DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS
60

ROBUST QSAR METHODS FOR THE PREDICTION OF PROPERTIES FROM MOLECULAR STRUCTURE

DIMITAR ATANASOV DOBCHEV

TARTU UNIVERSITY
PRESS

Department of Chemistry, University of Tartu, Estonia

Dissertation is accepted for the commencement of the degree of Doctor of Phylosophy in Chemistry on 2 November, 2006, by the Doctoral Committee of the Department of Chemistry, University of Tartu.

Supervisor: Professor Mati Karelson

Opponents: Dr. Mihkel Kaljurand, Professor of Analytical Chemistry, Head of Department of Chemistry, Tallinn University of Technology, Tallinn, Estonia
Dr. Emilio Benfenati, Head Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche "Mario Negri" Milano, Italy

Commencement: 18 December, 2006 at 18 Ülikooli St., room 204, 15:00 h.

ISSN 1406–0299
ISBN 9949–11–487–X (trükis)
ISBN 9949–11–488–8 (PDF)

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr 598

CONTENTS

LIST OF ORIGINAL PUBLICATIONS 6

LIST OF ABBREVIATIONS 7

INTRODUCTION 8

1. LITERATURE OVERVIEW 10
   1.1. QSAR general methodology 10
   1.2. Qantum-chemical descriptors – methods applied and improvements needed 14
   1.3. (B)MLR approach, variable selection 15
   1.4. ANN approach, general scheme 17
   1.5. Substructural molecular fragment method (ISIDA) 19

2. APPLICATION OF ROBUST QSAR METHODS ON DIVERSE BIOLOGICAL AND PHYSICOCHEMICAL PROPERTIES 21
   2.1. Reparameterization of AM1 method and its application on QSPR-s of liquid properties. [Article I] 21
   2.2. Comparison of (B)MLR and ANN approaches on the example of different properties. [Articles II–V] 22
      2.2.1. QSAR of anti-platelet derived growth factor 22
      2.2.2. QSAR of glycogen synthase kinase 3 24
      2.2.3. QSAR of anti-cancer activity 25
      2.2.4. Neural network convergence in QSARs 26
   2.3. Applicability of the fragment approach (ISIDA) [Article VI] 27

3. CONCLUSIONS 29

REFERENCES 30

ACKNOWLEDGMENTS 32

SUMMARY IN ESTONIAN. Üldised QSAR meetodid omaduste ennustamiseks molekulaarstruktuurist 33

PUBLICATIONS 35

LIST OF ORIGINAL PUBLICATIONS

The present thesis consists of the six articles listed below. All papers are denoted in the text by roman numerals I–VI.

I. Dobchev, Dimitar A.; Karelson, Mati. **Reparameterized Austin Model 1 for quantitative structure-property relationships in liquid media**. *Journal of Molecular Modeling* (2006), 12(4), 503–512.

II. Katritzky, Alan R.; Dobchev, Dimitar A.; Fara, Dan C.; Karelson, Mati. **QSAR studies on 1-phenylbenzimidazoles as inhibitors of the platelet-derived growth factor**. *Bioorganic & Medicinal Chemistry* (2005), 13(24), 6598–6608.

III. Katritzky, Alan R.; Pacureanu, Liliana M.; Dobchev, Dimitar A.; Fara, Dan C.; Duchowicz, Pablo R.; Karelson, Mati. **QSAR modeling of the inhibition of Glycogen Synthase Kinase-3.** *Bioorganic & Medicinal Chemistry* (2006), 14(14), 4987–5002.

IV. Katritzky, Alan R.; Kuanar, Minati; Dobchev, Dimitar A.; Vanhoecke, Barbara W. A.; Karelson, Mati; Parmar, Virinder S.; Stevens, Christian V.; Bracke, Marc E. **QSAR modeling of anti-invasive activity of organic compounds using structural descriptors**. *Bioorganic & Medicinal Chemistry (*2006), 14(20), 6933–6939.

V. Karelson, Mati; Dobchev, Dimitar A.; Kulshyn, Oleksandr V.; Katritzky, Alan R. **Neural Networks Convergence Using Physicochemical Data.** *Journal of Chemical Information and Modeling* (2006), 46(5), 1891–1897.

VI. Katritzky, Alan R.; Dobchev, Dimitar A.; Fara, Dan C.; Huer, Evrim; Taemm, Kaido; Kurunczi, Ludovic; Karelson, Mati; Varnek, Alexandre; Solov'ev, Vitaly P**. Skin permeation rate as a function of chemical structure.** *Journal of Medicinal Chemistry (*2006), 49(11), 3305–3314.

Author's contribution

Publication I, II, V: The author is responsible for the data sets, calculations, programing, preparation of the manuscript and interpretation of the results.
Publication III, IV: The author is responsible for the calculations, preparation and writing of the manuscript and interpretation of the results.
Publication VI: The author is responsible for the calculations and interpretation of the results.

LIST OF ABBREVIATIONS

AM1	Austin Model 1
ANN	Artificial Neural Network
BMLR, (B)MLR	Best Multilinear Regression
BFGS	Broyden-Fletcher-Goldfarb-Shanno method
CODESSA	COmprehensive DEscriptors for Structural and Statistical Analysis
FIT	Kubinyi's statistical criterion
GA	Genetic Algorithm
ISIDA	In Silico Design and Data Analysis
LMO	Leave-Many-Out cross-validation
LOO	Leave-One-Out cross-validation
MLP	Multilayer Perceptron
MLR	Multilinear Regression
MNDO	Modified Neglect of Diatomic Overlap
MO	Molecular Orbital
MOPAC	Molecular Orbital PACkage
PCA	Principal Component Analysis
PLS	Partial Least Squares
PM3	Parametric Method Number 3
(S)PRESS	(Squared) Standard Deviation of Predictions
QSAR	Quantitative Structure – Activity Relationship(s)
QSPR	Quantitative Structure – Property Relationship(s)
RMSPE	Root-mean Squared Error of Prediction
RMS	Root-mean Squared Error
SCF	Self consistent Field Method
SMF	Substructural Molecular Fragments
SVD	Singular Value Decomposition

INTRODUCTION

All details of the chemical, biological and physical properties of a compound are encoded within the structural formula of that compound. This fundamental principle of chemistry leads to establishment of the basic concept of Quantitative Structure Activity/Property Relationships (QSAR/QSPR)[1–5] *i.e.* to relate the structure of a compound expressed in terms of descriptors which can be calculated directly from the structure (e.g. number of carbon atoms or molecular weight) to a property/activity of interest (e.g. boiling point or biological activity). If such a relationship (correlation) can be established, activity/property values can be predicted for other, even not yet existing, compounds with a certain degree of confidence.

QSAR/QSPR studies are unquestionably of great importance in modern chemistry, biochemistry and drug design. One of the main objectives of these studies is to transform searches for compounds with desired properties or activities using chemical intuition and experience into a mathematically quantified and computerized form. Once a correlation between structure and activity is found, any number of compounds can be screened on the computer in order to select structures with the required properties/activities. It is then possible to select the most promising compounds to synthesize and test in the laboratory. Thus, the QSAR approach conserves resources and accelerates the process of development of new molecules for use as drugs, materials or any other purpose.

The development of robust and predictive statistical models is one of the fundamental tasks for a quantitative structure-activity relationship modeler. The statistical and machine learning literature provides a wide variety of methods to choose from. These include techniques such as multilinear regression (MLR) models, principal component analysis (PCA), partial least squares techniques (PLS) as well as more complex nonlinear techniques such as artificial neural networks (ANN) along with genetic algorithms (GA). These methods are the basis of the modern 2D, 3D, 4D and fragment QSAR development. The modeling techniques differ in a number of ways such as complexity, flexibility, accuracy, and speed.

Generally, the multilinear regression approach assumes that the biological activity or property can be modeled as a linear function of several molecular descriptors. This approach is the oldest and most widely used in the QSAR/QSPR area. One of the robust modifications of this modeling algorithm is so called best multilinear regression BMLR. It combines MLR and variable (descriptor) selection among a large descriptor space. The results from BMLR equations are easy to interpret and handy to use. However, most of the biological activities and physicochemical properties naturally possess nonlinear relationships with the molecular descriptors and thus in some cases (B)MLR equations fail to obtain robust predictive QSAR models. In this case, one of the

solutions for this problem is to use nonlinear algorithms as ANN. The ANN models are capable of finding strong nonlinearities between the property under investigation and the descriptors. Usually the neural network models consist of larger number of adjustable parameters (weights) compared with the MLR models. These parameters carry the nonlinear information between the property and the input variables. One of the biggest advantages of the ANN modeling procedure is that the obtained QSAR models are always externally validated *i.e.* the optimization (training) of the ANN parameters takes instantly into account compound properties that were not included in the QSAR modeling. However, in contrast with the MLR equations, the ANN models are harder to interpret from the physicochemical point of view because of their complex mathematical nature.

Another modeling procedure in the QSAR area is based on substructural molecular fragments (SMF) techniques. This approach represents the molecule as molecular graph with certain topology. It allows splitting of the molecular graphs into fragments and it is able to calculate their contributions to the property under investigation. The fragment approach combines linear (MLR) and nonlinear relations and can "handle" most of the practical studies and give direct insights for important molecular structural features.

A requirement for building robust predictive QSAR/QSPR models is to have reliable data objects *i.e.* molecular descriptors and experimental properties. Therefore, to obtain significant correlation, it is crucial that appropriate descriptors be employed, whether they are theoretical, empirical or derived from readily available experimental characteristics of the structures. Recent progress in computational hardware and development of efficient algorithms has assisted the routine development of molecular quantum mechanical calculations. Semi-empirical methods as MNDO, AM1, PM3 supply realistic quantum-chemical molecular quantities in a relatively short computational time. They are thus an attractive source of molecular descriptors, which can, in principle, express all of the electronic and geometric properties of molecules and their interactions. However, these semiempirical methods are based on limited experimental parameters and thus in some practical cases they obtain unsatisfactory results. Therefore, the need of improvement of these semiempirical optimizations is mandatory.

This Thesis presents a compilation of our work on the prediction of several biological activities and properties of diverse sets of organic compounds by using BMLR, ANN and SMF approaches. Also, we improved the AM1 parameterization method for liquid media, a requirement for robust QSAR model. In Chapter 1, an overview of the QSAR modeling algorithm is described as well as the different robust methods used. Chapter 2 presents our own results of the QSAR modeling tasks that are of potential use in various domains of chemistry, medicinal, pharmaceutical area.

1. LITERATURE OVERVIEW

1.1. QSAR general methodology

QSAR attempts to correlate structural molecular features (descriptors) with physicochemical properties, biological activities, toxicities, etc. for a set of compounds by means of statistical methods. As a result, a mathematical relationship is established:

$$A = f(\text{structural molecular or fragment descriptors})$$

where A is the above mentioned activity or property. Thus, based on numerical representation of the chemical structure, the QSAR/QSPR aims i) to understand how structural variation affects the biological activity or physicochemical property of a set of compounds and ii) to build a predictive equation which could be used to predict and describe (novel) compounds.

The QSAR/QSPR paradigm is based on the assumption that there is an underlying relation between molecular structure and biological activity/ property. In other words, the factors governing the events in a biological or physicochemical system are represented by descriptors characterizing the compounds, whose biological activity/property is expressed via the same mechanism[6].

The QSPR model development and validation involves several major steps, as presented in Figure 1:

A) Data selection

The selection of appropriate chemical sets to develop QSAR models is of great importance to obtain valid results.[7] A suitable set should consist of those chemicals that exert a given activity effect via a common mechanism that can be modeled by single QSAR equation. A large amount of experimental values lead to more valid QSAR models. Two are the general requirements for the data selection:

1) Experimental property data
   Before to start the QSAR modeling, it is necessary to analyze the experimental data for the applicability of statistical approach. Biological data can be distributed on a continuous scale or it can be classified on a discrete scale. High quality and reliable biological data are required. Biological activities/properties should be measured in a consistent manner, by using well standardized assays with a clear and unambiguous endpoint.[8] The data source should come from the same protocol and, if possible, the same laboratory.

2) Compound data set

The selection of the set of compounds must span the chemical domain of interest, according to the definition of the chemical space.[9] Therefore, a wide range of noncongeneric series lead to larger applicability domains of the QSAR models. However, the practice shows that the QSAR models for congeneric data sets are more statistically robust.

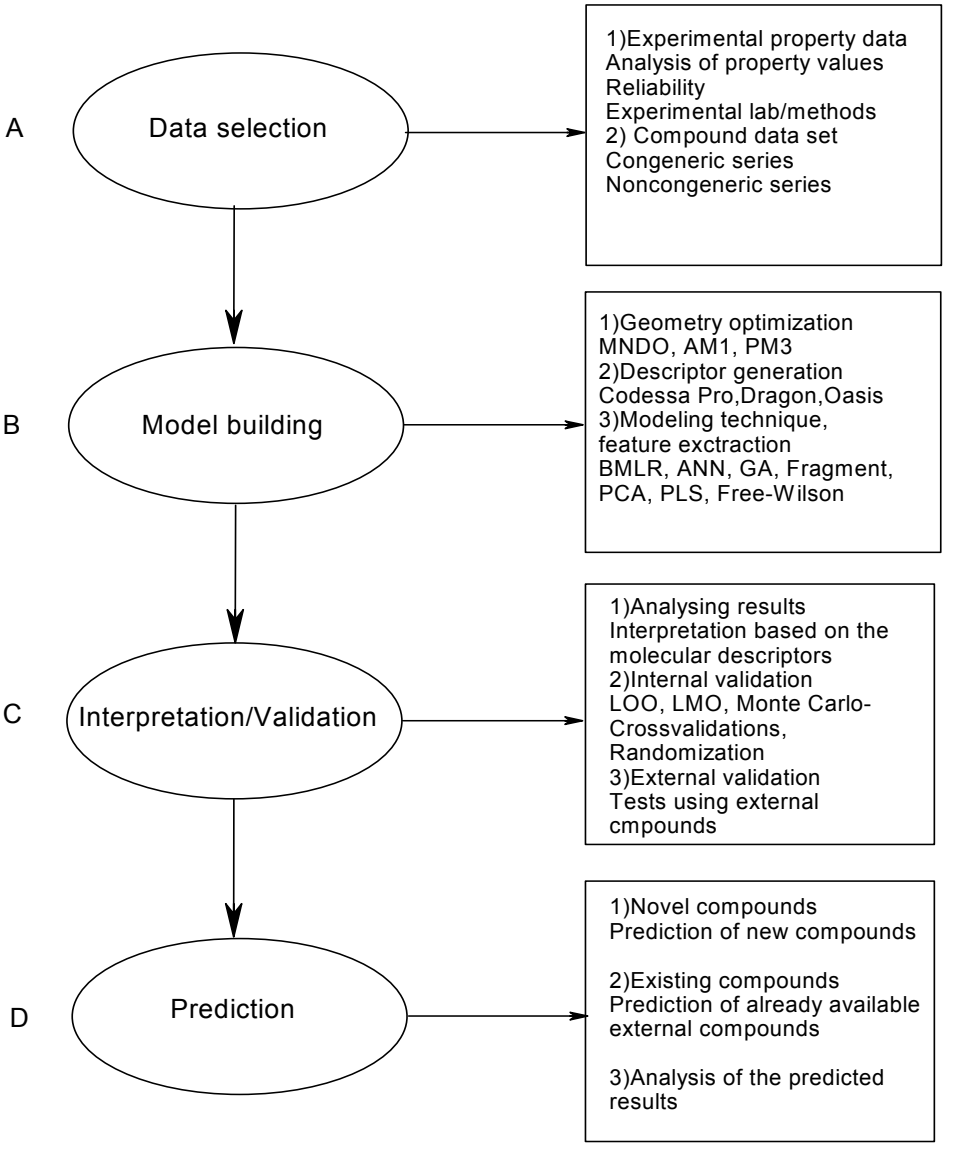

Figure 1. Scheme of a QSAR model generation

B) Model building

This stage is one of the most important steps in the general QSAR modeling. It includes:

1) Optimization of molecular structures
   Once the compounds from step A are selected they need to be subjected to geometry optimization of their structures. The use of the correct 3D molecular structures is vitally important to predict correctly the molecular properties or the biological activity. Therefore, the geometry of the molecules needs to be optimized to obtain the correct shape and conformation of the molecule. A variety of molecular modeling programs are available that employ different molecular mechanics algorithms, *ab initio* and semiempirical quantum mechanical calculations as Hyperchem,[10] MOPAC,[11] AMPAC [12] etc. that utilize methods as MM+,[10] OPLS,[10] AM1,[13] PM3,[14] MNDO.[15]
2) Descriptor generations
   The generation of the molecular features is based on previous presented computational methods (*ab initio*, semi-empirical) and involves the calculation of molecular structural descriptors in order to encode the compounds. Those include the constitutional, topological, geometrical, electronic, and quantum chemical classes of descriptors. The constitutional descriptors are fragment additive and reflect mostly the general properties of the compound reflected in their structures. The topological descriptors are calculated using the mathematical graph theory applied to the scheme of atom connections of the structure. The geometrical, electronic and quantum-chemical descriptors are usually derived from the results of empirical schemes or molecular orbital calculations and they encode the molecule's ability to participate in polar interactions or hydrogen bonding (donor, acceptor). Available software programs that can calculate these descriptors are CODESSA PRO,[16] Oasis,[17] Dragon.[18]
3) Modeling technique, feature extraction
   Feature extraction of descriptors involves finding the most informative subsets of descriptors from those in the descriptor pool due to the fact that models with too many variables lead to bad predictions because of overfitting. Various robust methods, like the classical forward selection, backward elimination and stepwise regression,[19] or more recently, the genetic algorithms (GA) can be used for the selection of the best combination of descriptors. The genetic algorithms have been shown to be very effective in performing descriptor selection in the case of large descriptor pools.[19, 20]
   Model building involves the application of the descriptors using (i) multilinear regression analysis, such as best multilinear regression (BMLR) methods and the Heuristic[21], (ii) multivariate statistical methods, such as principal component analysis (PCA) and partial least squares

(PLS)[22,23], or (iii) non-linear methods, such as artificial neural networks[24,25] (ANNs) and iv) additivity fragment approaches (Free-Wilson)[26] to build mathematical models linking the descriptors directly to the chemical property under investigation.

C) Interpretation/Validation

1) The analysis of the obtained models involves (i) the explanation of how each of the descriptors from the selected "best " subset contributes to the property of interest, (ii) the assessment of physicochemical meaning to descriptors, especially for those provided by multivariate statistical analysis.
   The validation of the models developed is an important aspect of any QSAR study. Once a model is obtained, it is important to determine its reliability and statistical significance. Several procedures are available to assist in this. These can be used to check that the size of the model is appropriate for the data available, as well as to provide some estimate of how well the model can predict activity for new molecules. The most widely used procedures are:
2) Internal validation – uses the data set from which the model is derived and checks for the internal consistency. Variations of this technique are Leave-One-Out (LOO), and Leave-Many-Out (LMO).[27] The idea behind the internal validation is to predict the property value for a compound from the data set, which is in turn predicted from the regression equation calculated from the data for all other compounds or group of compounds. For evaluation, predicted values can be used for *PRESS*, *RMSPE*, and squared crossvalidated correlation coefficient (R^2_{cv}). In addition, randomization tests are also used with conjunction to the LMO or LOO. These tests consist of repeated elaboration and random shuffling of the data for which the model equations are tested. Due to a factorial increase in time of permutations, Monte-Carlo method is often used for producing randomization test.[28]
3) External validation
   One of the most widely spread method of the correlation testing is the use of an external validation set. An external validation evaluates how well the equation generalizes the data. In this method, the correlation is employed to predict a property/activity value for a chemical structure that was not used in the creation of the QSAR model; some test statistics are calculated for the external dataset; the difference between the test statistics in the training and validation datasets is a measure of the reliability of the correlation.

D) Prediction

This stage of the QSAR modeling involves the use of successfuly built models to estimate the activity/property of interest for:

1) unknown compounds or newly synthesized compounds
2) already available compounds not involved in the QSAR modeling.
3) The results from 1) and 2) should be analyzed and assessed by the validation criteria. If badly predicted activity/property values (outliers) are observed the reasons have to be explored more thouroughly.

The general QSAR/QSPR methodology can be applied, in principle, to any property/activity that can be related to the molecular structure so that careful performance of the scheme could lead to good molecular modeling by using robust methods.

1.2. Quantum-chemical descriptors – methods applied and improvements needed

Quantum-chemical methods and molecular modeling techniques enable the definition of a large number of molecular and local quantities characterizing the reactivity, shape and binding properties of a complete molecule as well as of molecular fragments and substituents. Because of the large well-defined physical information content encoded in many theoretical descriptors, their use in the design of a training set in a QSAR study presents two main advantages: (a) the compounds and their various fragments and substituents can be directly characterized on the basis of their molecular structure only; and (b) the proposed mechanism of action can be directly accounted for in terms of the chemical reactivity of the compounds under study.[29] Consequently, the derived QSAR models will include information regarding the nature of the inter-molecular forces involved in determining the biological or other activity of the compounds in question.

A wide variety of *ab intio* methods beyond Hartree-Fock have been developed and coded to account for electron correlations in the molecule. These include configuration interaction (CI),[30] multiconfigurational self-consistent field (MC SCF),[31] correlated pair many-electron theory (CPMET)[32] and its various coupled-cluster approximations,[33] and perturbation theory (e.g. Møller-Plesset theory of various orders, MP2, MP3, MP4).[34] Most of these methods are extremely time consuming and require large CPU memories and are therefore impractical for the calculation of extended sets of relatively large molecules (i.e., more than 40 atoms).

As an alternative to *ab initio* methods, semiempirical quantum-chemical methods (MNDO, AM1, PM3 etc.) are therefore widely used for the calculation

of molecular descriptors. These methods have been developed within the mathematical framework of the molecular orbital theory (SCF MO), but based on simplifications and approximations introduced into the computational procedure, which dramatically reduce the computational time. Experimental data on atoms and prototype molecular systems have often been used to estimate values of quantities used in the calculations as parameters. For this reason, these procedures are widely known as semiempirical methods.[35] However, as mentioned above, these methods are partly based on limited data (atomic bond lengths, heat of atomization etc.) taken from the experiment.[36] Thus, the internal parameters encoded in these theoretical techniques are optimized[37] with respect to the experimental data which some times lead to undesired deviations of the calculated characteristics, in particular, quantum mechanical descriptors. Therefore, a constant improvement of these semiempirical methods is needed. In addition, most work employing quantum chemical descriptors has been carried out in the field of QSAR rather than QSPR, *i.e.* the descriptors have been correlated with biological activities such as enzyme inhibition activity, hallucinogenic activity, etc.[38] In part this has been because, historically, the search for quantitative relationships with chemical structure has been used early in the development of theoretical drug design methods. Thus, most of the processes in which certain biological activity is involved, take place in liquid media. Hence, the fact that these semiempirical parameterizations were developed by fitting data on properties of isolated molecules or those in the gas phase (not in liquid) could influence in a negative direction on the quantum-mechanical descriptors,[39] and consequently on the results of the QSAR methods.

1.3. (B)MLR approach, variable selection

Multilinear regression (MLR) method has been used throughout the classical QSAR work. The basic idea of this method is essentially the solution of a multilinear regression problem by means of least squares techique.[40] It can be expressed compactly and conveniently using matrix notation. Suppose that there are n property values in **Y** and n associated calculated values for each k molecular descriptor in **X** columns. Then Y_i, X_{ik}, and e_i can represent the ith value of the **Y** variable (property), the ith value of each of the **X** descriptors, and the ith unknown residual value, respectively. Thus, the MLR correlates the dependent variable **Y** by linear dependence of independent variable **X**:

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e} \tag{1}$$

where **b** is a column vector of coefficients and k is the number unknown regression coefficients for the descriptors. The goal of multiple regression is to minimize the sum of the squared residuals (2):

$$\min_{b} \|e\|^2 \tag{2}$$

Regression coefficients **b** in (1) that satisfy criterion (2) are found by solving the system of linear equations (3):

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3}$$

Several statistical characteristics give information about the "goodness " of the model: R^2 – squared correlation coefficient, R^2_{cv} – squared cross-validated correlation coefficient, F – Fisher criterion value, s^2 – squared standard error.[41]

Usually the QSAR study deals with a large number of molecular descriptors. The search of the best multiple linear regression model among such a large descriptor space is not a trivial task. Various regression techniques were proposed for the selection of the "best set " of regression predictors as backward elimination, forward selection, ridge regression. [23,42] In the current Thesis we have used another powerful algorithm called Best Multilinear Regression (BMLR)[21,41] which combines MLR and descriptor selection procedure. This method is encoded in CODESSA PRO software[16] and it is fully automatic. It is a modification of the so-called stepwise algorithms[42] and the main steps regarding equation building and variable selection are:

1) In a given descriptor space BMLR searches and selects all orthogonal descriptor pairs
2) In the created descriptor subspace in step 1), it builds all two-parameter regression equations and selects these descriptors which show highest correlation coefficient (R) in the respective two-parameter regressions
3) By using the best two-parameter equations in 2), it creates three-parameter regressions by adding additional noncollinear descriptor. Then Fisher criterion F and the crossvalidated coefficient (R_{cv}) are assessed and if there is no improvement over the best two-parameter regressions in 2), the procedure is halted. Otherwise, BMLR continues with the selection of the best three-descriptor regressions according to R.
4) The higher rank regression equations *e.g.* four-, five-, etc. descriptor regressions are recursively built based on step 3) together with simultaneous reduction of the number of descriptors in the initial descriptor pool and selection of the best multilinear regression equations with highest F, R_{cv} and R criteria. According to these statistical criteria, the final model is considered as the best representation of the activity/property in the given descriptor space.

The BMLR method is able to find the "best " regression for a short computational time in a descriptor space of hundreds of variables. The researcher can manage all the criteria so that certain chemical space can be explored more precisely for best correlations. However, the number of descriptors in the equation should not exceed certain limits because it could lead to overfitted

model.[23] This problem could be overcome by using simple techniques as investigation of the improvement of the R and R_{cv} with respect of the number of descriptors as well as chemical intuition.[43]

1.4. ANN approach, general scheme

Neural networks were originally designed as a model for the activity of the human brain. Recently, computational neural networks have been employed as nonlinear models for QSAR/QSPRs.[44–47] There are numerous types of ANN as multilayer perceptron (MPL), Kohenen, probabilistic, radial basis, entropy machines networks that differ by their ideology, topology, optimization algorithms etc. For a comprehensive review of the networks the reader can be referred to [48].

An artificial neural network (ANN) model can be thought of simply as a nonlinear regression model when applied to QSAR studies. However, since the model is nonlinear, the regression coefficients cannot be found in one step, and an iterative process must be used to determine the coefficients. The most widely used ANN type in the QSAR area is the MPL feed-forward neural network with backpropagation of the errors.[49]

A neural network architecture commonly used in a typical QSAR study is shown in Figure 2. The scheme depicts a fully connected, feed-forward, three-layer neural network. The ANN begins by performing a linear transformation on the input layer (of *p* descriptors) from its original range usually to the interval [0,1]. The transformed values are then passed to the hidden layer. The input value of a hidden layer neuron (circles) is the summation of the products of the weights and the corresponding outputs of the previous input layer plus a bias term β. The first neuron in the hidden layer has an input of $In_{(2,1)}$,. The output of the neuron $Out_{(2,1)}$, a sigmoidal transformation of the input, is also shown in Figure 2. The output in the last layer A_{out} represents a calculated estimate of the experimental biological activity A_{exp}. The weights (w) and biases (β) are adjusted iteratively to minimize the sum-squared error for prediction of the target values (biological activity):

$$\text{Error} = \frac{1}{2}\sum_i (A^i{}_{\text{exp}} - A^i{}_{out})^2 \qquad (4)$$

The ANN method is much more computationally intensive than linear regression since the nonlinear regression coefficients (weights and biases) must be changed iteratively, which requires repeated evaluation of the network outputs. However, the greater mathematical flexibility found in ANNs often leads to models that are superior to MLR models, so the additional time required is often worth the effort.

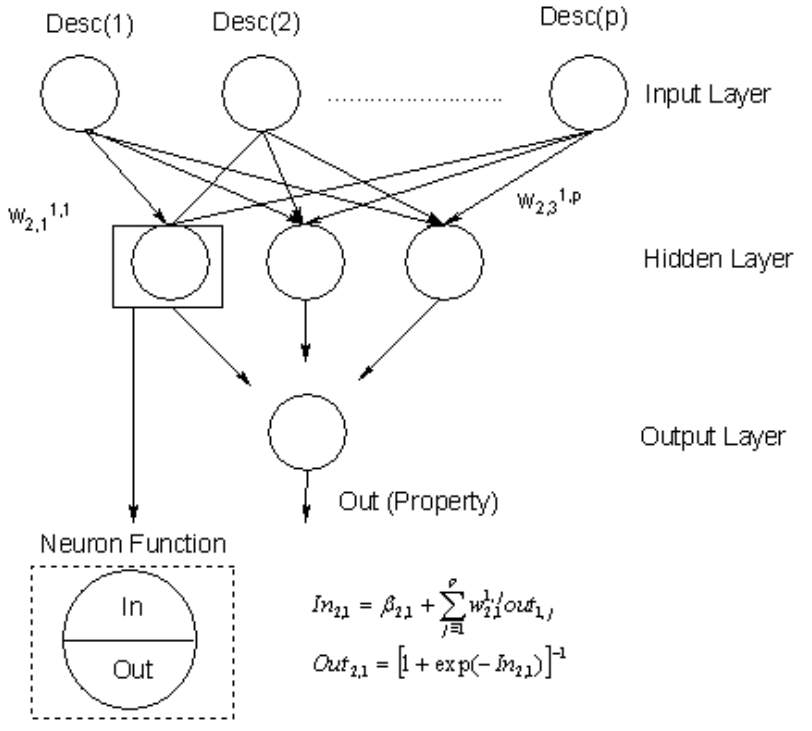

Figure 2. MLP feed-forward neural network

Adjusting the weights and biases to fit target values is known as training the network. Because of the increased mathematical flexibility of ANNs and the large number of adjustable parameters, it is possible to obtain apparently good fits by chance or to overtrain the neural network. To avoid this situation the data set is split into a training set and a validation set. The weights and biases are adjusted based on the RMS error of the training set members and the RMS error of the validation set is calculated periodically throughout the training. Overtraining is believed to occur when the RMS error of the validation set begins to rise. If training is stopped when the validation error is at a minimum, then the network may be used with reasonable confidence for future predictions.

Training the network is nothing more than an optimization problem. One of the simplest methods for optimization of the weights during the training is the delta rule. It propagates back the changes of (4) with respect to the weights on each iteration (epoch) so that the ANN uses supervised learning based on the experimental activity. When a large data set is used, or when the network has a large number of weights and biases, the task can become quite expensive computationally. It is worth then to utilize an efficient optimization method for training as BFGS, conjugate gradient, Levenberg-Marquardt algorithms.[50, 51]

Once the ANN is trained, it can be used in QSAR/QSPR predictions and analysis of novel compounds. The interpretation of the ANN results from the physicochemical point of view is hard because of the mathematical complexity of the ANN. However, there exist procedures based on the sensitivity and response surface analysis of the weights, which could elucidate the important descriptors of the net leading to desirable property values.[49]

Important aspects of the ANN QSAR modeling are the choice of the architecture and the selection of the input descriptors. The practice shows that ANN architectures with one or two hidden layers are enough for good modeling. In the modern state-of-art QSAR studies the input variable selections of the ANN is performed by genetic algorithms (GA)[52]. The GA represents the descriptors in a given pool as binary strings and based on the rules of mutation and crossover it is able to find significant set of descriptors which could be used as inputs in the ANN.

1.5. Substructural molecular fragment method (ISIDA)

Recently, a new robust method called substructural molecular fragment (SMF) has been developed to model the relationships between the structure of organic molecules and their thermodynamic parameters of complexation or extraction.[53] The method is based on the splitting of a molecule into fragments, and on calculations of their contributions to a given property. It can be considered as an extension of the Free-Wilson approach[26] applying molecular fragments as variables in a multiple regression analysis.

The success of the fragment approach in QSAR/QSPR studies[54–56] depends on the diversity of structural fragments as well as on the flexibility of atomic classification. In this sense, the SMF method represents a flexible structure-property tool because it generates a large number (e.g. 49) different types of fragments (atom/bond sequences and augmented atoms), then builds QSAR models involving linear and nonlinear fitting equations. The SMF is implemented in the software package In Silico Design and Data Analysis (ISIDA).[57]

The SMF method is based on the representation of a molecular graph as a superposition of fragments (subgraphs), and on the calculation of their contributions to a given property $\boldsymbol{Y}$. Two different classes of fragments are used: "sequences" and "augmented atoms". The sequences may contain atoms and bonds, atoms only, or bonds only. Only shortest paths from one atom to the other are used. For each type of sequences, the minimal (n_{min}) and maximal (n_{max}) number of constituted atoms is defined. In the current version of ISIDA, $n_{min} \geq 2$ and $n_{max} \leq 15$. An "augmented atom" represents a selected atom with its environment including either neighboring atoms and bonds, or atoms only, or bonds only. Atomic hybridization can be also taken into account.

Once a molecular graph is split into constitutive fragments, any corresponding quantitative physical or chemical property ***Y*** is calculated from the fragments contributions using linear (5) or non-linear (6) and (7) fitting equations.

$$\boldsymbol{Y} = a_o + \sum_i a_i N_i + \Gamma \quad (5)$$

$$\boldsymbol{Y} = a_o + \sum_i a_i N_i + \sum_i b_i (2N^2{}_i - 1) + \Gamma \quad (6)$$

$$\boldsymbol{Y} = a_o + \sum_i a_i N_i + \sum_{ki} b_{ik} N_i N_k + \Gamma \quad (7)$$

where, a_i and b_i (b_{ik}) are fragment contributions, N_i is the number of fragments of i type. The a_o term is fragment independent. The a_o, a_i and b_i (b_{ik}) are obtained by the multilinear regression procedure using the training set of compounds. An extra term $\Gamma = \Sigma c_m D_m$ can be used to describe any specific feature of the compound using external descriptors D_m (*e.g.,* topological, electronic, etc.); by default $\Gamma = 0$.

At the training stage, ISIDA can build up to 147 structure-property models involving 49 types of fragment descriptors and the 3 linear and non-linear fitting equations (5)–(7). If some fragments are linearly dependent, they are treated as one extended fragment. Using the singular value decomposition method (SVD)[23] the program fits the a_i and b_i terms in Eq. (5)–(7) and calculates the corresponding statistical characteristics (correlation coefficient (R), standard deviation (s), Fischer's criterion (F), cross-validation correlation coefficient (Q), standard deviation of predictions (sPRESS), Kubinyi's criterion (FIT), RH-factor of Hamilton and matrix of pair correlations (correlation matrix) for the terms a_i and b_i and performs statistical tests to select the best models.

The results and predictions of the SMF method can be analyzed on the basis of statistical criteria as well as on fragment descriptors. Thus, the QSAR models based on this approach can indicate important molecular structural characteristics, which could highlight important interactions behind the property in question. However, since the SMF is a modification of the Free – Wilson approach, it allows in some cases large number of orthogonal fragment descriptors to be involved. Hence, results are difficult to explain. Nevertheless, this is a straightforward method, which gives good insights for important structural features of the compounds.

2. APPLICATION OF ROBUST QSAR METHODS ON DIVERSE BIOLOGICAL AND PHYSICOCHEMICAL PROPERTIES

2.1. Reparameterization of AM1 method and its application on QSPR-s of liquid properties. [Article I]

Article I presents the results on an improvement of the parameterization of one of the most widely used semiempirical methods in the QSAR area, namely Austin Model 1.

The main object of this work was to use a nonlinear optimization technique toward the quantumchemical parameterization for the AM1 model so that an improvement of the QSPR models of molecular properties in the liquid phase can be obtained. The original AM1 parameterization had been developed by fitting the data on properties of isolated molecules or those in the gas phase (heats of formation, ionization energies, dipole moments etc.). However, most of the QSPR/QSAR models are built on the data on molecular properties in condensed media (liquids, solutions, and membranes). Thus, this new parameterization could lead to improvement of the QSAR/QSPR modeling.

In general, the task was an unconstrained optimization of a function defined by equation 8 in article I based on QSPR model for the boiling points previously developed in [21]. This model (Eq. 1 in article I) is based on only two molecular descriptors, namely the cubic root of the gravitation index $G_I^{1/3}$ and the charged surface area of hydrogen-donor atoms *HDSA(2)*. Moreover, from the physical point of view, both descriptors have obvious meaning; G_I is connected with the dispersion and cavity-formation effects in liquid, *HDSA(2)* reflects the hydrogen-bonding ability of compounds. In addition, it had an excellent correlation (R^2 = 0.954) between the predicted and the experimental normal boiling points of 298 diverse compounds. Hence, this good correlation backed up with the physicochemical meaning encouraged us to use equation 1 as a rule in the optimization problem.

The choice of optimization technique was related to the mathematical nature of the problem. In this work we chose the Levenberg-Marquardt optimization procedure, since Eq. 8 is a sum of squares related nonlinearly to the optimization parameters. Levenberg-Marquardt optimization is one of the best techniques to tackle with such type of problems. Also, it makes use of the Jacobian matrix of the derivatives with respect to the function (8) and the optimization parameters, which "gathers " more correct information.

The optimization parameters were α, β_s and β_p where α is core-core repulsion atomic parameter and β_s, β_p – one-electron resonance integral parameters (for s and p states, respectively). As can be noted from equations 5, 6 and 7 in article I, these parameters are related to the atomic distances and distances of

interactions occurring at Van der Waals radii. Thus, it makes them important for processes that take place in condensed media.

The number of optimization parameters was seventeen related to the atoms H, C, N, O, Cl and Br since these atoms are often used in the molecules. The derived new set of parameters after the optimization are shown in Table 1 of article I. A significant difference was found in α parameter for oxygen and chlorine atoms as compared with the original parameters of reference [5]. The new β parameter values for H and C atoms were close to the original data. A large difference in β_p values was found for chlorine atoms. For the oxygen atom, the values for β_s and β_p were set equal to each other in the original work. However, in our case, these parameters are different. The variance in the parameterization of O and Cl as compared to the original parameters indicates the higher electronegativity of these atoms with respect to the electrostatic molecular interactions in solute-solvent (solute-solute) systems.

To prove the usability of the new parameters, we have tested two data sets of compounds *i.e.* i) 9 inorganics with measured boiling points and ii) 165 organic compounds with measured critical temperatures. For both sets QSPR models had been built previously. These models were compared with those that used the new optimized semiempirical parameters in Table 1. The results from this comparison showed a significant improvement in the statistical parameters of the QSPR model for set ii) $R^2 = 0.868$, s = 28.7 K, F = 546 – old vs. $R^2 = 0.902$, s = 25.04 K, F = 742.73 – new. Also, the average deviation of the predicted values for inorganics was improved 17 vs. 22 K in the original work.

Taking into account these promising results, it can be expected that this reparameterization would improve the prediction of QSAR models for organic compounds in various liquid media.

2.2. Comparison of (B)MLR and ANN approaches on the example of different properties. [Articles II–V]

2.2.1. QSAR of anti-platelet derived growth factor

Article 2 presents Quantitative Structure-Activity Relationship (QSAR) modeling of the biological activity (IC_{50}) of 123 1-phenylbenzimidazoles as inhibitors of platelet-derived growth factor receptor (PDFG). Biological studies of the tumor angiogenesis have led to the identification of various molecules, which promote tumor development. Of particular interest are such factors as the platelet-derived growth factor (PDGF). These inhibitors are significant selective molecules of PDGF, with clear evidence of the relationship between their molecular features and inhibitory activity. Hence, development of QSAR models could help in revealing some important molecular features responsible

for the inhibition of the PDGF as well as prediction of IC_{50} of new anti-cancer drugs.

The QSAR modeling of article II employs two robust methods for the logarithms of the effectiveness of 1-phenylbenzimidazoles as inhibitors of the platelet-derived growth factor, log(1/IC50):

(i) QSAR modeling by best multilinear regression method performed with the CODESSA-PRO program which was applied to more than 800 different constitutional, geometrical, topological, electrostatic, quantum-chemical, and thermodynamic molecular descriptors. The resulted model was seven-descriptor equation shown in Table 2 of article II according to the BMLR and the breaking rule. The statistical characteristics as the squared correlation coefficient R^2, squared standard deviation s^2 and the Fisher criterion F were 0.71, 0.176 and 24.33, respectively. In addition, this model was developed for only 78 1-phenyl-benzimidazoles out of 123. The reason for this was because of the statistical invalidation of some experimental data and that the BMLR methodology requires numerical experimental values.

The multilinear equation was internally validated by division of the data in three subsets consisting of the 2/3 of the data as training sets and the corresponding 1/3 of the data as validation sets. Generally, this validation indicated relative predictive stability of the BMLR model (cf. Table 4, article II) with average R^2(fit 2/3) =0.684 and R^2 (pred 1/3) = 0.698.

(ii) The nonlinear modeling performed using artificial neural network (ANN) method with backpropagation learning algorithm and sigmoid activation function developed in house was used. This QSAR modeling was based on a classification problem since 45 compounds had only the upper limit of $\log(1/IC_{50})$ values. In addition, the experimental error of the data was reported to be 15%. Thus, we divided the data into 8 classes.

The experimental data points were randomly divided into validation (41) and training sets (82). The resulted ANN model after sensitivity analysis, had architecture 5–4–1 i.e. five input descriptors, 4 hidden neurons and one output neuron for the $\log(1/IC_{50})$. The RMS errors after training the ANN for training and validation sets were 0.77 and 1.54, respectively. The results from the correct prediction of the training set for the different classes were as follows: class 1–78%, class 2–40%, class 3–40%, class 4–66%, class 5–80%, class 6–57%, class 7–71%, class 8–100%. Notably, class 8 was predicted 100% for the 2 compounds that fall in this class (Table 5 and Figure 5 in article II). The percentage of the accurate predicted classes for the validation set not used in the training of the ANN were: class 1–81%, class 2–66%, class 3–100%, class 4–100%, class 5–33%, class 6–25%, class 7–100%, class 8–100% (Table 5 and Figure 6 in article II). Only for 8 compounds out of 39 (20%) does the predicted class differ by more than one from the correct experimental class. Therefore, the prediction that a compound belongs within the one class was achieved for 80% of compounds, which was a very good result.

In both treatments (BMLR and ANN) all descriptors used are derived solely from molecular structure and do not require experimental data or expensive theoretical calculations to be obtained. A comparison of the descriptors between the linear model in Table 2 and the nonlinear shows that there were two descriptors (*Min e-e repulsion for C-C bond*, *Relative negative charge (QMNEG/QTMINUS) (MOPAC PC)*) which appeared in both models. Therefore, two different approaches selected 2 common molecular features indicating that they are important for the property in question.

A general comparison in sense of prediction shows that the nonlinearities of the ANN model with the linear improved the prediction of the data. (in terms of coefficient of determinations (R^2) of the training set of the ANN and the linear model 0.81 and 0.71, respectively). Also, ANN model had less number of input variables and it was built on a bigger number of compounds.

2.2.2. QSAR of glycogen synthase kinase 3

Article III presents QSAR modeling of biological activity (pIC_{50}) for 277 inhibitors of glycogen synthase kinase 3 (GSK-3) enzyme. The GSK-3 is a multifunctional serine/threonine kinase usually found in mammalian tissues. It is involved in multiple physiological processes related to the cause of cancers, type-2 diabetes, neurodegenerative disorders (Alzheimer, bipolar disorders), proliferation of protozoan parasites, and viral infections such as HIV, cytomegalo virus, and herpes virus.

In this study we used two methods for QSAR modeling namely, best multilinear regression method and artificial neural network as it was in the case of Article II. For all 277 compounds program more than 900 different constitutional, geometrical, topological, electrostatic, quantum chemical, and thermodynamic molecular descriptors were calculated using CODESSA-PRO.

Before to build the BMLR equations we divided the compound set into four different subsets according to the different derivatives (Figure 1 in article III) so that it was possible to explore more precisely the chemical space of these derivatives. In addition, this division could help to elucidate the important descriptors for each set that would appear in the BMLR equations.

The resulted BMLR in conjunction with the breaking rule led to the following statistics for the four classes of compounds: Class I – N=74, n=6, R^2=0.896, R^2_{cv}=0.874, F=96.683, s^2=0.019, Class II – R^2 = 0.677, R^2_{cv}= 0.593, F = 21.900, s^2 = 0.374, n = 7, N = 81, Class III – N = 61, n = 5, R^2 = 0.745, R^2_{cv} = 0.703, F = 32.240, s^2 = 0.224 and Class IV – N = 6, n = 61, R^2 = 0.507, R^2_{cv} = 0.370, F = 9.245, s^2 = 0.535, where N is the number of data points; n is the number of descriptors; R^2 is the squared correlation coefficient; R^2_{cv} is the squared cross-validated correlation coefficient; F is the Fisher's criterion; and s^2 is the squared standard error of the model. All the models had satisfactory statistical characteristics bearing in mind the complexity of the phenomenon

behind the inhibitor and GSK-3 interactions. Only for Class II and IV the results were somewhat poorer. Also, Class II possessed the biggest number of compounds. Generally, the main tendency was that increasing the number of compounds in the models lead to lower statistical quality of the multilinear equations. This conjecture was proven at the beginning of the work where we wanted to build a common multilinear model for all 277 compounds. Speaking statistically, the variability in the data was difficult to describe by multilinear dependence. Thus, the second stage of the project was to use nonlinear ANN modeling.

The ANN modeling started of reselection of the initial CODESSA PRO descriptor space so that significant input descriptors to be found. The reselection was based on chemical intuition (exploring scatter plots between descriptors and property, criteria for intercorrelations) and sensitivity analysis of building simple 1-1-1 ANNs. The final model had topology of 6-6-6-1. The RMS errors for the training (187) and the validation (90) sets were 0.67 and 1.54.

Rough comparison can be done in terms of R^2 between the ANN and BMLR models since the models differ by number of compounds and methodology. First, the $R^2 = 0.782$ for the training set of the ANN is larger compared with all BMLR models except for Class I ($R^2 = 0.896$). Secondly, ANN model is trained on 187 compounds vs 74 congeneric compounds of model for Class 1. The number of compounds for the ANN model is much bigger than the number of compounds in Class 1. Therefore, ANN is able to describe with good accuracy the larger variability of the data than the BMLR. The same holds for the validation set of the ANN consisting of 90 randomly selected compounds with a prediction in terms of $R^2 = 0.68$. In other words, this set is larger than all congeneric classes and the prediction is superior to the BMLR models for classes II and IV.

Finally, the results from all models showed that the ANN approach has better prediction than BMLR for larger data sets. However, BMLR results shed light on the GSK-3 interactions with the inhibitors based on the descriptors in the correlations showing that electrostatic interaction are important.

2.2.3. QSAR of anti-cancer activity

Article IV provides the result of a QSAR investigation of anti-invasive activity score index of 139 drugs. Anti-invasive activity of a drug is a measure of tumor cell activity, related to the final outcome of cancer. The most active compounds are used as anti-cancer drugs.

The QSAR modeling in this study was a classification task since in the common practice the data for the score indexes are given in only five concentrations (cf Table 1, article IV). Thus, we classified the activity score index from 1 to 5 classes. This peculiarity of the experimental data encouraged

us to use as a main method backpropagation artificial neural network together with the best multilinear regression.

CODESSA PRO software was used to calculate 863 descriptors for each drug. In such a large descriptor space, a robust algorithm is needed to find significant descriptors as inputs to the ANN. We used BMLR approach to find the best descriptors in a descriptor space extended with several nonlinear functions of the primary descriptors. These nonlinearities included square, squared root and binomial terms of the descriptors. The BMLR method resulted in a 7-parameter equation whose descriptors were chosen as inputs for the ANN model. The descriptors were: *HA-dependent HDCA-2 (Zefirov PC)*, *WNSA-3 Weighted PNSA (PNSA3*TMSA/1000)(MOPAC PC)*, *(minimum atomic orbital electronic population) x (minimum e–e repulsion for atom H)*, *maximum resonance energy for bond C–C*, *Max e–n attraction for bond C–C*, *maximum Coulombic interaction for bond H–C*, and *minimum e–e repulsion for atom H*.

The ANN search resulted in 7-6-1 model with the above mentioned descriptors from the BMLR equation as the input units. After the training of the ANN model by Levenberg-Marquardt optimizer, the obtained RMS errors of the training and the validations set were 0.568 and 0.569, respectively. The results from the ANN classification (Figure 2 and 3 in article IV) of the classes is as follows: class 1–57% predicted correctly, class 2–63%, class 3–82%, and class 4–78%. Consequently, the average percentage to predict the exact class is 71% for the training set, which is quite significant. With respect to the validation set (which serves as an external set, Fig. 3 in article IV) the exactly predicted probabilities are: class 1–80%, class 2–60%, class 3–71%, and class 4–100%. From both figures (Figure 2 and 3 in article IV), it is noticeable that the largest number of compounds exactly predicted are situated on the main left diagonal of the confusion matrices. It can be noted that class 4 for the validation set is predicted exactly that is the goal of the most practical cases where one needs to know which compound is most active among the novel compounds synthesized.

It can be concluded that the BMLR selected descriptors served as good input variables for the ANN model. Therefore, combination of these two methods led to predictive QSAR for classification of the anti-invasive score index.

2.2.4. Neural network convergence in QSARs

Article V deals with the neural network convergence and prediction of three optimization algorithms implemented in the network. The results of this work are important for the QSAR/QSPR modeling where artificial neural networks are used as a method of choice.

In QSAR/QSPR modeling the neural network models possess certain architecture with input neurons in the range 2–12 and hidden layers from 1 to 3. Consequently, the main goal of this work was to simulate a practical neural

network structure that usually is used in QSAR/QSPR modeling and to explore this structure for fast convergence and prediction ability. For all neural network simulations the molecular descriptors used as inputs were calculated for all chemical compounds by CODESSA PRO software. The descriptors were selected by the Heuristic method implemented in CODESSA PRO as those appearing in the respective best multilinear equations.

Three optimization algorithms were implemented in the simulated ANNs based on large experimental sets of eight physicochemical properties, namely 411 vapor pressures, 298 boiling points, 60 carcinogenic activities, 115 milk/plasma ratios, 137 organic compounds with measured ozone tropospheric degradation rates, 158 skin permeation rates, 57 p-glycoprotein inhibitors, 115 log blood-brain partition coefficients. The investigated for convergence algorithms were Levenberg-Marquardt, Conjugate gradient and delta rule. In addition, we tested these algorithms on validation sets not used in the training of the networks. The convergence was assessed as related to the RMS-epoch of the training set at which the RMS of the validation set started increasing. Thus, the models were monitored so that to avoid overfitting problems.

The obtained results in Table 1 of article V clearly indicated that Levenberg-Marquardt algorithm is the fastest in a practical ANN QSAR/QSPR modeling (Figures 2–9 and Table 1 of article V) and provides better prediction abilities in most of the cases. Also, this conjecture is invariant with respect to the compound data sets as well as the experimental properties used.

2.3. Applicability of the fragment approach (ISIDA) [Article VI]

The penetration of exogenous drug chemicals through the human skin is of significance in many disciplines, ranging from the pharmaceutical and cosmetic industries, where control of permeation is essential for the topical application of lotions, creams, and ointments, to toxicological risk assessments of materials from the environment and in specific occupations. In these fast developing areas the use of good exploratory QSAR models of the skin penetration is mandatory.

Article VI presents a QSAR investigation of the skin permeation rate Kp, more precisely logKp, by means of three approaches: best multilinear regression (BMLR), MLP artificial neural network (ANN) and substructural molecular fragment (SMF). All these methods were applied on 143 diverse compounds with experimental logKp. The computational software used for the algorithms was CODESSA PRO for BMLR, ISIDA for SMF and an in-house built program for the ANN. Two large descriptor spaces were generated i) theoretical molecular descriptors by CODESSA PRO and ii) fragment descriptors by ISIDA. Also, the calculated and experimental octanol/water values LogP were separately included in these large descriptor spaces since the practice shows that LogP parameter correlates well with logKp.

The results of the QSAR modeling by using BMLR led to five-parameter multilinear equation (Eq 7 in article VI) with good statistical parameters: N = 143, k = 5, R^2 = 0.800, R^2_{cv} = 0.781, F = 109.6, and s = 0.54. In these notations, N is the number of data points; k is the number of descriptors; R^2 is the squared correlation coefficient; R^2_{cv} is the squared cross-validated correlation coefficient; F is the Fisher's criterion; and s is the standard error of the model.

The second QSAR model based on ANN resulted in small 4-4-1 architecture. This model requires division of the data into two subset i.e. training (103) and validation (40) as it was the same requirement in the modeling procedures for ANN in articles II–V so that to avoid over training problems. The obtained correlation in terms of R^2 for the training and validation sets resulted in high values of 0.813 and 0.721, respectively. The training of the net was stopped at RMS of the validation set 0.661, of course, resulted in a smaller RMS for the training set 0.519.

The third method used to build QSAR for the logKp was SMF, which is based on structural descriptors calculated by ISIDA program. The statistical parameters of this multilinear model (Eq 2 and Table 4 of article VI) are as follows: N = 143, k = 41, R^2 = 0.907, s = 0.43, R_{cv}^2 = 0.812, F = 24.8.

After analyzing the three models, the best predictive results in terms of R^2 were achieved by SMF method (0.907 vs. 0.800 and 0.832 for BMLR and ANN, respectively). However, the ANN model was built on different number of compounds than BMLR and SMF, and therefore the direct comparison is not straightforward. In addition, these methods differ in methodological aspects. SMF algorithm is a modification of the Free-Wilson approach which allows inclusion of large number of fragment descriptors as in our case k = 41 (Table 4 in article VI).

SMF obtained better prediction (based on LOO) according to the crossvalidation coefficient (R_{cv}^2) as compared to the BMLR. The significant difference between R^2 and R_{cv}^2 of the SMF results indicates that logKp is very sensitive to the fragment descriptors whereas the same comparison for the BMLR shows relative stability. The test of these two methods by the modified LMO validation showed superiority of the BMLR method over the SMF (Table 5 of article VI). The actual prediction for the ANN model was assessed by the validation set of 40 compounds that served as an external validation producing R^2 = 0.72, which of course is a good result.

The advantage of the SMF method over the BMLR and the ANN approaches is that this method shows directly which structural molecular features are the most important based on the fragment descriptors. In this work the most important characteristics were the quaternary carbon atom having as neighbors either three carbons and one oxygen atom or two carbons, and the carbon atom connected to one nitrogen. Both contribute significantly to LogKp. Regarding the BMLR and ANN molecular descriptors, LogP was the most significant descriptor.

3. CONCLUSIONS

The general scheme of building QSAR was applied on various physisco-chemical properties and biological activities. It was shown that to build reliable QSAR models the use of robust mathematical methods is mandatory. Approaches as BMLR, ANN and SFM can be applied in a broad spectra of QSAR tasks ranging from tropospheric rates to the anti-cancer indices. When one method fails, another could success. The combined use of these methods is a powerful tool for QSAR modeling based on theoretical descriptors. Also, fundamental semiempirical quantum chemical methods as AM1 are not perfect so that their improvement to the QSAR and especially to the quantum chemical descriptors is necessary.

To summarize, the reported QSAR models based on the above methods could have importance for various domains of research such as computational chemistry, biochemistry, medicine, and pharmacy.

REFERENCES

1. Hansch, C.; Leo, A. *Exploring QSAR, Fundamentals and Applications in Chemistry and Biology*; ACS: Washington, 1995.
2. Abraham, M. H. *New Solute Descriptors for Linear Free Energy Relationships and Quantitative Structure-Activity Relationships, In Quantitative Treatments of Solute/Solvent Interactions*; Politzer, P.; Murray, J. S.; Eds; Elsevier: Amsterdam, 1994; pp 83–133.
3. Abraham, M. H.; Chadha, H. S.; Dixon, J. P.; Rafols, C.; Treiner, C. *J. Chem. Soc. Perkin Trans*. 2 1995, 887–894.
4. Kier , L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; John Wiley & Sons: New York, 1986.
5. Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, 279–287.
6. Kubinyi, H.; Sadowski, J. *QSAR, 3D QSAR and beyond*. 217th ACS National Meeting, Anaheim, Calif., March 21–25, (1999).
7. Pleiss, M. A.; Unger, S. H.; *The design of test series and the significance of the QSAR relationships*. In *Quantitative Drug Design*. Ramsden, C. A. (Ed.) Pergamon Press: Oxford, 1990, 561–567.
8. Cronin, M. T. D.; Schults, T. W.; Pitfals in QASR. *J. Mol. Struct*., **2003**, 633, 39–51.
9. Kubinyi, H. QSAR and 3D QSAR in drug design. *Research Focus*, **1997**, *2*, 457.
10. http://www.hyper.com.
11. Stewart. J. J. P. *Quant. Chem. Prog. Exch.*, 10:86, **1990**.
12. http://www.semichem.com/ampac/default.php
13. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Model. *J. Am. Chem. Soc.*, **1985**, 107:3902–3909.
14. Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods IV: Extension of MNDO, AM1, and PM3 to more Main Group Elements, *J. Mol. Model*. **2004**,10, 155–164.
15. Dewar, M. J. S.; Thiel, W. Ground States of Molecules, 38. The MNDO Method. Approximations and Parameters. *J. Am. Chem. Soc.*, **1977,** 99:4899–4907.
16. http://www.codess-pro.com
17. Mekenyan, O. G.; Karabunarliev, S. H.; Ivanov, J. M.; Dimitrov, D. N. *Comput. Chem.* **1994**, *18*, 173.
18. http://www.talete.mi.it/download.htm
19. Hocking, R. Biometrics **1976**, *32*, 1–49.
20. Deaven, D. M.; Ho, K. M. *Phys. Rev. Lett.* **1995**, *75*, 288.
21. Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. *J. Phys. Chem*. **1996**, *100*, 10400
22. Gabrielsson, J.; Lindberg, N.-O.; Lundstedt, T. *J. Chemom.* **2002**, *16*, 141–160.
23. Hoskuldsson, A. *Chemom. Intel. Lab. Sys.* **2001**, *55*, 23–38.
24. Svozil, D.; Kvasnicka, V.; Pospichal, J. *Chemom. Intell. Lab. Sys.* **1997**, *39*, 43–62.
25. Ajay, A. *J. Med. Chem*. **1993**, *36*, 3565.
26. Free, S.; Wilson, J. *J. Med. Chem*. **1964**, *7*, 395.
27. Eriksson, L.; Johansson, E.; Muller, M.; Wold, S. On the selection of the training set in environmental QSAR analysis when compounds are clustered *J. Chemometrics* **2000**, *14*, 599–616.

28. Edington E. S. *Randomization tests*: Marsel Dekker, Inc.: New York and Basel, 1980, pp.195–216.
29. Cocchi, M.; Menziani, M. C.; De Benedetti, P. G.; Cruciani, G. *Chemom. Intell. Lab. Sys.* **1992**, *14*, 209.
30. Diercksen, G. H. F., Wilson, S. *Methods in Computational Molecular Physics*; Eds.; D. Reidel Publ. Co.: Dordrecht, 1983.
31. Goddard, J. D.; Handy, N. C.; Schaefer, H. F., III. *J. Chem. Phys.* **1979**, *71*, 1525.
32. Kucharski, S. A.; Bartlett, R. J. *Adv. Quantum Chem.* **1986**, *18*, 281.
33. Pople, J. A.; Krishnan, R.; Schlegel, H. B.; Binkley, J. S. *Int. J. Quantum Chem.* **1978**, *14*, 545.
34. M[illegible]ller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
35. Dewar, M. J. S. *Science* **1975**, *187*, 1037.
36. Stewart, J. J. P. Optimization of Parameters for Semi-Empirical Methods III–Extension of PM3 to Be, Mg, Zn, Ga, Ge, As, Se, Cd, In, Sn, Sb Te, Hg, Tl, Pb, and Bi. *J. Comp. Chem.*, **1991**, *12*, 320–341.
37. Stewart, J. J. P. Optimization of Parameters for Semi-Empirical Methods I-Method. *J. Comp. Chem.*, **1989**, *10*, 209–220.
38. Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984; pp 115–123.
39. Karelson, M.; Lobanov, V.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027–1043.
40. Darlington, R. B. *Regression and linear models*. New York: McGraw-Hill, 1990.
41. Karelson, M. *Molecular descriptors in QSAR/QSPR*. Wiley-Interscience: New York, 2000.
42. Drapper, N. R.; Smith, H. *Applied Regression Analysis*. Wiley, New York, 1981.
43. Katritzky, A; Fara, F.; Yang, H.; Karelson, M.; Suzuki, T.; Solov'ev, V.; Varnek, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 529–541.
44. Burns, J. A.; Whitesides, G. *Chem. Rev*. **1993**, *93*, 2583.
45. Mazzatorta, P.; Benfenati, E.; Neagu, C. D.; Gini, G. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 513–518.
46. Sild, S.; Karelson, M. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 360–367.
47. Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: an Introduction*; VCH-Verlag: Weinheim, 1993; pp 213–228.
48. Haykin, S. *Neural Networks. A comprehensive foundation*; Pearson Ed, 1999.
49. Baskin, I. I.; Ait A. O.; Halberstam, N. M.; Palyulin, V.A.; Zefirov, N. S. *SAR QSAR Environ. Res*. **2002**, *13*, 35.
50. Fletcher, R. *Practical Methods of Optimization, Vol. 1 , Unconstrained Optimization*; Wiley: New York, 1980.
51. Hagan, M.; Menhaj, M. *IEEE Trans. Neur. Net.* **1994**, *5*, 989.
52. Goldberg, D. E. *Genetic Algorithms*. Reading, MA: Addison Wesley, 1989.
53. Solov'ev, V.; Varnek, A.; Wipff, G. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847–858.
54. Solov'ev, V.; Varnek, A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1703–1719.
55. Solov'ev, V.; Varnek, A. *J. Comp. Aid. Mol. Design*. **2005**, *19*, 693–703.
56. Casalegno, M.; Benfenati, E.; Sello, G. *Chem. Res. Toxicol.* **2005**, *18*, 740–746.
57. ISIDA, http://infochim.u-strasbg.fr/recherche/isida/index-php.

ACKNOWLEDGMENTS

I would like to thank my doctoral advisor Professor Mati Karelson for his excellent guidance throughout my research.

I would like to express my sincere gratitude to Kenan Professor Alan R. Katritzky for his support.

SUMMARY IN ESTONIAN

Üldised QSAR meetodid omaduste ennustamiseks molekulaarstruktuurist

Käesolevas töös rakendati kvantitatiivsete struktuur-aktiivsus sõltuvuste (QSAR) üldist skeemi erinevate keemiliste ühendite füüsiko-keemiliste omaduste ja bioloogiliste aktiivsuste modelleerimisel. Näidati, et töökindla QSAR mudeli tuletamisel on vajalik rakendada erinevaid matemaatilisi lähenemisi. Sellisteks lähenemisteks on nn. parim multilineaarne regressioon (BMLR), tehisnärvivõrgud (ANN) ja molekulaarfragmentide meetod (SFM) mida rakendati QSAR ülesannete laias spektris alates troposfääris toimuvate reaktsioonide kiiruskonstantidest ning lõpetades vähivastaste indeksitega. Leiti, et parimaks lahenduseks on erinevate lähenemiste kombineeritud kasutmine, ühe lähenemise ebaõnnestumisel võib teine, alternatiivne lähenemine töötada. Käesolevas töös täiustati ka poolempiirilisi kvantkeemilisi meetodeid nagu AM1, parandamaks nende rakendatavust QSAR-is vajalike kvantkeemiliste deskriptorite arvutamisel.

Töös antud QSAR mudelid, mis baseeruvad eelpool mainitud lähendustel, omavad tähtsust erinevates teadusvaldkondades, nagu arvutikeemia, biokeemia, meditsiin ja farmaatsia.

PUBLICATIONS

Dobchev, Dimitar A.; Karelson, Mati. Reparameterized Austin Model 1 for quantitative structure-property relationships in liquid media. *J. Mol. Model.* **2006**, *12*, 503–512.

Katritzky, Alan R.; Dobchev, Dimitar A.; Fara, Dan C.; Karelson, Mati. QSAR studies on 1-phenylbenzimidazoles as inhibitors of the platelet-derived growth factor. *Bioorg. Med. Chem.* **2005**, *13(24)*, 6598–6608.

Katritzky, Alan R.; Pacureanu, Liliana M.; Dobchev, Dimitar A.; Fara, Dan C.; Duchowicz, Pablo R.; Karelson, Mati. QSAR modeling of the inhibition of Glycogen Synthase Kinase-3. *Bioorg. Med. Chem.* **2006**, *14(14)*, 4987–5002.

Katritzky, Alan R.; Kuanar, Minati; Dobchev, Dimitar A.; Vanhoecke, Barbara W. A.; Karelson, Mati; Parmar, Virinder S.; Stevens, Christian V.; Bracke, Marc E. QSAR modeling of anti-invasive activity of organic compounds using structural descriptors. *Bioorg.Med. Chem.* **2006**, *14*(20), 6933–6939.

Karelson, Mati; Dobchev, Dimitar A.; Kulshyn, Oleksandr V.; Katritzky, Alan R. Neural Networks Convergence Using Physicochemical Data. *J. Chem. Inf. Mod.* **2006**, *46(5)*, 1891–1897.

Katritzky, Alan R.; Dobchev, Dimitar A.; Fara, Dan C.; Hur, Evrim; Tamm, Kaido; Kurunczi, Ludovic; Karelson, Mati; Varnek, Alexandre; Solov'ev, Vitaly P. Skin permeation rate as a function of chemical structure. *J. Med. Chem*. **2006**, *49(11*), 3305–3314.

CURRICULUM VITAE

DIMITAR ATANASOV DOBCHEV

Born: 6 March 1977
Citizenship: Bulgarian
Marital Status: Single
Address: University of Florida
Department of Chemistry
Center of Heterocyclic Compounds
CRB 229, Gainesville 32611, FL, USA
tel: (352) 392–9865
E-mail: dobchev@chem.ufl.edu

Education

1995–1999	Student, Faculty of Physics, Sofia University, Bulgaria. B.Sc. in 1999.
1999–2001	Graduate student, Faculty of Physics, Sofia University, Bulgaria. M.Sc. in 2001.
2004–2006	Ph.D. student, Department of Chemistry, University of Tartu, doctoral advisor Prof. Mati Karelson.

Professional experience

1999–2002	Young researcher, Institute for Nuclear Research and Nuclear Energy, Bulgarian Academy of Science, Sofia, Bulgaria
2003–2004	Young Researcher, IMAGETOX, 5th European Framework Program, Tartu, Estonia
2004-present	Visiting Scholar, Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, USA.

Publications

1. Katritzky, Alan R.; Dobchev, Dimitar A.; Fara, Dan C.; Karelson, Mati. **QSAR studies on 1-phenylbenzimidazoles as inhibitors of the platelet-derived growth factor**. *Bioorganic & Medicinal Chemistry* (2005), 13(24), 6598–6608.
2. Katritzky, Alan R.; Pacureanu, Liliana M.; Dobchev, Dimitar A.; Fara, Dan C.; Duchowicz, Pablo R.; Karelson, Mati. **QSAR modeling of the inhibition of Glycogen Synthase Kinase-3.** *Bioorganic & Medicinal Chemistry* (2006), 14(14), 4987–5002.

3. Katritzky, Alan R.; Dobchev, Dimitar A.; Fara, Dan C.; Huer, Evrim; Taemm, Kaido; Kurunczi, Ludovic; Karelson, Mati; Varnek, Alexandre; Solov'ev, Vitaly P**. Skin permeation rate as a function of chemical structure.** *Journal of Medicinal Chemistry* (2006), 49(11), 3305–3314.
4. Katritzky, Alan R.; Kuanar, Minati; Dobchev, Dimitar A.; Vanhoecke, Barbara W. A.; Karelson, Mati; Parmar, Virinder S.; Stevens, Christian V.; Bracke, Marc E. **QSAR modeling of anti-invasive activity of organic compounds using structural descriptors**. *Bioorganic & Medicinal Chemistry* (2006), 14(20), 6933–6939.
5. Katritzky, Alan R.; Dobchev, Dimitar A.; Tulp, Indrek; Karelson, Mati; Carlson, David A. **QSAR study of mosquito repellents using Codessa Pro**. *Bioorganic & Medicinal Chemistry Letters* (2006), 16(8), 2306–2311
6. Katritzky, Alan R.; Kuanar, Minati; Slavov, Svetoslav; Dobchev, Dimitar A.; Fara, Dan C.; Karelson, Mati; Acree, William E.; Solov'ev, Vitaly P.; Varnek, Alexandre. **Correlation of blood-brain penetration using structural descriptors.** *Bioorganic & Medicinal Chemistry* (2006), 14(14), 4888–4917.
7. Katritzky, Alan R.; Kulshyn, Oleksandr V.; Stoyanova-Slavova, Iva; Dobchev, Dimitar A.; Kuanar, Minati; Fara, Dan C.; Karelson, Mati. **Antimalarial activity: A QSAR modeling using CODESSA PRO software.** *Bioorganic & Medicinal Chemistry* (2006), 14(7), 2333–2357.
8. Katritzky, Alan R.; Dobchev, Dimitar A.; Hur, Evrim; Fara, Dan C.; Karelson, Mati. **QSAR treatment of drugs transfer into human breast milk.** *Bioorganic & Medicinal Chemistry* (2005), 13(5), 1623–1632.
9. Dobchev, Dimitar A.; Karelson, Mati. **Reparameterized Austin Model 1 for quantitative structure-property relationships in liquid media**. *Journal of Molecular Modeling* (2006), 12(4), 503–512.
10. Katritzky, Alan R.; Dobchev, Dimitar A.; Karelson, Mati. **Physical, chemical, and technological property correlation with chemical structure: the potential of QSPR**. *Zeitschrift fur Naturforschung, B: Chemical Sciences* (2006), 61(4), 373–384.
11. Karelson, Mati; Dobchev, Dimitar A.; Kulshyn, Oleksandr V.; Katritzky, Alan R. **Neural Networks Convergence Using Physicochemical Data.** *Journal of Chemical Information and Modeling* (2006), 46(5), 1891–1897.
12. Katritzky, Alan R.; Pacureanu, Liliana M.; Slavov, Svetoslav; Dobchev, Dimitar; Karelson, Mati. **QSAR study of antiplateled agents.** *Bioorganic & Medicinal Chemistry* (2006), in press.
13. Katritzky, Alan R.; Slavov, Svetoslav H.; Dobchev, Dimitar A.; Karelson, Mati **Comparison between 2D and 3D-QSAR approaches for a series of Indole Amide Hydroxamic Acids.** *QSAR & Combinatorial Science*, (2006), in press.
14. Katritzky, Alan R.; Slavov, Svetoslav H.; Dobchev, Dimitar A.; Karelson, Mati **Rapid QSPR model development technique for prediction of vapor pressure of organic compounds**. *Computers & Chemical Engineering*, (2006), in press.

ELULOOKIRJELDUS

DIMITAR ATANASOV DOBCHEV

Sündinud: 6. märts 1977
Kodakondsus: Bulgaaria
Perekonnaseis: vallaline
Aadress: University of Florida
Department of Chemistry
Center of Heterocyclic Compounds
CRB 229, Gainesville 32611, FL, USA
tel: (352) 392–9865
E-mail: dobchev@chem.ufl.edu

Haridus

1995–1999	Sofia Ülikool, Bulgaaria, Füüsika teaduskond, B.Sc. 1999.
1999–2001	Magistriõpe, Sofia Ülikool, Bulgaaria, Füüsika teaduskond, M.Sc. 2001.
2004–2006	Doktoriõpe, Tartu Ülikool, Keemilise Füüsika Instituut, juhendaja prof. Mati Karelson.

Teenistuskäik

1999–2002	Nooremteadur, Tuumauuringute ja Tuumaenergia Isntituut, Bulgaaria Teaduste Akadeemia, Sofia, Bulgaaria.
2003–2004	Nooremteadur, IMAGETOX, 5. Euroopa Raamprogramm, Tartu, Eesti
2004–present	Külalisteadur, Heterotsükliliste Ühendite Keskus, Keemia teaduskond, Florida Ülikool, USA.

Publikatsioonid

1. Katritzky, Alan R.; Dobchev, Dimitar A.; Fara, Dan C.; Karelson, Mati. **QSAR studies on 1-phenylbenzimidazoles as inhibitors of the platelet-derived growth factor**. *Bioorganic & Medicinal Chemistry* (2005), 13(24), 6598–6608.
2. Katritzky, Alan R.; Pacureanu, Liliana M.; Dobchev, Dimitar A.; Fara, Dan C.; Duchowicz, Pablo R.; Karelson, Mati. **QSAR modeling of the inhibition of Glycogen Synthase Kinase-3.** *Bioorganic & Medicinal Chemistry* (2006), 14(14), 4987–5002.

3. Katritzky, Alan R.; Dobchev, Dimitar A.; Fara, Dan C.; Huer, Evrim; Taemm, Kaido; Kurunczi, Ludovic; Karelson, Mati; Varnek, Alexandre; Solov'ev, Vitaly P**. Skin permeation rate as a function of chemical structure.** *Journal of Medicinal Chemistry* (2006), 49(11), 3305–3314.
4. Katritzky, Alan R.; Kuanar, Minati; Dobchev, Dimitar A.; Vanhoecke, Barbara W. A.; Karelson, Mati; Parmar, Virinder S.; Stevens, Christian V.; Bracke, Marc E. **QSAR modeling of anti-invasive activity of organic compounds using structural descriptors**. *Bioorganic & Medicinal Chemistry* (2006), 14(20), 6933–6939.
5. Katritzky, Alan R.; Dobchev, Dimitar A.; Tulp, Indrek; Karelson, Mati; Carlson, David A. **QSAR study of mosquito repellents using Codessa Pro**. *Bioorganic & Medicinal Chemistry Letters* (2006), 16(8), 2306–2311
6. Katritzky, Alan R.; Kuanar, Minati; Slavov, Svetoslav; Dobchev, Dimitar A.; Fara, Dan C.; Karelson, Mati; Acree, William E.; Solov'ev, Vitaly P.; Varnek, Alexandre. **Correlation of blood-brain penetration using structural descriptors.** *Bioorganic & Medicinal Chemistry* (2006), 14(14), 4888–4917.
7. Katritzky, Alan R.; Kulshyn, Oleksandr V.; Stoyanova-Slavova, Iva; Dobchev, Dimitar A.; Kuanar, Minati; Fara, Dan C.; Karelson, Mati. **Antimalarial activity: A QSAR modeling using CODESSA PRO software.** *Bioorganic & Medicinal Chemistry* (2006), 14(7), 2333–2357.
8. Katritzky, Alan R.; Dobchev, Dimitar A.; Hur, Evrim; Fara, Dan C.; Karelson, Mati. **QSAR treatment of drugs transfer into human breast milk.** *Bioorganic & Medicinal Chemistry* (2005), 13(5), 1623–1632.
9. Dobchev, Dimitar A.; Karelson, Mati. **Reparameterized Austin Model 1 for quantitative structure-property relationships in liquid media**. *Journal of Molecular Modeling* (2006), 12(4), 503–512.
10. Katritzky, Alan R.; Dobchev, Dimitar A.; Karelson, Mati. **Physical, chemical, and technological property correlation with chemical structure: the potential of QSPR**. *Zeitschrift fur Naturforschung, B: Chemical Sciences* (2006), 61(4), 373–384.
11. Karelson, Mati; Dobchev, Dimitar A.; Kulshyn, Oleksandr V.; Katritzky, Alan R. **Neural Networks Convergence Using Physicochemical Data.** *Journal of Chemical Information and Modeling* (2006), 46(5), 1891–1897.
12. Katritzky, Alan R.; Pacureanu, Liliana M.; Slavov, Svetoslav; Dobchev, Dimitar; Karelson, Mati. **QSAR study of antiplateled agents.** *Bioorganic & Medicinal Chemistry* (2006), in press.
13. Katritzky, Alan R.; Slavov, Svetoslav H.; Dobchev, Dimitar A.; Karelson, Mati **Comparison between 2D and 3D-QSAR approaches for a series of Indole Amide Hydroxamic Acids.** *QSAR & Combinatorial Science*, (2006), in press.
14. Katritzky, Alan R.; Slavov, Svetoslav H.; Dobchev, Dimitar A.; Karelson, Mati **Rapid QSPR model development technique for prediction of vapor pressure of organic compounds**. *Computers & Chemical Engineering*, (2006), in press.