SAR and QSAR in Environmental Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100694

QNA-based 'Star Track' QSAR approach

D. A. Filimonov [a]; A. V. Zakharov [a]; A. A. Lagunin [a]; V. V. Poroikov [a]
[a] Institute of Biomedical Chemistry of Russian Academy of Medical Sciences, Moscow, Russia

Online publication date: 17 December 2009

To cite this Article Filimonov, D. A., Zakharov, A. V., Lagunin, A. A. and Poroikov, V. V.(2009) 'QNA-based 'Star Track' QSAR approach', SAR and QSAR in Environmental Research, 20: 7, 679 — 709
To link to this Article: DOI: 10.1080/10629360903438370
URL: http://dx.doi.org/10.1080/10629360903438370

SAR and QSAR in Environmental Research
Vol. 20, Nos. 7–8, October–December 2009, 679–709

Taylor & Francis
Taylor & Francis Group

QNA-based 'Star Track' QSAR approach†

D.A. Filimonov*, A.V. Zakharov, A.A. Lagunin and V.V. Poroikov

Institute of Biomedical Chemistry of Russian Academy of Medical Sciences, Moscow, Russia

(*Received 6 July 2009; in final form 1 October 2009*)

In the existing quantitative structure–activity relationship (QSAR) methods any molecule is represented as a single point in a many-dimensional space of molecular descriptors. We propose a new QSAR approach based on Quantitative Neighbourhoods of Atoms (QNA) descriptors, which characterize each atom of a molecule and depend on the whole molecule structure. In the 'Star Track' methodology any molecule is represented as a set of points in a two-dimensional space of QNA descriptors. With our new method the estimate of the target property of a chemical compound is calculated as the average value of the function of QNA descriptors in the points of the atoms of a molecule in QNA descriptor space. Substantially, we propose the use of only two descriptors rather than more than 3000 molecular descriptors that apply in the QSAR method. On the basis of this approach we have developed the computer program GUSAR and compared it with several widely used QSAR methods including CoMFA, CoMSIA, Golpe/GRID, HQSAR and others, using ten data sets representing various chemical series and diverse types of biological activity. We show that in the majority of cases the accuracy and predictivity of GUSAR models appears to be better than those for the reference QSAR methods. High predictive ability and robustness of GUSAR are also shown in the leave-20%-out cross-validation procedure.

Keywords: QNA; QSAR; biological activity; toxicity; GUSAR

1. Introduction

Quantitative structure–activity relationships (QSARs) have been employed in numerous areas from drug design to the assessment of chemical toxicity. Many QSAR methods have been developed over the past years. These methods differ by the particular molecular descriptors used to extract structural information in the form of a digital representation that is suitable for model development, and by the mathematical approaches used for finding the best predictive QSAR model. From a general point of view, the estimate y_{pred} of activity for an organic molecule can be represented as

$$y_{pred} = a_0 + \sum\nolimits_i a_i f_i(S), \tag{1}$$

where a_0, a_1, ... are the variable coefficients, $f_1(S)$, $f_2(S)$, ..., $f_i(S)$, ... are independent from the coefficients a_0, a_1, ... different functions of organic molecule's structure S.

*Corresponding author. Email: dmitry.filimonov@ibmc.msk.ru
† Presented at CMTPI 2009: Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (Istanbul, Turkey, 4–8 July 2009).

ISSN 1062–936X print/ISSN 1029–046X online

DOI: 10.1080/10629360903438370
http://www.informaworld.com

In classic QSAR methods, the functions $f_1(S)$, $f_2(S), \ldots$ represent physical-chemical parameters or other quantitative characteristics of molecular structure, and the coefficients a_0, $a_1, \ldots$ are determined using multiple linear regression (MLR), partial least squares (PLS) analysis, or support vector regression (SVR), etc. [1]. QSAR methods based on the similarity between a certain molecule S_i with known biological activity and the molecule S use the value $f_i(S)$ of their similarity [2–4].

More than 3000 molecular descriptors are currently used in QSARs [5–7]. Some authors have proposed universal descriptors [8–10] while others have used unique descriptors [7].

Earlier, we developed the uniform Multilevel Neighbourhoods of Atoms (MNA) descriptors [11] for prediction of the biological activity spectra for substances [12,13]. Recently, we have proposed Quantitative Neighbourhoods of Atoms (QNA) descriptors, which reflect better the nature of intermolecular interactions [14,15]. It appears that the specific nature of QNA descriptors requires the appropriate algorithm for their efficient application.

In this paper, we describe a QNA-based 'Star Track' QSAR approach, which differs significantly from other known methods. In these methods any molecule is represented as a single point in a many-dimensional space of molecular descriptors, where $f_1(S), f_2(S), \ldots, f_i(S), \ldots$ are the coordinates of this point (Equation (1)). On the contrary, in the 'Star Track' methodology any molecule is represented as a set of points in two-dimensional (2D) space of QNA descriptors. In this space biological activity can be considered as a 'potential' whose value, averaged through the points corresponding to the atoms of a certain molecule, gives the estimation of the biological activity of this molecule. In this study 2D Chebyshev polynomials are used for approximation of this 'potential' of biological activity.

To create the QSAR models, we applied the self-consistent regression (SCR) method which we developed earlier [15,16]. It has been demonstrated that SCR provides the selection of the optimal set of descriptors for creation of a reliable QSAR model [16].

The QNA-based 'Star Track' approach described in this paper is implemented in the computer program GUSAR. We have compared GUSAR with several widely used QSAR methods including CoMFA, CoMSIA, Golpe/GRID, HQSAR and others, using ten data sets representing various chemical series and diverse types of biological activity. Since in the majority of cases the accuracy and predictivity of GUSAR models appeared to be better than for the reference QSAR methods, GUSAR can be recommended as a tool for QSAR problem solving.

2. Methods

The 2D structural formula represents the atomic composition and the structure of the molecule, but it is in practice some abstraction of reality. On the other hand, the traditional 2D structural formula forms the basis for any calculation in molecular mechanics or quantum chemistry. Many characteristics of chemical compounds can be calculated on the basis of structural formula [1,5,17–21]. Hence, it can be concluded that the structural formula uniquely determines the properties of a molecule. Moreover, the 2D structural formula is the only information available in the early stages of research.

Neighbourhoods of Atoms descriptors (QNA as well as MNA) are calculated based on the structural formula representation, which, according to the valences and charges

of atoms, includes explicitly all hydrogen atoms and does not specify the type of bonds. This form of structural formula is uniquely determined, e.g. it does not depend on alternative methods of aromatic structure presentation.

The intermolecular interaction determines molecular recognition, the major cause of biological activity of organic molecules. In fact, interatomic and intermolecular forces are electrical in nature according to the Hellman–Feynman theorem [22]. On this fundamental basis we have developed topoelectrical indices [23] and, later, QNA descriptors [14,15].

2.1 *QNA descriptors*

QNA descriptors are calculated based on the connectivity matrix (**C**) and the standard values of the ionization potential (*IP*) and electron affinity (*EA*) of atoms in a molecule. For any given atom *i*, the QNA descriptors are calculated as follows:

$$P_i = B_i \sum_k \left(\mathrm{Exp}\left(-\frac{1}{2}\mathbf{C} \right) \right)_{ik} B_k, \tag{2}$$

$$Q_i = B_i \sum_k \left(\mathrm{Exp}\left(-\frac{1}{2}\mathbf{C} \right) \right)_{ik} B_k A_k, \tag{3}$$

where $A_k = \frac{1}{2}(IP_k + EA_k)$, $B_k = (IP_k - EA_k)^{-\frac{1}{2}}$. The values of *EA* and *IP* collected from many different sources and used in this work are represented in Appendix 1 (Table A1). Although the value $\mu P - Q$ can be considered by convention as the partial atomic charge, where μ is the chemical potential, in general the *P* and *Q* values are not the estimate of partial atomic charges or hardness, etc.

The QNA descriptors describe each of the atoms in a molecule and, at the same time, each of the *P* and *Q* values depend on the whole composition and structure of a molecule (Figure 1).

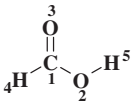
From Figure 1(c) it is clear that any atom influences the others, although the influence decreases with the increase of the distance between them; for example, components of matrix $\mathrm{Exp}(-\frac{1}{2}\mathbf{C})$ for atom 1 (C) are: 1.40 for atom 1 itself, −0.59 for its immediate neighbour atom 2 (O), −0.57 for atoms 3 (O) and 4 (H), and 0.14 for atom 5 (H).

The algorithm of the QNA descriptor calculation is really very simple due to the uselessness of the matrix $\mathrm{Exp}(-\frac{1}{2}\mathbf{C})$ itself, the fact that the product of $\mathrm{Exp}(-\frac{1}{2}\mathbf{C})$ by a vector is needed only, and the fact that the matrix **C** consists of 0 and 1 only. Appendix 1 includes the listing (in the Delphi 5.0 language) of the QNA descriptor calculation procedure.

2.2 *The 'Star Track' approach*

The main feature of QNA descriptors is that they represent a molecule as a set of *P* and *Q* values, or, in another words, as a 'constellation' in QNA descriptor space. It is important to emphasize that in this approach each of the atoms in a molecule is peering to the others. Each 'star' (an atom of a molecule) has a fixed unique position in QNA descriptor space, which does not depend on the method of the structural formula presentation. Such 'constellations' are shown in Figure 2 for acetylsalicylic acid and sulfathiazol in the normalized QNA descriptor space.

(a)



(b)

$$\mathbf{C} = \begin{vmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{vmatrix}$$

(c)

$$\mathrm{Exp}\left(-\tfrac{1}{2}\mathbf{C}\right) = \begin{vmatrix} 1.40 & -0.59 & -0.57 & -0.57 & 0.14 \\ -0.59 & 1.27 & 0.14 & 0.14 & -0.54 \\ -0.57 & 0.14 & 1.13 & 0.13 & -0.02 \\ -0.57 & 0.14 & 0.13 & 1.13 & -0.02 \\ 0.14 & -0.54 & -0.02 & -0.02 & 1.13 \end{vmatrix}$$

(d)

	EA	*IP*	*A*	*B*	*P*	*Q*
C	1.263	11.26	6.262	0.316	–0.00218	–0.1820
O	1.461	13.62	7.541	0.287	0.02944	0.3019
O	1.461	13.62	7.541	0.287	0.06199	0.5297
H	0.754	13.60	7.177	0.279	0.05812	0.4706
H	0.754	13.60	7.177	0.279	0.05304	0.3533

Figure 1. Example of the QNA descriptor calculation for a molecule of formic acid: (a) structural formula; (b) connectivity matrix; (c) exponent of the connectivity matrix; (d) electron affinities (*EA*), ionization potentials (*IP*), variables of Equations (2) and (3), and *P* and *Q* values for each of the atoms of formic acid molecule.

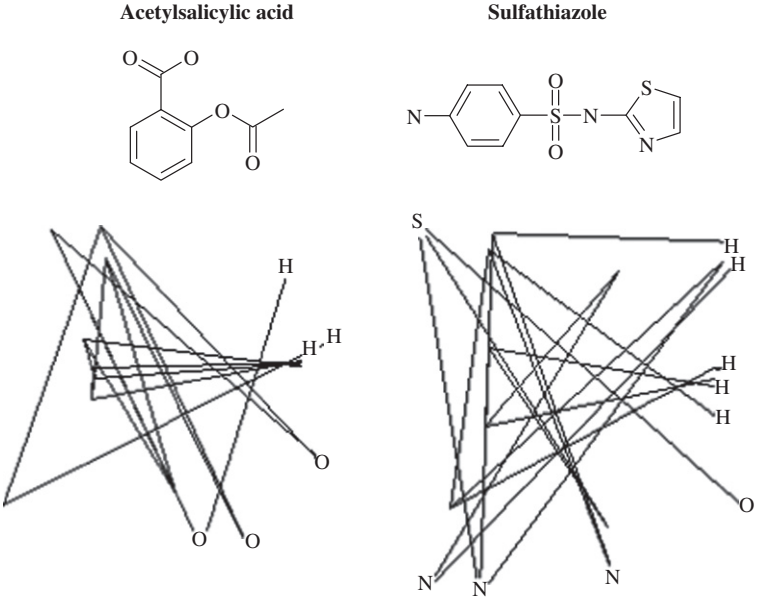


Figure 2. The acetylsalicylic acid and sulfathiazole 'constellations' in the normalized QNA descriptor space. Heteroatoms and hydrogens are explicitly presented, carbons are presented as points only, and bond types are not specified.

This 'stellar' feature of QNA descriptors means that they differ from those commonly used in QSAR descriptors, such as universal 4D fingerprints [8]. On the other hand, this feature of the QNA descriptors creates a problem of their use in Equation (1). Previously we have used QNA descriptors in another QSAR method [15] with QNA quantile functions. We showed that the QSAR method based on QNA quantile functions and SCR provided a reasonably accurate prediction for three training sets of acute aquatic toxicity [15]; nevertheless, this method appeared to be less accurate for some other biological activities compared to traditional QSAR methods. Therefore, we changed the algorithm of the QNA application. We propose to calculate each $f_i(S)$ function of the structure of a molecule in Equation (1) as the average value of the $g_i(P, Q)$ function of the P and Q variables for those m molecule atoms that have two or more immediate neighbours:

$$f_i(S) = \frac{1}{m}\sum\nolimits_k g_i(P_k, Q_k). \tag{4}$$

After substitution of expression (4) into Equation (1) and interchange of summations we find

$$y_{pred} = a_0 + \sum\nolimits_i a_i \frac{1}{m}\sum\nolimits_k g_i(P_k, Q_k) = \frac{1}{m}\sum\nolimits_k \left(a_0 + \sum\nolimits_i a_i g_i(P_k, Q_k)\right). \tag{5}$$

According to Equation (5) the estimate y_{pred} for a molecule can be interpreted as an average of the values predicted for particular atoms in a molecule. Formally, QNA descriptors represent a molecule structure by two descriptors only (P and Q), in contrast to the numerous traditional descriptors used in QSAR.

For all molecules from the data sets used in this work we have calculated 16,617 QNA descriptors for the atoms that have two or more immediate neighbours. In Figure 3(a) they are presented as points in QNA descriptor space: the white shading corresponds to the number of QNA descriptors in a cell (pixel). Figure 3(a) shows that the P and Q values are

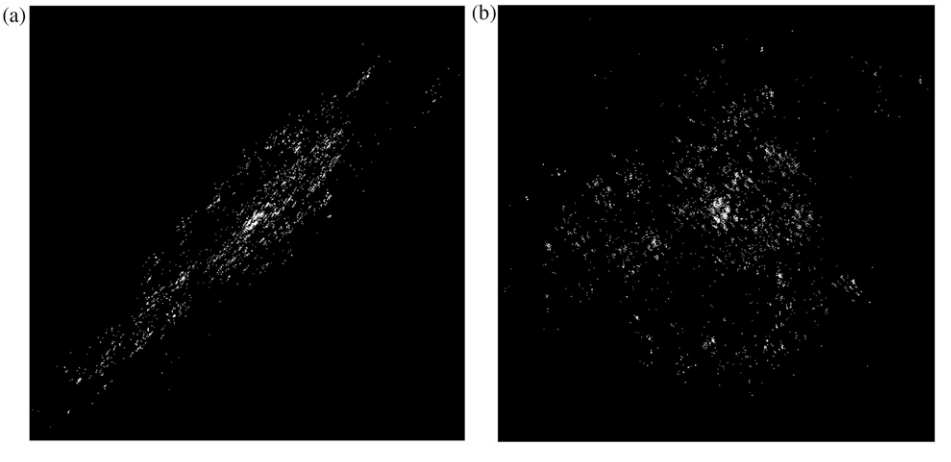


Figure 3. Distribution of the 16,617 QNA descriptors in 300×300 cells (pixels): (a) initial P (abscissa axis) and Q (ordinate axis) values within the boundaries $(-0.0579, 0.0784)$ for P and $(-0.581, 0.666)$ for Q; (b) normalized QNA within the boundaries $(-3, 3)$ for both U and V.

strongly correlated ($r = 0.903$). Since the P and Q values have different scales (the standard deviations are 0.023 and 0.208, respectively), we made the normalization to optimize a family of functions $g_i(P, Q)$.

Normalization has been performed by calculation of the average values (E_P and E_Q), the standard deviations (D_P and D_Q), and correlation between the P and Q values (R_{PQ}):

$$P' = \frac{P - E_P}{D_P}, \quad Q' = \frac{Q - E_Q}{D_Q}, \tag{6}$$

$$U = \frac{P' + Q'}{\sqrt{2(1 + R_{PQ})}}, \quad V = \frac{P' - Q'}{\sqrt{2(1 - R_{PQ})}}. \tag{7}$$

The orthonormal U and V have zero mean, unit variance, and they are uncorrelated, which is shown in Figure 3(b).

In this study we chose Chebyshev polynomials as the family of functions $g_i(P, Q)$, and the orthonormal U and V values have been additionally transformed by using a hyperbolic tangent, so the 'normalized QNA' vary from -1 to 1. After this, the functions $g_i(P, Q)$ in Equation (5) are represented using Chebyshev polynomials as

$$g_i(P, Q) = T_{uv}(P, Q) = \mathrm{Cos}(u \times \mathrm{ArcCos}(\mathrm{TanH}(U))) \times \mathrm{Cos}(v \times \mathrm{ArcCos}(\mathrm{TanH}(V))), \tag{8}$$

where the integers u, $v = 0, 1, 2, \ldots$ define the 2D Chebyshev polynomial degree. The final equation for estimate y_{pred} using QNA descriptors is

$$\begin{aligned} y_{pred} &= \frac{1}{m}\sum\nolimits_k \left(a_0 + \sum\nolimits_{uv} a_{uv} T_{uv}(P_k, Q_k)\right) = a_0 + \sum\nolimits_{uv} a_{uv} T_{uv}, \\ T_{uv} &= \frac{1}{m}\sum\nolimits_k T_{uv}(P_k, Q_k). \end{aligned} \tag{9}$$

QNA descriptors and their polynomial transformations (6)–(8) do not provide information on the shape and volume of a molecule although this information may be important for determination of the SARs. Therefore, these parameters were added to the QNA descriptors. The topological length of a molecule was calculated as the maximal distance between any two atoms and the volume of a molecule as the sum of each atom's volume, $\frac{4}{3}\pi R^3$, where R is the atomic radius (see Appendix 1, Table A1).

The Chebyshev polynomials are arranged in ascending order of their degrees $u + v$. For $u + v = 1$ they are $T_{1,0}$, $T_{0,1}$; for $u + v = 2$ they are $T_{2,0}$, $T_{1,1}$, $T_{0,2}$; for $u + v = 3$ they are $T_{3,0}$, $T_{2,1}$, $T_{1,2}$, $T_{0,3}$, etc. The first, second and third power of topological length and volume of a molecule were used. The number of initial variables equals the number of Chebyshev polynomials plus the number of the first, second and third power of topological length and volume of a molecule. It is significantly less comparing to the number of molecules in the training set. The number of final variables in the QSAR equation selected after self-consistent regression procedure is also significantly less comparing to the number of initial variables (see sections 2.3, 5.1 to 5.10).

2.3 *Self-consistent regression*

The classical MLR has a number of limitations. In particular, the number of objects in the training set should significantly exceed the number of independent variables, and it is important to use non-collinear variables only. To overcome these limitations we have

Table 1. Comparison of the prediction accuracy of GUSAR and other methods.

Methods	R^2	Q^2	R^2_{test}
CDK2 inhibitors [24]*			
GUSAR	**0.84**	**0.77**	**0.87**
CoMFA	0.94	0.56	0.86
DHFR inhibitors [25]			
GUSAR	**0.78**	**0.73**	**0.60**
CoMFA	0.79	0.65	0.59
CoMSIAbasic	0.76	0.63	0.52
CoMSIAextra	0.75	0.65	0.53
HQSAR	0.81	0.69	0.63
EVA/PLS	0.81	0.64	0.57
2D Cerius2/PLS	0.61	0.51	0.47
3D Cerius2/PLS	0.65	0.53	0.49
ACE inhibitors [25]			
GUSAR	**0.83**	**0.78**	**0.54**
CoMFA	0.80	0.68	0.49
CoMSIAbasic	0.76	0.65	0.52
CoMSIAextra	0.73	0.66	0.49
HQSAR	0.84	0.72	0.30
EVA	0.84	0.70	0.36
2D Cerius2/PLS	0.76	0.68	0.47
3D Cerius2/PLS	0.82	0.72	0.51
Alpha-2 adrenoreceptors [26]			
GUSAR	**0.82**	**0.71**	**N/A**
CoMFA	0.92	0.69	N/A
Estrogenic receptors [27]			
GUSAR	**0.93**	**0.89**	**N/A**
MLR/E-states descriptors	0.82	0.77	N/A
Vibrio fischeri toxicity [28]			
GUSAR	**0.88**	**0.84**	**N/A**
SCR-qQNA	0.91	0.87	N/A
CoMFA	0.92	0.79	N/A
PCR, MLR/ETA descriptors	0.84	0.73	N/A
PCR, MLR/different 2D descriptors	0.80	0.76	N/A
PCR, MLR/ETA and 2D descriptors	0.80	0.76	N/A
Factor score, PCR, MLR/ETA descriptors	0.89	0.82	N/A
Factor score, PCR, MLR/different 2D descriptors	0.87	0.83	N/A
Factor score, PCR, MLR/ETA and 2D descriptors	0.91	0.85	N/A
GFA/ETA descriptors	0.86	0.77	N/A
GFA/ different 2D descriptors	0.82	0.81	N/A
GFA/ETA and 2D descriptors	0.87	0.78	N/A
Chlorella vulgaris toxicity [29]			
GUSAR	**0.93**	**0.89**	**N/A**
SCR-qQNA	0.89	0.85	N/A
MLR/different 2D descriptors	0.84	0.82	N/A
PLS/different 2D descriptors	0.86	0.84	N/A
Tetrahymena pyriformis toxicity [30]			
GUSAR	**0.80**	**0.75**	**0.67**
SCR-qQNA	0.69	0.65	N/A

(*Continued*)

Table 1. Continued.

Methods	R^2	Q^2	R^2_{test}
MLR/different 2D descriptors	0.54	0.53	0.48
SWR1/different 2D descriptors	0.65	0.63	0.58
PLS/different 2D descriptors	0.76	0.75	0.64
GA single/different 2D descriptors	0.65	0.64	0.71
SWR2/different 2D descriptors	0.66	0.64	0.72
Neural Network/different 2D descriptors	0.71	N/A	0.73
CYP2A5 inhibitors [31]			
GUSAR	**0.90**	**0.88**	**0.93**
CoMFA	0.94	0.79	0.83
GRID/GOLPE	0.94	0.86	0.90
CYP2A6 inhibitors [31]			
GUSAR	**0.90**	**0.84**	**0.93**
CoMFA	0.97	0.81	0.77
GRID/GOLPE	0.93	0.78	0.76

R^2 is the square of the regression coefficient;
Q^2 is the cross-validated R^2;
R^2_{test} is the R^2 value for the test set, if it is available;
*– literature references;
N/A – Not available;
PCR – principal component regression;
MLR – multiple linear regression;
Factor score – factor scores were used as independent variables so that the backward stepwise regression method could be applied;
GFA – genetic function approximation;
SWR – stepwise regression.

employed the approach based on statistical regularization of ill-posed problems [15,16]. This resulted in a regularized least-squares method:

$$\boldsymbol{a} = \mathrm{ArgMin}\left[\sum_{i=1}^{n}\left(y_i - \sum_{k=0}^{m} x_{ik}a_k\right)^2 + \sum_{k=1}^{m} v_k a_k^2\right] \tag{10}$$

where $\boldsymbol{a}$ is the regression coefficients, n is the number of objects, y_i is the response value of the ith object, m is, here and below, the number of the independent variables, x_{ik} is the value of the kth independent variable of the ith object, a_k is the kth value of the regression coefficients, and v_k is the kth value of the regularization parameters. Equation (10) has the following solution:

$$\boldsymbol{a} = \mathbf{T}\mathbf{X}^T y, \quad \mathbf{T} = (\mathbf{X}^T\mathbf{X} + \mathbf{V})^{-1} \tag{11}$$

where $\mathbf{X}^T$ is the transposed regression matrix $\mathbf{X}$ and $\mathbf{V}$ is a diagonal matrix of the regularization parameters. The best regularization $\mathbf{V}$ satisfies the equations

$$v_k\left(a_k^2 + s^2 t_k\right) = s^2, \quad k = 1, \ldots, m \tag{12}$$

where s is the standard deviation of residuals and t_k is the kth diagonal element of matrix $\mathbf{T}$.

We called this method 'self-consistent regression' (SCR) because the same data samples (**X** and **y**) are used to estimate both the regression coefficients and the regularization parameters. Unlike the stepwise regression and other methods of combinatorial search, the initial SCR model includes all the regressors. Nevertheless, the final model may contain a few variables only, correctly representing the existing relationship.

3. Evaluation sets

The proposed method was validated using ten data sets and it was compared with several well-known QSAR methods: CoMFA, CoMSIA, HQSAR, EVA, GRID/GOLPE and others. The data sets were collected in such a way that they significantly varied in molecular flexibility, size, and structural heterogeneity for estimation of the proposed method. The types of biological endpoints and ranges in the activity measures also varied across the ten training sets. These sets represent the following types of biological activities: ligand–enzyme interactions (cyclin-dependent kinases 2 (CDK2) inhibitors [24], dihydrofolate reductase (DHFR) inhibitors and angiotensin-converting enzyme (ACE) inhibitors [25]), ligand–receptor interactions (alpha-2 adrenoreceptor ligands [26], estrogen receptor ligands [27]), acute toxicity (*Vibrio fischeri* [28], *Chlorella vulgaris* [29] and *Tetrahymena pyriformis* [30]) and interaction with drug-metabolism enzymes (CYP2A5 inhibitors and CYP2A6 inhibitors [31]). Brief descriptions of the sets are represented below.

3.1 *CDK2 inhibitors*

A training set of 29 and a test set of seven CDK2 inhibitors ($CDK2_{train}$ and $CDK2_{test}$, respectively) extracted from the literature [24] were used. These compounds are bisarylmaleimide derivatives. CDK2 is an enzyme belonging to the family of serine/threoinine kinases that play a key role in the regulation of the complex processes of cell division, apoptosis, transcription, and differentiation. CDK inhibitors are known to have a wide spectrum of applications ranging from protozoan infections (malaria, leishmania, trypanosomiasis), viral infections (human cytomegalovirus (HCMV), herpes simplex virus (HSV), human immunodeficiency virus (HIV), human papillomavirus (HPV)), reproduction disorders, cardiovascular diseases (atheroscelorosis, restenosis, cardiac hypertrophy), glomerulonephritis, cancers and nervous system diseases (Alzheimer's disease, stroke, amyotrophic disease, drug abuse) [32]. The experimental data are represented by IC_{50} values (50% inhibitory concentration) in mol/L (M), which are presented as $pIC_{50} = -\log IC_{50}$, and they varied from 5.057 to 8.194.

3.2 *DHFR inhibitors*

A training set of 237 and a test set of 124 DHFR inhibitors [IC_{50}, M] belonging to 11 structural classes were taken from Jeffrey et al. [25] ($DHFR_{train}$ and $DHFR_{test}$, respectively). DDHFR plays an important role in the biosynthesis of nucleic acids. Inhibition of the enzyme leads to the damage of DNA synthesis and cell death. The pIC_{50} values of these sets for rat liver enzymes ranged from 3.3 to 9.8.

3.3 *ACE inhibitors*

A training set of 76 and a test set of 38 ACE inhibitors [IC_{50}, M] were used (ACE_{train} and ACE_{test}, respectively). These compounds, taken from Jeffrey et al. [25], are carboxylic acid derivatives, including different heterocyclic groups (pyrrolidine, indole, imidazole, etc.). ACE involves in the action of the renin–angiotensin–aldosterone system. ACE inhibitors are effective antihypertensive agents. The pIC_{50} values of these sets ranged from 2.1 to 9.9.

3.4 *Alpha-2 adrenoreceptor ligands*

A training set $ADREN_{train}$ consisting of 30 structures with binding affinities to alpha-2 adrenoreceptor [Ki, μM] was taken from Lopez-Rodriguez et al. [26]. Adrenoreceptors are involved in the regulation of metabolism, secretion, contraction of muscles, and arterial pressure. Alpha-adrenoceptor agonists are effective analgesics and anxiolytics, and they have sedative and antihypertensive effects. The log (Ki) values of this set varied from 3.33 to 6.66.

3.5 *Estrogen receptor ligands*

A training set $ESTR_{train}$ consisting of 21 tetrahydroisoquinoline derivatives with binding affinities [IC_{50}, μM] to estrogenic receptor-β was obtained from Mukherjee et al. [27]. Estrogenic ligands are endocrine regulators of the male and female reproductive system. They also play a protective role in the tissues of bone, liver, and the cardiovascular system. Estrogen receptor ligands are used for breast cancer and osteoporosis treatment. The pIC_{50} values of this set varied from −0.567 to 0.983.

3.6 *Vibrio fischeri*

We used the training set $VIBRIO_{train}$ that consisted of 56 phenylsulfonyl carboxylates with acute toxicity to *Vibrio fischeri*. *Vibrio fischeri* is a marine bacterium which is used as a test system for assessment of acute aquatic toxicity of chemicals. Toxicity values were presented as log EC_{50} [15 min – EC_{50}, μM] [28] with the log EC_{50} values ranging from −0.44 to 2.28.

3.7 *Chlorella vulgaris*

A training set $ALGAE_{train}$ on acute toxicity to *Chlorella vulgaris* of 65 aromatic compounds [log ($1/EC_{50}$), mM] was taken from Netzeva et al. [29] with the log ($1/EC_{50}$) values ranging from −1.46 to 3.10. The set included phenols, anilines, benzaldehydes, and nitrobenzenes, as well as alkyl-substituted phenols, halogenated phenols and anilines, nitro-substituted phenols and anilines, and halogenated nitrobenzenes. *Chlorella vulgaris* is an alga which is used as a test system for the assessment of aquatic toxicity of chemicals.

3.8 *Tetrahymena pyriformis*

The training and test sets ($TETRA_{train}$ and $TETRA_{test}$, respectively) on acute toxicity to *Tetrahymena pyriformis* of 200 and 50 phenols [log ($1/IGC_{50}$), mM], respectively, were

taken from Cronin et al. [30] with the log ($1/IGC_{50}$) values ranging from −1.50 to 2.71. The compounds varied in structure from phenol itself, its relatively inert alkyl and halogen derivatives, through to reactive multisubstituted phenols. *Tetrahymena pyriformis* is a ciliate protozoan which is used as a test system for the assessment of aquatic toxicity of chemicals. Toxicity values were obtained in the population growth impairment assay on the ubiquitous freshwater ciliate *Tetrahymena pyriformis* (strain GL-C).

3.9 *CYP2A5 inhibitors*

A training set of 23 and a test set of five diverse competitive inhibitors of CYP2A5 [IC_{50}, M] ($CYP2A5_{train}$ and $CYP2A5_{test}$, respectively) were taken from Poso et al. [31]. Cytochrome P450 (CYPs) form a large superfamily of heme enzymes which catalyse the oxygenation of many endogenous and exogenous compounds. CYP2A5 inhibitors result in toxic effects and side effects in mice. The pIC_{50} values of these sets ranged from 1.73 to 5.68.

3.10 *CYP2A6 inhibitors*

A training set of 23 and a test set of five diverse competitive inhibitors of CYP2A6 [IC_{50}, M] ($CYP2A6_{train}$ and $CYP2A6_{test}$, respectively) were taken from Poso et al. [31]. CYP2A6 inhibitors may lead to toxic effects and side effects in humans. The pIC_{50} values of these sets ranged from 0.46 to 4.52.

4. Evaluation of predictive accuracy

The QSAR models were built by the program GUSAR for all the training sets. The accuracy of prediction was calculated by the leave-one-out cross-validation procedure. Moreover, the models were evaluated by prediction of the appropriate test sets (CDK2, DHFR, ACE, *Tetrahymena pyriformis*, CYP2A5, and CYP2A6). Some original authors do not provide any test sets for certain data sets (*Vibrio fischeri*, *Chlorella vulgaris*, alpha-2 adrenoreceptor, estrogenic receptor). However, it has been shown that to obtain a predictive QSAR it is necessary to use an external evaluation set [33]. Therefore, we decided to leave a part of the initial data as an external test set, to be used to estimate the performance of the model [34]. Random selection of compounds was performed by splitting the initial data into the external test and the training sets in the proportion 20% and 80%, respectively. Data splitting was repeated 20 times to obtain an objective assessment of the predictive accuracy and the robustness of the developed method for all data sets. For each splitting the training set was used to build the model and the external set for the assessment of model predictivity.

4.1 *Y-randomization test*

Several data sets used in this study contained less than 25 compounds (estrogenic receptor, CYP2A5, and CYP2A6). Therefore, we used a *Y*-randomization technique to ensure that the developed method did not suffer from overfitting. In this test the dependent-variable vector, the *Y*-vector, is randomly shuffled and a new QSAR model is developed using the

original independent-variable matrix [35]. The process was repeated 20 times. It is expected that the resulting models should generally have low Q^2 values.

5. Results and discussion

All the QSAR models were obtained using the GUSAR program and compared with the other above-mentioned methods.

The regression equation obtained by GUSAR contained the following values:

- T_{uv}, the Chebyshev polynomial in Equation (9) (see Equations (6)–(8));
- L, the topological molecule length;
- V, the sum of each atom's volume.

The quality of the models was estimated by the following parameters:

- n is the number of compounds in the training set;
- R^2 is the square of the regression coefficient;
- Q^2 is the cross-validated R^2;
- F is the value of the Fisher statistics;
- SD is the standard deviation;
- D is the number of variables in the final regression equation.

Cross-validated Q^2 values are typically smaller than the usual R^2 values, and the Q^2 values are considered to be more indicative of the predictive ability of a model. Therefore, the value of Q^2 is more important than the value of R^2. Whereas R^2 is a measure of

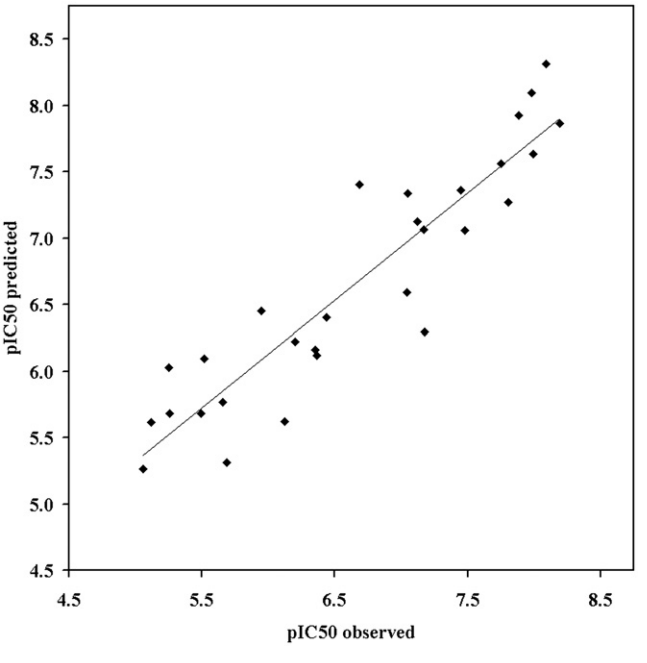


Figure 4. $CDK2_{train}$ data set, GUSAR predicted versus observed values.

goodness of fit, Q^2 is a measure of prediction accuracy. It should also be noted that an accurate QSAR model should have close Q^2 and R^2 values.

5.1 *CDK2 inhibitors*

The QSAR equation obtained by GUSAR was as follows:

$$\mathrm{pIC}_{50} = 12.8T_{3,1} + 8.5T_{2,2} + 3.07T_{2,0} + 5.42T_{1,3} - 3.14T_{1,2} + 7.95,$$

$$n = 29,\ R^2 = 0.84,\ F = 24.776,\ SD = 0.486,\ Q^2 = 0.77,\ D = 5.$$

This equation contains five variables represented by Chebyshev polynomials T_{uv}. This means that the CDK2 inhibition is well described by the QNA descriptors only. Figure 4 presents a plot of the predicted versus observed pIC_{50} values of the CDK2 inhibitors.

Comparison of the prediction accuracy of GUSAR with that of CoMFA [24] applied to the same data is presented in Table 1. The best CoMFA results were obtained with steric and electrostatic fields using the CoMFA_STD scaling option. The CoMFA model showed a high R^2 value (0.94) which was better than that of the GUSAR model, but the Q^2 (0.56) value of the CoMFA model was less than the Q^2 value of the GUSAR (0.77) model. The CoMFA model R^2 value for the test set was lower compared with the GUSAR model R^2 value: 0.86 and 0.87, respectively. Thus, GUSAR showed better accuracy of prediction compared to the CoMFA method on the $\mathrm{CDK2}_{\mathrm{train}}$ and $\mathrm{CDK2}_{\mathrm{test}}$ sets.

5.2 *DHFR inhibitors*

The QSAR equation obtained by GUSAR was as follows:

$$\begin{aligned}\mathrm{pIC}_{50} = {} & 4.87T_{0,6} - 5.45T_{1,8} + 0.000312L^3 + 4.29T_{4,9} - 2.16T_{0,13} - 3.81T_{6,7} + 2.7T_{6,0} \\ & + 3.1T_{5,6} - 3.28T_{11,3} + 3.91T_{4,7} - 3.49T_{2,8} - 2.53T_{0,5} - 3.05T_{0,2} + 2.28T_{4,0} \\ & - 1.38T_{12,0} - 4.22T_{2,5} - 2.29T_{7,6} - 1.84T_{5,7} - 2.17T_{3,0} - 4.08T_{7,1} - 1.4T_{0,11} \\ & - 2.17T_{6,4} + 2.85T_{2,3} + 1.86T_{9,4} - 1.87T_{7,4} + 0.96T_{3,8} + 0.942T_{1,10} - 1.34T_{5,1} + 6.1,\end{aligned}$$

$$n = 237,\ R^2 = 0.78,\ F = 26.825,\ SD = 0.663,\ Q^2 = 0.73,\ D = 28.$$

This equation contains 28 variables: 27 Chebyshev polynomials T_{uv} and third power of molecular length L. The $\mathrm{DHFR}_{\mathrm{train}}$ contains 237 compounds and many of the variables show the complex relationship between structure and activity. Figure 5 presents a plot of the predicted versus observed pIC_{50} values of the DHFR inhibitors.

Table 1 presents a comparison the prediction accuracy of GUSAR with the other QSAR methods applied to the same data. The $\mathrm{DHFR}_{\mathrm{train}}$ set was used for QSAR modelling by PLS regression on the basis of different descriptors calculated by the other authors [25] using CoMFA, CoMSIAbasic, CoMSIAextra, HQSAR, EVA, 2D and 3D descriptors of Cerius2. 2D descriptors were generated by the 'Combichem' defaults in Cerius2 (e.g., χ indices, counts of rotatable bonds, molecular weight, etc.) and also included E-state indices (both sums of indices and counts for each atom type). In addition, whole-molecule 3D descriptors, such as molecular volume and charged partial surface

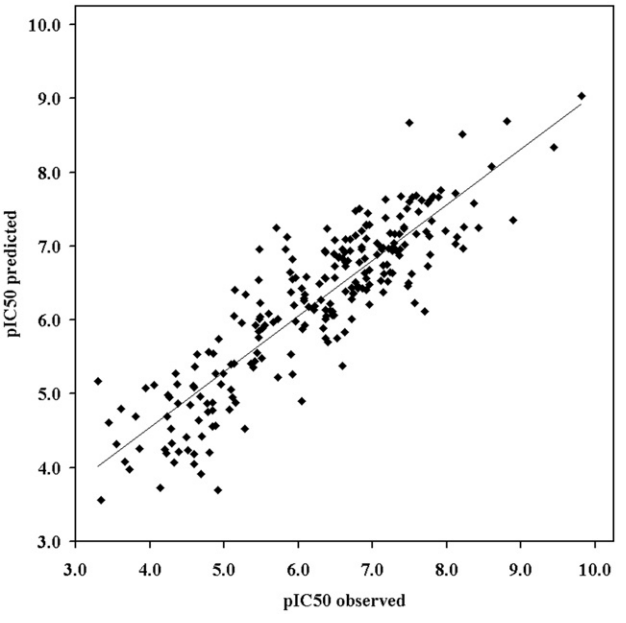


Figure 5. $DHFR_{train}$ data set, GUSAR predicted versus observed values.

area (CPSA) descriptors, were calculated using Gasteiger–Marsili charges implemented in Cerius2 (the Polygraph set) and the CORINA structures generated from SMILES strings. The CoMFA, EVA, and HQSAR methods have R^2 values slightly better (0.79, 0.81, and 0.81, respectively) than GUSAR (0.78), but the Q^2 values are worse (0.65, 0.64, and 0.69) than in GUSAR (0.72). The CoMSIAbasic, CoMSIAextra, 2D and 3D descriptors have poor statistical parameters for both R^2 (0.76, 0.75, 0.61, and 0.65, respectively) and for Q^2 (0.63, 0.65, 0.51, and 0.53, respectively). Table 1 shows that GUSAR provides accurate prediction for the test set, which is maximum dissimilar to the training set: $DHFR_{test} - R^2 = 0.60$. For the heterogeneous sets ($DHFR_{train}$ and $DHFR_{test}$) the GUSAR predictive accuracy was comparable to the accuracy of CoMSIAbasic, CoMSIAextra, EVA, and CoMFA. The accuracy was less than that of HQSAR and better than accuracy of 2D and 3D Cerius2 descriptors.

5.3 *ACE inhibitors*

The QSAR equation obtained by GUSAR was as follows:

$$\begin{aligned} \mathrm{pIC}_{50} &= 0.139V + 5.54T_{4,3} - 4.98E\text{-}06V^3 + 3.43T_{1,0} + 3.95T_{3,1} \\ &\quad - 1.95T_{0,6} + 2.58T_{1,5} + 1.82T_{5,0} - 1.15T_{4,0} - 0.545T_{1,4} + 0.00695, \\ n &= 76,\ R^2 = 0.83,\ F = 32.032,\ SD = 1.083, \\ Q^2 &= 0.78,\ D = 10. \end{aligned}$$

This equation contains ten variables: eight Chebyshev polynomials T_{uv} and the first and third powers of the molecular volume V. Figure 6 presents a plot of the predicted versus observed pIC_{50} values of the ACE inhibitors.

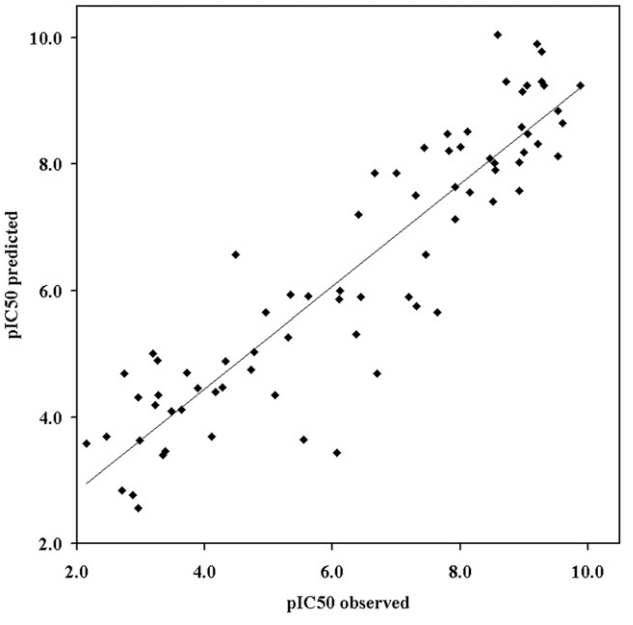


Figure 6. ACE_{train} data set, GUSAR predicted versus observed values.

Table 1 shows that ACE_{train} was used for QSAR modelling by PLS analysis on the basis of different descriptors calculated by other authors [25] using CoMFA, CoMSIAbasic, CoMSIAextra, HQSAR, EVA, 2D and 3D descriptors of Cerius2. The GUSAR model has the best statistical parameters of the correlation compared to other methods (Table 1). The difference between the value of R^2 of the GUSAR model and that of the other methods is considerable for one case: CoMSIAextra. The difference between the value of Q^2 of the GUSAR model and the other methods is considerable in four cases: CoMFA, CoMSIAbasic, CoMSIAextra, and 2D descriptors of Cerius2, where such a difference exceeds 0.1. Table 1 shows that for the test set, which is maximum dissimilar to the training set, GUSAR has reasonable predictive accuracy. Thus, the GUSAR model had comparable or better prediction accuracy than the other methods.

5.4 *Alpha-2 adrenoreceptor ligands*

The QSAR equation obtained by GUSAR was as follows:

$$\log(\mathrm{Ki}) = 11.9T_{1,2} - 6.7T_{1,3} + 3.06T_{5,0} + 2.48T_{2,1} + 0.821T_{4,0} - 0.969T_{3,1} + 5.31,$$

$$n = 30,\ R^2 = 0.82,\ F = 17.924,\ SD = 0.471,$$

$$Q^2 = 0.71,\ D = 6.$$

This equation contains six variables represented by Chebyshev polynomials T_{uv}. Figure 7 presents a plot of the predicted versus observed log (Ki) values of the $ADREN_{train}$ data set.

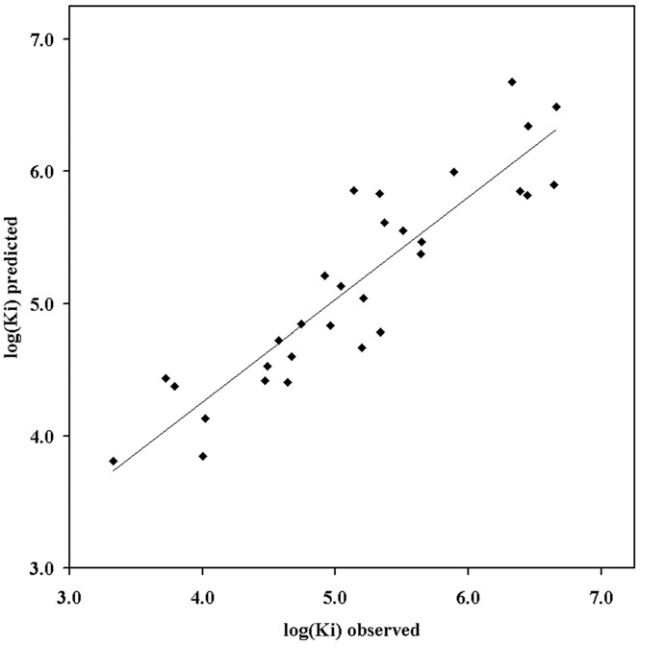


Figure 7. $ADREN_{train}$ data set, GUSAR predicted versus observed values.

The accuracy of GUSAR and CoMFA for the same set of compounds is presented in Table 1. The CoMFA analysis was performed with a C_{sp3} probe atom [26]. The following results were obtained by CoMFA: $R^2 = 0.92$ and $Q^2 = 0.69$. The results obtained by GUSAR are comparable with those of CoMFA. The value of R^2 of GUSAR is less than that of CoMFA, but at the same time the value of Q^2 of GUSAR is higher than that of CoMFA.

5.5 *Estrogen receptors ligands*

The QSAR equation obtained by GUSAR was as follows:

$$\mathrm{pIC}_{50} = -0.000289L^3 - 7.14T_{2,1} + 3.15T_{5,0} - 3.78T_{3,2} + 2.5T_{4,1} + 0.5,$$

$$n = 21,\ R^2 = 0.93,\ F = 42.800,$$

$$SD = 0.155,\ Q^2 = 0.89,\ D = 5.$$

This equation contains five variables: four Chebyshev polynomials T_{uv} and third power of the molecular length L. Figure 8 presents a plot of the predicted versus observed pIC_{50} values of the $ESTR_{train}$ data set.

The results obtained by GUSAR were compared with the 2D QSAR model based on E-states descriptors and MLR [27]. Table 1 shows that the value of $R^2 = 0.82$ for 2D QSAR is less than that of GUSAR; at the same time the value of Q^2 of 2D QSAR (0.77) is less than that of GUSAR. The difference between R^2 and Q^2 of the GUSAR model

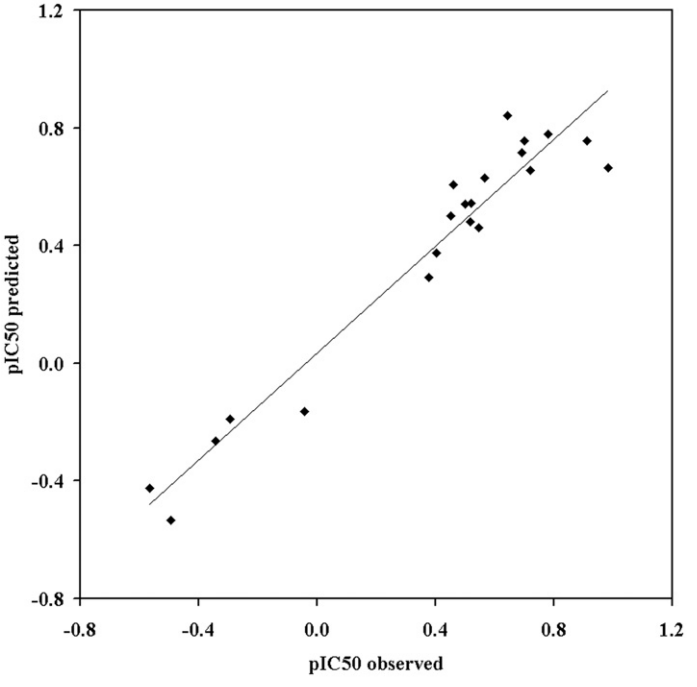


Figure 8. $ESTR_{train}$ data set, GUSAR predicted versus observed values.

and models obtained with 2D QSAR methods is significant. The results obtained by GUSAR are better than those obtained by another 2D QSAR model.

5.6 *Vibrio fischeri*

The QSAR equation obtained by GUSAR was as follows:

$$\log \mathrm{EC}_{50} = -0.0406V + 2.49T_{1,5} + 1.34T_{4,2} + 1.42T_{1,2} - 1.78T_{2,4} - 0.693T_{4,0} + 0.573T_{3,4} + 2.94,$$

$$n = 56,\ R^2 = 0.88,\ F = 48.185,\ SD = 0.186,$$

$$Q^2 = 0.84,\ D = 7.$$

This equation contains seven variables: six Chebyshev polynomials T_{uv} and the molecular volume V. Figure 9 presents a plot of the predicted versus observed log (EC_{50}) values of the $VIBRIO_{train}$ data set.

The studied set of compounds was analysed by other authors (Table 1) using CoMFA [36], MLR analysis, principal component regression (PCR) analysis, and the genetic function approximation (GFA) on the basis of extended topochemical atom (ETA) indices and non-ETA (physicochemical) parameters [28,37]. Non-ETA parameters included topological indices such as the Wiener, Hosoya Z, molecular connectivity, kappa shape, Balaban J, and E-State parameters, as well as physicochemical parameters such as the AlogP98, MolRef, and H-bond-acceptor. Factor scores were used as independent

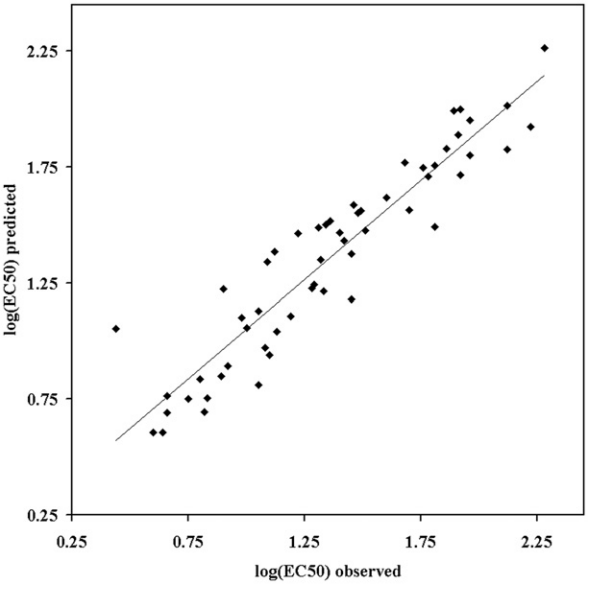


Figure 9. $\text{VIBRIO}_{\text{train}}$ data set, GUSAR predicted versus observed values.

variables for the backward stepwise regression method (Factor score, PCR, MLR) on the basis of a combination of ETA indices and non-ETA descriptors. Table 1 shows that the R^2 and Q^2 values of the GUSAR model are close to the best values obtained by the backward stepwise regression method (Factor score, PCR, MLR) on the basis of a combination of ETA indices and non-ETA descriptors. The difference between the R^2 and Q^2 values obtained by these methods is not considerable. The results obtained by GUSAR are similar to those obtained by our previous SCR-qQNA method [15].

5.7 *Chlorella vulgaris*

We obtained a reasonable correlation between the observed and predicted values of acute toxicity. The following regression equation was obtained by GUSAR:

$$\begin{aligned}\log(1/\text{EC50}) = {} & 0.183V + 3.01T_{3,1} - 2.55T_{3,0} - 2.79T_{1,0} - 2.28T_{1,4} \\ & + 1.38T_{2,2} + 0.787T_{7,0} + 1.98T_{0,1} + 1.03T_{6,1} \\ & - 1.92E\text{-}05V^3 - 0.399T_{0,6} + 0.668T_{1,5} - 4.93, \\ n = 65,\ R^2 = 0.93,\ F = 60.131,\ SD = 0.336, \\ Q^2 = 0.89,\ D = 12.\end{aligned}$$

This equation contains 12 variables: 10 Chebyshev polynomials T_{uv} and the first and third powers of the molecular volume V. Figure 10 presents a plot of the predicted versus observed log $(1/\text{EC}_{50})$ values of the $\text{ALGAE}_{\text{train}}$ data set.

The set of compounds tested on *Chlorella vulgaris* toxicity was used for QSAR modelling by MLR and PLS analysis on the basis of 102 molecular descriptors calculated by ClogP, MOPAC93, TSAR 3.3 (Oxford Molecular Limited, Oxford, England) and QSARis version 1.1 software (SciVision – Academic Press, San Diego, CA). MLR was

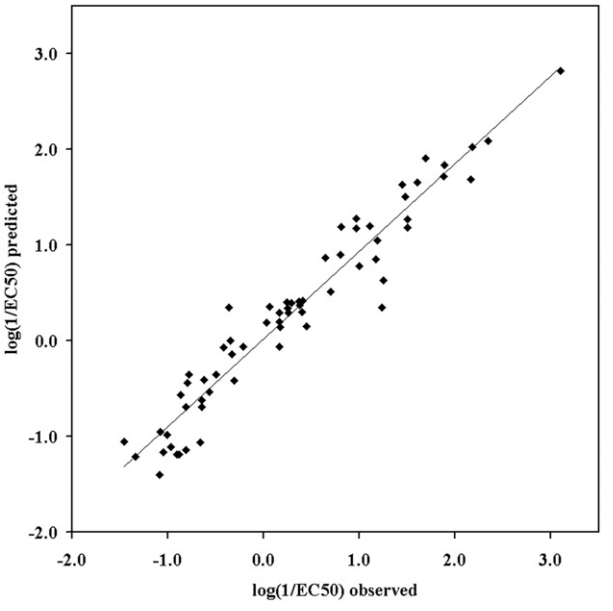


Figure 10. $ALGAE_{train}$ data set, GUSAR predicted versus observed values.

carried out by MINITAB version 13.1 (Minitab Inc., State College, PA) and PLS analysis was carried out in SIMCA-P version 9.0 (Umetrics AB, Umeå, Sweden) [29]. The statistical characteristics of GUSAR are better than those achieved by MLR and PLS analysis. The difference between the values of R^2 and Q^2 of the GUSAR model and those of other models is not significant in all cases (Table 1). The results obtained by GUSAR are comparable with those obtained by our previous SCR-qQNA method [15].

5.8 *Tetrahymena pyriformis*

The QSAR equation obtained by GUSAR was as follows:

$$\begin{aligned}\log(1/\mathrm{IGC50}) = {} & 0.0975V - 2.61T_{1,6} - 1.67T_{4,5} + 3.14T_{0,1} + 1.48T_{0,3} - 1.17T_{2,9} \\ & + 1.08T_{5,0} + 1.16T_{2,3} - 1.05T_{7,1} + 0.766T_{5,7} - 0.603T_{2,5} - 0.819T_{4,1} \\ & - 0.655T_{2,0} + 1.08T_{3,1} - 0.719T_{5,1} + 0.653T_{4,8} + 0.512T_{8,0} \\ & + 0.551T_{3,9} + 0.489T_{8,3} - 0.557T_{1,2} - 0.192T_{0,12} - 2.44, \\ n = 200,\ R^2 = 0.80,\ F = 33.427, & \\ SD = 0.413,\ Q^2 = 0.75,\ D = 21. & \end{aligned}$$

This equation contains 21 variables: 20 Chebyshev polynomials T_{uv} and the molecular volume V. Figure 11 presents a plot of the predicted versus observed log $(1/\mathrm{IGC}_{50})$ values of the TETRA_{train} data set.

Table 1 shows a comparison of the GUSAR accuracy with those achieved by the other QSAR methods. The set of compounds tested on *Tetrahymena pyriformis* toxicity was used for QSAR modelling by different QSAR approaches. At first this set was studied [30] by MLR, PLS analysis, and stepwise regression (SWR1) on the basis of 108 physicochemical descriptors calculated by ACD/Labs software, Chem-X version 2000.1, MOPAC version

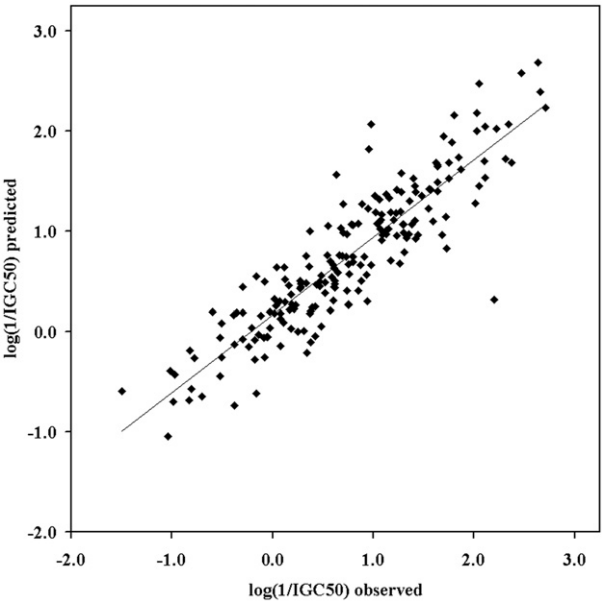


Figure 11. $TETRA_{train}$ data set, GUSAR predicted versus observed values.

6.49, TSAR 3.3, and QSARis version 1.1 software. MLR and stepwise regression were carried out by MINITAB version 13.1, and PLS analysis was carried out in TSAR 3.3 [30]. Second the same data set was also analysed by a genetic algorithm (GA) on the basis of different descriptor types calculated by TSAR V3.3, HYBOT v2.1.0.706, Dragon Professional v5.3, and ACD Labs V9.08. The GA was calculated by MOBYDIGS software and used to develop a large number of models [38]. The authors used the best model and top 10 models for consensus prediction. GUSAR could also create the different models and make the consensus prediction, but the aim of this investigation was the comparison of separate models only. Thus, in this paper we used statistical characteristics for the best GA model. In another article the authors used neural network and stepwise regression (SWR2) based on 168 descriptors obtained from ACD/Labs software, Chem-X version 2000.1, MOPAC version 6.49, TSAR 3.3, and QSARis version 1.1 software for the analysis of the same data set [39]. The statistical characteristics of these QSAR methods are represented in Table 1. This table shows that our results are better than those achieved by the MLR, SWR1, and PLS models on the heterogeneous set $TETRA_{train}$. The results obtained by the GA single models, SWR2, and neural network are slightly better than the GUSAR results. The results obtained by GUSAR are better than those obtained by our previous SCR-qQNA method [15].

5.9 *CYP2A5 inhibitors*

The QSAR equation obtained by GUSAR was as follows:

$$\mathrm{pIC}_{50} = 0.461V - 0.000149V^3 - 4.37,$$
$$n = 23,\ R^2 = 0.90,\ F = 94.484,$$
$$SD = 0.452,\ Q^2 = 0.88,\ D = 2.$$

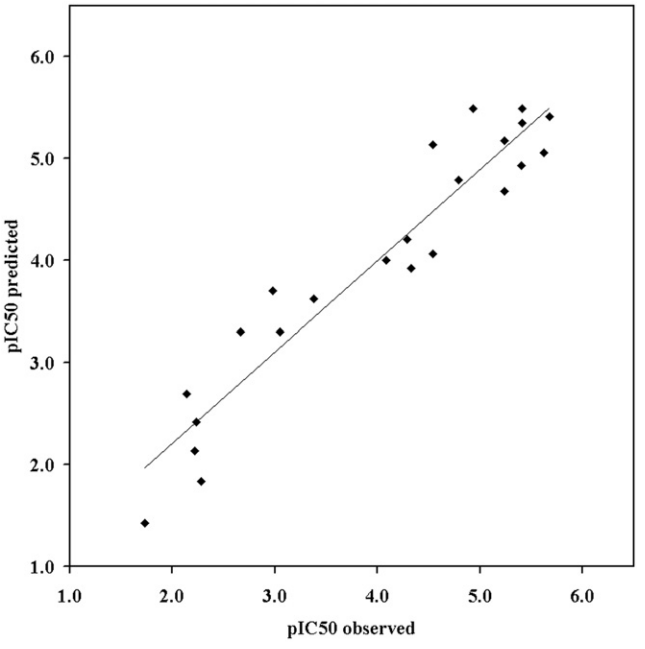


Figure 12. $CYP2A5_{train}$ dataset, GUSAR predicted versus observed values.

This equation contains two variables: first and third powers of molecular volume V only; it does not contain the Chebyshev polynomial of QNA descriptors. Figure 12 presents a plot of the predicted versus observed pIC_{50} values of the $CYP2A5_{train}$ data set.

A comparison of GUSAR accuracy with that of the 3D QSAR methods [31] (CoMFA and GRID/GOLPE) used for the same set of compounds is presented in Table 1. The differences between the R^2 and Q^2 values of CoMFA, GRID/GOLPE, and GUSAR were not considerable. All models had Q^2 values over 0.7. R^2 values for the test set ($CYP2A5_{test}$) by the CoMFA and GRID/GOLPE models were 0.83 and 0.90, respectively, whereas for GUSAR it was 0.93. The results obtained by GUSAR are close to those of the 3D-QSAR analysis.

5.10 *CYP2A6 inhibitors*

The QSAR equation obtained by GUSAR was as follows:

$$pIC_{50} = 0.252V - 0.0028L^3 - 1.94T_{0,2} - 0.426T_{3,0} - 0.992T_{2,2} + 0.135T_{2,3} - 1.71,$$

$$n = 23,\ R^2 = 0.90,\ F = 25.278,$$

$$SD = 0.439,\ Q^2 = 0.84,\ D = 6.$$

This equation contains six variables: four Chebyshev polynomials T_{uv}, first power of molecular volume V and third power of molecular length L. This model, in contrary to the model obtained for the $CYP2A5_{train}$ data set, does not contain descriptors of the topological length and the volume of the molecule. Thus, these descriptors are not

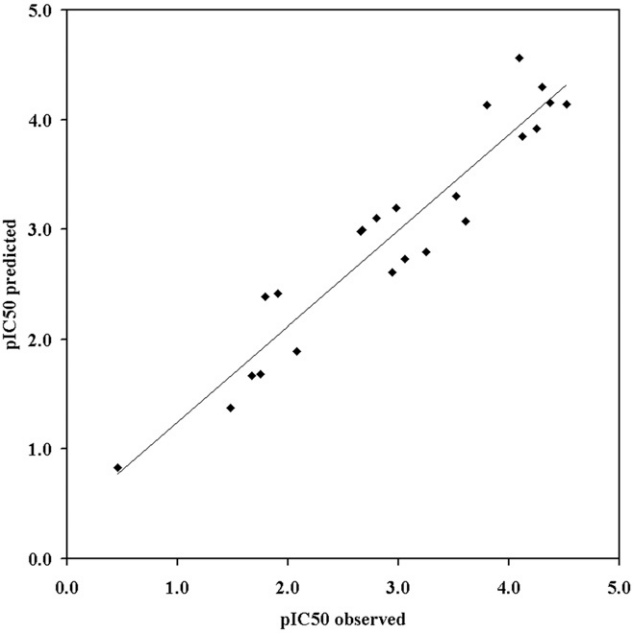


Figure 13. $CYP2A6_{train}$ data set, GUSAR predicted versus observed values.

important for modelling of the $CYP2A6_{train}$ data set. Figure 13 presents a plot of the predicted versus observed pIC_{50} values of the $CYP2A6_{train}$ data set.

A comparison of GUSAR accuracy with that of 3D QSAR methods [31] (CoMFA and GRID/GOLPE) used for the same set of compounds is presented in Table 1. The R^2 value of GUSAR is less than the R^2 values of CoMFA and GRID/GOLPE. The differences between the Q^2 values of CoMFA, GRID/GOLPE, and GUSAR were not considerable. All models had Q^2 values over 0.7. The R^2 values for the test set ($CYP2A6_{test}$) obtained by the CoMFA and GRID/GOLPE models were 0.77 and 0.76, respectively, whereas for GUSAR it was 0.93. The results obtained by GUSAR are better than those of the 3D-QSAR analysis (for the Q^2 value and the R^2 of the test set value).

5.11 *Statistical validation of comparison*

The analysis of the above-mentioned models showed that GUSAR has comparable or better accuracy of prediction for the studied sets. We compared R^2_{test} and Q^2 values of GUSAR models with other models. There are 38 Q^2 values and 25 R^2_{test} values. If GUSAR does not differ from other methods and all these methods are taken from the same parent entity, then we can expect that GUSAR will have the best Q^2 (R^2_{test}) value in half of the cases and the worst value in half of the cases: that is the zero hypothesis. Thus, the probability of obtaining m or higher success numbers in n comparisons at the zero hypothesis condition is given by:

$$\Pr(m|n) = 2^{-n} \sum_{k=m}^{n} \frac{n!}{k!(n-k)!}.$$

For Q^2 there are 36 successes in 38 comparisons and $\Pr(36|38) = 0.00000000971$. For R^2_{test} we have $\Pr(21|25) = 0.000455$. Thus both probabilities are less than 0.0005. Therefore, these estimations show that the accuracy of GUSAR certainly exceeded the accuracy of the other 2D and 3D QSAR methods used for comparison.

It is interesting to note that for R^2 there are 25 successes in 39 comparisons and $\Pr(25|39) = 0.054$. This fact can be explained by considering that GUSAR is less overfitted in comparison with other methods.

5.12 *Y-randomization test*

We used a *Y*-randomization technique as described above to ensure that the developed method did not have overfitting. The results of this test are shown in Table 2. The average Q^2 value for all ten data sets is 0.08. The maximum value $Q^2 = 0.28$ was obtained for the $\text{ADREN}_{\text{train}}$ data set. These results demonstrate that the developed method is free of any overfitting effects.

5.13 *Random multiple splitting procedure*

At present many authors usually apply only one test set for evaluation of the predictive ability of a QSAR method. We believe that one test set is not enough for the assessment of predictivity for the obtained QSAR model and for evaluation of the QSAR method in general. One of the most common approaches for model validation in the QSAR area is the leave-one-out cross-validation procedure (LOO CV). Vapnik [40] proved the theorem of unbiasedness and the justifiability of the LOO CV criterion, but this is so if only the correct LOO CV procedure is used. 'Correct' means the requirement to recalculate the number of components for the PLS analysis or reselect the number of descriptors after elimination of each compound for the GA stepwise regression, etc. Some authors do not do this [41,42]. In the current research we decided to perform an alternative validation procedure which allows us to assess not only the predictive ability of the method but also its robustness. For this purpose a multiple splitting procedure was performed, because such a validation obligates the recalculation of the model using the same approach for each split. Six training sets ($\text{CDK2}_{\text{train}}$, $\text{DHFR}_{\text{train}}$, $\text{ACE}_{\text{train}}$, $\text{TETRA}_{\text{train}}$, $\text{CYP2A5}_{\text{train}}$, and $\text{CYP2A6}_{\text{train}}$) were combined with the appropriate test sets ($\text{CDK2}_{\text{test}}$, $\text{DHFR}_{\text{test}}$, ACE_{test}, $\text{TETRA}_{\text{test}}$, $\text{CYP2A5}_{\text{test}}$, and $\text{CYP2A6}_{\text{test}}$). After that, the random

Table 2. *Y*-randomization procedure.

Set	*Average* Q^2	*Min* Q^2	*Max* Q^2
CDK2 inhibitors	0.07	−0.07	0.21
DHFR inhibitors	0.06	0.02	0.12
ACE inhibitors	0.06	−0.03	0.19
Vibrio fischeri	0.07	−0.04	0.20
Chlorella vulgaris	0.07	−0.02	0.17
Tetrahymena	0.07	0.01	0.11
Alpha-2 adrenoreceptor	0.12	−0.07	0.28
Estrogen	0.06	−0.10	0.26
CYP2A5	0.09	−0.08	0.26
CYP2A6	0.08	−0.11	0.26
Average value	0.08	−0.05	0.21

Table 3. Assessment of the prediction accuracy of GUSAR by a random multiple splitting procedure.

Sets	*Average* R^2_{pred}	*Min* R^2_{pred}	*Max* R^2_{pred}	$Q^2_{whole\ set}$
CDK2 inhibitors	0.81	0.60	0.97	0.86
DHFR inhibitors	0.65	0.52	0.79	0.71
ACE inhibitors	0.70	0.59	0.83	0.75
Vibrio fischeri	0.83	0.67	0.94	0.84
Chlorella vulgaris	0.81	0.58	0.95	0.89
Tetrahymena	0.66	0.53	0.80	0.75
Alpha-2 adrenoreceptor	0.73	0.51	0.89	0.71
Estrogen	0.86	0.56	0.99	0.89
CYP2A5	0.94	0.83	0.99	0.89
CYP2A6	0.83	0.54	0.96	0.87
Average value	0.78	0.59	0.91	0.82

multiple splitting procedure described above was performed for each of the ten training sets. Table 3 shows the average, maximum, and minimum R^2_{pred} values obtained in the prediction for the test sets. These values are compared to the $Q^2_{whole\ set}$ values of the models obtained before multiple splitting for the ten training sets. The average R^2_{pred} values for three data sets exceeded 0.6, for two sets exceeded 0.7, and for the remaining five data sets exceeded 0.8. Thus, these results showed high predictivity of GUSAR. In addition, the difference between the average value of R^2_{pred} (0.78) and the average value of $Q^2_{whole\ set}$ (0.82) was only 0.04, which indicates the high robustness of the developed method.

5.14 *Approximation of length and volume by QNA*

As stated above, QNA descriptors are local by construction and, for this reason, do not provide information on the shape and volume of a molecule. Therefore, in GUSAR the topological length and the volume of a molecule were added to the QNA descriptors. The results of the study show the validity of this addition. The volume descriptors are included in six of the ten final models, and the topological length descriptors are included in three final models. The addition of these descriptors lead to an increase of the model accuracy.

However, fundamentally, are the topological length and the volume of a molecule essentially different descriptors or might they also be approximated by QNA descriptors? To investigate this interesting question we have built the QSAR models for the topological length and the volume of a molecule based on QNA descriptors only using all unique molecules from the data sets used in this study. The obtained models have the following parameters:

$$\text{topological length}: n = 989,\ R^2 = 0.950,\ SD = 1.070,\ Q^2 = 0.931,\ D = 142,$$

$$\text{volume of a molecule}: n = 989,\ R^2 = 0.947,\ SD = 4.708,\ Q^2 = 0.925,\ D = 151.$$

Therefore, the use of these additional descriptors is insignificant for QSAR modelling; it may be convenient for QSAR model building only, especially in the case of a small data set size such as that for $CYP2A5_{train}$ and $CYP2A6_{train}$ with $n = 23$.

6. Conclusions

We have proposed a new QNA-based 'Star Track' QSAR approach, in which any molecule is represented as a set of points in 2D space of QNA descriptors. Our approach significantly differs from other known methods. In contrast to the classic QSAR methods it does not require the selection of the best set of descriptors among numerous descriptors used in QSAR.

The new 'Star Track' QSAR approach is realized in a computer program GUSAR, which is based on self-consistent regression, QNA descriptors, and the topological length and volume of a molecule. GUSAR predicts the quantitative values of biological activity of chemical compounds on the basis of their structural formulae and does not require the use of information about the 3D structure of ligands and/or target proteins. We compared GUSAR with different 3D and 2D QSAR methods using ten sets of molecules from different chemical classes having diverse kinds of biological activity. It was shown that the predictivity of GUSAR was comparable or better than that of the other QSAR methods both on heterogeneous (DHFR inhibitors, ACE inhibitors, *Tetrahymena pyriformis*) and on homogeneous (the other) data sets. The method does not use the selection of the model by the values of Q^2. In addition, GUSAR showed high prediction ability and robustness on the basis of a random multiple splitting procedure. Thus, GUSAR can be easily applied to different routine QSAR tasks, for building many models, and for prediction by these models of the different quantitative values simultaneously.

Acknowledgments

This work was supported in part by European Commission FP6 grant LSHB-CT-2007-037590 'Net2Drug', FP7 grant 200787 (OpenTox) and ISTC grant 3777.

References

[1] J. Gasteiger (ed.), *Handbooks of Cheminformatics*: *From Data to Knowledge*, Wiley-VCH, Weinheim, 2003.

[2] W. Zheng and A. Tropsha, *Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle*, J. Chem. Inf. Comput. Sci. 40 (2000), pp. 185–194.

[3] P. Itskowitz and A. Tropsha, *k nearest neighbors QSAR modeling as a variational problem: Theory and applications*, J. Chem. Inf. Model. 45 (2005), pp. 777–785.

[4] L. Peltason and Ju. Bajorath, *Molecular similarity analysis in virtual screening*, in *Chemoinformatics Approaches to Virtual Screening*, A. Varnek and A. Tropsha, eds., RSC Publishing, Cambridge, 2008, pp. 182–216.

[5] R. Todeschini and V. Consonni (eds.), *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.

[6] M. Olah, C. Bologa, and TI. Oprea, *An automated PLS search for biologically relevant QSAR descriptors*, J. Comput.-Aided Mol. Design 18 (2004), pp. 437–449.
[7] H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, and W. Tong, *Mold^2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics*, J. Chem. Inf. Model. 48 (2008), pp. 1337–1344.
[8] L. Senese, J. Duca, D. Pan, A. Hopfinger, and Y. Tseng, *4D-fingerprints, universal QSAR and QSPR descriptors*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 1526–1539.
[9] H. Sun, *A universal molecular descriptor system for prediction of LogP, LogS, LogBB, and absorption*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 748–757.
[10] F.R. Burden, M.J. Polley, and D.A. Winkler, *Toward novel universal descriptors: charge fingerprints*, J. Chem. Inf. Model. 49 (2009), pp. 710–715.
[11] D. Filimonov, V. Poroikov, Yu. Borodina, and T. Gloriozova, *Chemical similarity assessment through multilevel neighborhoods of atoms: Definition and comparison with the other descriptors*, J. Chem. Inf. Comput. Sci. 39 (1999), pp. 666–670.
[12] V. Poroikov, D. Filimonov, Yu. Borodina, A. Lagunin, and A. Kos, *Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds*, J. Chem. Inf. Comput. Sci. 40 (2000), pp. 1349–1355.
[13] D. Filimonov and V. Poroikov, *Probabilistic approaches in activity prediction*, in *Chemoinformatics Approaches to Virtual Screening*, A. Varnek and A. Tropsha, eds., RSC Publishing, Cambridge, 2008, pp. 182–216.
[14] D. Filimonov, A. Lagunin, and V. Poroikov, *Prediction of activity spectra for substances using new local integrative descriptors*, in *Proceedings of the 15th European Symposium on Structure-Activity Relationships* (*QSAR*) *and Molecular Modelling*, E. Aki and I. Yalcin, eds., CADD&D Society in Turkey, Ankara, 2005, pp. 98–99.
[15] A. Lagunin, A. Zakharov, D. Filimonov, and V. Poroikov, *A new approach to QSAR modelling of acute toxicity*, SAR QSAR Environ. Res. 18 (2007), pp. 285–298.
[16] D. Filimonov, D. Akimov, and V. Poroikov, *The method of self-consistent regression for the quantitative analysis of relationships between structure and properties of chemicals*, Pharm. Chem. J. 1 (2004), pp. 21–24.
[17] A. Bender, H.Y. Mussa, and R.C. Glen, *Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 170–178.
[18] L. Xing and R.C. Glen, *Novel methods for the prediction of logP, pKa, and logD*, J. Chem. Inf. Comput. Sci. 42 (2002), pp. 796–805.
[19] W. Guba, *Representation of chemicals*, in *Predictive Toxicology*, C. Helma, ed., Marcel Dekker, New York, 2003, pp. 11–35.
[20] A. Varnek, D. Fourches, F. Hoonakker, and V.P. Solov'ev, *Substructural fragments: An universal language to encode reactions, molecular and supramolecular structures*, J. Comp.-Aided Mol. Design 19 (2005), pp. 693–703.
[21] I. Baskin and A. Varnek, *Fragment descriptors in SAR/QSAR/QSPR studies, molecular similarity analysis and in virtual screening*, in *Chemoinformatics Approaches to Virtual Screening*, A. Varnek and A. Tropsha, eds., RSC Publishing, Cambridge, 2008, pp. 1–43.
[22] R. Feynman, *Forces in molecules*, Phys. Rev. 56 (1939), pp. 340–343.
[23] Yu. Borodina, D. Filimonov, and V. Poroikov, *Computer-aided estimation of synthetic compounds similarity with endogenous bioregulators*, Quant. Struct.-Act. Relat. 17 (1998), pp. 459–464.
[24] N. Dessalew and P. Bharatam, *3D-QSAR and molecular docking study on bisarylmaleimide series as glycogen synthase kinase 3, cyclin dependent kinase 2 and cyclin dependent kinase 4 inhibitors: An insight into the criteria for selectivity*, Eur. J. Med. Chem. 42 (2007), pp. 1014–1027.
[25] J. Jeffrey, A. Lee, and F. Donald, *A comparison of methods for modeling quantitative structure-activity relationships*, J. Med. Chem. 47 (2004), pp. 5541–5554.

[26] M. Lopez-Rodriguez, L. Rosado, B. Benhamu, J. Morcillo, E. Fernandez, and K.J. Schaper, *Synthesis and structure-activity relationships of a new model of arylpiperazines. 2. Three-dimensional quantitative structure-activity relationships of hydantoin-phenylpiperazine derivatives with affinity for 5-HT1A and α1 receptors. A comparison of CoMFA models*, J. Med. Chem. 40 (1997), pp. 1648–1656.

[27] S. Mukherjee, A. Sahaa, and K. Roy, *QSAR of estrogen receptor modulators: Exploring selectivity requirements for ERa versus ERb binding of tetrahydroisoquinoline derivatives using E-state and physicochemical parameters*, Bioorg. Med. Chem. Lett. 15 (2005), pp. 957–961.

[28] K. Roy and G. Ghosh, *QSRT with extended topochemical atom indices. 4. Modeling the acute toxicity of phenylsulfonyl carboxylates to* Vibrio fischeri *using principal component factor analysis and principal component regression analysis*, QSAR Comb. Sci. 23 (2004), pp. 526–535.

[29] T. Netzeva, J. Dearden, R. Edwards, A. Worgan, and M. Cronin, *QSAR analysis of the toxicity of aromatic compounds to* Chlorella vulgaris *in a novel short-term assay*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 258–265.

[30] M. Cronin, A. Aptula, J. Duffy, T. Netzeva, P. Rowe, I. Valkova, and T. Schultz, *Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to* Tetrahymena pyriformis, Chemosphere 49 (2002), pp. 1201–1221.

[31] A. Poso, J. Gynther, and R. Juvonen, *A comparative molecular field analysis of cytochrome P450 2A5 and 2A6 inhibitors*, J. Comput.-Aided Mol. Design 15 (2001), pp. 195–202.

[32] M. Knockaert, P. Greengard, and L. Meijer, *Pharmacological inhibitors of cyclin-dependent kinases*, Trends Pharmacol. Sci. 23 (2002), pp. 417–425.

[33] A. Golbraikh and A. Tropsha, *Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection*, J. Comput.-Aided Mol. Design 16 (2002), pp. 357–369.

[34] I.V. Tetko, Iu. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, and A. Varnek, *Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection*, J. Chem. Inf. Model. 48 (2008), pp. 1733–1746.

[35] A. Tropsha, P. Gramatica, and V. Gombar, *The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models*, QSAR Comb. Sci. 22 (2003), pp. 69–77.

[36] X. Liu, Z. Yang, and L. Wang, *CoMFA of the acute toxicity of phenylsulfonyl carboxylates to Vibrio fischeri*, SAR QSAR Environ. Res. 14 (2003), pp. 183–190.

[37] K. Roy and G. Ghosh, *QSTR with extended topochemical atom indices. Part 5: Modeling of the acute toxicity of phenylsulfonyl carboxylates to Vibrio fischeri using genetic function approximation*, Bioorg. Med. Chem. 13 (2005), pp. 1185–1194.

[38] M. Hewitt, M. Cronin, J. Madden, P. Rowe, C. Johnson, A. Obi, and S. Enoch, *Consensus QSAR models: Do the benefits outweigh the complexity?*, J. Chem. Inf. Model. 47 (2007), pp. 1460–1468.

[39] S. Enoch, M. Cronin, T. Schultz, and J. Madden, *An evaluation of global QSAR models for the prediction of the toxicity of phenols to* Tetrahymena pyriformis, Chemosphere 71 (2008), pp. 1225–1232.

[40] V. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, New York, 1982.

[41] H. Hong, H. Fang, Q. Xie, R. Perkins, D.M. Sheehan, and W. Tong, *Comparative molecular field analysis (CoMFA) model using a large diverse set of natural, synthetic and environmental chemicals for binding to the androgen receptor*, SAR QSAR Environ. Res. 14 (2003), pp. 373–388.

[42] A. Afantitis, G. Melagraki, H. Sarimveis, P. Koutentis, J. Markopoulos, and O. Igglessi-Markopoulou, *A novel QSAR model for evaluating and predicting the inhibition activity of Dipeptidyl Aspartyl Fluoromethylketones*, QSAR Comb. Sci. 10 (2006), pp. 928–935.

Appendix 1

Table A1. The electron affinity (*EA*) and first ionization potential (*IP*), in electronvolts, and the atomic radius (*AR*), in angstroms, used in this work.

Atom	*EA*	*IP*	*AR*
H	0.75	13.60	0.46
He	0.08	24.59	1.22
Li	0.62	5.39	1.55
Be	−0.20	9.32	1.13
B	0.28	8.30	0.91
C	1.26	11.26	0.77
N	0.44	14.53	0.71
O	1.46	13.62	0.73
F	3.45	17.42	0.71
Ne	0.00	21.57	1.60
Na	0.55	5.14	1.87
Mg	−0.31	7.64	1.60
Al	0.30	5.99	1.43
Si	1.39	8.15	1.34
P	0.75	10.49	1.30
S	2.00	10.36	1.04
Cl	3.61	12.97	0.99
Ar	−0.37	15.76	1.92
K	0.50	4.34	2.36
Ca	−0.19	6.11	1.97
Sc	0.19	6.56	1.64
Ti	0.33	6.82	1.46
V	0.53	6.74	1.34
Cr	0.67	6.77	1.27
Mn	−0.17	7.43	1.30
Fe	0.50	7.90	1.26
Co	0.66	7.86	1.25
Ni	1.16	7.64	1.24
Cu	1.23	7.72	1.28
Zn	−0.44	9.39	1.39
Ga	0.30	6.00	1.39
Ge	1.39	7.90	1.39
As	0.80	9.79	1.48
Se	2.02	9.75	1.17
Br	3.45	11.81	1.14
Kr	−0.42	14.00	1.98
Rb	0.49	4.18	2.48
Sr	−0.15	5.69	2.15
Y	0.31	6.22	1.81
Zr	0.33	6.84	1.60
Nb	0.51	6.88	1.45
Mo	0.68	7.09	1.39
Tc	0.54	7.23	1.36
Ru	1.10	7.37	1.34
Rh	1.14	7.46	1.34
Pd	1.11	8.34	1.37
Ag	1.22	7.58	1.44

(*continued*)

Table A1. Continued.

Atom	*EA*	*IP*	*AR*
Cd	−0.43	8.99	1.56
In	0.31	5.79	1.66
Sn	1.39	7.34	1.58
Sb	0.90	8.64	1.61
Te	1.97	9.01	1.70
I	3.23	10.45	1.53
Xe	−0.25	12.13	2.18
Cs	0.47	3.89	2.66
Ba	−0.15	5.21	2.23
La	0.30	5.59	1.87
Ce	0.25	5.54	1.83
Pr	0.20	5.47	1.83
Nd	0.20	5.53	1.82
Pm	0.20	5.58	1.81
Sm	0.20	5.64	1.80
Eu	0.20	5.67	2.04
Gd	0.20	6.15	1.80
Tb	0.20	5.86	1.78
Dy	0.20	5.94	1.77
Ho	0.20	6.02	1.78
Er	0.20	6.11	1.76
Tm	0.20	6.18	1.75
Yb	0.20	6.25	1.94
Lu	0.20	6.15	1.75
Hf	0.33	7.50	1.59
Ta	0.40	7.89	1.46
W	0.67	7.98	1.40
Re	0.23	7.88	1.37
Os	1.44	8.73	1.35
Ir	1.57	9.10	1.35
Pt	1.10	8.96	1.38
Au	1.25	9.23	1.44
Hg	−0.19	10.44	1.57
Tl	0.31	6.11	1.71
Pb	1.39	7.42	1.75
Bi	0.97	7.29	1.82
Po	1.97	8.42	1.56
At	2.90	9.20	1.48
Rn	−0.15	10.75	2.27
Fr	0.48	3.98	2.80
Ra	−0.15	5.28	2.35
Ac	0.80	5.20	2.03
Th	0.80	6.10	1.80
Pa	0.84	6.00	1.62
U	0.82	6.19	1.53
Np	0.82	6.20	1.50
Pu	0.84	6.06	1.62
Am	0.85	6.00	1.70
Cm	0.85	6.09	1.55
Bk	0.82	6.23	1.49
Cf	0.84	6.27	1.42
Es	0.86	6.47	1.43

(*continued*)

Table A1. Continued.

Atom	*EA*	*IP*	*AR*
Fm	0.86	6.60	1.38
Md	0.83	6.68	1.38
No	0.79	6.58	1.47
Lr	0.85	6.69	1.30
Db	0.46	6.43	1.14
Jl	0.50	6.78	1.01

Algorithm of the QNA descriptor calculation (in the Delphi 5.0 language):

```
const
  MaxAtom = 1000; // Maximal size of molecule
  LastTerm = 31; // Last term of exponent series
type
  TAtomRecord = record
    Z   : Integer; // Atomic number
    P, Q: Real; // QNA descriptor values
  end;
  TBondRecord = record
    N1 : Integer; // Number of 1-st binding atom
    N2 : Integer; // Number of 2-nd binding atom
  end;
var
  NA : Integer; // Atoms count
  AtomsList : array[1..MaxAtom] of TAtomRecord;
  NB : Integer; // Bonds count
  BondsList : array[1..MaxAtom] of TBondRecord;
const
  CAtomPQ : array[1..105, 1..2] of Real =
{H}((0.2790, 2.0024),
{He}(0.2020, 2.4911),
{Li}(0.4578, 1.3751),
{Be}(0.3241, 1.4779),
{B}(0.3531, 1.5149),
{C}(0.3163, 1.9804),
{N}(0.2665, 1.9951),
{O}(0.2868, 2.1625),
{F}(0.2675, 2.7914),
     ...
{Jl}(0.3989, 1.4529));

procedure SetQNA;
var
    i, n : Integer;
    x : Real;
    SP, SQ, T : array[1..MaxAtom] of Real;
begin
  for i := 1 to NA do with AtomsList[i] do begin
    P := CAtomPQ[Z, 1]; SP[i]:= P;
    Q := CAtomPQ[Z, 2]; SQ[i]:= Q;
  end;
  for n := MaxStep downto 1 do begin
    x := 0.5/n;
```

```
    for i := 1 to NA do T[i]:= 0;
    for i := 1 to NB do with BondsList[i] do begin
      T[N1] := T[N1] + SP[N2];
      T[N2] := T[N2] + SP[N1];
  end;
  for i := 1 to NA do SP[i]:= AtomsList[i].P - x*T[i];
  for i := 1 to NA do T[i]:= 0;
  for i := 1 to NB do with BondsList[i] do begin
     T[N1] := T[N1] + SQ[N2];
     T[N2] := T[N2] + SQ[N1];
  end;
  for i := 1 to NA do SQ[i]:= AtomsList[i].Q - x*T[i];
  end;
  for i := 1 to NA do
  with AtomsList[i] do begin
    P := CAtomPQ[Z, 1]*SP[i];
    Q := CAtomPQ[Z, 1]*SQ[i];
  end;
end;
```