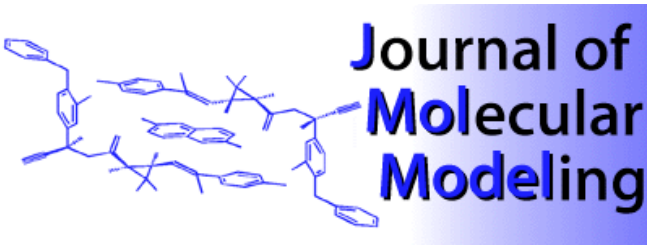
J. Mol. Model. **2000**, *6*, 517 – 526



FULL PAPER

Lattice Model for QSAR Studies

Victor E. Kuz'min[1], Anatoly G. Artemenko[1], Nikolay A. Kovdienko[1], Igor V. Tetko[2], and David J. Livingstone[3]

[1]O.V. Bogatsky Physico-Chemical Institute of the National Academy of Sciences of Ukraine, Lutsdorfskaya doroga 86, Odesa, 270080, Ukraine. Tel.: +380(482) 225127; Fax: +380(482) 652012. E-mail: victor@farlep.net

[2]Institute of Bioorganic and Petroleum Chemistry, Murmans'ka 1, Kyiv-660, 253660, Ukraine

[3]ChemQuest, Delamere House, 1, Royal Crescent, Sandown, Isle of Wight, PO36 8LZ and Centre for Molecular design, University of Portsmouth, Portsmouth, Hants, PO1 2EG, U.K.

Received: 13 December 1999/ Accepted: 15 March 2000/ Published: 16 August 2000

Abstract A system of lattice models that takes into account the structures of molecules, their form, stereochemical features and their interaction with the enclosing space, is proposed. The local, integral and field structural parameters of molecules (more than 20 thousand per compound) are estimated within the proposed framework. An investigation of the utility of these parameters in Quantitative Structure-Activity Relationships (QSAR) has been made using several statistical methods (multiple regression analysis, partial least squares (PLS), trend - vector procedure). The efficiency of the proposed approach has been examined using a data set derived from the formation of charge-transfer complexes of monosubstituted bezens with 1,3,5-trinitrobenzene.

Keywords QSAR, Lattice models, Partial Least Squares, Trend-vector

Introduction

A large number of QSAR methods [1, 2, 3, 4, 5, 6] are available now for use by medicinal chemists. Many of these methods, however, use only restricted, one-sided structural information that does not adequately describe all the relevant properties of the analyzed molecules. For example, in simple models only specific structural fragments (descriptors) of molecules (e.g., the Free-Wilson method [2]) or physicochemical parameters of the molecular fragments such as lipophilicity, charges, etc. (see for example the Hansch approach [1]) are analyzed. In other approaches both sets of these parameters together with various topological indices of the molecules are considered.[7, 8] The Hopfinger model [9] considers only parameters that describe the shape of a molecule, while the Cramer principal component based approach utilizes only physicochemical characteristics (B, C, D, E, F-parameters).[10] These and many other well-known QSAR models used by medicinal chemists do not, as a rule, consider the stereochemical peculiarities of molecules.

Two more recent and apparently more complex approaches such as CoMFA [11] and HASL [12, 13] utilize a more elaborate description of the molecules and consider parameters reflecting peculiarities of the intermolecular interaction of the compounds analyzed and their spatial structure. The approach proposed here uses similar parameters as applied in these powerful methods. However, in addition the molecule properties are described with a variety of com-

Correspondence to: V. E. Kuz'min

plementary parameters. The whole set of parameters generated ranges from the most simple, such as presence or absence of particular atoms in the molecular structure, to more sophisticated parameters that could be used to take into account stereochemistry of the analyzed molecule and its interaction with the environment. We show that analysis of a large number of parameters generated by our model could provide a pertinent description of the molecules and can be very important for successful QSAR modeling. In order to analyze the generated parameters, which include up to tens of thousands of descriptors per molecule, we apply methods [14, 15, 16] developed for the processing of large arrays of data without essential loss of reliability of the calculated model.

Representation of a molecule

The description of the compounds includes several steps. In the first, the spatial structure of the analyzed molecules is obtained from experimental data (i.e., X-ray analysis) or from molecular or quantum mechanical calculations. In the case of flexible molecules, it is necessary to select one of the stable conformations. This may be achieved using some conformational search procedure or making use of some complementary information regarding the biologically active conformation of the molecule. The conformation of each molecule is placed into a lattice of cubic cells.[17] The size of a cell, **h**=2Å (it can be varied), approximately corresponds to the average van der Waals radius of an organogenic atom. The invariant disposition of the molecule in the lattice is achieved by superposition of the center of mass of the molecules with the origin of the coordinates. In addition the principal axes of inertia of the molecule are also superimposed with the coordinate axes of the lattice. If the analyzed structures contain a large common structural fragment, their alignment is carried out mainly according to this fragment.

A broken spiral curve (SC) is constructed within the lattice.[18] This curve passes over the center of all lattice cells (both occupied and vacant) and it represents a complex line consisting of coaxial fragments of spirals, embedded one into another (Figure 1).[18] Two types of SC, one that turns entirely to the right and one that turns to the left, i.e. left and right SC are usually considered in the analysis.

The SCs are used to calculate molecular codes (MC) of the analyzed structures. This code consists of a sequence of real values b_i representing the atomic characteristics (for example, atomic number, lipophilicity, atomic refraction, etc.) and integer values a_i corresponding to the distance between atoms ***i*** and ***i+1*** measured as the number of empty cells along the SC. This code includes all information about shape and stereochemistry of the analyzed molecule in a compressed form.[18] Two types of MC, i.e. left (LMC) and right (RMC) molecular code corresponding to the two types of broken lines, are calculated for each molecule. The spatial structure of a molecule is easily restored from its MCs and each molecule is characterized by its unique MC. Sensitivity of the MC for description of the spatial structure of the molecule depends on the size **h** of the lattice cell. The "conformational sensitivity" of the models is decreased with an increase of the cell size **h**. For example, the rotation of the C-C bond in H_3C-CH_2 OH changes the MC only for each 15º and 30º using the size of the lattice cell **h**=1Å and **h**=2Å respectively. The ability to vary the length of the cell in the lattice makes it possible to describe the molecular structures with varying degrees of precision. A model with lower precision can be especially useful in the analysis of flexible molecules, especially if there are difficulties in the determination of the biologically active conformation of the molecules. The use of large lattice cells also allows the consideration in the analysis of a set of conformations.

MC provides the possibility to estimate a structural similarity/dissimilarity of different compounds. Let us consider RMC of two arbitrary molecules M_1 and M_2:

$$\text{RMC}(M_1) = a_1\, b_1\, a_2\, b_2\, a_3\, b_3 \ldots a_i\, b_i \ldots a_{n1}\, b_{n1} \quad (1)$$
$$\text{RMC}(M_2) = a'_1\, b'_1\, a'_2\, b'_2\, a'_3\, b'_3 \ldots a'_i\, b'_i \ldots a'_{n2}\, b'_{n2};$$

where $n_1 \le n_2$.

A measure of structural dissimilarity (SD) can be calculated using Euclidean distance in the space of parameters (a_i, a'_i) and (b_i, b'_i)

$$R_a(M_1, M_2) = \sqrt{\sum_{i=1}^{n_2} (a_i - a'_i)^2}$$
$$R_b(M_1, M_2) = \sqrt{\sum_{i=1}^{n_2} (b_i - b'_i)^2} \quad (2)$$

where $a_i=b_i=0$ for $i>n_1$. The distances $R_a(M_1,M_2)$ and $R_b(M_1,M_2)$ estimate the geometrical structural dissimilarity

Table 1 *Distribution of atoms (fragments) by cells*

Number of cell	1	2	3	4	…	38	…	45	46	47	48	…	57
Molecule **a**	C	C	CH	CH		H		F		CH	CH		
Molecule **b**	C	C	CH	CH		H		CH	H	CH	CH		H

Molecular codes are:
a: 0.**12**.1.**12**.1.**13**.1.**13**.34.**1**.7.**19**.2.**13**.1.**13**.0. **0**. 0.**0**.
b: 0.**12**.1.**12**.1.**13**.1.**13**.34.**1**.7.**13**.1. **1**. 1.**13**.1.**13**.9.**1**.

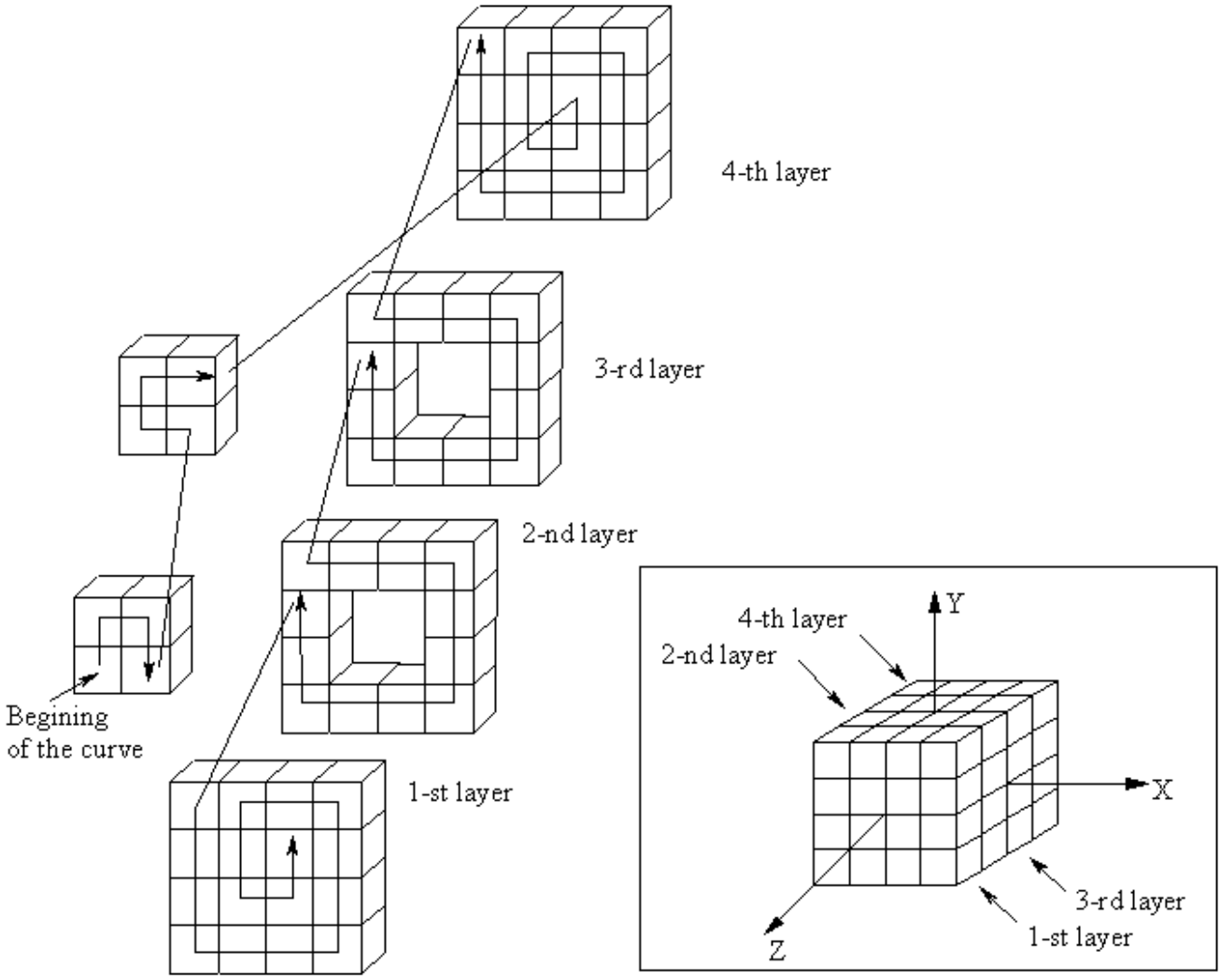


Figure 1 *The broken spiral curve in the spatial lattice*

and the structural dissimilarity of the atom characteristics of the molecules, respectively. The comparison of RMC and LMC is used to estimate the chirality level of molecules [18] that is introduced as χ = R (RMC, LMC), where RMC and LMC are the left and the right molecular code of the analyzed molecule.

Let us disregard the nature of the values a_i and b_i and formally consider all terms from Eq. (1) as an array. This makes it possible to calculate a structural similarity (SS) and dissimilarity (DS) of molecules using the coefficient of correlation, R [19], and Camber distance [20]

$$D(X,Y) = \sum_{i=1}^{N} \frac{|X_i - Y_i|}{|X_i + Y_i|} \qquad (3)$$

The molecules fluorobenzene (a) and toluene (b) were used for illustration of the procedure of a construction of a molecular code (Figure 2). The distribution of atoms (fragments) by cells is shown in Table 1.

The first digit ”0” in molecular codes means the spiral begins in the filled cell. The underlined zeroes were added to have identical lengths of molecular codes for both molecules in SS/DS calculations.

Structural similarity/dissimilarity (SS/DS):

R_a (**a, b**) = 9,110 R_b (**a, b**) =18,708

R (**a, b**) = 0,850 D (**a, b**) = 5,378

The parameters R_a, R_b, R, D reflect SS/DS and can be used in QSAR studies. These parameters reflect the widely held principle, that substances with similar structures also have similar properties. RMC is usually used to calculate the structural similarity of compounds.

(a) (b)

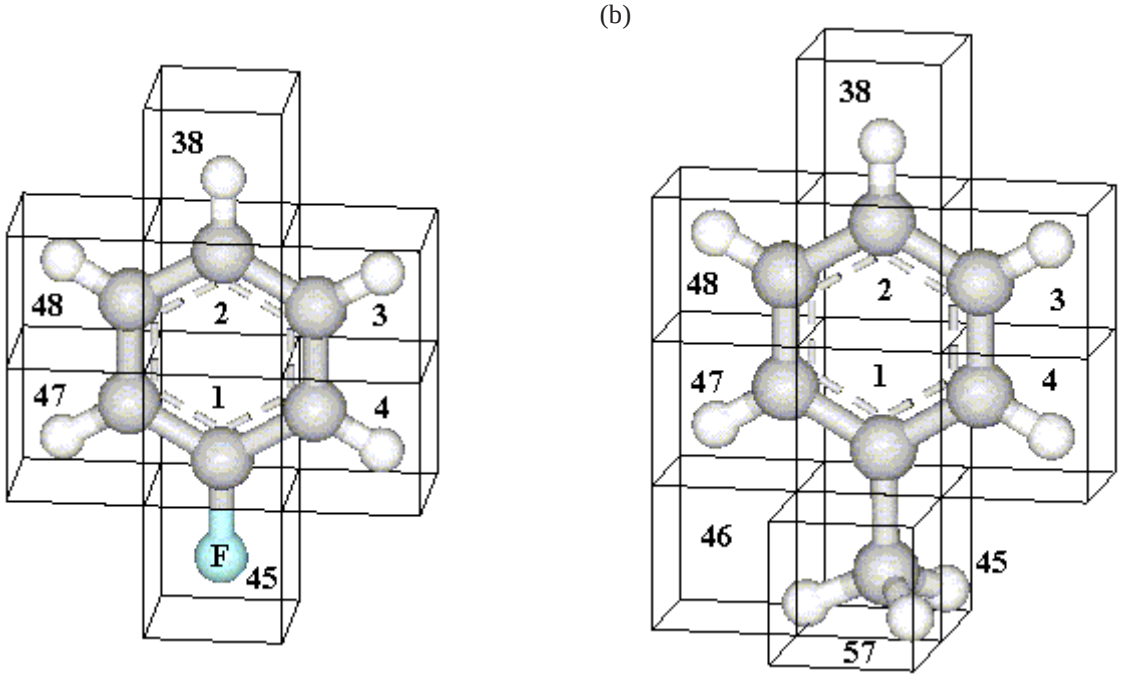


Figure 2 *The lattice models of molecules of fluobenzene (a) and toluene (b)*

Structural parameters of molecules

We propose to classify the structural parameters considered in this study as follows:

1. Integral parameters describing properties of the whole molecular structure;
2. Local parameters describing the separate fragments of the molecule;
3. Field parameters describing the influence of the molecule on the enclosing space.

Integral parameter

Integral parameters include characteristics of inertia ellipsoid, dipole moment, molecular refraction, lipophilicity, parachor, and average polarizability. Several parameters were originated from the MC. They include the length of the left and the right codes $L\ (M_1) = 1 + \sum_{i=1}^{n_1} a_i$, the chirality levels and the parameters of structural similarity of molecules Ra, Rb, R, D. If available, some information about the environment and mutual disposition of the pharmacophores, can be also included into the analysis.[21]

A number of parameters calculated by Fourier transform of atom property distributions along the SC were also included in this group of parameters. The Fourier transform describes the integral parameters of the analyzed structure. The high-frequency harmonics characterize small fragments while the low-frequency harmonics correspond to the global molecule properties. The Fourier transform of a discrete function of parameters P(i):

$$p(i) = \frac{c_0}{2} + \sum_{m=1}^{M-1} \left\{ c_m \cos\left[2\pi m(i-1) \Big/ N \right] + d_m \sin\left[2\pi m(i-1) \Big/ N \right] \right\} + \frac{c_n}{2} \cos\left[\pi(i-1) \right] \tag{4}$$

where

$$c_m = \frac{2}{N \sum_{i=1}^{N} p_i \cos\left[2\pi m(i-1) \Big/ N \right]} \tag{5}$$

$$d_m = \frac{2}{N \sum_{i=1}^{N} p_i \sin\left[2\pi m(i-1) \Big/ N \right]}$$

or an alternative form

$$P(i) = \frac{q_0}{2} + \sum_{m=1}^{M-1} q_m \sin\left[\frac{2\pi m(i-1)}{N + \psi_m} \right] + \frac{q_n}{2} \cos\left[\pi(i-1) \right] \tag{6}$$

where m is the number of harmonic, q_m is the amplitude, ψ_m is the phase angle, N is the total number of cells, M= int(N-1)/2 is the total number of harmonics, c_m and d_m is the coefficients of transform, $q_{n/2}$ =0 for odd N, $q_m = \sqrt{c_m^2 + d_m^2}$, $ø_m$=arctan(c_m/d_m) is the phase angle. The values of the amplitudes q_m were used as the parameters for a QSAR study.

Local parameters

Local parameters were used to describe the properties of cells occupied by atoms. They include parameters corresponding to the presence or absence of some atoms in the cell (i.e., presence of C or O), average lipophilicity, refraction, polarizability, electrostatic charge and electronegativity of fragments and atoms. All charge characteristics were calculated using the method of smoothing of electronegativity according to Jolly-Perry.[22 , 23]

Field parameters

Field parameters described characteristics of vacant cells. They include

1) An electrostatic potential in the vacant cell

$$EP_i = \sum_j^{n_1} \frac{q_j}{r_{ij}}, \qquad (7)$$

where ***i*** is the number of the cell, ***j*** is the number of the atom, $\boldsymbol{q_j}$ is the charge of the atom ***j***,[22, 23], $\boldsymbol{r_{ij}}$ is the distance between the atom ***j*** and the cell ***i.***

2) A lipophilicity potential [24] in the vacant cell

$$LP_i = \sum_j \frac{f_j}{(1+r_{ij})}, \qquad (8)$$

where ***i*** is the number of the cell, ***j*** is the number of the atom, $\boldsymbol{f_j}$ is the lipophilicity of the atom (group), $\boldsymbol{r_{ij}}$ is the distance between the atom ***j*** and the cell ***i.***

3) A probability of an occupancy of a vacant cell by different atoms *i*,*k* ("probe-atoms") or probability to be empty:

$$P_k = \left\{1 + \sum_{i \neq k} EXP\left(\frac{-(E_i - E_k)}{RT}\right)\right\}^{-1}, \quad \sum_k P_k = 1, \qquad (9)$$

where $\boldsymbol{E_i}$ or $\boldsymbol{E}_k$ is the energy of interaction between the molecule and the corresponding probe-atom *i* or *k* in the analyzed cell. A set of atoms C_{sp^3}, N_{sp^3}, O_{sp^3}, C_{sp^2}, N_{sp^2}, O_{sp^2} Cl, H and absence of any atom ("vacuum") were used as probes. Let us note that CoMFA [11] uses energy attributes to characterize the analyzed cells. In the method described here the probabilities of occupancy of a cell represents a different approach for the description of interactions between the molecule and the biological target. It might be argued that a probability based scheme offers improvements over an energy based method.

4) A possibility of a presence of a donor (or an acceptor) of a hydrogen bond in the cell. It is assumed that such a hydrogen bond can be formed between this donor or this acceptor and the analyzed molecule.

All structural parameters, i.e. integral, local and field parameters contain an exhaustive description of the molecular structure. The thousands of parameters (their exact number depends on the parameters of the lattice) are generated within the proposed approach for each analyzed molecule. This reduces the probability of missing the most significant parameters required to correlate activity of the analyzed molecules with their structure. The analysis of such large numbers of parameters requires an application of specialized methods, such as the trend-vector approach described in the next paragraph.

Due to the large set of structural parameters, in most cases, a chance to construct a few approximately equivalent models of "structure – property" appears. We suppose, that this fact results to the best interpretability of such kind dependences.

Data analysis

The trend-vector (T-vector) procedure [15, 16] does not concretize the form of a corresponding dependence and can use a great number of structural parameters. However, this method can predict properties of analyzed molecules only in a rank scale, i.e. it forecasts that the molecule is, let us say, more active than molecule A and less than B. This is not a crucial limitation for QSAR tasks.

The T-vector method is based on the fundamental idea of the pattern recognition theory. It divides n analyzed objects into two classes relative to the average value of their activity ($\overline{A}$). The data samples i with positive $A_i - \overline{A} > 0$ form one class and the data samples with negative values $A_i - \overline{A} < 0$ form another class. It is possible to consider $Ai - \overline{A}$ as charges. Hence, similarly to the dipole moment vector, the T-vector characterizes a division of charges (corresponding to active and inactive classes) in the multi-dimensional space of structural parameters S_{ij} ($i = \overline{1,n}$ -no of molecule, $j = \overline{1,m}$ - no of structural parameter). Each component of a T-vector is determined as

$$T_j = \frac{1}{n} \cdot \sum_{i=1}^{n} (A_i - \overline{A}) \cdot S_{ij}, \qquad (10)$$

and reflects a degree and direction of influence of the j-th structural parameter on the magnitude of a property A. The inverse problem is solved using the following relation:

$$rank(\hat{A}_i) = rank\left(\sum_{j=1}^{m} T_j S_{ij}\right). \qquad (11)$$

It is important to note that each component of the T-vector is calculated independently from the others and its contri-

Table 1 The structure-property analysis for monosubstituted benzenes (R-Ph)

No	R	measu-red k	rank of k	predicted values of k				
Method of the analysis				**model I (rank) T-vector**	**model II (rank) T-vector**	**model III (rank) T-vector**	**model IV (value) MLR**	**model V (value) PLS**
1	-H	0.00	5.5	14.0	10.0	10.0	0.23	0.22
2	$-CH_3$	0.11	11	10.0	8.0	12.0	0.10	0.10
3	$-C_2H_5$	0.13	12	9.0	13.0	9.0	0.11	0.11
4	$-C_3H_7$	0.04	8	3.0	6.0	7.0	-0.05	-0.04
5	$-CH(CH_3)_2$	0.07	9.5	6.0	7.0	8.0	0.04	0.04
6	$-C_4H_9$	0.07	9.5	11.0	9.0	11.0	0.12	0.13
7	$-N(NH_3)_3$	-0.07	3	7.0	11.0	1.0	0.06	0.00
8	$-C_6H_5$	0.45	21	19.0	18.0	19.0	0.35	0.36
9	-CHO	0.32	15	15.0	17.0	18.0	0.30	0.34
10	$-CO-CH_3$	0.48	22.5	21.0	23.0	24.0	0.46	0.52
11	$-CO-OCH_3$	0.48	22.5	23.0	24.0	21.0	0.46	0.48
12	$-COOC_2H_5$	0.55	24	27.0	29.0	23.0	0.65	0.70
13	$-OCH_3$	0.44	20	24.0	19.0	22.0	0.53	0.54
14	$-OC_2H_5$	0.39	16.5	16.0	21.0	14.0	0.40	0.38
15	-OH	0.40	18.5	20.0	16.0	16.0	0.44	0.40
16	$-SCH_3$	0.40	18.5	18.0	12.0	20.0	0.37	0.41
17	$-CF_3$	-0.09	2	4.0	2.0	2.0	-0.13	-0.15
18	-CN	0.23	13	12.0	15.0	15.0	0.28	0.24
19	-Br	0.00	5.5	5.0	4.0	5.0	-0.03	-0.04
20	-Cl	-0.01	4	8.0	5.0	3.0	0.01	0.00
21	-I	0.01	7	2.0	3.0	6.0	-0.08	-0.09
22	-F	-0.16	1	1.0	1.0	4.0	-0.15	-0.16
23	$-NO_2$	0.26	14	13.0	14.0	13.0	0.28	0.26
24	$-NH_2$	0.66	26	25.0	25.0	27.0	0.65	0.65
25	$-NHCH_3$	0.73	27	28.0	26.0	26.0	0.74	0.77
26	$-NHC_2H_5$	0.79	28	31.0	28.0	30.0	0.87	0.86
27	$-CH_2CN$	0.39	16.5	17.0	20.0	17.0	0.31	0.34
28	$-N(CH_3)_2$	0.90	30	29.0	27.0	25.0	0.79	0.82
29	$-N(C_2H_5)_2$	0.81	29	26.0	30.0	29.0	0.76	0.83
30	$-CH_2OH$	0.59	25	22.0	22.0	28.0	0.55	0.57
31	$-CO-N(CH_3)_2$	1.31	33	34.0	35.0	33.0	1.28	1.29
32	$-CO-N(C_2H_5)_2$	1.31	33	33.0	32.0	31.0	1.31	1.21
33	$-SO_2-N(CH_3)_2$	1.24	31	32.0	31.0	34.0	1.19	1.15
34	$-SO_2-N(C_2H_5)_2$	1.31	33	30.0	34.0	32.0	1.20	1.42
35	$-SO_2-N(C_3H_7)_2$	1.33	35	35.0	33.0	35.0	1.43	1.22
36	$-O-CH(CH_3)_2$	0.51	24	23.0	24.0	17.0	0.54	0.51
37	-O-Ph	0.43	20	22.0	17.0	23.0	0.52	0.50
38	$-O-CH_2$-Ph	0.60	26	29.0	29.0	25.0	0.78	0.76
Correlation between measured and predicted k values (R^2)				0.928	0.923	0.959	0.936	0.974
Cross-validated R^2 (Q^2)				0.924	0.738	0.751	0.921	0.890

Table 2 *The structural parameters of monosubstituted benzenes*

No	Structural parameter	Component of T-vector	VIP[a]
Integral			
P_1	Distribution of lipophilicity potential (70)[b]	-6.613	1.745 (-)[c]
P_2	Distribution of lipophilicity potential (105)	4.824	1.143 (+)
P_3	Distribution of lipophilicity potential (29)	-3.521	0.948 (-)
P_4	Distribution of lipophilicity potential (332)	3.264	0.954 (+)
P_5	Distribution of carbon atoms (99)	5.889	1.586 (+)
P_6	Distribution of Nsp^3-carbon atoms (833)	5.641	1.408 (+)
P_7	Distribution of hydrogen atoms (757)	4.663	1.216 (+)
P_8	Distribution of fluorine atoms (839)	-3.177	0.675 (-)
P_9	Polarizability of molecule (models III.V)	5.852	1.731 (+)
P_{10}	Lipophilicity of molecule (models II. III.V)	-3.627	0.926 (-)
P_{11}	Parameter of structural similarity (Rb) by electronegativity of atom to structure 32[d] (model II)	-5.162	1.486 (-)
Local			
P_{12}	Average atom charge in the cell 411	-3.997	1.219 (-)
P_{13}	Average atom charge in the cell 383	2.140	0.808 (+)
P_{14}	Average lipophilicity in the cell 46	-2.760	0.927 (-)
P_{15}	Average polarizability in the cell 384	1.427	0.275 (+)
P_{16}	Average polarizability in the cell 356	0.507	0.086 (-)
P_{17}	Average electronegativity in the cell 46	-0.735	0.167 (-)
P_{18}	Enable of fluorine in the cell 45	-2.874	0.615 (-)
Field			
P_{19}	Electrostatic potential in the cell 1425	-2.333	0.429 (-)
P_{20}	Electrostatic potential in the cell 734	-0.683	0.490 (+)
P_{21}	Electrostatic potential in the cell 33	-1.065	0.243 (-)
P_{22}	Lipophilicity potential in the cell 45	-2.087	0.597 (-)
P_{23}	A probability that cell 55 is occupied by nitrogen	1.884	0.572 (+)
P_{24}	A probability of the cell 102 to be empty	2.979	0.828 (+)

[a] VIP is the sum over all model dimensions of the contributions VIN (variable influence). For a given PLS dimension, a, $(VIN_k^a)^2$ *is equal to the squared PLS weight* $(w_{ak})^2$ *of that term, multiplied by the percent explained dispersion by that PLS dimension. The accumulated (over all PLS dimensions) value*

$$VIP_k = \sum_a (VIN_k^a)^2$$

[b] This is the number of harmonic for Fourier transformation of distributions of atom properties along SC
[c] This is the sign of weight (w_{ak}) *for last dimension (a=2).*
[d] The structure 32 has one from the largest k values

bution to a model is not adjusted. Thus, the influence of the number of used structural parameters on the reliability of the model is not so critical, as in the case of the regression methods.

A quality of the structure - property relationship was estimated by the Spearman rank correlation coefficient calculated between ranks of the experimental and calculated activities Ai. The estimation of a reliability of a model was done using K series of the training set compounds (K=30 is usually enough to reproduce results) with randomly shuffled activities. The same analysis and Spearman rank correlation coefficients $\rho_{cr}^{rand} = f(A_i, \hat{A}_i)$ were calculated for all random series. The calculated model is considered to be statistically reliable, if

$$\rho(A_i, \hat{A}_i) >> \rho_{rand}(A_i, \hat{A}_i) + \varepsilon ,$$

$$\text{where } \overline{\rho}_{rand} = 1/K \sum_{k=1}^{K} \rho_{cr}^{rand} , \qquad (12)$$

where the confidence intervals are calculated at level of significance p=α=0.99.[15, 25] Each model was also tested by leave-one-out cross validation and, at the final step, it was used to predict the test set compounds.

Table 3 *The PLS-method results calculated for different groups of structural parameters*

parameters	R^2	R^2 cross-validated	number of PLS latent variables	total number of parameters N
Harmonic	0.953	0.844	2	14470
Integral	0.558	0.487	1	513
Local	0.809	0.638	2	1593
Integral and local	0.934	0.754	3	2106
Selected by T-vector (model IV)	0.973	0.891	2	23

Results and discussion

The efficiency of the proposed approach was examined for a data set derived from the formation of charge-transfer complexes between mono-substituted benzenes and 1,3,5-trinitrobenzene.[26] As a target property the formation constants of the charge-transfer complexes (Table 2) were analyzed. The integral characteristics of molecules, i.e. dipole moment, parameters of an inertia ellipsoid, calculated logP and a molar refraction, quantum-chemical characteristics of some atoms of substituent R, were used in the original work.[26] The authors applied multi-dimensional regression analysis, principal component regression (PCR) and partial least squares (PLS) to study the structure-property relationship. A high correlation coefficient R^2=0.95 was calculated for the training data set of 35 compounds. The calculated model was used to predict formation constants of 3 compounds from the test set with an average error of about 27 %. The same training and test sets as in the original study [26] were used in the current analysis (Table 2).

A cubic lattice with a cell size of 1.8 Å was used. The molecules were superimposed according to their aromatic rings. Each molecule was represented with about 20,000 structural parameters calculated as indicated in the Method section. The regression analysis, PLS-method [14] and trend-vector procedure [15, 16] were applied to detect ”structure-property” relationships. It is clear that regression analysis and the trend-vector procedure can not be applied for a set of 20 thousand structural parameters. Therefore, at the beginning of the analysis highly correlated parameters were ex-

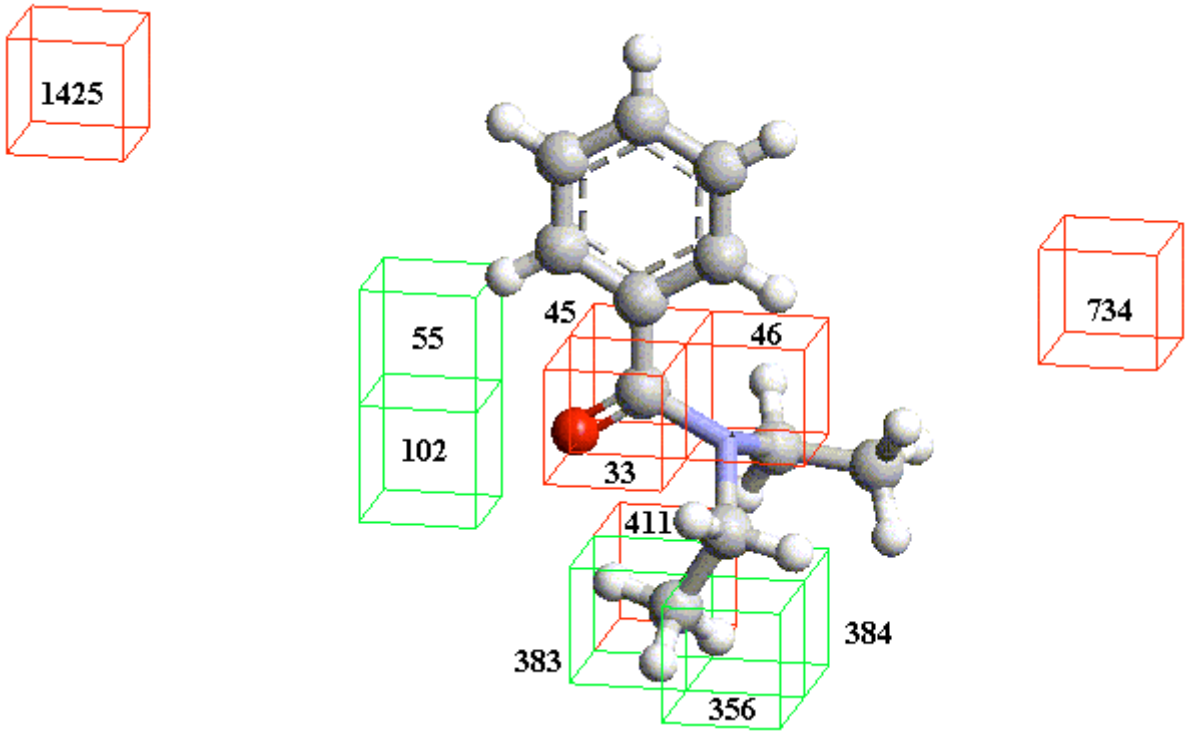


Figure 3 *Cubic cells of the monosubstituted benzenes, that influence the complexing ability of molecules. Red (green) color indicate that the corresponding cell have negative (positive) influence for the formation of the charge-transfer complexes*

cluded (at the level R=0.7) in each of harmonic, integral, local and field groups. This procedure decreased the total number of parameters to 125, 35, 79 and 68 in each group respectively. The analysis was carried out with three sets of parameters, namely

1) harmonic parameters;
2) local and integral parameters;
3) field and integral parameters.

As a result three approximately equivalent models (I, II, III) (Table 2, 3) were calculated with the trend-vector method. The first (I) model (Table 3) contained parameters (P_1-P_8) calculated by the Fourier transform of distributions of atom properties along SC. The structural characteristics of molecules, related to lipophilicity (P_1-P_4), and parameters of the shape of molecules (P_5-P_8) were found to have the strongest influence on the complex formation constants of the compounds. The first group of these parameters reflects the ability of molecules to form intermolecular associations. The second group of parameters (P_5-P_7) represents the steric factors of the complexing process. To some degree the reactivity of substituted benzene depends on the distribution of the electronegative fluorine atoms.

The model II reflects the influence of integral and local structural characteristics (P_{10}-P_{18}) of molecules on their ability to form the charge-transfer complexes (Figure 3). The results of calculation (Table 3) show that electronegative substituents (P_{17}-P_{18}) reduce electron donor properties of the aromatic ring, and this prevents the formation of the charge-transfer complexes. The easily polarized substituents (P_{15}, P_{16}) increase the reactivity of the molecules. The charge characteristics (P_{12}, P_{13}) and electronegativity (P_{17}) of the substituents also have a very important influence on the complex formation constant. The integral parameters (P_9, P_{10}) also show that higher polarizability and lower lipophilicity of the molecule increase their ability to form the charge-transfer complexes. The molecules characterized by polarity (electronegativity of atoms, see Parameter P_{11} in Table 3) that was close to that of one of the active molecules (structure 32, Table 2) were characterized by high reactivity.

The integral and field structural parameters of monosubstituted benzenes (P_9, P_{10}, P_{19}-P_{24}) were used in the model III. As was expected, the parameters of an electrostatic field of a molecule (P_{19}-P_{21}) were important for the process of formation of charge-transfer complexes. In addition, the considered intermolecular interaction depended on a lipophilicity field (P_{22}) that accounted for the contribution of hydrophilic/hydrophobic interactions.

It is important to mention that the target receptor for the charge-transfer complex formation is a molecule of trinitrobenzene. The analysis of probabilities of an occupancy of lattice cells with various probe atoms (fragments of a receptor) in space around the monosubstituted benzenes, demonstrated, that the presence of a nitrogen atom in the cell 55 (P_{23}) increased the activity of compounds. The probability that this region contains nitrogen atoms of ($–NO_2$) group of the trinitrobenzene is very high. On the contrary, the cell 102 (P24) is probably unavailable during the complexing process due to presence of a steric barrier.

Satisfactory results were obtained by the multiple linear regression method (model IV):

$$k_{calc}=0.206-0.041\ \mathbf{P_{11}}-0.410\ \mathbf{P_{10}}+0.104\ \mathbf{P_9},$$
$$R=0.967;\ F=151;\ S=0{,}12 \qquad (13)$$

This equation was calculated by stepwise regression using parameters of all three groups after excluding cross-correlated terms (R>0.70). The results calculated by this model were in perfect agreement with those calculated by the T-vector procedure. The decrease of lipophilicity and increase of polarizability of a molecule increases its complexing ability.

The best results in the present work were calculated using the PLS method, as shown for the model (V). It is important to mention, that the preliminary selection of structural parameters essentially influences the quality of the calculated model. For example, if highly correlated terms were not eliminated from the analysis, the best PLS models were in the range R^2 [0.934, 0.953], while the cross-validated result measured by Q^2 were only [0.754-0.844] (Table 4). The structural parameters pre-selected by the T-vector procedure calculated two-dimensional PLS- model with R^2=0.974 and Q^2=0.890, indicating its high efficiency for data description and high generalization ability. The prediction of this given model for the data from the test set was also quite satisfactory (Table 2).

It should be pointed out here that the calculated results logically reflect the physico-chemical peculiarities of the formation of charge-transfer complexes and are in agreement with the conclusions of the previous study.[26] From the results shown here, the use of the lattice model for QSAR tasks in a combination with various statistical methods represents a new approach to the construction of QSAR models. Further studies are in progress to confirm this.

Acknowledgment This study was supported in part with INTAS - UA 95-0060 grant.

References

1. Hansch, C.; Fujita, T. *J. Am. Chem. Soc*. **1964**, *86*, 1616-1626.
2. Free, S.; M., Wilson, J. M. *J. Med. Chem*. **1964**, *7*, 395-399.
3. Höltje, H.-D.; Kier, L. B. *J. Med. Chem*. **1974**, *17*, 814-819.
4. Goodford, P. J. *J. Med. Chem*. **1985**, *28*, 849-857.
5. Kubinyi, H. (Ed.) *3D QSAR in Drug Design. Theory, Methods and Applications* ; ESCOM: Leiden, 1993.
6. Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function* ; A Wiley-Interscience Publication, 1979.
7. Suchachev, D. V.; Pivina, T. S.; Shliapochnikov, V. A.; Petrov, E. A.; Paliulin, V. A.; Zefirov, N. S. *Dokl. RAN* (in Russ), **1993**, *328*, 50-57.

8. Rozenblit, A. B.; Golender, V. E. *The Logic – Combinatorial Methods in Drug Design*; Zinatne: Riga, 1983 (in Russ).
9. Walters, D. E.; Hopfinger, A. J. *J. Mol. Structure (Theochem)*, **1986**, *134*, 317-323.
10. Cramer, R. D. *J. Am. Chem. Soc*. **1980**, *102*, 1837-1859.
11. Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc*. **1988**, *110*, 5959-5967.
12. Doweyko, A.M. *J. Math. Chem.* **1988**, *31*, 1396-1406.
13. Doweyko, A.M. *J. Math. Chem*. **1991**, *7*, 273-285.
14. Glen, W. G.; Dunn, W. J.; Scott, D. R. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349.
15. Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci*. **1985**, *25*, 64-73.
16. Vitiuk, N. V.; Kuz'min, V. E. *Zh. Anal. Khimii* (in Russ), **1994**, *49*, 65-167.
17. Kuz'min, V. E.; Krutius, S. V. *Khim.-Pharm. Zhurn* (in Russ), **1986**, *7*, 791-794.
18. Kuz'min, V. E.; Artemenko, A. G. *Zh. Struct. Khimii* (in Russ), **1998**, *39*, 537-542.
19. Ferster, E.; Renz, B. *Methoden der Korrelations und Regressionanalyse* ; Verlag Die Wirtschaft:Berlin, 1979.
20. Faure, A. *Perception et reconnaissance des formes*; Editests, 1985.
21. Kuz'min, V.E.; Artemenko, A.G.; Kovdienko, N.A.; Zheltvay A.I. *Khim.-Pharm. Zhurn* (in Russ), **1999**, *9*, 14-20.
22. Jolly, W.L.; Perry, W.B. *J. Am. Chem. Soc*. **1973**, *95*, 5442.
23. Kuz'min, V. E.; Beresteckaja, E. L. *Zh. Struct. Khimii* (in Russ), **1983**, *24*, 187-188.
24. Croizet, F.; Langlois, M. H.; Dubost, J. P.; Braquet, P.; Audry, E.; Dallet, P.; Colleter, J. C. at al. *J.Mol.Graphics*, **1990**, *8*, 153-155.
25. Likeš, J.; Laga, J. *Zakladni Statisticke Tabulky* ; SNTL: Praha, 1978.
26. Livingstone, D. J.; Evans, D. A.; Saunders, M. R. *J.Chem.Soc.Perkin Trans*, **1992**, *2*, 1545-1550.