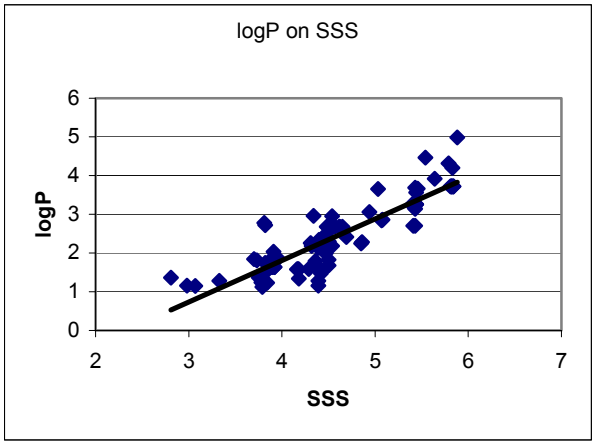

Figure 6. Correlation line between calculated and literature log *P* of Basak *et al*. data set.

3.4 Discussion

Considering that the interest towards the modeling of solvent effects is presently very high and its complexity is still challenging we propose a new approach. Its main characteristic is the use of a single general descriptor to evaluate the similarity between a compound and a solvent. A second attribute is the a priori definition of the environment we are modeling. This point is seldom present in current literature, where the environment definition is more a result than a hypothesis. Clearly, the definition need is a limit, because the possibility of finding a relation by chance is impossible; but it is also an advantage, because it calls for a better understanding. In addition, it can be effectively used to better define the model that is developing.

On the other hand, the choice of using a single descriptor is definitely a plus, both for the greater insight that it permits and for its simplicity. It is not our aim to criticize the multivariate analyses; on the contrary, we would like to emphasize that complex problems usually needs complex answers. However, our need of simplification for understanding is powered by the dissection of problems into pieces that can be separately studied.

If a complete, powerful system for the evaluation of the similarity between solutes and solvents was available, we probably could be in the position to effectively challenge the solvation problem. But, this is not the case, at least at the moment. There are difficulties inherent in the definition of similarity, as well as there are real obstacles in the definition of solvent effects. What we can expect is a preliminary step that can stimulate further work.

Let us now make a list of the disadvantages and of the advantages of the approach. The first problem is the necessity of the correct definition and choice of the model. For example, a popular problem is the evaluation of the water solubility of solid compounds. In this case it is not sufficient to consider the compound to water similarity to have a good response, because part (probably the most important) of the effect relates to the solid stability. On the other hand, the definition of the

model is highly recommended in all the cases where it is possible. A second problem is the precision of the value quantification. The system always permits the ordering of the objects, i.e. the most similar solvent will be always in the first position, etc. The absolute numbers are, however, less universal and more difficult to establish. For example, it is hard to imagine a similarity measure that gives a number you can use outside the single comparison. What is possible is to get a series of numbers in a series of comparisons to one reference. Then, we must decide if the comparison of two series of values has a meaning. If we assume that this is possible, then the measure will be acceptable. We must be always very careful in our similarity application.

Connected to the two mentioned problems is the presence of the outliers. Let we concede that there can be some outliers, nevertheless if the outliers are a compound class (as the polyhalogenated derivatives in our examples) we have a problem. Presently, we don't know where the problem is; the fact is that the electronegativity similarity between oxygen and fluorine or chlorine atoms has not an experimental counterpart. Consequently, the calculated values are always overestimated. Because we aim at the maximum of generality we don't want to introduce descriptors to care for this anomaly; on the contrary, we are still searching for the general solution inside the model.

Coming to the benefits, they are those that we expected: the model is general, flexible, and understandable. General, because we use the same common definition to solve different problems, only tuning the mathematical manipulation to conform to the current background. Flexible, because the model can be potentially applied to any problem that concerns solvent–solute interactions. It is up to us to choose the correct representation of the problem. Understandable, because this quality is present in the philosophy of the approach and represents its main strength. It is always possible to dissect the result to the point where the model limits appear clear.

Finally, let us comment our application results. We tested the approach by modeling two different problems: PAH dermal penetration; solvent effects on aromatic amines carcinogenicity and mutagenicity.

PAH dermal penetration is an experimentally measured quantity, and it is quite approximated, as demonstrated by the sameness of many experimental values. This not withstanding, Gute *et al*. [16] obtained an acceptable correlation (r^2 = 0. 695) selecting a single descriptor that describes the size/geometry of the compounds and Ivanciuc *et al*. [20] obtained an even better result (r^2 = 0. 711) using an electronic descriptor. The use of our method gives a comparable correlation (r^2 = 0.669) using a model based on the similarity to octanol. It is interesting that also the correlation between the *SSS* and the log *P* is quite good (r^2 = 0.882). This result shows that an extremely simple model (similarity to octanol) is able to capture the importance of water/octanol partition coefficient in modeling this biological effect. Part of the success is clearly due to the homogeneity of the set; nevertheless, the method is as effective as others and is easier to understand.

The aromatic amine sets were chosen as representative of multi component studies concerning an

extremely complex biological activity. Here, the solvent effect is only used to represent one of the variables affecting this activity: the compound transport. Thus, a direct correlation with carcinogenicity or mutagenicity is out of question. The only possible comparison is with the calculated log *P*. The use of the weighted similarity of each compound with water gives an acceptable result, for both data sets. It is clear that the agreement between the two variables (*SSS* and log *P*) is not good enough to allow for their mutual exchange; but the choice of either of them could improve the meaning of the entire study. However, the calculation of the log *P* for the same compound sometimes depends on the used method, as demonstrated by the compounds that the two sets of aromatic amines have in common.

4 CONCLUSIONS

The interaction between solvents and solutes is definitely a fundamental attribute of the complex interactions that influence the behavior of a multi component system, and the modeling of the perturbation it brings to the chemical activity is a highly desired objective. Due to the complexity of the problem and to the different expertise of the people interested in its study, there have been many proposals towards an effective approach to its understanding. Nevertheless, there is still need of new contributions to broaden our knowledge. In this perspective, we have introduced a slightly different approach to solute–solvent interaction modeling; it is based on a general use of the concept of similarity in solvation that can be applied to different problems without changing the basic principles of the system. We thus propose the use of a single descriptor, the solvent – solute similarity, inside a group of models built to consider the specific aspects of each problem. In a sense, we propose to move the description complexity from the compound to the environment in the conviction that this should allow for a better understanding. It is clear that inside this scheme it will be possible to change both the molecular descriptor and the model definition; but we are confident that the general philosophy can provide many benefits to the problem solution.

Acknowledgment

Partial financial support by the Consiglio Nazionale delle Ricerche, and by the Ministero dell'Universita' e della Ricerca Scientifica e Tecnologica, is gratefully acknowledged.

5 REFERENCES

[1] Y. Ran and S. H. Yalkowsky, Prediction of Drug Solubility by the General Solubility Equation (GSE), *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354–357.

[2] J. W. McFarland, A. Avdeef, C. M. Berger, and O. A. Raevsky, Estimating the Water Solubilities of Crystalline Compounds from Their Chemical Structures Alone, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1355–1359

[3] D. Yaffe, Y. Cohen, G. Espinosa, A. Arenas, and F. Giralt, A Fuzzy ARTMAP Based on Quantitative Structure–Property Relationships (QSPRs) for Predicting Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177–1207.

[4] N. R. McElroy and P. C. Jurs, Prediction of Aqueous Solubility of Heteroatom–Containing Organic Compounds from Molecular Structure, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237–1247.

[5] I. V. Tetko, V. Y. Tanchuk, and A. E. P. Villa, Prediction of n–Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E–State Indices, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407–1421
[6] P. Bruneau, Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605–1616.
[7] R. Liu and S.–S. So, Development of Quantitative Structure–Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633–1639.
[8] G. Klopman and H. Zhu, Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439–445.
[9] M. Durante and G. Sello, The Prediction of Organic Reaction Products: Determining the Best Reaction Conditions, *J. Chem Inf. Comput. Sci.* **2000**, *40*, 221–235.
[10] L. Pauling, *The Nature of the Chemical Bond 3rd ed.*, Cornell University, Ithaca NY, 1960.
[11] L. Baumer, G. Sala, and G. Sello, Residual Charges on Atoms in Organic Structures: A New Algorithm for Their Calculation. *Tetrahedron Comp. Met.* **1989**, *2*, 37–46.
[12] L. Baumer, G. Sala, and G. Sello, Residual Charges on Atoms in Organic Structures: A New Method for the Identification of Conjugated Systems and the Evaluation of Atomic Charge Distribution on Them. *Tetrahedron Comp. Met.* **1989**, *2*, 93–103.
[13] L. Baumer, G. Sala, and G. Sello, Residual Charges on Atoms in Organic Structures: Molecules Containing Charged and Backdonating Atoms. *Tetrahedron Comp. Met.* **1989**, *2*, 105–118.
[14] G. Sello, Empirical Atomic Charges: a 3–D Approach. *Teochem.* **1995**, *340*, 15–28.
[15] G. Sello, Similarity Measures: Is It Possible to Compare Dissimilar Structures? *J. Chem. Inf. Comput. Sci.***1998**, *38*, 691–701.
[16] B. D. Gute, G. D. Grunwald, and S. C. Basak, Prediction of the Dermal Penetration of Polycyclic Aromatic Hydrocarbons (PAHs): A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1999**, *10*, 1–16.
[17] R. Franke, A. Gruska, A. Giuliani, and R. Benigni, Prediction of Rodent Carcinogenicity of Aromatic Amines: a Quantitative Structure–Activity Relationships Model. *Carcinogenesis* **2001**, *22*, 1561–1571.
[18] S. C. Basak, D. R. Mills, A. T. Balaban, and B. D. Gute, Prediction of Mutagenicity of Aromatic and Heteroaromatic Amines from Structure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 671–678.
[19] A. Leo, D. Weininger, *CLOGP Version 3.2 User Reference Manual*. Medicinal Chemistry Project. Pomona College, Claremont, CA, USA. **1984**.
[20] T. A. Roy, W. Neil, J. J. Yang, A. J. Krueger, A. M. Arroyo, and C. R. Mackerer, SAR Models for Estimating the Percutaneous Absorption of Polynuclear Aromatic Hydrocarbons. *SAR QSAR Environ. Res.* **1998**, *9*, 171–185.
[21] O. Ivanciuc, T. Ivanciuc, and A. T. Balaban, QSAR Models for the Dermal Penetration of Polycyclic Aromatic Hydrocarbons, *Internet Electron. J. Mol. Des.* **2002**, *1*, 559–571, http://www.biochempress.com.
[22] TSAR. Oxford Molecular, Oxford, UK.
[23] B. N. Ames, Mutagenesis and Carcinogenesis: Endogenous and Exogenous Factors. *Environ. Mol. Mutagen.* **1989**, *14 (Suppl. 16)*, 66–77.
[24] A. K. Debnath, G. Debnath, A. J. Shusterman, and C. Hansch, A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.* **1992**, *19*, 37–52.