Simple Linear QSAR Models Based on Quantum Similarity Measures

Lluís Amat and Ramon Carbó-Dorca*

Institute of Computational Chemistry, University of Girona, Catalonia, 17071 Spain

Robert Ponec

Institute of Chemical Process Fundamentals, Czech Academy of Sciences, Prague 6, Suchdol 2, 165 02 Czech Republic

Received May 5, 1999

A novel QSAR approach based on quantum similarity measures was developed and tested in this paper. This approach consists of replacing the usual physicochemical parameters employed in QSAR analysis, such as octanol−water partition coefficient or Hammett σ constant, by appropriate quantum chemical descriptors. The methodological basis for this substitution is found in recent theoretical studies [*J. Comput. Chem.* **1998**, *19*, 1575−1583, *J. Comput.-Aided Mol. Des.* **1999**, *13*, 259−270], in which it was demonstrated that both molecular hydrophobic character and electronic substituent effect can be modeled by appropriately chosen quantum self-similarity measures (QS-SM). The most important aim of this study was to prove that selected QS-SM descriptors can be advantageously used in empirical QSAR analysis instead of classical descriptors. For this purpose several QSAR correlations are proposed, in which empirical descriptors such as Hammett σ constants or log P values are replaced by the appropriate QS-SM. These examples involve: (i) a set of benzenesulfonamides which bind to human carbonic anhydrase, (ii) a set of benzylamines as competitive inhibitors of the enzyme trypsin, and (iii) a set of indole derivatives which are benzodiazepine receptor inverse agonist site ligands. Simple linear QSAR models were developed in order to obtain mathematical relationships between the biological activity and the pertinent quantum chemical descriptors. The validity of the obtained QSAR models is supported by comparison of the observed and predicted values of the biological activity and by a statistical analysis based on a randomization test.

Introduction

In the past few years much effort has been devoted to applying the idea of quantum similarity measures (QSM) to rational drug design.[1−15] Because of its importance, this area of chemistry has experienced rapid growth. The mathematical background for this new expanding field was formulated some time ago by Carbó et al.,[16] who introduced the concept of QSM. Since then, great progress has been made not only in basic methodology but also in the formulation of robust computational schemes.[17−28] The basic idea of the above similarity approach to QSAR is to replace the traditional parameters in empirical QSAR analysis by selected theoretical descriptors based on QSM.

In keeping with this general philosophy, the present article reports an attempt to develop simple linear QSAR models based on quantum mechanical descriptors instead of empirical physicochemical parameters, characterizing the molecular hydrophobicity and electronic substituent effect in classical QSAR. The study is based on previous reports which described how the quantum self-similarity measure (QS-SM) of the whole molecule could be used as a descriptor of molecular hydrophobicity (log P).[5,6] Similarly, the electronic substituent effect may be appropriately modeled by fragment QS-SM corresponding to a functional group (re)active in a given process.[6,7] The fact that electronic phenomena such as the substituent effect can be replaced by means of QS-SM attached to local molecular regions can be explained through the recently reported holographic electron density theorem.[29] This theorem states that all the information contained in the total electron density of the whole molecule is also contained in the density of any local fragment of the molecule. In consequence, the QS-SM characterizing the functional group (re)active in a given process can be used as an appropriate descriptor.

Several molecular sets were examined in this study: (i) a series of benzenesulfonamides that show some affinity to binding to the human carbonic anhydrase; (ii) a series of benzylamine derivatives as competitive inhibitors of the proteolytic enzyme trypsin; (iii) a set of indole derivatives which are benzodiazepine receptor inverse agonists. Indole derivatives are able to displace [^{3}H]flunitrazepam from binding to bovine cortical membranes.

As will be shown, the theoretical QSAR models using QSM descriptors show statistical reliability comparable to descriptors derived from empirical correlations.[30−32]

Theoretical Framework

The idea of QSM arises from the incorporation of the intuitive concept of molecular similarity into the framework of quantum mechanics. According to this mechanics, all the information concerning a quantum object (QO) is contained in the associated electron density

* To whom correspondence should be addressed. E-mail: director@iqc.udg.es. Phone: 34 972 418359. Fax: 34 972 418356.

10.1021/jm9910728 CCC: $18.00
Published on Web 11/25/1999

function obtained from the corresponding wave function square module. From this point of view, electron density can be regarded as an ultimate molecular descriptor. In consequence, the similarity of any two QO can be assessed quantitatively by comparing the similarity of the corresponding electron density clouds.[27]

In this way, a consistent expression of QSM between two QO *A* and *B*, described by the first-order density functions $\{\rho_A(r), \rho_B(r)\}$, is defined by the integral

$$Z_{AB}(\Omega) = \int\int \rho_A(r_1)\, \Omega(r_1, r_2)\, \rho_B(r_2)\, dr_1\, dr_2 \quad (1)$$

where Ω is a positive definite operator. The most commonly used form of the two-electron operator Ω is a Dirac δ function: $\Omega(r_1,r_2) = \delta(r_1 - r_2)$. By replacing this operator into eq 1, a general definition of the so-called overlap-like QSM is obtained:

$$Z_{AB} = \int \rho_A(r)\, \rho_B(r)\, dr \quad (2)$$

The values of similarity measures Z_{AB} defined by eq 2 depend on the relative translation and orientation of the compared QO *A* and *B* in 3D space. This implies that in order to get a meaningful and unique value of QSM, the relative mutual position of both QO has to be optimized so that a maximal value of the integral in eq 2 is reached.[28] However, such optimization depending of the relative position of both QO is no longer necessary when *A* and *B* are identical. In this case, definition 2 yields an invariant quantum self-similarity measure (QS-SM):

$$Z_{AA} = \int |\rho_A(r)|^2\, dr \quad (3)$$

The fact that the relative position optimization becomes irrelevant for self-similarity measures is crucial from a computational point of view, not only because of substantial reduction of similarity integral measure computation time, but also because it reduces the conformational dependence of the results. As a consequence, because the 3D alignment procedure is avoided, QS-SM acquire special relevance as molecular descriptors in building up quantum mechanical equivalents of empirical QSAR Hansch-like models.

The procedure of generating these theoretical QSAR models is as follows: First, the appropriate QS-SM are computed for the whole series of compounds belonging to the studied set. These self-similarity measures are then arranged in the form of column vectors, **z**, entering into linear regression analysis. But before this, each column vector **z** is standardized so as to give new scaled variables with zero mean and unit variance. The idea underlying such statistical standardization is to ensure comparable weights of individual molecular descriptors in the final QSAR model. This standardization is described in the usual way

$$\theta = s^{-1}(\mathbf{z} - \langle \mathbf{z} \rangle 1) \quad (4)$$

where *s* and $\langle \mathbf{z} \rangle$ are the standard deviation and the arithmetic mean of the original descriptors, respectively.

A great number of methodologies and computational algorithms have been developed for the practical implementation of QSM, which opens up the possibility of their application in many areas of theoretical chemistry. Among these techniques, the most widely used is what is known as atomic shell approximation (ASA).[24−26] Since this approximation is also employed for the calculation of QS-SM in this study, we consider the basic idea of the ASA approach worth re-stating.

ASA density functions are constructed as a linear combination of spherical functions, with the restriction that all coefficients of expansion have to be real and positive. Thus, this constraint enables the statistical meaning of a correct probability distribution to be preserved. In addition, a *promolecular* model is employed, based on a plausible description of molecular density functions as a sum of individual atomic contributions. Then, the first-order density function under the *promolecular* ASA form for a molecular QO *A* may be expressed as

$$\rho_A^{\mathrm{ASA}}(r) = \sum_{a \in A} P_a \rho_a^{\mathrm{ASA}}(r) \quad (5)$$

where the coefficient P_a represents the atom *a* total charge, and $\rho_A^{\mathrm{ASA}}(r)$ the density function. In the present study, QS-SM were computed using weighting factors P_a equal to total valence atomic charge on individual atoms. Density for a given atom *a* is expressed as a linear combination of square normalized 1S-type GTO

$$\rho_a^{\mathrm{ASA}}(r) = \sum_{i \in a} w_i |S_i(r - R_a; \zeta_i)|^2 \quad (6)$$

where the sum in eq 6 is performed over the functions associated to the atomic shells. Within this promolecular ASA model, only the coefficients w_i and exponents ζ_i are needed to construct the density function. In this paper, one function is used to modulate density on H atoms, three functions are needed for C, O, and N atoms, and four functions are needed for Cl atoms. Coefficients w_i and exponents ζ_i used here can be downloaded from a World Wide Web site.[33]

Simple Linear QSAR Models Using QS-SM

Around 1960, Hansch and co-workers[34,35] introduced a new approach which proved to be especially fruitful in the field of rational drug design. Their approach was based on the application of linear-free energy relationships (LFER) to correlate biological activities with appropriate physicochemical descriptors. Since then, the application of what is known as QSAR became a respectable and widely used methodology in pharmacochemical research. A wealth of empirical descriptors relating to various physicochemical properties was introduced in consequence. The fundamental idea of the Hansch approach consists of the design of suitable QSAR models in the form of a multiple linear regression (MLR) between physicochemical descriptors and biological activities:

$$\text{biological activity} = f(\text{molecular or fragmental contributions}) = f(\log P, \sigma, E_s) \quad (7)$$

The most usual factors determining the biological activity are the hydrophobic character, characterized by log *P* values, and the substituent electronic and steric effect represented by the Hammett and Taft constants.

On the basis of these parameters, the traditional drug design consists of combining these molecular descriptors in the form of an MLR so as to get the best statistical description of the biological data.

In order to place the above empirical process on a safer theoretical footing and so to provide a theoretical interpretation of the origins of QSAR, a mathematical formalism, based on the combination of the idea of molecular similarity with quantum mechanical postulates, has been proposed in a recent study.[23] In the following part, the basic idea of this rationalization will be briefly explained.

According to quantum mechanics, any observable property of a quantum system I, π_I, for which the density function $\rho_I(r)$ is known, can be calculated as the expectation value of an associated hermitean operator $\Omega(r)$

$$\pi_I = \langle \omega \rangle = \int \Omega(r)\, \rho_I(r)\, \mathrm{d}r \tag{8}$$

Equation 8 represents a continuous description of an observable property. However, such a continuous description is considerably different from the intrinsically discrete form of empirical QSAR. A clue to the resolution of this difference and to the theoretical formulation of QSAR lies precisely in the application of the idea of molecular similarity. For this purpose, given a set of molecules (A, B, C, ... M) whose properties will be studied, first pairwise QSM for all possible molecular couples is calculated. These QSM can be conveniently arranged in the form of a matrix $\mathbf{Z} = \{Z_{IJ}\}$, which can be considered as a hypermatrix formed by column vectors as elements, $\mathbf{Z} = \{\mathbf{z}_I\}$. Using this symmetric matrix, the molecular property π_I can be approximated according to the general equation

$$\pi_I \approx \mathbf{a}^{\mathrm{T}} \mathbf{Z}_I = \sum_K a_K Z_{KI} \tag{9}$$

where **a** is an n-dimensional vector associated with the discrete representation of the unknown operator Ω. This equation represents the discrete counterpart of eq 8. The unknown coefficients **a**, characterizing the operator Ω, can be determined in a least-squares manner.

Although eq 9 represents the most general form of theoretical QSAR models, in some cases the form of the correlation equation can be further simplified. Such a simplification is typical of a situation in which it is possible to extract the QS-SM, Z_{II}, from the rest of the elements of the similarity matrix, $\{Z_{KI}, K \neq I\}$, leading to the result:

$$\pi_I \approx a_I Z_{II} + \sum_{K \neq I} a_K Z_{KI} \tag{10}$$

In some cases, particularly when homogeneous series of QO are studied, the terms $\alpha = a_I$ and $\beta = \sum_{K \neq I} a_K Z_{KI}$ can be considered as constants. Then, a simple linear equation may be expressed by means of the QS-SM, which represents the theoretical counterpart of the simple one-parameter QSAR model, like the Hammett equation:

$$\pi_I \approx \alpha Z_{II} + \beta \tag{11}$$

The situation with the correlation of biological data is, however, more complex since the final biological effect is usually due to the combination of several different factors. This suggests that in this case the theoretical QSAR models should have the form of multilineal correlation equations. Another important factor, which plays an important role when correlating biological data, is that the majority of the processes responsible for the observed activity are usually restricted to certain more or less localized regions of the molecule (pharmacophore, binding site, etc). As a consequence, in some cases it is possible and more useful to focus just on the comparison of these active molecular regions, R. Under these circumstances the original β constant in eq 11 can be approximately rewritten in the form of a set of fragment self-similarities. A justification of this procedure may be obtained when inspecting the definition of β in eq 11 as follows

$$\beta \approx \sum_{K \neq I} a_K \sum_{b \in K} \sum_{a \in I} P_b^K P_a^I Z_{ba}^{KI} \tag{12}$$

where $\{Z_{ba}^{KI}\}$ are interatomic similarity contributions involving molecules K and I to the global molecular QSM, Z_{KI}, when comparing molecular densities written in the ASA approach as in eq 5. Reordering terms

$$\beta \approx \sum_{a \in I} P_a^I \vartheta_a^I = B \tag{13}$$

and the symbol in the sum is defined as:

$$\vartheta_a^I = \sum_{K \neq I} a_K \sum_{b \in K} P_b^K Z_{ba}^{KI} \tag{14}$$

while the whole result can be further approximated using the fact that B only depends explicitly on atomic contributions of molecule I, that is,

$$B \approx \sum_{\mathrm{R}} \alpha_{\mathrm{R}} Z_{I,\mathrm{RR}} + \gamma \tag{15}$$

where R are groups of atoms present in molecule I as well as in the common skeleton shared by the rest of the studied molecular set. To smooth the successive approximation source of errors, the new coefficients, $\{\alpha_{\mathrm{R}}, \gamma\}$ in eq 15 are to be adapted to a specific molecular set, using a conventional fitting procedure.

Because $\beta \approx B$, eq 15, when substituted in eq 11, can be regarded as an alternative multilineal theoretical QSAR model. In this equation, $\{Z_{I,\mathrm{RR}}\}$ are the appropriate self-similarity measures for each individual fragment R contributing to the biological response. The fragment self-similarities constitute a simple way to take into account in some cases the variability of the supposed constant β. Moreover they provide information about the relevant parts of the common skeleton which can be taken as responsible for biological activity.

The basic goal of the present article is to demonstrate that in view of the analogy between empirical QSAR and theoretical equations, the physicochemical descriptors employed in classical QSAR studies can be replaced by appropriate theoretical descriptors based on QS-SM. In the following section some examples of such replacement will be presented, with the aim of showing that

theoretical QSAR models can be used to correlate biological data at least as successfully as the classical QSAR models.

Results and Discussion

Having introduced the necessary theoretical background, its application to the construction of theoretical QSAR models for the series of biologically active molecules studied will be reported in this section. These molecular series involve three well-defined cases: (i) benzenesulfonamides which show binding affinity with human carbonic anhydrase (HCA); (ii) benzylamine derivatives as competitive inhibitors of the proteolytic enzyme trypsin; (iii) indole derivatives which are benzodiazepine receptor inverse agonists.

1. Preliminary Considerations. Before presenting the results, some general remarks concerning the computation of QS-SM are summarized together with the statistical aspects of the QSARs obtained:

•Molecular geometry of all involved molecules was fully optimized using the AMPAC program[36] and semiempirical AM1 Hamiltonian.[37]

•The following QS-SM were calculated for each series of molecules: (a) QS-SM for the whole molecule as an alternative descriptor to log *P*;[5] (b) the variation of the electronic structure of the fragment presumably responsible for the biological activity, induced by the systematic variation of substituents, was modeled by QS-SM for appropriate molecular fragments.

•For the statistical analysis of the QSAR models, two regression coefficients were calculated: conventional squared regression coefficient (r^2) and the cross-validated (CV) coefficient for prediction (q^2). This latter coefficient, which permits evaluation of the predictive power of the model, is defined as $q^2 = (1 - \text{PRESS/SD})$, where PRESS (predictive residual sum of squares) is the sum of squared errors of predictions in a leave-one-out (LOO) CV analysis, and SD is the squared sum of the difference of the observed values from their mean. A statistically reasonable QSAR model usually requires the q^2 value to be greater than 0.6.[38]

•Using a nested summation symbol (NSS) algorithm,[39,40] all possible combinations of the computed QS-SM were generated and subsequently employed in QSAR models. Using this approach, the corresponding optimal QSAR model, in which a QS-SM set yields a maximal value of the q^2 coefficient, was chosen. In this way, the study focused on determining which QS-SM descriptors produced the linear regression model with the best predictability.

•Finally, to verify that the results of the QSAR models designed are not due to accidental correlations or to over-parametrization of the model, a randomization test[41] was performed. This test consists of randomly rearranging the order of the components of the vector of biological activity data and correlating these rearranged vectors with the vector of QS-SM. This procedure was repeated 100 times for each chosen QS-SM set, keeping the coefficients r^2 and q^2 for each random run and recording all the obtained (r^2, q^2) pairs as points on a graph at the end. A consistent QSAR model is obtained when only the original arrangement of the activities produces a satisfactory regression model.

Chart 1. Common Molecular Structure for Substituted Benzenesulfonamides

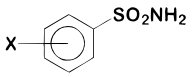


Table 1. Inhibitor Constants for the Binding of X-$C_6H_4SO_2NH_2$ to HCA

	X	observed log K^a
1	H	6.69
2	4-CH_3	7.09
3	4-C_2H_5	7.53
4	4-C_3H_7	7.77
5	4-C_4H_9	8.30
6	4-C_5H_{11}	8.86
7	4-CO_2CH_3	7.98
8	4-$CO_2C_2H_5$	8.50
9	4-$CO_2C_3H_7$	8.77
10	4-$CO_2C_4H_9$	9.11
11	4-$CO_2C_5H_{11}$	9.39
12	4-$CO_2C_6H_{13}$	9.39
13	4-$CONHCH_3$	7.08
14	4-$CONHC_2H_5$	7.53
15	4-$CONHC_3H_7$	8.08
16	4-$CONHC_4H_9$	8.49
17	4-$CONHC_5H_{11}$	8.75
18	4-$CONHC_6H_{13}$	8.88
19	4-$CONHC_7H_{15}$	8.93
20	3-CO_2CH_3	5.87
21	3-$CO_2C_2H_5$	6.21
22	3-$CO_2C_3H_7$	6.44
23	3-$CO_2C_4H_9$	6.95
24	3-$CO_2C_5H_{11}$	6.86
25	2-CO_2CH_3	4.41
26	2-$CO_2C_2H_5$	4.80
27	2-$CO_2C_3H_7$	5.28
28	2-$CO_2C_4H_9$	5.76
29	2-$CO_2C_5H_{11}$	6.18

[a] From ref 30.

2. Results. The first two examples presented in this paper refer to QSAR studies of enzyme−ligand interactions, and the third one deals with the prediction of the ability of substituted indole derivatives to displace [^{3}H] flunitrazepam from binding to bovine cortical membranes. The traditional approach to the description of such biological activity data is based on the construction of classical QSAR models using as descriptors hydrophobicity parameters (log *P*), Hammett substituent constant, etc. As was explained earlier, the aim of this study was to propose a new universal methodology, based on the use of theoretical QS-SM based descriptors, which could serve as an alternative procedure for designing new QSAR models.

(a) Benzenesulfonamide Derivatives. A set of 29 substituted benzenesulfonamides, with a common structure shown in Chart 1 and substituents listed in Table 1, was studied as HCA inhibitors. Traditional studies have shown that the HCA inhibitory activity of substituted benzenesulfonamides is predominantly influenced by two basic factors: the hydrophobic interactions of these molecules with enzyme−receptor cavity, and the electronic structure of the active SO_2NH_2 group reflecting the systematic variation of substituents within the series. On the basis of these findings, Hansch proposed empirical QSAR models[30] using log *P* and Hammett's substituent constant σ as appropriate descriptors. For the series of 19 *para*-substituted benzenesulfonamides

Table 2. QS-SM (Z_{AA}) and Scaled QS-SM (θ_{AA}) Used To Derive Eqs 17−19 and 22−24 for the Binding of X-$C_6H_4SO_2NH_2$ to HCA[a]

	Z_{AA}	θ_{AA}	$Z_{AA}^{SO_2NH_2}$	$\theta_{AA}^{SO_2NH_2}$	$Z_{AA}^{SO_2}$	$\theta_{AA}^{SO_2}$	$Z_{AA}^{NH_2}$	$\theta_{AA}^{NH_2}$	Z_{AA}^{m-C}	θ_{AA}^{m-C}
1	283.7310	−2.41624	189.7814	1.29909	151.9538	1.05833	37.8017	0.20978	15.0628	1.31094
2	298.9168	−2.13996	189.7945	1.56123	151.9679	1.29930	37.8008	0.15125	15.0831	1.44959
3	314.2117	−1.86171	189.7960	1.59287	151.9706	1.34686	37.7996	0.07627	15.0795	1.42482
4	329.5206	−1.58319	189.7943	1.55763	151.9689	1.31674	37.7996	0.07627	15.0860	1.46941
5	344.8184	−1.30488	189.7960	1.59231	151.9706	1.34650	37.7996	0.07544	15.0846	1.45950
6	360.1176	−1.02654	189.7960	1.59265	151.9706	1.34679	37.7996	0.07544	15.0853	1.46446
7	409.2965	−0.13184	189.6808	−0.72517	151.8467	−0.78089	37.8078	0.60322	14.6662	−1.39857
8	424.7788	0.14983	189.6858	−0.62495	151.8517	−0.69590	37.8079	0.60680	14.6690	−1.37903
9	440.0616	0.42787	189.6858	−0.62457	151.8516	−0.69606	37.8079	0.60782	14.6690	−1.37903
10	455.3495	0.70600	189.6857	−0.62594	151.8516	−0.69638	37.8079	0.60488	14.6697	−1.37414
11	470.6511	0.98438	189.6858	−0.62501	151.8517	−0.69568	37.8079	0.60571	14.6690	−1.37903
12	485.9479	1.26267	189.6857	−0.62582	151.8516	−0.69616	37.8079	0.60488	14.6690	−1.37903
13	391.2925	−0.45938	189.7171	0.00466	151.8892	−0.05090	37.8016	0.20620	14.9045	0.22969
14	406.7968	−0.17731	189.7189	0.04043	151.8909	−0.02171	37.8017	0.21252	14.9189	0.32820
15	422.0553	0.10028	189.7189	0.04081	151.8910	−0.02103	37.8017	0.21124	14.9204	0.33805
16	437.3506	0.37855	189.7188	0.04014	151.8910	−0.02113	37.8017	0.20933	14.9197	0.33312
17	452.6488	0.65687	189.7188	0.04039	151.8909	−0.02149	37.8017	0.21124	14.9197	0.33312
18	467.9466	0.93518	189.7188	0.04031	151.8909	−0.02140	37.8017	0.21124	14.9197	0.33312
19	483.2430	1.21346	189.7189	0.04083	151.8910	−0.02094	37.8017	0.21124	14.9197	0.33312
20	409.3961	−0.13003	189.6384	−1.57724	151.7981	−1.61525	37.8141	1.00305	14.9262	0.37747
21	424.8777	0.15163	189.6422	−1.50194	151.8031	−1.53016	37.8129	0.92608	14.9254	0.37254
22	440.1588	0.42964	189.6422	−1.50160	151.8031	−1.52985	37.8129	0.92608	14.9254	0.37254
23	455.4468	0.70777	189.6422	−1.50194	151.8031	−1.53020	37.8129	0.92608	14.9254	0.37254
24	470.7482	0.98615	189.6422	−1.50204	151.8031	−1.53018	37.8129	0.92608	14.9254	0.37254
25	409.2360	−0.13294	189.7274	0.21275	151.9359	0.75057	37.7646	−2.15843	14.7378	−0.90951
26	424.8852	0.15177	189.7396	0.45832	151.9468	0.93840	37.7658	−2.07770	14.7435	−0.87034
27	440.1338	0.42918	189.7381	0.42789	151.9453	0.91278	37.7658	−2.07936	14.7435	−0.87034
28	455.4154	0.70720	189.7381	0.42698	151.9453	0.91251	37.7658	−2.08140	14.7442	−0.86544
29	470.7187	0.98561	189.7401	0.46693	151.9473	0.94657	37.7658	−2.08127	14.7435	−0.87034

[a] Standardized values are obtained from eq 4.

(**1**−**19** in Table 1), the following correlation equation was found:

$$\log K = 1.55\sigma + 0.65 \log P + 6.93$$
$$n = 19; \quad r^2 = 0.943 \qquad (16)$$

The present approach to the design of alternative theoretical QSAR models arises from previously reported findings[5−7] that, in a series of structurally related molecules, both hydrophobic parameter log *P* and the effect of the systematic substitution can be modeled by appropriate theoretical descriptors. Such quantum similarity descriptors have been chosen as the QS-SM θ_{AA}, instead of log *P*, and the fragment QS-SM θ_{AA}^{R} replacing the substituent constant. Equation 17 is an example of such replacement: it describes the correlation of Hammett substituent constant with $\theta_{AA}^{SO_2NH_2}$ (listed in Table 2) in a series of 19 *para*-substituted benzenesulfonamides. The correlation is fairly good.

$$\sigma = -0.2779\theta_{AA}^{SO_2NH_2} + 0.3160$$
$$n = 19; \quad r^2 = 0.966 \qquad (17)$$

A slightly more complex situation arises when considering the correlation of log *P* with θ_{AA}, which in the same series splits into two regression lines with the same slope and different intercepts, as is shown in Figure 1. This specific form of correlation suggests that all these data can be described by a single regression line of the form: $\log P = a\theta_{AA} + bI + c$, where the variable *I* is the Boolean parameter, introduced to distinguish between alkyl- and nonalkyl-substituted derivatives ($I = 0$ for molecules **1**−**6** and $I = 1$ otherwise). The existence of this splitting may suggest that

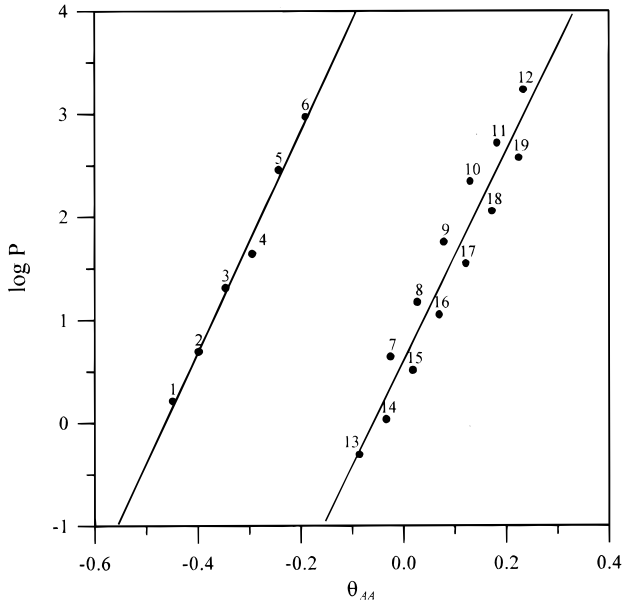


Figure 1. Linear correlation between log *P* and θ_{AA} for a series of 19 *para*-substituted benzenesulfonamides.

the basic assumption in deriving eq 11, namely the requirement of constancy of the term β, is not apparently satisfied within the whole series. It thus seems quite plausible to regard this splitting as an indirect indication of the fact that the studied molecular set apparently does not form a homogeneous series and that there are in fact two different series, formed by the molecules **1**−**6** and **7**−**19**. The actual form of the general MLR equation is given by

$$\log P = 1.9205\theta_{AA} - 4.2624I + 4.8523$$
$$n = 19; \quad r^2 = 0.943; \quad q^2 = 0.925 \qquad (18)$$

Using such a unified correlation equation, the activity of the whole set of 19 *para*-substituted benzenesulfonamides can be described by eq 19, which can be regarded as a theoretical counterpart of the original empirical eq 16 given by Hansch:

$$\log K = 1.2511\theta_{AA} - 0.4879\theta_{AA}^{SO_2NH_2} - 2.7968I + 10.6089$$

$$n = 19; \quad r^2 = 0.947; \quad q^2 = 0.906 \qquad (19)$$

Boolean variables were also used by Hansch when extending the applicability of his QSAR model to *ortho*- and *meta*-substituted benzenesulfonamides.[30] The corresponding Hansch QSAR equations take the form

$$\log K = 1.55\sigma + 0.62 \log P - 2.07I_m + 6.98$$

$$n = 24; \quad r^2 = 0.964 \qquad (20)$$

$$\log K = 1.55\sigma + 0.64 \log P - 2.07I_m - 3.28I_o + 6.94$$

$$n = 29; \quad r^2 = 0.982 \qquad (21)$$

where two additional Boolean parameters indicating the presence of *meta*- (I_m) and *ortho*-substituents (I_o) are included. While keeping the Hansch results, the original theoretical eq 19 can be generalized within the QS-SM procedure for the set of 24 *meta*- and *para*-derivatives, using the form

$$\log K = 1.2175\theta_{AA} - 0.4927\theta_{AA}^{SO_2NH_2} - 2.7321I - 2.6301I_m + 10.5585$$

$$n = 24; \quad r^2 = 0.971; \quad q^2 = 0.950 \qquad (22)$$

and for the whole set of 29 *ortho*-, *meta*-, and *para*-substituted derivatives using

$$\log K = 1.2739\theta_{AA} - 0.4536\theta_{AA}^{SO_2NH_2} - 2.7847I - 2.5796I_m - 2.8896I_o + 10.5957$$

$$n = 29; \quad r^2 = 0.984; \quad q^2 = 0.973 \qquad (23)$$

In addition to eqs 22 and 23, straightforwardly derived from the Hansch empirical equations, an alternative theoretical equation, which does not rely on mixing Boolean and real variables, can be proposed. This alternative equation, found by using systematically the NSS algorithm over the available QS-SM set, has the form

$$\log K = 1.2000\theta_{AA} + 2.3157\theta_{AA}^{SO_2NH_2} + 2.5149\theta_{AA}^{NH_2} - 0.6264\theta_{AA}^{m-C} + 7.4441$$

$$n = 29; \quad r^2 = 0.984; \quad q^2 = 0.976 \qquad (24)$$

This equation points out the splitting of the fragment SO_2NH_2, expected to be responsible for the activity, into two independent subfragments: SO_2 and NH_2. This splitting could suggest that the SO_2NH_2 group binds to the receptor site in the pocket of the enzyme in two points—by oxygen of the SO_2 fragment and by hydrogen bonding to the NH_2 fragment. The existence of multisites as responsible for the receptor/ligand binding has been proposed in several hypothesized pharmacophore models, for example, in the binding of indole derivatives to benzodiazepine receptor. More detailed discussion of this phenomenon, together with its possible consequences for the construction of theoretical QSAR models, will be given in section c. As it will be shown there, the systematic NSS procedure permits the localization of individual interaction sites responsible for the biological activity in this particular case.

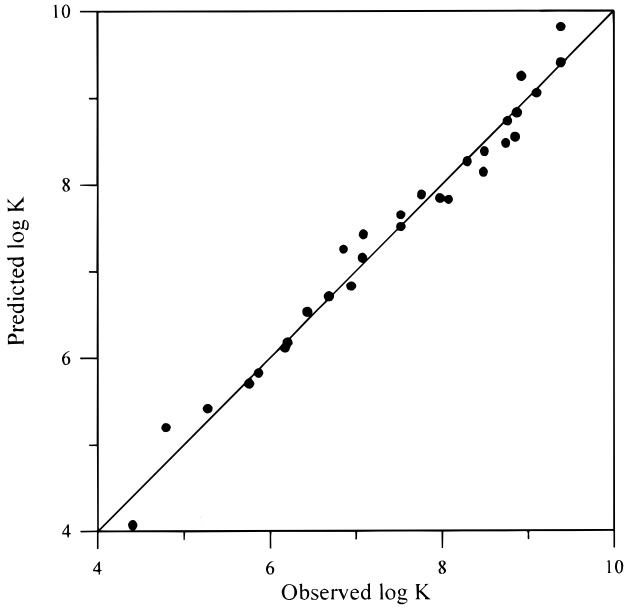


Figure 2. Observed versus predicted log *K* values for benzenesulfonamide compounds obtained from a LOO CV analysis.

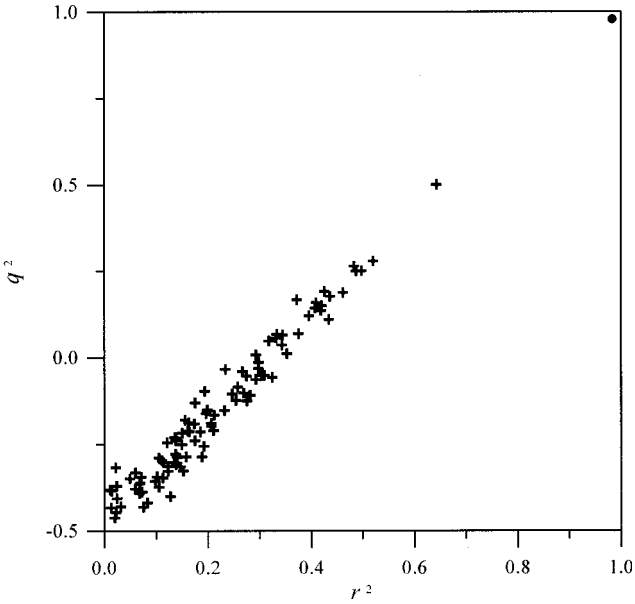


Figure 3. Representation of the r^2 vs q^2 statistical coefficients obtained from a random reordering test for the set of 29 benzenesulfonamides. Real (●) and random (+) QSAR models.

To verify the predictive power of the reported QSAR model corresponding to eq 24, a correlation between observed and predicted values of the inhibition constant log *K* is shown in Figure 2. The predicted values are computed employing a LOO CV analysis, yielding a q^2 value of 0.976. Additionally, Figure 3 shows the results for a random reordering test of the vector containing HCA inhibition activities. As can be observed from this illustration, the correct arrangement of the vector

Chart 2. Common Molecular Structure for Benzylamine Derivatives

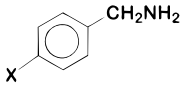


Table 3. Inhibitor Constants for the Binding of X-$C_6H_4CH_2NH_2$ to the Enzyme Trypsin

	X	observed log $1/K_i$[a]
1	H	0.523
2	CH_3	−0.176
3	Cl	0.155
4	OCH_3	0
5	$OCH_2C_5H_6$	0.398
6	NH_2	0.301
7	COOH	−0.301
8	$COOCH_3$	−0.362
9	$COOCH_2CH_3$	−0.447
10	$COO(CH_2)_2CH_3$	−0.301
11	$COO(CH_2)_3CH_3$	−0.041
12	$COO(CH_2)_4CH_3$	0.155
13	$COO(CH_2)_5CH_3$	0.523
14	$COOCH_2-C_6H_5$	1.523
15	$COOCH_2$-p$-C_6H_4Cl$	1.523
16	$COO(CH_2)_2C_6H_5$	0.222
17	$COO(CH_2)_3C_6H_5$	0.301
18	$CONH_2$	−0.398
19	$CONHC_6H_5$	0.699
20	$CONHCH_2C_6H_5$	0.398
21	$CONH(CH_2)_2C_6H_5$	0.523
22	$CONHC_{10}H_7$ (naphthalene)	1

[a] From ref 42.

containing activity values, which is depicted with a filled circle, corresponds to the best QSAR model.

In fact, to conclude this first example, it can be said that a satisfactory correlation between QS-SM and the binding constant K to HCA for this set of 29 benzenesulfonamides was found by means of eq 24. Results from these studies are comparable with those in a previous classical QSAR study. However, it must be stressed that in the present QS-SM model, which uses the same number of variables as the Hansch model, no mixing between Boolean and real variables is needed; in addition, the role of a pharmacophore position can easily be guessed, and the search for best fragment similarities is fully automated.

(b) Benzylamine Derivatives. In this example, the QSM theoretical approach was used to study the action of 22 benzylamine derivatives (Chart 2) as competitive inhibitors of the proteolytic enzyme trypsin. The activity of these derivatives was studied by Markwardt et al.,[42] and the corresponding data are summarized in Table 3. The biological activity of this series of substituted benzylamines was quantitatively studied by Hansch,[31] who reported the existence of empirical correlation with log P and Hammett σ constants as descriptors of a subset of nine molecules (**8**−**17** except phenyl derivative **14**):

$$\log 1/K_i = 0.41 \log P - 0.45\sigma - 1.07$$

$$n = 9; \quad r^2 = 0.955 \qquad (25)$$

Such a form of empirical QSAR suggests again constructing the alternative theoretical QS-SM model in such a way that both traditional descriptors are replaced by their corresponding theoretical counterparts θ_{AA} and $\theta_{AA}^{CH_2NH_2}$, respectively, listed in Table 4. Using this replacement, the initial empirical equation can be rewritten in the form

$$\log 1/K_i = 0.6685\theta_{AA} - 3.7872\theta_{AA}^{SO_2NH_2} + 2.6124$$

$$n = 9; \quad r^2 = 0.964; \quad q^2 = 0.901 \qquad (26)$$

which has a statistical importance similar to eq 25.

The ability to reproduce alternatively the traditional QSAR models is interesting, but certainly not the most important result of the present QS-SM approach. Its main advantage over traditional approaches is that the required theoretical QS-SM descriptors can be calculated easily. This is also true in situations where the traditional descriptors are either difficult to determine or completely unknown (for example, σ constants for some special substituents). Such is the situation with the whole series of 22 substituted benzylamines, where the lack of traditional descriptors restricted Hansch to studying only the subset of nine substituted derivatives. This limitation does not exist for the present theoretical approach and, in fact, a fairly good QSAR model describing the activity of the series of 21 derivatives has been found. It should be noted that the strongly deviating point **14** was excluded from the model, as it was by Hansch.[31] The resulting QSAR model takes the form of three-parameter correlation, with an additional descriptor, the fragment QS-SM $\theta_{AA}^{C_6H_4}$:

$$\log 1/K_i = 0.5572\theta_{AA} - 0.2608\theta_{AA}^{CH_2NH_2} + 0.2771\theta_{AA}^{C_6H_4} + 0.2220$$

$$n = 21; \quad r^2 = 0.828; \quad q^2 = 0.689 \qquad (27)$$

The reason for the presence of this additional parameter is not yet completely clear. However, a plausible explanation could be proposed invoking possible specific interactions of the benzene ring with the enzyme cavity, which can become important in determining inhibition activity. This equation enables the experimental values of biological activity to be confronted with LOO CV values, as shown in Figure 4.

Figure 5 shows the results for the random reordering test performed over the set of 21 molecules in order to reject accidental correlation. As can be seen in this figure, the best model, with the highest values of r^2 and q^2, does indeed correspond to the correct arrangement of log $1/K_i$ values, so that accidental correlation can be excluded.

(c) Indole Derivatives. In this last example, a quantitative study of the relationships between the structure of a group of indole derivatives and their capacity to displace [^{3}H] flunitrazepam from binding to bovine cortical membranes is presented. Molecular structures and experimental biological activity[43] for a set of 23 *N*-(indol-3-ylglyoxylyl)benzylamine derivatives (Chart 3) are listed in Table 5. These data were analyzed using the traditional QSAR approach,[32] providing some linear correlations with the main result that the biological activity for this molecular set does not depend on the hydrophobic parameter log P. In view of this, the decisive role in influencing biological activity in this series of substituted indole derivatives is presumably to be played by the substituent-induced variation in electronic structure of the active fragment, whose

Table 4. QS-SM (Z_{AA}) and Scaled QS-SM (θ_{AA}) Used To Derive Eqs 26 and 27 for the Binding of X-$C_6H_4CH_2NH_2$ to the Enzyme Trypsin[a]

	Z_{AA}	θ_{AA}	$Z_{AA}^{CH_2NH_2}$	$\theta_{AA}^{CH_2NH_2}$	$Z_{AA}^{C_6H_4}$	$\theta_{AA}^{C_6H_4}$
1	134.9215	−1.87870	44.9282	−0.59640	89.8533	3.35123
2	149.9472	−1.70216	44.9221	−0.90111	89.4114	1.45386
3	301.8829	0.08296	44.9265	−0.68188	89.1607	0.37739
4	197.2198	−1.14674	44.8990	−2.03526	88.8875	−0.79540
5	286.0141	−0.10348	44.8990	−2.03561	88.8637	−0.89796
6	164.5553	−1.53053	44.8916	−2.40006	89.1915	0.50949
7	246.9712	−0.56221	44.9583	0.88469	88.8978	−0.75155
8	260.8403	−0.39925	44.9559	0.76648	88.9174	−0.66717
9	276.3086	−0.21751	44.9553	0.73528	88.9282	−0.62070
10	291.5922	−0.03795	44.9554	0.73765	88.9283	−0.62062
11	306.8786	0.14166	44.9553	0.73563	88.9283	−0.62057
12	322.1767	0.32140	44.9542	0.68066	88.9276	−0.62353
13	337.4751	0.50114	44.9542	0.68066	88.9275	−0.62366
14	349.6811	0.64455	44.9563	0.78464	88.9144	−0.67993
15	516.6186	2.60593	44.9558	0.75723	88.9015	−0.73549
16	365.1546	0.82635	44.9560	0.76939	88.9268	−0.62683
17	380.4667	1.00626	44.9560	0.76796	88.9245	−0.63660
18	228.6429	−0.77755	44.9422	0.09151	89.2038	0.56246
19	315.9116	0.24779	44.9416	0.05874	89.2277	0.66494
20	331.6791	0.43304	44.9425	0.10554	89.2149	0.61030
21	347.0398	0.61352	44.9411	0.03719	89.2330	0.68766
22	374.1008	0.93147	44.9415	0.05707	89.2318	0.68267

[a] Standardized values are obtained from eq 4.

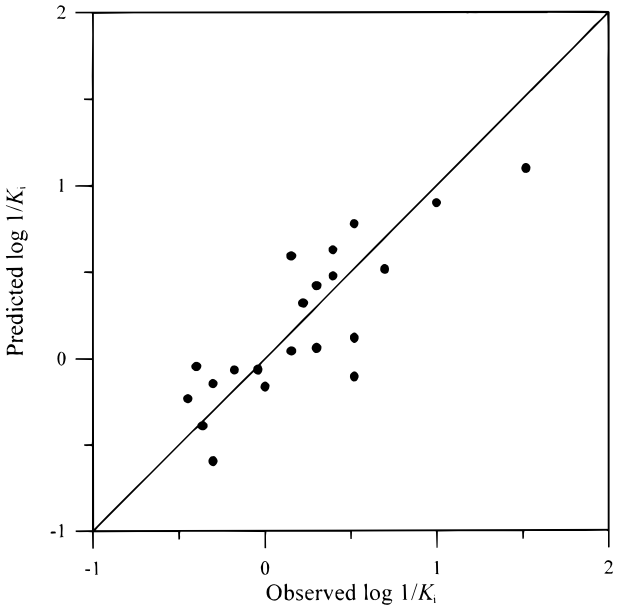


Figure 4. Observed versus predicted log $1/K_i$ values for benzylamine compounds obtained from a LOO CV analysis.

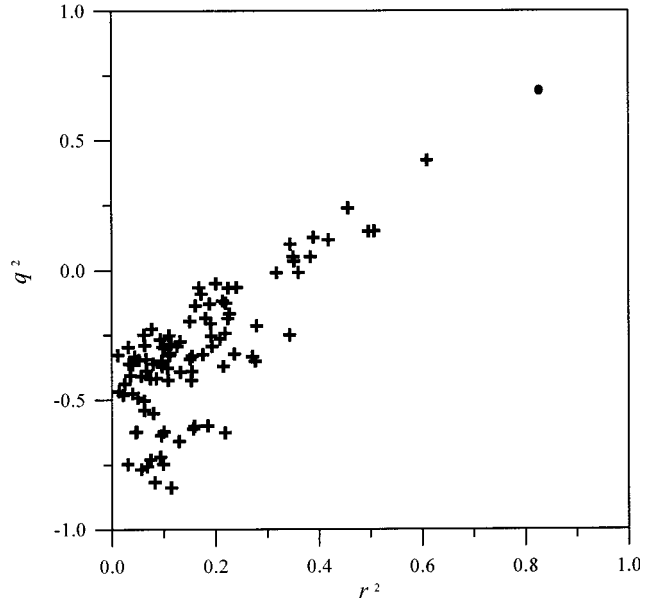


Figure 5. Representation of the r^2 vs q^2 statistical coefficients obtained from a random reordering test for the set of 21 benzylamines. Real (●) and random (+) QSAR models.

structure is not known. In keeping with this expectation, Hadjipavlou-Litina and Hansch proposed the correlation of the biological activity for a set of 20 derivatives (points **2**, **13**, and **14** were excluded) with the Hammett substituent constant σ of the substituent R:[32]

$$\log 1/K_i = 1.00\sigma + 6.60$$

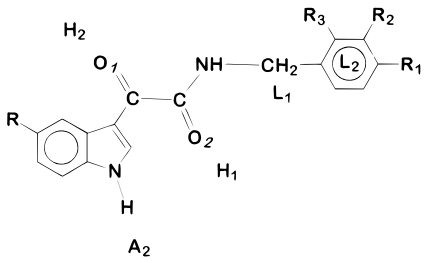
$$n = 20; \quad r^2 = 0.498 \qquad (28)$$

This correlation is not very satisfactory, but the graphical form of this dependence, shown in Figure 6, suggests that the description of the activity of the indole derivatives could be improved by adding two Boolean variables, I_2 and I_3, as was proposed by Hadjipavlou-Litina and Hansch:[32]

$$\log 1/K_i = 1.01\sigma + 0.60I_2 - 0.40I_3 + 6.56$$

$$n = 20; \quad r^2 = 0.810 \qquad (29)$$

Chart 3. Common Molecular Structure for Indole Derivatives



In this equation, σ is related to the substituent R, Boolean variable I_2 is defined as 1 when both R_1 and R_2 are the CH_3O group and as 0 otherwise, while I_3 = 1 for the cases R_2 = OH/R_1 = H and 0 otherwise.

Although eq 29 provides a reasonable description of the biological data, it is not completely satisfactory from a theoretical point of view. Namely, it is clear that,

Table 5. Benzodiazepine Receptor Affinity of Indole Derivatives

	R	R_1	R_2	R_3	observed log $1/K_i$[a]
1	H	H	H	H	6.92
2	Cl	H	H	H	6.31
3	NO_2	H	H	H	6.93
4	H	OCH_3	H	H	6.79
5	Cl	OCH_3	H	H	6.97
6	NO_2	OCH_3	H	H	7.28
7	H	H	OCH_3	H	6.54
8	Cl	H	OCH_3	H	6.79
9	NO_2	H	OCH_3	H	7.42
10	H	OCH_3	OCH_3	H	7.03
11	Cl	OCH_3	OCH_3	H	7.52
12	NO_2	OCH_3	OCH_3	H	7.96
13	H	Cl	H	H	7.17
14	H	H	H	Cl	5.59
15	H	OH	H	H	6.37
16	Cl	OH	H	H	6.82
17	NO_2	OH	H	H	7.92
18	H	H	OH	H	6.09
19	Cl	H	OH	H	6.24
20	NO_2	H	OH	H	7.19
21	H	OH	OH	H	6.46
22	Cl	OH	OH	H	6.74
23	NO_2	OH	OH	H	7.32

[a] From ref 43.

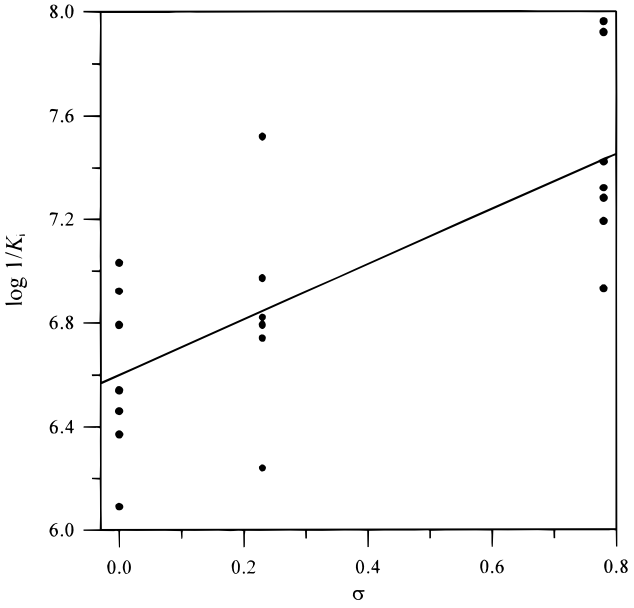


Figure 6. Dependence of Hammett constant σ for a series of indole derivatives on observed log $1/K_i$ values.

whatever the biologically active fragment may be, its structure will certainly be affected by the cumulative effect of all substituents (R, R_1, and R_2). Thus, it would be much more realistic to seek the QSAR model in the form of a linear combination of σ constants for all substituents, rather than focusing on the effect of the single isolated substituent R. In analyzing such possible equations, an excellent correlation was found between the linear combination of σ constants of substituents R, R_1, and R_2 and the fragment QS-SM $\theta_{AA}^{COCONHCH_2}$

$$\theta_{AA}^{COCONHCH_2} = -3.0455\sigma_p(R) - 0.1714\sigma_p(R_1) - 0.5838\sigma_m(R_2) + 0.9458$$

$$n = 23; \quad r^2 = 0.991; \quad q^2 = 0.987 \qquad (30)$$

which indicates that the biological activity of these molecules could be due to the presence of the $COCONHCH_2$ fragment. But the correlation of experimental activity with this selected theoretical descriptor is not very satisfactory:

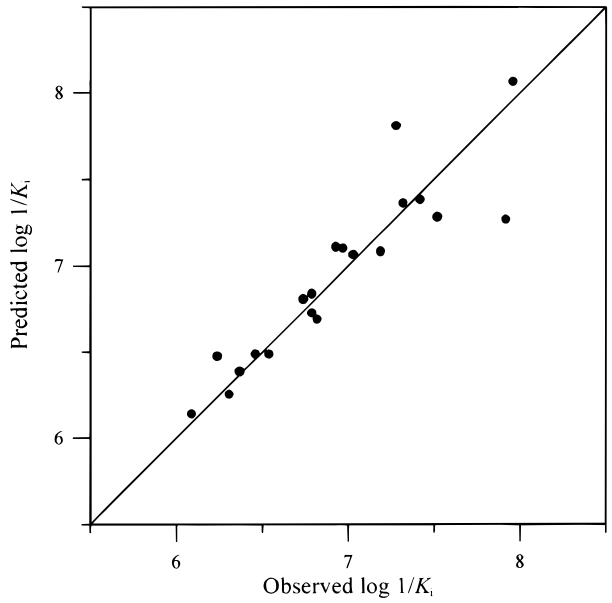


Figure 7. Observed versus predicted log $1/K_i$ values for indole derivatives obtained from a LOO CV analysi.

$$\log 1/K_i = -0.3900\theta_{AA}^{COCONHCH_2} + 6.8856$$

$$n = 23; \quad r^2 = 0.491; \quad q^2 = 0.393 \qquad (31)$$

This result shows that the problem of determining the fragment responsible for biological activity is more complex. A solution to this problem seems to be offered by a recent study,[44] in which a mechanism of benzodiazepine receptor (BzR) activity was proposed. According to this study, the biological activity of BzR ligands relies on the presence of four interaction sites: (i) a hydrogen bond acceptor site (A_2), (ii) a hydrogen bond donor site (H_1), (iii) a "bifunctional" hydrogen bond donor/acceptor site (H_2/A_3), and (iv) three lipophilic pockets (L_1, L_2, and L_3). In particular, for the series of *N*-(indol-3-ylglyoxylyl)benzylamine derivatives, these (re)active sites were identified as[43] follows: A_2 = NH indole group, H_1 = C=O_2 group, H_2 = C=O_1 group, L_1 = CH_2 group, and L_2 = phenyl ring. All these crucial molecular fragments are depicted in Chart 3. It is worth mentioning that three of these sites are also present in the $COCONHCH_2$ fragment.

According to the previously described pharmacophore model, five QS-SM fragments were selected and tested as possible molecular descriptors for modeling the action of indole derivatives to the BzR: $\theta_{AA}^{NH}(A_2)$ = QS-SM for the indole group NH, $\theta_{AA}^{C=O_2}(H_1)$ = QS-SM for the group C=O_2, $\theta_{AA}^{C=O_1}(H_2)$ = QS-SM for the group C=O_1, $\theta_{AA}^{CH_2}(L_1)$ = QS-SM for the group CH_2, and $\theta_{AA}^{Ph}(L_2)$ = QS-SM for the phenyl ring plus R_1, R_2, and R_3 substituents.

The corresponding QS-SM are listed in Table 6. On the basis of the above list, the theoretical QSAR model was searched for in the form of a linear combination of theoretical descriptors corresponding to individual interaction sites which give the best statistical description

Table 6. QS-SM (Z_{AA}) and Scaled QS-SM (θ_{AA}) Used To Derive Eqs 30−33 for the BzR Affinity of Indole Derivatives

	$Z_{AA}^{COCONHCH_2}$	$\theta_{AA}^{COCONHCH_2}$	Z_{AA}^{NH}	θ_{AA}^{NH} [a]	$Z_{AA}^{C=O_2}$	$\theta_{AA}^{C=O_2}$ [a]	$Z_{AA}^{C=O_1}$	$\theta_{AA}^{C=O_1}$ [a]	$Z_{AA}^{CH_2}$	$\theta_{AA}^{CH_2}$ [a]	Z_{AA}^{Ph}	θ_{AA}^{Ph} [a]
1	169.8903	0.88904	28.7523	1.32205	63.3717	0.42105	61.8345	0.55964	14.0570	0.72942	89.7885	−1.75575
2	169.8128	0.19186	28.7319	−1.21936	63.3677	0.20432	61.7738	−0.11068	14.0590	0.84100	89.7804	−1.75593
3	169.6365	−1.39478	28.7407	−0.12213	63.3725	0.46295	61.6323	−1.67314	14.0650	1.17221	89.7585	−1.75641
4	169.8982	0.96021	28.7523	1.32005	63.3928	1.57312	61.8559	0.79688	14.0155	−1.55821	152.1434	−0.37822
5	169.8338	0.38107	28.7330	−1.07917	63.3970	1.80317	61.8034	0.21634	14.0189	−1.37179	152.1318	−0.37847
6	169.6342	−1.41534	28.7407	−0.12238	63.3876	1.28778	61.6547	−1.42621	14.0243	−1.07726	152.1108	−0.37894
7	169.8915	0.90029	28.7523	1.32043	63.3677	0.20416	61.8301	0.51089	14.0621	1.01270	152.1492	−0.37809
8	169.8245	0.29744	28.7330	−1.08104	63.3720	0.43580	61.7744	−0.10384	14.0648	1.15974	152.1419	−0.37825
9	169.6223	−1.52232	28.7407	−0.12325	63.3590	−0.27081	61.6264	−1.73895	14.0750	1.72239	152.1219	−0.37869
10	169.8694	0.70160	28.7523	1.32255	63.3749	0.59325	61.8392	0.61172	14.0272	−0.91306	213.7586	0.98298
11	169.7975	0.05443	28.7330	−1.07854	63.3741	0.54976	61.7832	−0.00648	14.0292	−0.80324	213.7489	0.98277
12	169.6082	−1.64916	28.7407	−0.12362	63.3719	0.43165	61.6377	−1.61417	14.0366	−0.39542	213.7180	0.98208
13	169.8838	0.83044	28.7433	0.20721	63.3219	−2.29971	61.8627	0.87116	14.0551	0.62545	256.7471	1.93268
14	169.9046	1.01837	28.7522	1.31568	63.3489	−0.82597	61.9016	1.30078	14.0599	0.88910	257.1127	1.94075
15	169.9007	0.98283	28.7500	1.03828	63.3813	0.94529	61.8694	0.94556	14.0196	−1.33638	138.1157	−0.68811
16	169.8339	0.38151	28.7308	−1.35969	63.3840	1.09236	61.8134	0.32730	14.0209	−1.26456	138.1048	−0.68836
17	169.6404	−1.35953	28.7385	−0.40240	63.3780	0.76457	61.6653	−1.30869	14.0248	−1.04521	138.0710	−0.68910
18	169.8903	0.88955	28.7500	1.04415	63.3540	−0.54369	61.8386	0.60486	14.0662	1.23641	138.1384	−0.68761
19	169.8191	0.24885	28.7308	−1.35881	63.3551	−0.48802	61.7804	−0.03794	14.0682	1.34496	138.1280	−0.68784
20	169.6321	−1.43456	28.7384	−0.40365	63.3513	−0.69142	61.6364	−1.62845	14.0756	1.75498	138.1006	−0.68845
21	169.9051	1.02277	28.7478	0.76201	63.3627	−0.07107	61.8726	0.98134	14.0350	−0.48704	186.5875	0.38272
22	169.8369	0.40908	28.7286	−1.63746	63.3647	0.03835	61.8159	0.35504	14.0363	−0.41506	186.5741	0.38242
23	169.6377	−1.38365	28.7362	−0.68279	63.3521	−0.65142	61.6685	−1.27304	14.0416	−0.11827	186.5384	0.38163

[a] Standardized similarity measure values are obtained from eq 4 and have been calculated together with QS-SM of molecules listed in Table 7.

Table 7. Benzodiazepine Receptor Affinity of Indole Derivatives

	R	R_1	R_2	R_3	observed log $1/K_i$[a]	predicted log $1/K_i$[b]
24	H	H	Cl	H	6.80	6.52
25	H	F	H	H	7.28	5.82
26	H	H	H	F	6.18	5.77
27	H	OH	OCH_3	H	6.85	6.74
28	Cl	OH	OCH_3	H	7.57	7.04
29	NO_2	OH	OCH_3	H	7.89	7.67

[a] From ref 43. [b] Calculated from eq 33 using scaled QS-SM described in Table 8.

of the data. Such a search can best be performed by means of the NSS algorithm. With this approach, the following QSAR models were obtained for a set of 20 compounds, where molecules **1**, **13**, and **14** are rejected:

$$\log 1/K_i = -0.4086\theta_{AA}^{C=O1} + 0.3541\theta_{AA}^{Ph} + 6.9237$$

$$n = 20; \quad r^2 = 0.751; \quad q^2 = 0.683 \qquad (32)$$

$$\log 1/K_i = 0.2767\theta_{AA}^{C=O2} - 0.4573\theta_{AA}^{C=O1} + 0.3697\theta_{AA}^{Ph} + 6.8085$$

$$n = 20; \quad r^2 = 0.751; \quad q^2 = 0.683 \qquad (33)$$

Not only do these equations provide a better statistical description of the biological activity than the original Hansch equation (eq 29) but also the physical meaning of individual descriptors is clearer than that for the Boolean variables I_2 and I_3. In addition, eq 33 suggests that the importance of potential interaction sites in determining biological activity is not the same for all fragments. The interactions for the hydrogen bond donator sites H_1, H_2, and L_2 probably dominate the rest. In conclusion, the presented theoretical procedure not only provides a satisfactory statistical description of the biological activities but also permits the localization and identification of the possible reaction sites responsible for the biological activity in a given series of compounds. Consequently, the reported theoretical approach can be a reasonable and reliable alternative to classical QSAR approaches.

The predictive q^2 value obtained from a LOO CV analysis in eq 33 is satisfactory. This assertion is confirmed by the representation of predicted and observed values of log $1/K_i$ for the set of 20 indole derivatives, shown in Figure 7.

Finally, an independent validation study of QSAR model[33] was carried out with the aim of predicting the activity for a new subset of compounds which were not considered in the construction of the model. In reference 43, the activities of six new indole derivatives (listed in Table 7) with the same common skeleton given in Chart 3 where reported. Fragment QS-SM for these compounds are listed in Table 8. Using eq 33, the theoretical values of log $1/K_i$ were calculated which are also presented in Table 7. As can be seen, most of the predictions are correct. The only deviation concerns the molecule **25**, which has a fluorine atom as a substituent. However, when predicted and observed values for the

Table 8. Benzodiazepine Receptor Affinity of Indole Derivatives

	Z_{AA}^{NH}	θ_{AA}^{NH} [a]	$Z_{AA}^{C=O_2}$	$\theta_{AA}^{C=O_2}$ [a]	$Z_{AA}^{C=O_1}$	$\theta_{AA}^{C=O_1}$ [a]	$Z_{AA}^{CH_2}$	$\theta_{AA}^{CH_2}$ [a]	Z_{AA}^{Ph}	θ_{AA}^{Ph} [a]
24	28.7433	0.20521	63.3266	−2.04092	61.8707	0.95939	14.0535	0.43682	256.8588	1.93514
25	28.7433	0.20209	63.3308	−1.81120	61.8700	0.95187	14.0378	−0.32852	162.4339	−0.15088
26	28.7533	1.45239	63.3359	−1.53684	61.8946	1.22355	14.0377	−0.33514	162.5210	−0.14895
27	28.7500	1.04265	63.3682	0.22929	61.8610	0.85266	14.0309	−0.71003	200.4366	0.68867
28	28.7308	−1.35657	63.3693	0.28769	61.8043	0.22669	14.0329	−0.60049	200.4240	0.68839
29	28.7384	−0.40390	63.3623	−0.09347	61.6594	−1.37409	14.0390	−0.26553	200.3974	0.68781

[a] Standardized similarity measures values are obtained from eq 4 and have been calculated together with QS-SM of molecules listed in Table 5.

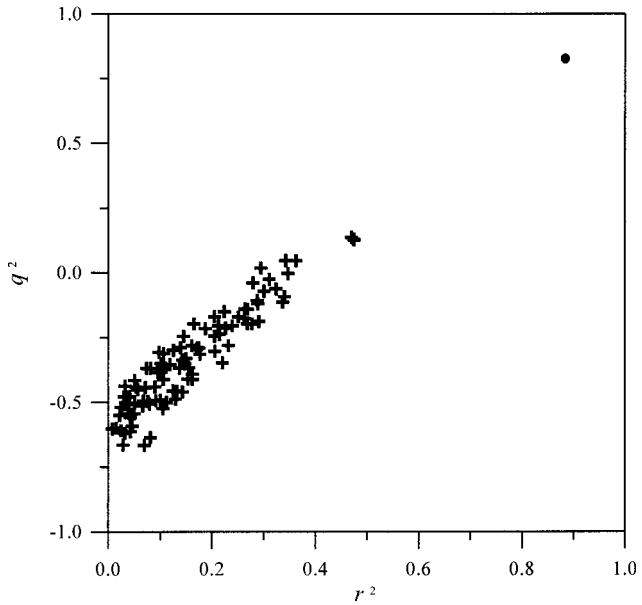


Figure 8. Representation of the r^2 vs q^2 statistical coefficients obtained from a random reordering test for the set of 20 indole derivatives. Real (●) and random (+) QSAR models.

rest of the five compounds are analyzed, a squared regression coefficient of 0.945 is obtained.

As in previous examples, a randomization test was carried out to estimate statistical reliability of the QSAR model given in eq 33. This validation test is presented in Figure 8. As can be seen, only the correct arrangement of biological data provides a satisfactory QSAR model.

Conclusions

The examples put forward here clearly show that QSSM for the whole molecule and the appropriate molecular fragments can advantageously be used as efficient descriptors for predicting biological and pharmacological activities. We can thus believe that because of its relative simplicity and complete generality the method opens a new interesting possibility to enrich the traditional QSAR approaches by describing a systematic procedure of constructing new theoretical QSAR models. Moreover, the possibility of identification of individual interaction sites responsible in each particular case for the observed biological activity could also be of considerable importance for the rational design of new biologically active molecules.

Acknowledgment. R.P. acknowledges the support of this work by the grant of the Grant Agency of the Czech Republic No. 203/00/1289. The research was also partly funded by a CICYT grant (SAF 96-0158), the Fundació Maria Francisca de Roviralta, and an European Commission contract (#ENV4-CT97-0508). The authors also thank the referees for their constructive criticism, which improved several aspects of this work.

References

(1) Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Application of Molecular Quantum Similarity to QSAR. *Quant. Struct.-Act. Relat.* **1997**, *16*, 25−32.
(2) Lobato, M.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Structure−Activity Relationships of a Steroid Family using Quantum Similarity Measures and Topological Quantum Similarity Indices. *Quant. Struct.-Act. Relat.* **1997**, *16*, 465−472.
(3) Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular Quantum Similarity Measures Tuned 3D QSAR: An Antitumoral Family Validation Study. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 624−631.
(4) Robert, D.; Amat, L.; Carbó-Dorca, R. Three-Dimensional Quantitative Structure−Activity Relationships from Tuned Molecular Quantum Similarity Measures: Prediction of the Corticosteroid-Binding Globulin Binding Affinity for a Steroid Family. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 333−344.
(5) Amat, L.; Carbó-Dorca, R.; Ponec, R. Molecular Quantum Similarity Measures as an Alternative to Log *P* Values in QSAR Studies. *J. Comput. Chem.* **1998**, *19*, 1575−1583.
(6) Ponec, R.; Amat, L.; Carbó-Dorca, R. Molecular basis of quantitative structure-properties relationships (QSPR): A quantum similarity approach. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 259−270.
(7) Ponec, R.; Amat, L.; Carbó-Dorca, R. Quantum Similarity Approach to LFER: Substituent and Solvent Effects on the Acidities of Carboxylic Acids. *J. Phys. Org. Chem.* **1999**, *12*, 447−454.
(8) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Similarity screening of molecular data sets. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 513−520.
(9) Good, A. C.; So, S.-S.; Richards, W. G. Structure−activity relationships from molecular similarity matrices. *J. Med. Chem.* **1993**, *36*, 433−438.
(10) Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. *J. Med. Chem.* **1993**, *36*, 2929−2937.
(11) Cooper, D. L.; Allan, N. L. A novel approach to molecular similarity. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 253−259.
(12) Measures, P. T.; Mort, K. A.; Allan, N. L.; Cooper, D. L. Applications of momentum-space similarity. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 331−340.
(13) Benigni, R.; Cotta-Ramusino, M.; Giorgi, F.; Gallo, G. Molecular similarity matrixes and quantitative structure−activity relationships: a case study with methodological implications. *J. Med. Chem.* **1995**, *38*, 629−635.
(14) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A molecular field-based similarity approach to pharmacophoric pattern recognition. *J. Mol. Graphics Modelling* **1997**, *15*, 114−121.
(15) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A molecular-field-based similarity study of nonnucleoside HIV-1 reverse transcriptase inhibitors. *J. Comput. Aided-Mol. Des.* **1999**, *13*, 79−93.
(16) Carbó, R.; Leyda, L.; Arnau, M. How Similar is a Molecule to Another? An Electron Density Measure of Similarity between Two Molecular Structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.
(17) Carbó, R.; Domingo, L. LCAO-MO Similarity Measures and Taxonomy. *Int. J. Quantum Chem.* **1987**, *23*, 517−545.
(18) Carbó, R.; Calabuig, B. Quantum molecular similarity measures and the *n*-dimensional representation of a molecular set: phenyldimethylthiazines. *J. Mol. Struct. (THEOCHEM)* **1992**, *254*, 517−531.
(19) Carbó, R.; Calabuig, B.; Vera, L.; Besalú, E. Molecular Quantum Similarity: Theoretical Framework, Ordering Principles, and Visualization Techniques. *Adv. Quantum Chem.* **1994**, *25*, 253−313.
(20) Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and Recent Developments on Molecular Quantum Similarity. *Topics Curr. Chem.* **1995**, *173*, 31−62.
(21) *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*; Carbó, R., Ed.; Kluwer Academic: Amsterdam, 1995.
(22) *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press Inc.: Greenwich, CT, 1996; Vol. 1. and 1998; Vol. 2.
(23) Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships (QSPR) *J. Math. Chem.* **1995**, *18*, 237−246.
(24) Constans, P.; Carbó, R. Atomic Shell Approximation: Electron Density Fitting Algorithm Restricting Coefficients to Positive Values. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046−1053.
(25) Amat, L.; Carbó-Dorca, R. Quantum Similarity Measures under Atomic Shell Approximation: First-Order Density Fitting using Elementary Jacobi Rotations. *J. Comput. Chem.* **1997**, *18*, 2023−2039.
(26) Amat, L.; Carbó-Dorca, R. Fitted Electronic Density Functions from H to Rn for use in Quantum Similarity Measures: *Cis*-diamminedichloroplatinum(II) complex as an Application Example. *J. Comput. Chem.* **1999**, *20*, 911−920.

(27) Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *J. Mol. Struct. (THEOCHEM)* **1998**, *451*, 11−23.
(28) Constans, P.; Amat, L.; Carbó-Dorca, R. Toward a Global Maximization of the Molecular Similarity Function: Superposition of Two Molecules. *J. Comput. Chem.* **1997**, *18*, 826−846.
(29) Mezey, P. G. The Holographic Electron Density Theorem and Quantum Similarity Measures. *Mol. Phys.* **1999**, *96*, 169−178.
(30) Hansch, C.; McClarin, J.; Klein, T.; Langridge, R. A Quantitative Structure−Activity Relationship and Molecular Graphics Study of Carbonic Anhydrase Inhibitors. *Mol. Pharmacol.* **1985**, *27*, 493−498.
(31) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(32) Hadjipavlou-Litina, D.; Hansch, C. Quantitative Structure−Activity Relationships of the Benzodiazepines. A Review and Reevaluation. *Chem. Rev.* **1994**, *94*, 1483−1505.
(33) ASA coefficients and exponents can be seen and downloaded from the following WWW site: http://iqc.udg.es/cat/similarity/ASA/funcset.html.
(34) Hansch, C.; Fujita, T. ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.
(35) Fujita, T.; Iwasa, J.; Hansch, C. A New Substituent Constant, π, Derived from Partition Coefficients. *J. Am. Chem. Soc.* **1964**, *86*, 5175−5180.
(36) AMPAC 6.01, Semichem, Inc., 7128 Summit, Shawnee, KS 66216. D.A.
(37) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(38) Clementi, S.; Wold, S. How to Choose the Proper Statistical Method. In: *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH Publishers Inc.: Weinheim, 1995; Vol. 2; pp 319−338.
(39) Carbó, R.; Besalú, E. Definition, mathematical examples an quantum chemical applications of nested summation symbols and logical Kronecker deltas. *Comput. Chem.* **1994**, *18*, 117−126.
(40) Carbó, R.; Besalú, E. Definition and quantum chemical applications of nested summations symbols and logical functions: Pedagogical artificial intelligence devices for formulae writing, sequential programming and automatic parallel implementation. *J. Math. Chem.* **1995**, *18*, 37−72.
(41) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In: *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH Publishers Inc.: Weinheim, 1995; Vol. 2; pp 309−318.
(42) Markwart, F.; Landmann, H.; Walsmann, P. Comparative Studies on the Inhibition of Trypsin, Plasmin, and Thrombin by Derivatives of Benzylamine and Benzamidine. *Eur. J. Biochem.* **1968**, *6*, 502−506.
(43) Da Settimo, A.; Primofiore, G.; Da Settimo, F.; Marini, A. M.; Novellino, E.; Greco, G.; Martini, C.; Giannaccini, G.; Lucacchini, A. Synthesis, Structure−Activity Relationships, and Molecular Modeling Studies of *N*-(Indole-3-ylglyoxylyl)benzylamine Derivatives Acting at the Benzodiazepine Receptor. *J. Med. Chem.* **1996**, *39*, 5083−5091.
(44) Zhang, W.; Koehler, K. F.; Zhang, P.; Cook, J. M. Development of a Comprehensive Pharmacophore Model for the Benzodiazepine Receptor. *Drug Des. Discovery* **1995**, *12*, 193−248.

JM9910728