

# Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points?

Alexandre Varnek\* and Natalia Kireeva

Laboratoire d'Infochimie, UMR 7551 CNRS, Université Louis Pasteur, 4, rue B. Pascal,  
Strasbourg 67000, France

Igor V. Tetko

GSF- Institute for Bioinformatics, Neuherberg D-85764, Germany, and Institute of Bioorganic &  
Petrochemistry, Kiev, Ukraine

Igor I. Baskin

Department of Chemistry, Moscow State University, Moscow 119992, Russia

Vitaly P. Solov'ev

Institute of Physical Chemistry, Russian Academy of Sciences, Leninskiy prospect 31a,  
Moscow 119992, Russia

Received November 4, 2006

Several popular machine learning methods—Associative Neural Networks (ANN), Support Vector Machines (SVM),  $k$  Nearest Neighbors ( $k$ NN), modified version of the partial least-squares analysis (PLSM), backpropagation neural network (BPNN), and Multiple Linear Regression Analysis (MLR)—implemented in ISIDA, NASAWIN, and VCCLAB software have been used to perform QSPR modeling of melting point of structurally diverse data set of 717 bromides of nitrogen-containing organic cations (*FULL*) including 126 pyridinium bromides (*PYR*), 384 imidazolium and benzoimidazolium bromides (*IMZ*), and 207 quaternary ammonium bromides (*QUAT*). Several types of descriptors were tested: E-state indices, counts of atoms determined for E-state atom types, molecular descriptors generated by the DRAGON program, and different types of substructural molecular fragments. Predictive ability of the models was analyzed using a 5-fold external cross-validation procedure in which every compound in the parent set was included in one of five test sets. Among the 16 types of developed structure – melting point models, nonlinear SVM, ASNN, and BPNN techniques demonstrate slightly better performance over other methods. For the full set, the accuracy of predictions does not significantly change as a function of the type of descriptors. For other sets, the performance of descriptors varies as a function of method and data set used. The root-mean squared error (RMSE) of prediction calculated on independent test sets is in the range of 37.5–46.4 °C (*FULL*), 26.2–34.8 °C (*PYR*), 38.8–45.9 °C (*IMZ*), and 34.2–49.3 °C (*QUAT*). The moderate accuracy of predictions can be related to the quality of the experimental data used for obtaining the models as well as to difficulties to take into account the structural features of ionic liquids in the solid state (polymorphic effects, eutectics, glass formation).

## 1. INTRODUCTION

Ionic liquids (IL) have received a great attention due to their green and tuneable properties. The negligible vapor pressures allow for their potential use as an alternative for organic volatile solvents.<sup>1,2</sup> Careful choice of cation/anion combination permits fabrication of IL with physical and chemical properties well fitted to a specific problem. One of the most important physical properties of IL, melting point (*mp*), was a subject of numerous studies (see book<sup>3</sup> and references therein). Melting point characterizing a passage from solid to liquid state has a very complex relationship with the structure of constituent ions because of many different factors.<sup>4</sup> Thus, both in solid and liquid phases,

various types of interactions between ions should be taken into account: electrostatic and van der Waals interactions, hydrogen bonds, and aromatic  $\pi$ – $\pi$ -stacking. The symmetry and conformational flexibility of individual species play an important role because they affect the crystal packing and, hence, melting points. Another problem is related to the phase content of the solids. Unlike high-melting salts, certain types of IL (i.e., halides of imidazolium cations<sup>5</sup>) melt from eutectic mixtures of several crystalline polymorphs. Usually, the eutectic temperature is considerably lower than melting points of individual polymorphs. One should not also exclude formation of glasses instead of crystalline phases which is quite typical low-melting IL.<sup>6</sup> In this case, *mp* represents the glass transition temperature which is rather different from melting point of the corresponding crystalline state.

\* Corresponding author e-mail: varnek@chimie.u-strasbg.fr; <http://infochem.u-strasbg.fr>.

Taking into account a variety of these factors, one can hardly estimate an impact of variation of structure of ions (size, symmetry, conformational flexibility, etc.) on *mp*. Any structural modification may lead to opposite trends of *mp*. For instance, an increase of the length of alkyl substituents reduces electrostatic interactions between ions thus leading to reduction of *mp*. Species with longer alkyl chains have more stable conformers. Therefore, corresponding IL have more chances to form polymorphs and their eutectic mixtures with low *mp*. However, it does not mean that *mp* always decreases with the length of alkyl groups because van der Waals attractive interactions between bulk alkyl radicals favor an increase of *mp*. Indeed, according to experimental observations, *mp* of ionic liquids oscillates with the size of alkyl groups.<sup>3</sup>

According to Katritzky et al.,<sup>2</sup> there exist approximately  $10^{18}$  combinations of ions that could lead to useful ionic liquids. Thus, there is a clear necessity to develop predictive computational tools allowing one to design new IL possessing desirable properties. In particular, this concerns *mp* which was a subject of several structure–property studies. Earlier, QSPR models were obtained for relatively small congener sets of bromides of some nitrogen-containing organic cations: pyridinium,<sup>1</sup> imidazolium,<sup>2</sup> benzoimidazolium,<sup>2</sup> and quaternary ammonium cations.<sup>7</sup> In most publications,<sup>2,7–11</sup> multilinear regression techniques were used; only one paper<sup>12</sup> reported application of the nonlinear methods (decision trees and neural networks) leading to models of better performance.

One problem with the compounds studied in ref 7 is that almost all those molecules melt above room temperature.<sup>2,7–12</sup> On the other hand, it is known that substitution of a small bromide anion with large anions such as  $\text{PF}_6^-$ ,  $\text{BF}_4^-$ ,  $(\text{CF}_3\text{SO}_2)_2\text{N}^-$ , and some others, leads to a significant decrease in *mp*.<sup>13</sup> Therefore, one could believe that the trend of variation of the melting point as a function of the cation's structure found for bromides will be similar for IL with larger anions.

Generally, the robustness of models depends on the method (linear, nonlinear) as well as descriptors used. The question arises as to what type of descriptors and methods could be recommended for QSPR studies of the melting point of IL. Is it possible to develop predictive models for large diverse sets of IL, or should one always look for congener sets? Could we predict the melting point of IL with reasonable accuracy?

The goal of this paper is to answer these questions. Here, QSPR modeling of the melting point for a structurally diverse data set of 717 bromides of nitrogen-containing organic cations was performed using different machine learning methods (Associative Neural Networks, Support Vector Machines, *k* Nearest Neighbors, modified version of the Partial Least-Squares analysis, Backpropagation Neural Network, and Multiple Linear Regression) and different types of descriptors (E-state indices, counts of atoms determined for E-state atom types, and molecular descriptors describing molecules as a whole and those based on different kinds of substructural fragments). With an influence of structural diversity on the quality of the models, the calculations were investigated on three congener subsets derived from the initial parent set. Predictive ability of the models was analyzed using 5-fold external cross-validation procedure and analysis

based on Arithmetic Average Model (see the Computational Procedure section).

Most of the calculations were performed using software developed by the authors: ISIDA (In Silico Design and Data Analysis),<sup>14</sup> NASAWIN,<sup>15,16</sup> and VCCLAB (Virtual Computational Chemistry Laboratory).<sup>17</sup> In some VCCLAB calculations, the molecular descriptors generated by the DRAGON program were used.

## 2. DATA PREPARATION

The calculations have been performed on the structurally diverse data set of 717 bromides of nitrogen-containing organic cations (*FULL*) containing 126 pyridinium bromides (*PYR*), 384 imidazolium and benzoimidazolium bromides (*IMZ*), and 207 quaternary ammonium bromides (*QUAT*). Experimental values of the melting point (*mp*, °C) for *PYR* were taken from ref 2, and those for subsets 2 and 3 were critically selected from the Beilstein database.<sup>18</sup> The data corresponding to glass transitions were not used. Typical structures in subsets 1–3 and corresponding ranges of melting points are given, respectively, in Figures 1 and 2.

The structure data files (SDF) containing the 2D structures of IL and the experimental values of their melting points have been prepared with the EdChemS and EdSDF programs included in the ISIDA package.<sup>14</sup> These data are available in the Supporting Information.

## 3. COMPUTATIONAL PROCEDURE

Three software packages ISIDA,<sup>14</sup> NASAWIN,<sup>15,16</sup> and VCCLAB<sup>17</sup> have been used to perform QSPR modeling of melting point. Each program uses its own descriptors, the procedure of variables selection, fitting techniques, and data analysis tools. Below we give some information concerning obtaining and validation of QSPR models using these three programs.

**3.1. ISIDA Software.** The ISIDA (In Silico Design and Data Analysis) software<sup>14</sup> is an ensemble of tools for computer-aided design of new compounds with desired properties using fragment descriptors. The QSPR/MLR module of the ISIDA program package has been used for structure–property modeling of the melting point for ionic liquids using multilinear regression analysis.

**3.1.1. Descriptors.** Substructural molecular fragments (SMF) have been used as descriptors in the calculations by the ISIDA program.<sup>19–23</sup> The SMF are subgraphs of molecular graph. All these descriptors were derived solely from chemical structure and did not require any experimental data to be calculated.

Two subclasses of SMF were used: “sequences” and “augmented atoms”.<sup>19</sup> Three subtypes were defined for each class. The sequences may include atoms and bonds, atoms only, or bonds only. For each type of sequences, the minimal ( $n_{\min} \geq 2$ ) and maximal ( $n_{\max} \leq 15$ ) number of constituent atoms is defined. Shortest or all paths from one atom to the other can be used. For the sequences with selected  $n_{\min}$  and  $n_{\max}$  the program generates all intermediate sequences involving  $n$  atoms ( $n_{\max} \geq n \geq n_{\min}$ ). An “augmented atom” represents a particular atom with its environment including either neighboring atoms and bonds, or atoms only, or bonds only. The atom hybridization can be taken into account or not for both subclasses of the fragments. The counts of

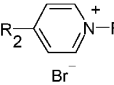
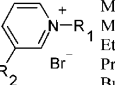
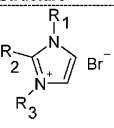
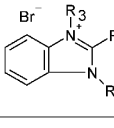
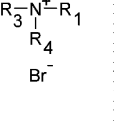
PYR: Pyridinium Bromides (126 compounds)					
Structure	R <sub>1</sub>	R <sub>2</sub>	Structure	R <sub>1</sub>	R <sub>2</sub>
	Me; Et; Pr; i-Pr; Bu; i-Bu Ph; Bz; 2-Py; OBz; Vinyl Me Me Me; Bu Me i-Pr; Allyl Et; Allyl i-Pr; OEt (CH <sub>2</sub> ) <sub>2</sub> C(O)OEt CH <sub>2</sub> EtC(O)OEt (CH <sub>2</sub> ) <sub>11</sub> C(O)OEt n-Oct (CH <sub>2</sub> ) <sub>9</sub> Me (CH <sub>2</sub> ) <sub>10</sub> Me; CH <sub>2</sub> C(O)NH <sub>2</sub> (CH <sub>2</sub> ) <sub>11</sub> Me (CH <sub>2</sub> ) <sub>12</sub> Me (CH <sub>2</sub> ) <sub>13</sub> Me Me; Et; (CH <sub>2</sub> ) <sub>2</sub> Ph; (CH <sub>2</sub> ) <sub>2</sub> CN 4-F-Bz; (CH <sub>2</sub> ) <sub>2</sub> OPh	H H Bz; (CH <sub>2</sub> ) <sub>3</sub> OH 4-Py; C(O)OMe C(O)OEt (CH <sub>2</sub> ) <sub>2</sub> C(O)OEt CH <sub>2</sub> OH CN OMe H H H Et; n-Pr H; C(O)OEt H H; Et; n-Pr H II; n-Hex Me H		Me Me Et Pr Bu (CH <sub>2</sub> ) <sub>2</sub> F (CH <sub>2</sub> ) <sub>2</sub> OH Vinyl Allyl Bz (CH <sub>2</sub> ) <sub>9</sub> Me (CH <sub>2</sub> ) <sub>13</sub> Me (CH <sub>2</sub> ) <sub>3</sub> Cl CH <sub>2</sub> COOH (CH <sub>2</sub> ) <sub>3</sub> COOH CH <sub>2</sub> C(O)OMe (CH <sub>2</sub> ) <sub>4</sub> C(O)OMe (CH <sub>2</sub> ) <sub>2</sub> CN (CH <sub>2</sub> ) <sub>3</sub> CN CH <sub>2</sub> CN	(CH <sub>2</sub> ) <sub>3</sub> OH; C(O)Me OH; 2-Py; C(O)OMe OH; C(O)NEt <sub>2</sub> ; NMe <sub>2</sub> C(O)NH <sub>2</sub> COOH C(O)OEt H; Me; OH H; OH CHO; C(O)NEt <sub>2</sub> Mc Am Am H H H H H OH H; Me; NH <sub>2</sub> H H
IMZ: Imidazolium and Benzoimidazolium Bromides (384 compounds)					
Structure	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>		
	Me Me; OII Me; Pr Me; i-Pr; Allyl; i-Bu; Bz Me; Vinyl Allyl; Bz; (CH <sub>2</sub> ) <sub>6</sub> CN Me; Am Me; Alk; Bz; Ar; OAlk; OAr; O(CO)Alk; O(CH <sub>2</sub> ) <sub>m</sub> CHClCH <sub>2</sub> ; CH <sub>2</sub> C(O)Ar	C(Me)CH <sub>2</sub> ; C(t-Bu)CH <sub>2</sub> ; C(Ph)CH <sub>2</sub> Me; Ph; Bz; CN H; i-Bu H H H; Me; (CH <sub>2</sub> ) <sub>8</sub> Me (CH <sub>2</sub> ) <sub>7-10</sub> Me H; Me; Alk	CH <sub>2</sub> C(O)Ph Me; OII Pr; i-Pr; Bz Pr; Bu; t-Bu; i-Bu; Am; Bz Et; Pr; Bu; C(O)Ph; (CH <sub>2</sub> ) <sub>3</sub> COOH (CH <sub>2</sub> ) <sub>9</sub> Me; (CH <sub>2</sub> ) <sub>m</sub> CN Bz Alk; Ar; CH <sub>2</sub> C(O)Ar; OCH <sub>2</sub> Ar; CH <sub>2</sub> C(O)OEt; Allyl		
	Et; (CH <sub>2</sub> ) <sub>9</sub> Me H; Me; Et; i-Pr; Ph; Bz; NH <sub>2</sub> ; Alk; Ar; (CH <sub>2</sub> ) <sub>m</sub> CN; Allyl; Ad; CH <sub>2</sub> C(O)Alk; (CH <sub>2</sub> ) <sub>6</sub> (CO)NH <sub>2</sub> ; NHBz; CH <sub>2</sub> OH; (CH <sub>2</sub> ) <sub>6</sub> OEt; (CH <sub>2</sub> ) <sub>6</sub> SiMe <sub>3</sub> ; (CH <sub>2</sub> ) <sub>6</sub> COOH	II H; Me; Et; Pr; Bu; CF <sub>3</sub> ; NH <sub>2</sub> ; Bz; Ar; CH <sub>2</sub> OH; CH(OH)Me	(CH <sub>2</sub> ) <sub>9</sub> Me; (CH <sub>2</sub> ) <sub>11</sub> Me; (CH <sub>2</sub> ) <sub>17</sub> Me Me; Et; CH <sub>2</sub> CN; CH <sub>2</sub> C(O)Ar; Ar; CH <sub>2</sub> C(O)Alk; CH <sub>2</sub> C(O)OAlk; Ad; (CH <sub>2</sub> ) <sub>6</sub> CN; Bz; (CH <sub>2</sub> ) <sub>6</sub> COOH; Allyl; (CH <sub>2</sub> ) <sub>6</sub> CH=CHMe; Alk; (CH <sub>2</sub> ) <sub>6</sub> OH; (CH <sub>2</sub> ) <sub>6</sub> Br		
QUAT: Ammonium Bromides (207 compounds)					
Structure	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	
	Me Me Me Me; Et; Bu Me Me Me Me; Et Me; CH <sub>2</sub> CCH Me Me Me Et Et Pr	Me Me; Bu Me Me; Et; Bu Me Me Me Me; Et Me; Allyl Et Me; (CH <sub>2</sub> ) <sub>2</sub> OH Et Et Pr	Me; (CH <sub>2</sub> ) <sub>2</sub> OH; (CH <sub>2</sub> ) <sub>m</sub> Me Me; Bu Me; Pr; (CH <sub>2</sub> ) <sub>6</sub> Ph (CH <sub>2</sub> ) <sub>6</sub> Me (CH <sub>2</sub> ) <sub>m</sub> Me; Ar; (CH <sub>2</sub> ) <sub>m</sub> Me Bz; Cl <sub>2</sub> C(O)Ph Bz Me; Et; (CH <sub>2</sub> ) <sub>9-11</sub> Me Ph; Bz; Allyl Et Bu; (CH <sub>2</sub> ) <sub>2</sub> OH; Ph Et; Bz Et Et Pr	Ar; OH; OEt; i-Pr; (CH <sub>2</sub> ) <sub>4-19</sub> Me i-Hex; Bu; (CH <sub>2</sub> ) <sub>6</sub> (c-CH <sub>2</sub> ) <sub>3-8</sub> Me (CH <sub>2</sub> ) <sub>2</sub> O(CH <sub>2</sub> ) <sub>2</sub> OH; (CH <sub>2</sub> ) <sub>6</sub> Ph (CH <sub>2</sub> ) <sub>m</sub> C(O)OMe; (CH <sub>2</sub> ) <sub>6</sub> OH (CH <sub>2</sub> ) <sub>m</sub> CHCH <sub>2</sub> ; (C <sub>6</sub> H <sub>4</sub> )OC <sub>n</sub> H <sub>2n+1</sub> (CH <sub>2</sub> ) <sub>m</sub> Me; Cl <sub>2</sub> (C <sub>6</sub> H <sub>4</sub> )OMe (CH <sub>2</sub> ) <sub>6</sub> OPh; (CH <sub>2</sub> ) <sub>6</sub> ONaphtyl (CH <sub>2</sub> ) <sub>6</sub> O(CH <sub>2</sub> ) <sub>m</sub> Me; (CH <sub>2</sub> ) <sub>6</sub> COOH Allyl; Ph i-Am; (CH <sub>2</sub> ) <sub>2</sub> Me (CH <sub>2</sub> ) <sub>6</sub> CH(OH)(CH <sub>2</sub> ) <sub>m</sub> Me (CH <sub>2</sub> ) <sub>15</sub> Me Bu; Hep; (CH <sub>2</sub> ) <sub>2</sub> OCH <sub>2</sub> Me; Bz CH <sub>2</sub> C(O)(CH <sub>2</sub> ) <sub>3</sub> Me; CH <sub>2</sub> CCH Bz	

Figure 1. Typical structures of quaternary nitrogen-containing organic bromides in the parent data set.

“sequences” and “augmented atoms” are used in ISIDA as descriptors. The hydrogen atoms were omitted. Large pools of SMF descriptors generated for the *FULL* and sets 1–3, are respectively 13 538, 11 741, 2722, and 2020.

**3.1.2. Variable Selection.** ISIDA uses three steps procedure of variables selection to build statistically significant models.

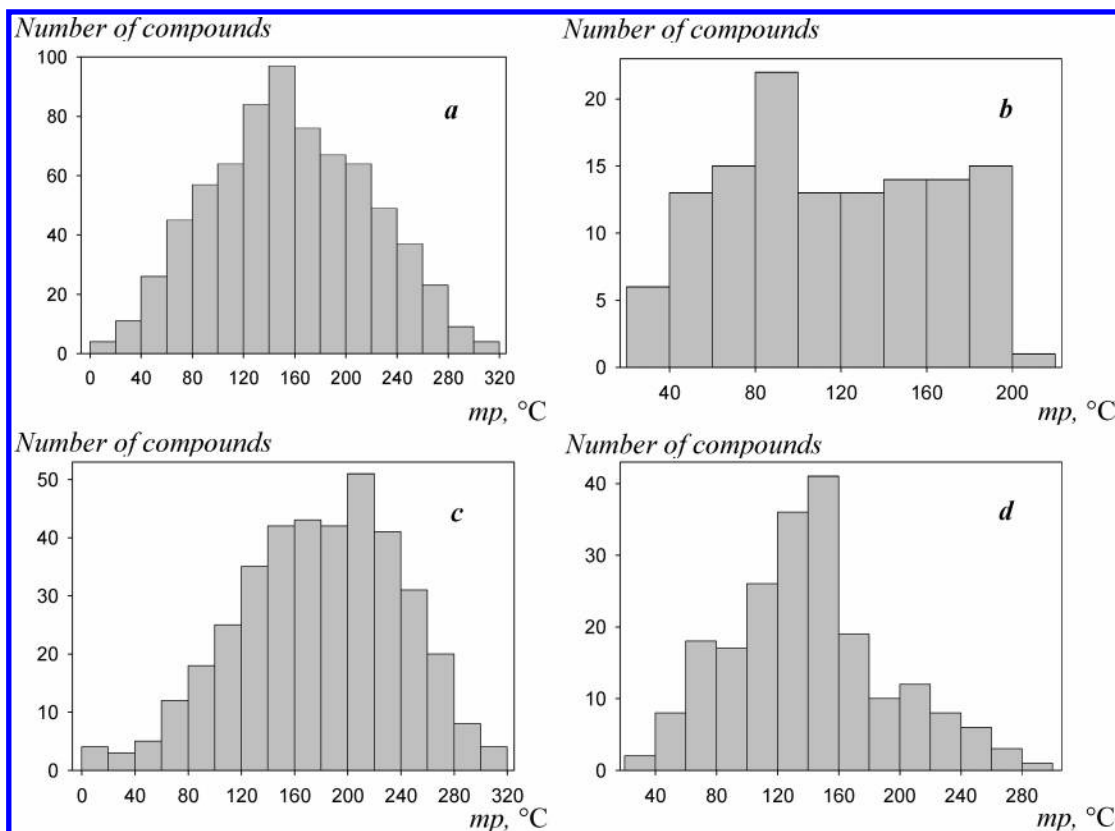
(1). *Filtering Stage.* The program eliminates variables  $X_i$  which have a small correlation coefficient with the property,  $R_{y,i} < R^0_{y,i}$ , and those highly correlated with other variables  $X_j$  ( $R_{i,j} < R^0_{i,j}$ ). In this work, the values  $R^0_{y,i} = 0.001$  and  $R^0_{i,j} = 0.99$  were used. Our experience shows that smaller threshold  $R^0_{i,j}$  does not always lead to improvement of the models. The fragments of “rare” occurrence (i.e., found in less than  $m$  molecules, here  $m \geq 2$ ) are excluded from the training set.

(2). *Forward Stepwise Preselection Stage.* This is an iterative procedure, on each step of which the program selects two variables  $X_i$  and  $X_j$  maximizing the correlation coefficient  $R_{y,ij}$  between  $X_i$  and  $X_j$  and dependent variable  $Y$ ,  $R_{y,ij} = (R^2_{y,i} + R^2_{y,j} - 2R_{y,i}R_{y,j}R_{ij})/(1 - R^2_{ij})$ . At the first step, the modeled experimental property  $Y_{\text{exp}}$  is taken as  $Y$ . At each next step,  $Y = Y - Y_{\text{calc}}$ , where  $Y_{\text{calc}} = a_0 + a_iX_i + a_jX_j$  is calculated

with the two-variables model found for  $Y$  using selected variables  $X_i$  and  $X_j$ . This loop is repeated until the number of variables  $N$  reaches the user’s defined value; by default,  $N$  is calculated as 60% of the number of molecules in the training set.

(3). *Backward Stepwise Selection Stage.* The final selection is performed using backward stepwise variable selection procedure based on  $t$  statistic criterion.<sup>22,24,25</sup> Here, the program eliminates the variables with low values of  $t$  statistic criterion  $t_i = a_i/s_i$  where  $s_i$  is the standard deviation for the coefficient  $a_i$  at the  $i$ th variable in the model. First, the program selects the variable with the smallest  $t < t_0$ , then it performs a new fitting excluding that variable. This procedure is repeated until  $t \geq t_0$  for selected variables or if the number of variables reaches the user’s defined value. The tabulated value of Student’s  $t_0$  criterion is a function of the number of data points, the number of variables, and the significance level.

**3.1.3. Structure–Property Modeling with ISIDA.** Selected descriptors are used by ISIDA to build multilinear correlation equation  $Y_{\text{exp}} = a_0 + \sum a_iX_i$ , where  $X_i$  is the value of  $i$ th descriptor,  $a_i$  is its contribution, and  $a_0$  is the descriptor



**Figure 2.** Distribution of melting point values in the data sets of bromides of quaternary nitrogen-containing organic cations used for QSPR modeling: (a) *FULL*: 717 compounds,  $mp = 5.5\text{--}319\text{ }^{\circ}\text{C}$ ; (b) *PYR*: 126 compounds,  $mp = 30\text{--}200\text{ }^{\circ}\text{C}$ ; (c) *IMZ*: 384 compounds,  $mp = 5.5\text{--}319\text{ }^{\circ}\text{C}$ ; and (d) *QUAT*: 207 compounds,  $mp = 39\text{--}281\text{ }^{\circ}\text{C}$ .

independent term. Using the Singular Value Decomposition method (SVD) the program fits the  $a_0$  and  $a_i$  terms and calculates the corresponding statistical characteristics: correlation coefficient ( $R$ ) and LOO cross-validation correlation coefficient ( $Q$ ), root mean squared error (RMSE), and mean absolute error (MAE) for training and test sets.

ISIDA possesses an interesting option to calculate a Consensus Model (CM) combining the information issued from several individual models originated from different pools of fragment descriptors. The program can generate several subsets of fragment descriptors corresponding either to sequences of particular length from  $n_{\min}$  to  $n_{\max}$  atoms containing atoms and bonds, atoms only or bonds only, or to augmented atoms. The total number of all possible types of sequences of different length is 315, that of augmented atoms types are four. Thus, ISIDA generates 319 sets of descriptors, each of them could be involved in one QSPR model. The idea is to use simultaneously  $M$  “best” models for which  $Q^2 \geq Q^2_{\text{lim}}$ , where  $Q^2_{\text{lim}} = 0.6$  is a user defined threshold. Thus, for each compound from the test set, the program computes the property as an arithmetic mean of values obtained with these  $M$  models excluding those leading to outlying values (Grubbs’s test is used<sup>26</sup>). Our experience shows<sup>22,27,28</sup> that such an ensemble modeling allows one to smooth inaccuracies of individual models.

**3.2. NASAWIN Program.** The NASAWIN (Neural Approach to Structure–activity studies for WINdows) software<sup>15,16</sup> allows constructing of QSPR/QSAR/SAR models using various neural and non-neural statistical approaches and different types of descriptors.

**3.2.1. Descriptors.** Two sets of descriptors were used in this study. The first one includes the fragment descriptors implemented in the FRAGMENT module of the NASAWIN software,<sup>15,16</sup> while the second one consists of the FRAGPROP descriptors computed by combining values of atomic properties within different substructural fragments.

The FRAGMENT descriptors<sup>29–31</sup> represent occurrences of different substructural fragments belonging to the following generic types: chains, branches, cycles, bicycles, tri-cycles, and arbitrary.<sup>30</sup> In this study, only chains, branches, and cycles were considered. In addition, each atom in the fragments can be described using up to four levels of classification in accordance with its neighborhood.<sup>29</sup> Specification of each fragment includes also sequence of bond orders. The FRAGPROP descriptors are computed by combining values of several atomic properties (the number of electrons and lone pairs, atomic radius, electronegativity, ionization potential, etc.) within different substructural fragments (chains containing up to 5 atoms) where each combination has some physical meaning. Hydrogen atoms are represented explicitly in FRAPROP but not in FRAGMENT type descriptors.

**3.2.3. Structure–Property Modeling.** Three statistical and machine learning methods implemented in NASAWIN were used in this study: the fast stepwise multiple linear regression analysis (FSMLR), a modified version of the partial least-squares analysis (PLSM), and the standard backpropagation neural network (BPNN).

FSMLR is an original method for stepwise construction of linear models. At the beginning, a parent data set is split



into three subsets: training set, internal validation set, and external test set. At each iteration, a new descriptor exhibiting the strongest correlation with the current error vector is added to the list of already selected descriptors; then the resulting model is used to calculate a new error vector which will be applied for selecting a new descriptor at the next step. This process of the stepwise model formation is stopped at achieving the best prediction performance on the internal validation set, whereas the overall performance of the model is assessed on the external test set.

PLSM is the modified method for performing the PLS-type analysis for the case of only one output variable. The method is based on stepwise construction of the orthogonal set of latent variables with projection of current error vector as well as descriptor vectors on corresponding perpendicular hyperplanes. The optimal number of the latent variables is chosen using the criterion of the maximal values of  $Q^2$  computed using a multifold cross-validation procedure. Although this method does not correspond exactly to the standard NIPALS procedure for PLS analysis reported in refs 32 and 33, it is very efficient and can be easily implemented.

BPNN represents standard multilayered feed-forward neural networks with backpropagation of errors.<sup>34</sup> The training of the networks is based on the resilient propagation algorithm.<sup>35,36</sup> The training procedure stops when the error of predictions for the compounds from the internal validation set reaches its minimum. Finally, the models are validated on the compounds from the external test set.<sup>37</sup>

**3.3. VCCLAB Suite.** The Virtual Computational Chemistry Laboratory software is available at <http://www.vcclab.org>. Besides, a number of Perl scripts and Java classes have been developed to automate the data processing.

**3.2.1. Descriptors.** Three types of descriptors were considered for the analysis: atom type E-state indices, atom type E-state counts, and molecular descriptors generated by the DRAGON program.

**Atom Type E-State Indices.** The electrotopological state (E-state) indices introduced by Hall and Kier<sup>38,39</sup> combine together both electronic and topological characteristics of the analyzed molecules. For each atom type in a molecule the E-state indices are summed and are used in a group contribution manner. In this study we used an extended set of atom-type E-state indices which was developed to better cover functional groups and neighborhood of nitrogen and oxygen atoms.<sup>40,41</sup>

**Atom Type E-State Counts.** Several recent studies indicated<sup>42–44</sup> that a use of atom counts corresponding to atom types determined for E-state indices could provide models with similar or even higher prediction ability<sup>27</sup> to the models developed using the E-state indices. Therefore we also considered counts of atoms corresponding to E-state indices as an additional set of descriptors. In order to distinguish these two types of descriptors we will refer to E-state indices as “E-state values” and to atom counts corresponding to them as “E-state counts”. The atom-type E-state indices and their counts were calculated using the program available at <http://www.vcclab.org/lab/pclient>.<sup>45</sup> Similarly to our previous studies,<sup>42,46</sup> molecular weight (MW) and the number of non-hydrogen atoms (NA) were used as additional descriptors.

**Dragon Descriptors.** The E-Dragon applet<sup>45</sup> available at <http://www.vcclab.org/lab/edragon> was used to generate 1666

descriptors for the parent data set. The E-Dragon provides WWW online interface to Dragon 5.4 software developed by Todeschini et al.<sup>47</sup> The calculated descriptors belonging to 20 major groups, covering topological, molecular, and 3D properties of molecules. Their full list is available at <http://www.disat.unimib.it/chm/Help/edragon/index.html>, and detailed calculation procedures are described in the book.<sup>47</sup> Prior to calculation of the descriptors, the molecules were converted from 2D structures in SD format to 3D structures using Corina program,<sup>48</sup> which is integrated in the E-Dragon applet. Because of limited maximum size of molecules for E-Dragon (150 atoms including hydrogens) and failure of Corina to perform 2D to 3D conversion, five molecules were excluded from calculations.

**Variables Pretreatment.** Prior to analysis with all methods, except Support Vector Machines, we excluded highly correlated ( $R > 0.95$ ) and near constant variables (each variable was required to have nonconstant values for at least 5 molecules). For the SVM method we excluded only constant variables. The variables were normalized to [0, 1] interval range. For the Associative Neural Network (ASNN) we also normalized the target values on [0.2, 0.8] interval.

**3.3.3. Structure–Property Modeling.** The open source LibSVM package<sup>49</sup> was used to perform *Support Vectors Machine* (SVM) calculations. The performance of SVM depends on several internal parameters of the algorithm ( $C$  and  $\epsilon$ ) and type of the kernel as well as parameters of the kernel. Here, only the RBF kernel was used. The grid search considered two parameters  $C = 2^{-5}, 2^{-3}, \dots, 2^{15}$  and  $\epsilon = 0.0001, 0.001, \dots, 10$  and width of the RBF kernel  $\gamma = 2^{-15}, 2^{-13}, \dots, 2^3$  as recommended in LibSVM manual. Thus for each data split set we performed  $11 \times 6 \times 9 = 594$  SVM runs. The parameters providing optimal performance of SVM for the cross-validation training set were applied to predict *mp* for the molecules from the corresponding test set. The internal cross-validation results are reported as the SVM accuracy for the training sets.

**The *k*-Nearest Neighbor Method (*k*NN)** method predicts activity of the target compounds as an average value of activities of its  $k$ -nearest neighbors in space of input descriptors. The Euclidian distance was used. The number of neighbors,  $k$ , was optimized in the range from 1 to 100, using corresponding cross-validation training sets and then applied to predict the corresponding cross-validation test sets. The *k*NN method was programmed in house. The Leave-One-Out results calculated in internal cross-validation were reported as the *k*NN accuracy for the training sets.

**Associative Neural Network (ASNN)** represents a combination of an ensemble of feed-forward neural networks and the *k*NN. This method uses the correlation between ensemble responses (each molecule is represented in space of neural network models as a vector of model predictions) as a measure of distance amid the analyzed cases for the nearest neighbor technique. Thus ASNN performs *k*NN in space of ensemble residuals. This provides an improved prediction by the bias correction of the neural network ensemble.<sup>50</sup> The neural networks ensemble of 50 networks with one hidden layer was used. After a few preliminary runs we fixed three hidden neurons for all data sets. The Efficient Partition Algorithm was used to train the neural network ensemble.<sup>51</sup> The calculations were performed using the program available at <http://www.vcclab.org/lab/asnn>. The leave-one-out results

**Table 1.** QSPR Modeling of Melting Point (*mp*) for 717 Compounds (*FULL*) Statistical Parameters of the Models at the Training and Validation Stages<sup>a</sup>

					training set		combined test set <sup>c</sup>		
no.	program	method	descriptors	no. of variables	$R^2(\text{fit})^d$	RMSE(fit) <sup>d</sup>	$R^2$	RMSE	MAE
1	VCCLAB	ASNN	E-state indices	110–113	$0.66 \pm 0.02$	$36 \pm 1$	0.61	39.1	30.2
2		KNN		110–113	$0.51 \pm 0.02$	$44 \pm 1$	0.53	42.8	33
3		MLR		110–113	$0.61 \pm 0.03$	$38 \pm 2$	0.54	42.4	33.8
4		SVM		208	$0.57 \pm 0.01$	$40.4 \pm 0.6$	0.59	39.9	31.2
5		ASNN	E-state counts	114–117	$0.65 \pm 0.01$	$36.5 \pm 0.7$	0.61	38.8	30.2
6		KNN		114–117	$0.53 \pm 0.01$	$42.8 \pm 0.8$	0.52	43	32.9
7		MLR		114–117	$0.6 \pm 0.02$	$39 \pm 1$	0.44	46.4	36.4
8		SVM		208	$0.59 \pm 0.01$	$39.9 \pm 0.6$	0.61	39.1	30
9	NASAWIN	ASNN	Dragon <sup>b</sup>	778–788	$0.71 \pm 0.01$	$33.2 \pm 0.7$	0.62	38.5	29.4
10		KNN		778–788	$0.52 \pm 0.01$	$43.1 \pm 0.5$	0.54	42.1	32.1
11		MLR		1435	$0.77 \pm 0.04$	$29 \pm 2$	0.55	41.7	32.9
12		SVM		778–788	$0.61 \pm 0.01$	$38.4 \pm 0.4$	0.63	37.5	28.9
13		PLSM	Fr-6–6 + Fragprop	3717	$0.75 \pm 0.01$	$31.2 \pm 1.0$	0.58	40.3	31.9
14		FSMLR		6–86	$0.71 \pm 0.09$	$32.8 \pm 5.5$	0.52	42.9	33.7
15		BPNN		2837	$0.77 \pm 0.04$	$30.0 \pm 2.9$	0.58	39.9	31.5
16		MLR-CM		35–149	$0.85 \pm 0.01$	$24.2 \pm 1.0$	0.53	42.6	32.3
no regression							0	61.9	50.6

<sup>a</sup> Statistical parameters: squared correlation coefficient  $R^2(\text{fit})$  and root-mean squared error, RMSE (at the fitting stage), and squared predictive correlation coefficient,  $R^2 = 1 - \Sigma(Y_{\text{exp}} - Y_{\text{pred}})^2 / \Sigma(Y_{\text{exp}} - \langle Y \rangle_{\text{exp}})^2$ , mean absolute error, MAE (at the validation stage). <sup>b</sup> VCCLAB and NASAWIN calculations: The number of compounds in the training sets depends on the program applied because of using of internal test sets in NASAWIN and ASNN programs (see Computational Procedure section). Five compounds could not be processed by Dragon/Corina, and thus the total number of 712 compounds was available for calculations with Dragon descriptors. ISIDA calculations: results were obtained using consensus model (MLR-CM) calculated as an arithmetic average of selected linear individual models with  $Q^2 > 0.6$ . <sup>c</sup> The test set represents a combination of all 5 external test sets. Thus, it contains all molecules from the initial parent set (see Computational Procedure section). <sup>d</sup> Arithmetic mean value and standard deviation for the five training sets.

calculated for neural networks as described elsewhere<sup>52</sup> were reported as the method accuracy for the training set.

The Singular Value Decomposition (MLR-SVD) algorithm can be efficiently applied to solve an overdetermined system of linear equations, i.e., when the number of variables is much larger than the number of data. The algorithm is well described in a number of books, see, e.g., Press et al.<sup>53</sup> During the SVD analysis of overdetermined matrices one has to set to zero a number of small diagonal elements of the SVD weight matrix  $w_j$  (which can be roughly considered as analogs of small and nonsignificant weights in multiple linear regression). While the ISIDA/SVD routine uses a fixed value  $w_{\text{cut}} = 10^{-12}$  to 0 the element of the inverse matrix, VCCLAB optimizes this value by a two-step grid search using internal 5-fold cross-validation. The first grid searches  $w$ -values in the range  $w = \{10^{-12}, 10^{-11}, \dots, 1\}$ . When an approximate value of the parameter  $w_0$  is found, the program performs the second grid search in the range  $w_0/5^* \{1, 1.5, 1.5^2, \dots, 1.5^8\}$  in order to determine the final  $w_{\text{cut}}$  value more precisely. The selected  $w_{\text{cut}}$  value is then used to predict the property values for the compounds from the test set.

**3.4. Validation of QSPR Models. Internal Validation.** All compounds in each initial data set were randomly shuffled to avoid possible artificial ordering due to data preparation. Fivefold cross-validation procedure has been applied to examine the efficiency of developed models. The parent data set was divided into five subsets: the first, sixth, eleventh, etc. entries form the first subset (#1); the second, seventh, twelfth, etc. entries form the second subset (#2); the third, eighth, thirteenth, etc. form the third subset (#3); the fourth, ninth, fourteenth, etc. form the fourth subset (#4); and the fifth, tenth, fifteenth, etc. form the fifth subset (#5). Five training sets were prepared as a combination of four subsets: set I (#1 - #4), set II (#1 - #3 and #5), set III (#1 - #2 and #4 - #5), set IV (#1 and #3 - #5), and set V (#2 -

#5). For each training set the remaining subset constituted the test set. Thus, each compound in the parent set was included in one of the test sets. The models were developed on a training set followed by prediction on the corresponding test set. Finally, all values calculated for five test sets were merged into one file to analyze overall linear correlations between experimental and predicted melting points. Statistical parameters for these correlations (predictive correlation coefficient ( $R^2 = 1 - \Sigma(Y_{\text{exp}} - Y_{\text{pred}})^2 / \Sigma(Y_{\text{exp}} - \langle Y \rangle_{\text{exp}})^2$ ), root-mean squared error (RMSE), and mean absolute error (MAE) for the combined test set are given in Tables 1–4.

**“No Regression” Calculations.** In order to provide evidence statistical significance of the models, a “no regression” (NR) approach<sup>51</sup> was used as a null hypothesis. In NR calculations, an average over a given set property’s value is taken as the predicted one for each compound. This “model” has no prediction power because the corresponding correlation coefficient with the target activity  $R^2 = 0$ . Thus, a given model was considered as significant if it leads to a significantly small ( $p < 0.05$ ) mean absolute error compared to the MAE of NR according to the bootstrap test based on 10 000 replicas (see ref 27 for details).

#### 4. RESULTS

Tables 1–4 contain statistical characteristics of the models of different approaches: arithmetic mean values of the correlation coefficient  $R^2(\text{fit})$  and root mean squared error RMSE(fit) for 5-fold cross-validation training sets, cross-validation correlation coefficient  $R^2$ , RMSE, and mean absolute error MAE for 5-fold external cross-validation test sets.

**4.1. ISIDA Calculations.** A large number of fragment descriptors has been generated: 13 538 (*FULL*), 2020 (*PYR*), 11 741 (*IMZ*), and 2722 (*QUAT*). Each initial pool has been

**Table 2.** QSPR Modeling of Melting Point (*mp*) for 126 Compounds (*PYR*): Statistical Parameters of the Models at the Training and Validation Stages<sup>a</sup>

no.	program	method	descriptors	no. of variables	training set		combined test set		
					$R^2(\text{fit})$	RMSE(fit)	$R^2$	RMSE	MAE
1	VCCLAB	ASNN	E-state indices	44–46	$0.7 \pm 0.04$	$27 \pm 2$	0.6	30.9	23.7
2		KNN		44–46	$0.52 \pm 0.04$	$34 \pm 1$	0.54	33.2	26.4
3		MLR		44–46	$0.69 \pm 0.03$	$27 \pm 1$	0.5	34.6	27.5
4		SVM		111	$0.64 \pm 0.02$	$29.2 \pm 0.7$	0.62	30.1	23.9
5		ASNN	E-state counts	48–50	$0.69 \pm 0.03$	$27 \pm 1$	0.61	30.5	23.4
6		KNN		48–50	$0.52 \pm 0.03$	$34 \pm 1$	0.51	34.2	26.5
7		MLR		48–50	$0.7 \pm 0.03$	$27 \pm 1$	0.53	33.6	26.6
8		SVM		111	$0.65 \pm 0.01$	$28.9 \pm 0.6$	0.61	30.6	24.5
9		ASNN	Dragon	647–667	$0.81 \pm 0.01$	$21.3 \pm 0.6$	0.71	26.4	19.1
10		KNN		647–667	$0.7 \pm 0.02$	$27 \pm 1$	0.69	27.5	22
11		MLR		647–667	$0.84 \pm 0.04$	$20 \pm 2$	0.67	28.3	21.3
12		SVM		1334	$0.7 \pm 0.05$	$27 \pm 2$	0.7	26.7	20.2
13	NASAWIN	PLSM	Fr-6–3 + Fragprop	787	$0.78 \pm 0.11$	$23.6 \pm 7.1$	0.55	32.5	26.0
14		FSMLR		1–15	$0.71 \pm 0.14$	$25.6 \pm 6.5$	0.48	34.8	
15		BPNN		787	$0.88 \pm 0.06$	$16.5 \pm 4.3$	0.71	26.2	20.2
16	ISIDA	MLR-CM	SMF	7–41	$0.92 \pm 0.02$	$14.4 \pm 1.6$	0.65	28.7	22.7
	no regression						0	48.6	41.7

<sup>a</sup> See footnotes for Table 1.**Table 3.** QSPR Modeling of Melting Point (*mp*) for 384 Compounds (*IMZ*): Statistical Parameters of the Models at the Training and Validation Stages<sup>a</sup>

no.	program	method	descriptors	no. of variables	training set		combined test set		
					$R^2(\text{fit})$	RMSE(fit)	$R^2$	RMSE	MAE
1	VCCLAB	ASNN	E-state indices	92–94	$0.62 \pm 0.02$	$37.5 \pm 0.7$	0.58	39.9	31.5
2		KNN		92–94	$0.47 \pm 0.02$	$45 \pm 1$	0.45	45.3	35.5
3		MLR		92–94	$0.59 \pm 0.05$	$39 \pm 2$	0.53	42.3	33.4
4		SVM		186	$0.54 \pm 0.02$	$41.6 \pm 0.8$	0.56	40.7	31.2
5		ASNN	E-state counts	94–98	$0.63 \pm 0.02$	$37.2 \pm 0.8$	0.58	39.8	31.5
6		KNN		94–98	$0.46 \pm 0.02$	$45 \pm 1$	0.46	45.1	34.3
7		MLR		94–98	$0.59 \pm 0.04$	$39 \pm 2$	0.49	43.9	35.2
8		SVM		186	$0.55 \pm 0.02$	$41 \pm 1$	0.57	40.5	31.8
9		ASNN	Dragon <sup>1</sup>	793–801	$0.64 \pm 0.02$	$37 \pm 1$	0.54	41.3	32.4
10		KNN		793–801	$0.45 \pm 0.03$	$45 \pm 1$	0.47	44.1	34.9
11		MLR		793–801	$0.69 \pm 0.03$	$34 \pm 2$	0.51	42.5	34
12		SVM		1418	$0.56 \pm 0.03$	$40 \pm 1$	0.54	41.3	32.3
13	NASAWIN	PLSM	FR-6–5 + Fragprop	3038	$0.80 \pm 0.01$	$27.5 \pm 0.5$	0.55	40.9	31.9
14		FSMLR		5–43	$0.73 \pm 0.10$	$31.4 \pm 5.6$	0.42	45.9	36.2
15		BPNN		3038	$0.75 \pm 0.04$	$30.5 \pm 2.3$	0.54	41.3	32.4
16	ISIDA	MLR-CM	SMF	25–97	$0.88 \pm 0.01$	$21.7 \pm 0.6$	0.50	43.3	32.1
	no regression						0	61.1	49.5

<sup>a</sup> See footnotes for Table 1. <sup>b</sup> VCCLAB calculations: 1–5 compounds could not be processed by Dragon/Corina, and thus the total number of 379 compounds was available for Dragon descriptors.

split into several subsets corresponding to a given type fragmentation. Each individual model involved descriptors from one subset. The number of descriptors per model varied as a function of the size and nature of compounds in the training set: 35–149 (*FULL*), 7–41 (*PYR*), 25–97 (*IMZ*), and 19–62 (*QUAT*). The *mp* values for each compound from the test set were assessed as an arithmetic average of selected individual models excluding outlying values according to Grubbs's statistics. Here the following number of individual models with  $Q^2 > 0.6$  were selected for consensus model calculations: 16–39 (*FULL*), 117–217 (*PYR*), 57–116 (*IMZ*), and 4–30 (*QUAT*). Consensus models calculations demonstrate reasonable predictive ability: for external test sets RMSE and MAE are, respectively, 42.6 and 32.3 (*FULL*), 28.7 and 22.7 (*PYR*), and 43.3 and 32.1 (*IMZ*). Only 4–30 individual models have been selected for consensus model calculations on *QUAT*, which led to larger error values (RMSE = 48.1 and MAE = 35.5) compared to other data sets (Tables 1–4).

Calculations reveal importance of long atom/bond sequences containing up to 9 atoms used as descriptors in QSPR models for *mp*.

**4.2. NASAWIN Calculations.** The FRAGPROP descriptors were chosen in this study since some of them have shown strong individual correlations with the modeled property. The “best” FRAGPROP descriptors exhibiting the strongest individual correlation with the melting point for all training sets were the mean value of the products of atomic radii of atoms separated by two, three, or four bonds. Preliminary studies with the variables selection on internal test sets show also the importance of branched 6-atoms fragments; therefore, a set of fragments containing from one to six atoms was chosen in this study. A threshold of the variance of the fragments, i.e., the number of compounds in the data set with nonzero values of their occurrence, was used in the PLSM-analysis: three for *PYR*, five for *IMZ*, four for *QUAT*, and six for the *FULL*. Additionally, in BPNN calculations, descriptors highly correlated with already

**Table 4.** QSPR Modeling of Melting Point (*mp*) for 207 Compounds (*QUAT*): Statistical Parameters of the Models at the Training and Validation Stages<sup>a</sup>

					training set		combined test set		
no.	program	method	descriptors	no. of variables	$R^2(\text{fit})$	RMSE(fit)	$R^2$	RMSE	MAE
1	VCCLAB	ASNN	E-state indices	51–57	$0.54 \pm 0.05$	$36 \pm 2$	0.45	38.9	30.9
2		KNN		51–57	$0.4 \pm 0.04$	$41 \pm 1$	0.42	39.9	32.1
3		MLR		51–57	$0.49 \pm 0.13$	$37 \pm 4$	0.3	43.7	34.5
4		SVM		89	$0.42 \pm 0.04$	$40 \pm 1$	0.41	40.2	31
5		ASNN	E-state counts	55–59	$0.52 \pm 0.04$	$36 \pm 2$	0.46	38.6	30.6
6		KNN		55–59	$0.43 \pm 0.03$	$40 \pm 1$	0.36	41.9	33.9
7		MLR		55–59	$0.44 \pm 0.05$	$39 \pm 1$	0.11	49.3	37.9
8		SVM		89	$0.43 \pm 0.03$	$40 \pm 1$	0.42	39.9	30.5
9	NASAWIN	ASNN	Dragon <sup>1</sup>	618–628	$0.61 \pm 0.02$	$32.6 \pm 0.5$	0.52	36.1	28.2
10		KNN		618–628	$0.39 \pm 0.05$	$41 \pm 1$	0.31	43.3	32.6
11		MLR		618–628	$0.67 \pm 0.07$	$29 \pm 3$	0.44	39.3	31.2
12		SVM		1342	$0.47 \pm 0.05$	$38 \pm 1$	0.57	34.2	26.5
13		PLSM	FR-6–4 + Fragprop	994	$0.64 \pm 0.14$	$30.9 \pm 5.3$	0.37	41.1	31.8
14		FSMLR		1–18	$0.59 \pm 0.18$	$32.8 \pm 7.1$	0.18	46.2	36.1
15		BPNN		994	$0.71 \pm 0.09$	$28.5 \pm 4.5$	0.46	38.0	30.3
16		MLR-CM		19–62	$0.88 \pm 0.03$	$17.8 \pm 1.8$	0.15	48.1	35.5
	no regression						0	52.1	40.6

<sup>a</sup> See footnotes for Table 1.

selected ones ( $R^2 > 0.99$ ) were excluded. The architecture with one hidden layer containing two neurons (and one bias pseudoneuron) was used in BPNN calculations. To obtain and validate the model with FSMLR and BPNN, 3 data sets are needed: the training and the internal validation and external test sets. The internal validation set is used either to select the most pertinent descriptors (FSMLR) or for early stopping (BPNN). In both cases, we used the internal and external test sets of the same size, so the overall splitting of a given parent set into training and internal validation and external test sets was in a ratio of 3:1:1. In the 5-fold cross-validation procedure, for each of the 5 external data sets 4 different internal test sets were considered in FSMLR; therefore, the obtained statistics was averaged over 20 models. For BPNN models for each external data set only one internal validation data set with the lowest prediction error on it was taken, and statistics were averaged over 25 models.

It should be noted that all BPNN calculations were performed on the initial pools of descriptors without any preliminary selection. Although the number of descriptors and, consequently, the number of adjustable parameters of neural network exceeded the number of compounds in the training sets, this has not deteriorated the predictive performance of the models. The point is that the use of internal validation sets for “stopping” the learning prevents overtraining and overfitting of models.<sup>52</sup> Moreover, our attempts to build neural networks based on small subsets of preselected with FSMLR descriptors resulted in considerable deterioration of the robustness of the models. The performance of BPNN is the best for relatively small subsets 1 and 3, whereas for larger *IMZ* and *FULL* this method is as good as PLSM (Tables 1–4). The variables selection procedure implemented in FSMLR is inefficient for relatively small data sets; however, for large data sets all methods display similar prediction ability (Tables 1–4).

**4.3. VCCLAB Calculations.** No variable selection has been performed except of the filtering of highly correlated and constant variables, as indicated in the methods section.

According to Wilcoxon matched-pairs signed-ranks test the ASNN and SVM nonlinear methods provided consistently

better prediction of the melting point of compounds from the external tests sets compared to MLR and *k*NN methods ( $p < 0.005$ ). No significant difference in the performance of ASNN vs SVM methods has been observed (Tables 1–4).

Unlike our previous study<sup>27</sup> we did not find any trend in the comparison of E-state indices and counts: in some cases models involving E-state indices were found more predictive; in other cases the opposite situation was observed (Tables 1–4).

The DRAGON descriptors provided better results compared to the models based on either the E-state count or indices ( $p < 0.05$ ) according to the Wilcoxon matched-pairs signed-ranks test.<sup>53</sup>

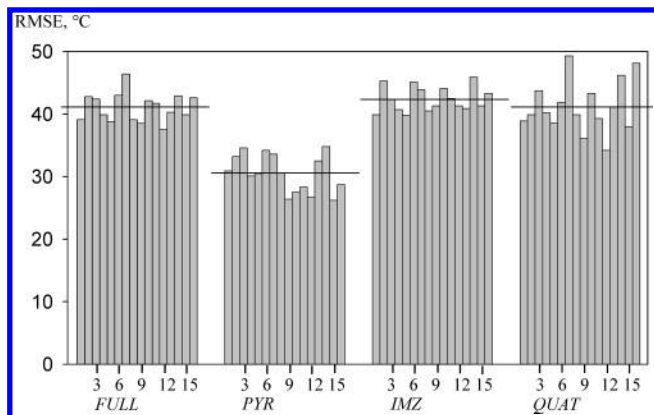
## 5. DISCUSSION

**5.1. Quality of Models as a Function of Machine Learning Methods and Descriptors.** A combination of different machine learning methods and various descriptor types resulted in 16 types of models for each data set. Statistical parameters obtained at the training and validation stages for *FULL*, *PYR*, *IMZ*, and *QUAT* sets are given, respectively, in Tables 1–4 and Figures 3–5. One can see that all models are significant since their RMSE and MAE are smaller than corresponding “no regression” values.

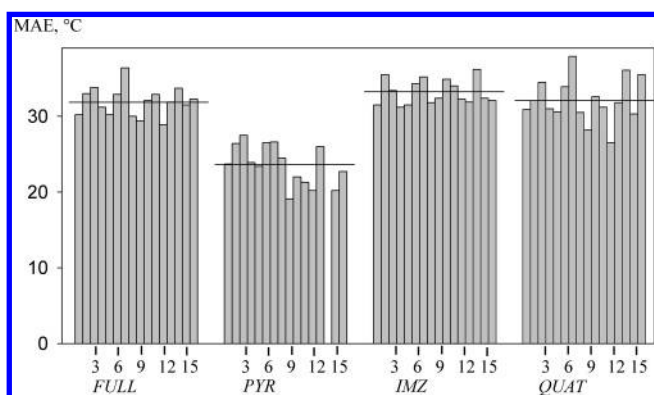
In order to compare the predictive performance of different approaches, the Student test for the standard deviation<sup>54</sup> (for RMSE and MAE) and the Fisher’s Z test for correlation coefficients at a significance level of 5%<sup>55</sup> (for  $R^2$ ) were used. This comparison shows that the list of the most efficient methods/descriptors combinations depends from the data set used. Thus, these are SVM, ASNN, BPNN, and PLSM methods for *FULL* (Table 1); SVM, ASNN, BPNN, *k*NN/DRAGON, MLR/DRAGON, and MLR-CM/ISIDA for *PYR* (Table 2); SVM, ASNN, BPNN, PLSM MLR/DRAGON, MLR/E-state indices, and MLR-CM/ISIDA for *IMZ* (Table 3); and SVM/DRAGON, ASNN, and BPNN for *QUAT* (Table 4).

In order to draw a conclusion concerning the performance of different machine learning methods, we applied a scoring function for RMSE, MAE, and  $R^2$  based on comparison

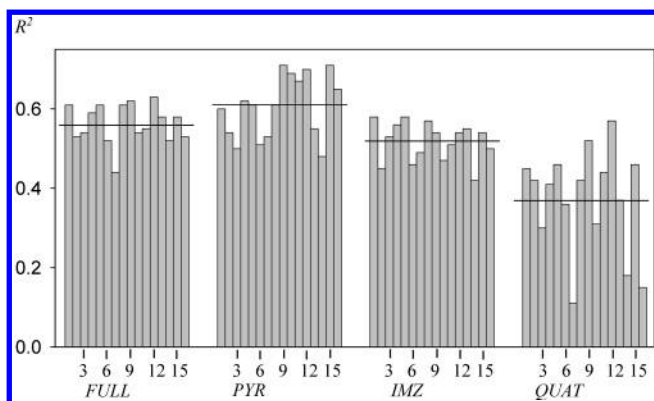




**Figure 3.** Root-mean squared error RMSE (°C) of prediction of melting point as a function of method for the full data set and its subsets, see Tables 1–4 for details. The numbers at the horizontal axis correspond to the methods numerations in Tables 1–4. Upper horizontal lines are average values of RMSE for each series.



**Figure 4.** Mean absolute error MAE (°C) of prediction of melting point as a function of method for the full data set and its subsets, see Tables 1–4 for details. The numbers at the horizontal axis correspond to the methods numerations in Tables 1–4. Upper horizontal lines are average values of MAE for each series.



**Figure 5.** Predictive correlation coefficient  $R^2 = 1 - \frac{\sum(Y_{\text{exp}} - Y_{\text{pred}})^2}{\sum(Y_{\text{exp}} - \langle Y \rangle_{\text{exp}})^2}$  as a function of the method the full data set and its subsets, see Tables 1–4 for details. The numbers at the horizontal axis correspond to the methods numerations in Tables 1–4. Upper horizontal lines are average values of  $R^2$  for each series.

of the given statistical parameter with its average value (RMSE<sub>av</sub>, MAE<sub>av</sub>, and  $R^2_{\text{av}}$ , respectively). The score “0” corresponds to the model for which RMSE < RMSE<sub>av</sub>, or MAE < MAE<sub>av</sub>, or  $R^2 > R^2_{\text{av}}$ ; otherwise the score is “1”. The score corresponding to a given statistical parameter sums up the scores for the full data set and its subsets (Table 5). The “total” score is a sum of these scores. The smaller the

**Table 5.** Comparison of Efficiency of Different Methods To Predict Melting Points of Ionic Liquids<sup>a</sup>

method	program	$R^2$	RMSE	MAE	total
BPNN	NASAWIN	0	0	0	0
ASNN/E-counts	VCCLAB	0	0	0	0
ASNN/Dragon	VCCLAB	0	0	0	0
SVM/Dragon	VCCLAB	0	0	0	0
SVM/E-state	VCCLAB	0	0	1	1
SVM/E-counts	VCCLAB	0	1	1	2
ASNN/E-state	VCCLAB	1	1	1	3
PLSM	NASAWIN	2	1	2	5
MLR/Dragon	VCCLAB	2	2	2	6
MLR-CM/SMF	ISIDA	3	3	2	8
kNN/Dragon	VCCLAB	3	3	3	9
FSMLR	NASAWIN	4	4	4	12
kNN/E-state	VCCLAB	3	3	4	10
MLR/E-state	VCCLAB	3	4	4	11
kNN/E-counts	VCCLAB	4	4	4	12
MLR/E-counts	VCCLAB	4	4	4	12

<sup>a</sup> Score for statistical parameters obtained for the external test sets. For any individual data set, the score “0” corresponds to the model for which  $R^2$  (external test) >  $R^2_{\text{av}}$  or RMSE < RMSE<sub>av</sub> or MAE < MAE<sub>av</sub>, otherwise the score is “1”. The average values of statistical parameters  $R^2_{\text{av}}$ , RMSE<sub>av</sub>, and MAE<sub>av</sub> (Table 4) are calculated for the ensemble of models in Tables 1–4. The score corresponding to a given statistical parameter sums up the scores for the full set and its subsets. The “total” score is a sum of these scores.

total score is, the more predictive is a given model. According to the data from Table 5, the most efficient models with the total score from 0 to 3 were obtained with neural networks (ASNN, BPNN) and support vector machine methods whatever descriptors were used. However, the difference between the performance of “good” and “less good” methods is relatively small. Thus, the differences of RMSE<sub>av</sub> and MAE<sub>av</sub> values obtained by averaging over the ensemble of all 16 models and those obtained for 7 “best” models does not exceed 3 °C.

Comparison of the descriptors’ performance is limited to VCCLAB calculations where three types of descriptors—E-state indices, E-state counts, and DRAGON’s descriptors—were systematically used together with ASNN, SVM, kNN, and MLR methods. These calculations resulted in 12 different types of models (Tables 1–4). Obtained results show that the descriptors’ performance depends both on the data set and on the machine learning method. Thus, for *FULL*, only for the MLR method E-state counts were found less efficient than other types of descriptors, whereas for ASNN, SVM, and kNN no preference was found. For *PYR*, for all methods, DRAGON’s descriptors lead to the most predictive models. No descriptors preferences were found in *IMZ* calculations whatever method was used. For *QUAT*, DRAGON’s descriptors were found more efficient in SVM and MLR calculations, E-state indices were the best in kNN calculations, whereas no preferences were found in ASNN calculations.

**5.2. Comparison with Previous QSPR Studies of Ionic Liquids.** Compared to previous QSPR studies of *mp* of ionic liquids, our calculations were performed on a much more structurally diverse data set. Below, we list some previous QSPR works in this field. Thus, Katritzky et al.<sup>1</sup> performed QSPR modeling on the set of 126 pyridinium bromides (*PYR* in this work) using multilinear regression analysis with CODESSA descriptors. Later on Carrera et al.<sup>12</sup> obtained the models for this set using regression trees and neural

networks and DRAGON descriptors.<sup>50</sup> Linear QSPR models with the CODESSA PRO program for small sets of 59, 29, and 19 substituted imidazolium bromides as well as for 45 benzoimidazolium bromides were reported in a paper by Katritzky et al.<sup>2</sup> Eike et al.<sup>7</sup> performed linear structure - *mp* modeling for 75 tetraalkylammonium bromides and 34 (*n*-hydroxyalkyl)trialkylammonium bromides using CODESSA PRO. Trohalaki et al.<sup>11</sup> used CODESSA PRO to establish linear correlations between some quantum and thermodynamic descriptors and *mp* for 13 1-substituted 4-amino-1,2,4-triazolium bromide and nitrate salts. In most of these publications, the quality of the models was assessed according to statistical parameters obtained at the training stage. External test sets in refs 12 and 7 contained very few compounds.

Unlike previous publications, we used a thorough 5-fold cross-validation procedure, where each compound from the initial parent set took part of one of the external test sets. So, instead of doing an assessment of the quality of fitting, we estimated the quality of predictions. It should be noted that 3-fold cross-validation in ref 12 has been performed using descriptors selected in preliminary QSPR modeling on the parent set. Thus, three test sets used in refs 2 and 12 were not independent, and the calculated statistical parameters could be overfitted.<sup>53</sup> Using ISIDA we performed the calculations following the protocol suggested in refs 2 and 12 which led to significant improvement of predictions. Thus, for *PYR* (126 pyridinium bromides) RMSE decreases from 28.3° (Table 2) to 26°, whereas for *QUAT* (207 ammonium bromides) more remarkable reduction of RMSE from 48.1° (Table 4) to 32.0° has been achieved.

**5.3. Splitting into Subsets: Does it Improve the Models' Performance?** Results reported in ref 56 show that QSPR modeling on congener subsets of the given parent set could represent a promising way to improve the robustness of predictions. Sometimes, this is the only way to obtain the models. Thus, in order to obtain reasonable structure - *mp* correlations Katritzky et al.<sup>2</sup> had to split the parent set of 104 substituted imidazolium and benzoimidazolium bromides into four subsets.

In this work, 5-fold cross-validation calculations were performed both on full set as well as on its three subsets containing bromides of substituted pyridines and of imidazolium and quaternary ammonium, respectively. Calculations show that this splitting in most of the cases does not improve the robustness of the models. Thus, calculations on *IMZ* (384 imidazolium and benzoimidazolium bromides) and *QUAT* (207 quaternary ammonium bromides) subsets did not decrease the error of prediction compared to the full set: the RMSE<sub>av</sub> values are 41.1 °C, 42.4 °C, and 41.2 °C, for *FULL*, *IMZ*, and *QUAT*, respectively. Only for *PYR* (26 pyridinium bromides) was a better performance of models (RMSE<sub>av</sub> = 30.6 °C) achieved.

The main problem here is a way of obtaining subsets from the initial parent set. Intuitive division of the parent set into different chemical classes does not always lead to models improvement. The "divide and conquer" approach based on joint application of clustering and QSPR techniques<sup>25</sup> might be useful as a valuable alternative to that purely chemical split.

**5.4. How accurately Can We Predict the Melting Point of Ionic Liquids?** The 5-fold cross-validation calculations led

to pretty large RMSE values (30–40 °C) whatever the method or descriptors used. This means that obtained QSPR models could be more useful to observe a trend of *mp* for the series of compounds than to accurately predict the properties of new materials. The question arises could one predict melting point with better accuracy.

In fact, the accurate estimation of melting points of ionic liquids faces two main problems. The first one concerns the quality of available experimental data. Sometimes, in the literature, different *mp* values are reported for the same compound. One should also avoid using the temperature of decomposition or glass transition instead of melting points. The IUPAC Ionic Liquids database<sup>57</sup> which has recently become available provides still only a small ensemble of validated data.

Another problem is related to the difficulties of depicting the phenomenon of melting which is a process of the passage from the solid to the liquid state of the species. Thus, descriptors used for the modeling should be able to describe quantitatively the energy of a crystal lattice, on one hand, and molecular interactions in the liquid phase, on the other hand. The main problem is related to the fact that a compound can be crystallized in several different forms corresponding to a different arrangement of atoms in the crystal (polymorphs) whose melting points could be very different. The situation becomes even more complicated if eutectic mixtures are formed. Even for purely organic compounds the difference between melting points for different polymorphs could exceed 30 °C.<sup>58,59</sup> Therefore, it is not surprising that RMSE for large diverse sets of organic compounds is 35–39 °C<sup>60,61</sup> and even 41–50 °C.<sup>62,63</sup> For ionic species studied in this work this difference could be larger because of long-range electrostatic interactions between the ions. In order to improve an efficiency of QSPR models, structural information about different polymorphs and related energy characteristics should be taken into account. However, these data are hardly available which imposes a limit of accuracy of predictions of *mp* for ionic liquids.

Generally speaking, one may assume that for any property *Y* there exists such a limit of prediction accuracy, which is determined by the nature of *Y* as well as by the quality and the amount of available experimental data. As soon as this limit is reached, any adequate machine learning method in combination with reasonable descriptors would provide the models having close predictive performances. In that case, the prediction accuracy of QSPR models is driven by data used for the learning rather than by machine learning methods.<sup>64</sup> Is it possible to overcome this "natural" limit of prediction accuracy of models imposed by available data? We believe that the multitask learning<sup>65</sup> might be a possible solution of this problem. The idea is to combine a data set for a given property *Y* with some other data sets for properties *Z<sub>i</sub>* closely related to *Y*.<sup>64</sup> Thus, the models for *Y* and *Z<sub>i</sub>* are developed simultaneously using common parameters of the fitting procedure like latent variables in PLS, hidden layers in neuron networks, etc. In such a way, information about *Z<sub>i</sub>* enriches the QSPR model for *Y*. Recently, this "data expansion" methodology has been successfully applied for predicting the boiling point and the enthalpy of evaporation for whose only low quality experimental data were available using a QSPR model for vapor pressure as a function of the

temperature built on high-quality experimental data.<sup>66</sup> An alternative way to improve the predictive performance of models could be based on the “gray-box” approach<sup>67</sup> in which machine learning techniques are combined with some fundamental physicochemical equations. It seems that this strategy is more appropriate for further QSPR studies of melting points of IL than for application of more and more advanced machine learning techniques.

## CONCLUSIONS

In this article we report an exhaustive structure - melting point study using different linear (MLR, MLR-CM) and nonlinear (SVM, ASNN, BPNN, kNN, and PLS) machine learning methods and different types of descriptors (molecular fragments, E-state indices, various molecular features) incorporated in ISIDA, NASAWIN, and VCCLAB programs. The models have been built on structurally diverse data set of 717 bromides of nitrogen-containing organic cations as well as on its three subsets containing bromides of substituted pyridines and of imidazolium and quaternary ammonium, respectively. A thorough 5-fold external cross-validation procedure used to assess the predictive ability of the models shows the slight preference of nonlinear SVM, ASNN, and BPNN techniques over other methods whatever descriptors were used. For the full set, the accuracy of predictions does not significantly change as a function of the type of descriptors. For other sets, the performance of descriptors varies as a function of method and data set used. Splitting the initial data set on several congener subsets does not always improve the robustness of the models. Practically, all calculations led to rather converging values of statistical parameters at the training and validation stages. Thus for the full data set, RMSE of prediction calculated on combination of five independent test sets is in the range 37.5–46.4 °C. This moderate accuracy of predictions can be related to the quality of the experimental data used for obtaining the models as well as to difficulties to take into account the structural features of IL in the solid state (polymorphic effects, eutectics, glass formation). It seems that further improvement of the performance of the models for *mp* of ionic liquids could be achieved using multitask<sup>68</sup> or/and gray-box<sup>69</sup> approaches rather than the application of more and more advanced machine learning techniques.

## ACKNOWLEDGMENT

We thank Dr. Isabelle Billard, Dr. Gilles Marcou, and Dr. Olga Klimchuk for fruitful discussions and help with the data preparation and Louis Pasteur University for the Invited Professor's position to I.V.T. GDR PARIS is acknowledged for the support. A part of this work has been performed in the framework of GDRE SupraChem and the ARCUS “Alsace-Russia/Ukraine” project.

**Supporting Information Available:** Structures and melting points of the initial parent set (717 compounds) and 3 of its subsets. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## REFERENCES AND NOTES

- Katritzky, A. R.; Lomaka, A.; Petrukhin, R.; Jain, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. QSPR Correlation of the Melting Point for Pyridinium Bromides, Potential Ionic Liquids. *J. Chem. Inf. Comput. Sci.* **2002**, 42 (1), 71–74.
- Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. Correlation of the Melting Points of Potential Ionic Liquids (Imidazolium Bromides and Benzimidazolium Bromides) Using the CODESSA Program. *J. Chem. Inf. Comput. Sci.* **2002**, 42 (2), 225–231.
- Ionic Liquids in Synthesis*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, 2002.
- Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; Karelson, M. Perspective on the Relationship between Melting Points and Chemical Structure. *Cryst. Growth Des.* **2001**, 1, 261–265.
- Holbrey, J. D.; Reichert, W. M.; Nieuwenhuyzen, M.; Johnston, S.; Seddon, K. R.; Rogers, R. D. Crystal Polymorphism in 1-Butyl-3-methylimidazolium Halides: Supporting Ionic Liquid Formation by Inhibition of Crystallization. *Chem. Commun.* **2003**, 1636–1637.
- Xu, W.; Cooper, E. I.; Angell, C. A. Ionic Liquids: Ion Mobilities, Glass Temperatures, and Fragilities. *J. Phys. Chem. B* **2003**, 107, 6170–6178.
- Eike, D.; Brennecke, J.; Maginn, E. Predicting Melting Points of Quaternary Ammonium Ionic Liquids. *Green Chem.* **2003**, 5, 323–328.
- Abraham, M.; Zissimos, M.; Huddleston, J.; Willauer, H.; Rogers, R. D.; Acree, W. Some Novel Liquid Partitioning Systems: Water-Ionic Liquids and Aqueous Biphasic Systems. *Ind. Eng. Chem. Res.* **2003**, 42, 413–418.
- Mathieu, D.; Becker, J.-P. Improved Evaluation of Liquid Densities Using van der Waals Molecular Models. *J. Phys. Chem.* **2006**, 110, 17182–17187.
- Trohalaki, S.; Pachter, R.; Drake, G.; Hawkins, T. Quantitative Structure-Property Relationships for Melting Points and Densities of Ionic Liquids. *Energy Fuels* **2005**, 19, 279–284.
- Trohalaki, S.; Pachter, R. Prediction of Melting Points for Ionic Liquids. *QSAR Comb. Sci.* **2005**, 24, 485–490.
- Carrera, G.; Aires-de-Sousa, J. Estimation of Melting Points of Pyridinium Bromide Ionic Liquids with Decision Trees and Neural Networks. *Green Chem.* **2004**, 7, 20–27.
- Matsumoto, H.; Kageyama, H.; Miyazaki, Y. Room Temperature Molten Salts Based on Tetraalkylammonium Cations and Bis-(trifluoromethylsulfonyl)imide. *Chem. Lett.* **2001**, 30 (2), 182.
- ISIDA (In Silico Design and Data Analysis) Software. <http://infochim.u-strasbg.fr/recherche/isida/index.php> (accessed 2006).
- Halberstam, N. M.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. NASAWIN - A Program Simulator of Neural Networks for Structure-Activity Relationship Studies. In *International Symposium CACR-96, Moscow, Russia, December 17–18, 1996*; pp 37–38.
- Baskin, I. I.; Halberstam, N. M.; Artemenko, N. V.; Palyulin, V. A.; Zefirov, N. S. NASAWIN – a Universal Software for QSPR/QSAR Studies. In *EuroQSAR 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions*, 2003; Ford, M., Ed.; Blackwell Publishing: 2003; pp 260–263.
- The Virtual Computational Chemistry Laboratory Software. <http://www.vcclab.org> (accessed 2006).
- Beilstein Information System, Version 4; GmbH: 1995–1998.
- Solov'ev, V. P.; Varnek, A.; Wipff, G. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. *J. Chem. Inf. Comput. Sci.* **2000**, 40 (3), 847–858.
- Varnek, A.; Wipff, G.; Solov'ev, V. P.; Solotnov, A. F. Assessment of the Macrocyclic Effect for the Complexation of Crown-Ethers with Alkali Cations Using the Substructural Molecular Fragments Method. *J. Chem. Inf. Comput. Sci.* **2002**, 42 (4), 812–829.
- Solov'ev, V. P.; Varnek, A. Anti-HIV Activity of HEPT, TIBO, and Cyclic Urea Derivatives: Structure-Property Studies, Focused Combinatorial Library Generation, and Hits Selection Using Substructural Molecular Fragments Method. *J. Chem. Inf. Comput. Sci.* **2003**, 43 (5), 1703–1719.
- Solov'ev, V. P.; Varnek, A. A. Structure-Property Modeling of Metal Binders Using Molecular Fragments. *Rus. Chem. Bull.* **2004**, 53 (7), 1434–1445.
- Varnek, A.; Solov'ev, V. P. “In Silico” Design of Potential Anti-HIV Actives Using Fragment Descriptors. *Comb. Chem. High Throughput Screening* **2005**, 8 (5), 403–416.
- Katritzky, A. R.; Kuanar, M.; Fara, D. C.; Karelson, M.; Acree, W. E.; Solov'ev, V. P.; Varnek, A. QSAR Modeling of Blood:Air and Tissue:Air Partition Coefficients Using Theoretical Descriptors. *Bioorg. Med. Chem.* **2005**, 13 (23), 6450–6463.
- Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput.-Aided Mol. Des.* **2005**, 19 (9–10), 693–703.
- Grubbs, F. E. Procedures for Detecting Outlying Observations in Samples. *Technometrics* **1969**, 11 (1), 1–21.



- (27) Tetko, I. V.; Solov'ev, V. P.; Antonov, A. V.; Yao, X. J.; Fan, B. T.; Hoonakker, F.; Fourches, D.; Lachiche, N.; Varnek, A. Benchmarking of Linear and Non-Linear Approaches for Quantitative Structure-Property Relationship Studies of Metal Complexation with Organic Ligands. *J. Chem. Inf. Model.* **2006**, *46* (2), 808–819.
- (28) Solov'ev, V. P.; Kireeva, N. V.; Tsivadze, A. Y.; Varnek, A. A. Structure-Property Modelling of Complex Formation of Strontium with Organic Ligands in Water. *J. Struct. Chem.* **2006**, *47* (2), 298–311.
- (29) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Computer Program for Generating Sets of Subgraphs for Molecular Graphs. In *Proceedings of the Conference 'Molecular Graphs in Chemistry Studies'*, Kalinin, 1990; Kalinin, 1990; p 5.
- (30) Artemenko, N. V.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Prediction of Physical Properties of Organic Compounds by Artificial Neural Networks in the Framework of Substructural Approach. *Dokl. Akad. Nauk SSSR (Russ.)* **2001**, *381* (2), 203–206.
- (31) Artemenko, N. V.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Artificial Neural Network and Fragmental Approach in Prediction of Physicochemical Properties of Organic Compounds. *Russ. Chem. Bull.* **2003**, *52* (1), 20–29.
- (32) Geladi, P.; Kowalski, B. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17.
- (33) Gustaffson, M. G. A Probabilistic Derivation of the Partial Least-Squares Algorithm. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 288–294.
- (34) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Internal Representations by Error Propagation. In *Parallel Data Processing*; Rumelhart, D. E., McClelland, J. L., Eds.; M.I.T. Press: Cambridge, 1986; Vol. 1, pp 318–362.
- (35) Patnaik, L. M.; Rajan, K. Target Detection Through Image Processing and Resilient Propagation Algorithms. *Neurocomputing* **2000**, *35*, 123–135.
- (36) Riedmiller, M.; Braun, H. In *A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm*, Proceedings of the IEEE International Conference on Neural Networks, 1993; 1993; pp 586–591.
- (37) Baskin, I. I.; Skvortsova, M. I.; Palyulin, V. A.; Zefirov, N. S. Quantitative Chemical Structure-Property/Activity Relationship Studies Using Artificial Neural Networks. *Found. Comput. Decision Sci.* **1997**, *22* (2), 107–116.
- (38) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: London, 1999.
- (39) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types - a Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (6), 1039–1045.
- (40) Huuskonen, J. J.; Livingstone, D. J.; Tetko, I. V. Neural Network Modeling for Estimation of Partition Coefficient Based on Atom-Type Electrotopological State Indexes. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (4), 947–955.
- (41) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (6), 1488–1493.
- (42) Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. In Silico Approaches to Prediction of Aqueous and DMSO Solubility of Drug-like Compounds: Trends, Problems and Solutions. *Curr. Med. Chem.* **2006**, *13* (2), 223–241.
- (43) Butina, D. Performance of Kier-Hall E-state Descriptors in Quantitative Structure Activity Relationship (QSAR) Studies of Multifunctional Molecules. *Molecules* **2004**, *9*, 1004–1009.
- (44) Taskinen, J.; Yliruusi, J. Prediction of Physicochemical Properties Based on Neural Network Modelling. *Adv. Drug. Delivery Rev.* **2003**, *55* (9), 1163–1183.
- (45) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual Computational Chemistry Laboratory - Design and Description. *J. Comput.-Aided Mol. Des.* **2005**, *19* (6), 453–463.
- (46) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. Prediction of n-Octanol/Water Partition Coefficients From PHYSPROP Database Using Artificial Neural Networks and E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1407–1421.
- (47) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; WILEY-VCH: Weinheim, 2000; p 667.
- (48) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (4), 1000–1008.
- (49) Chang, C. C.; Lee, C. J. *LIBSVM: a Library for Support Vector Machines*. <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (accessed 2005).
- (50) Tetko, I. V. Associative Neural Network. *Neural Process. Lett.* **2002**, *16* (2), 187–199.
- (51) Tetko, I. V.; Villa, A. E. P. Efficient Partition of Learning Data Sets for Neural Network Training. *Neural Networks* **1997**, *10* (8), 1361–1374.
- (52) Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural Network Studies. I. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (5), 826–833.
- (53) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd ed.; Cambridge, 2002; p 1002.
- (54) Doerffel, K. *Statistik in der analytischen Chemie*; Deutscher Verlag für Grundstoffindustrie GmbH: Leipzig, 1990; p 270.
- (55) Hawkins, D. L. Using U Statistics to Derive the Asymptotic Distribution of Fisher's Z Statistic. *Am. Statistician* **1989**, *43* (4), 235–237.
- (56) He, L.; Jurs, P. C. Assessing the Reliability of a QSAR Model's Predictions. *J. Mol. Graphics Modell.* **2005**, *23*, 503–523.
- (57) *Ionic Liquids Database- (ILThermo) NIST Standard Reference Database 147*. <http://ilthermo.boulder.nist.gov/ILThermo/mainmenu.uix> (accessed 2006).
- (58) Raevskii, O. A.; Solov'ev, V. P.; Govorkova, L. V. The Study of the Polymorphism of Dibenzo-24-Crow-8 by the Methods of Differential Scanning Calorimetry and IR Spectroscopy. *Zh. Obshchei Khimii (Russ.)* **1985**, *55* (6), 1381–1384.
- (59) Solov'ev, V. P.; Govorkova, L. V.; Raevskii, O. A. Determination of the Purity, Melting-Point and Heat of Melting of Cyclic Polyethers. *Bull. Acad. Sci. USSR Div. Chem. Sci.* **1986**, *35*, 632–633.
- (60) Blanchard, L. A.; Brennecke, J. F. Recovery of Organic Products from Ionic Liquids Using Supercritical Carbon Dioxide. *Ind. Eng. Chem. Res.* **2001**, *40*, 287–292.
- (61) Dutta, L. M. School of Chemistry and Molecular Sciences, Thesis, University of Sussex, 1994.
- (62) Nigsch, F.; Bender, A.; Buuren, B.; Tissen, J.; Nigsch, E.; Mitchell, J. B. O. Melting Point Prediction Employing k-Nearest Neighbor Algorithms and Genetic Parameter Optimization. *J. Chem. Inf. Model.* **2006**, *46* (6), 2412–2422.
- (63) Modarresi, H.; Dearden, J. C.; Modarress, H. QSPR Correlation of Melting Point for Drug Compounds Based on Different Sources of Molecular Descriptors. *J. Chem. Inf. Model.* **2006**, *46* (2), 930–936.
- (64) Clark, T. Modelling the Chemistry: Time to Break the Mould? In *EuroQSAR 2002. Designing Drugs and Crop Protectants: Processes, Problems and Solutions*, Ford, M., Livingstone, D., Dearden, J., Van de Waterbeemd, H., Eds.; Blackwell Science Inc.: Massachusetts, 2003; pp 111–121.
- (65) Caruana, R. Multitask Learning: A Knowledge-Based Source of Inductive Bias. *Machine Learning* **1997**, *28*, 41–75.
- (66) Chalk, A. J.; Beck, B.; Clark, T. A Temperature-Dependent Quantum Mechanical/Neural Net Model for Vapor Pressure. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (4), 1053–1059.
- (67) Oussar, Y.; Dreyfus, G. How to Be a Gray Box: Dynamic Semi-Physical Modeling. *Neural Networks* **2001**, *14* (9), 1161–1172.
- (68) Erhan, D.; L'Heureux, P.-J.; Yue, S. Y.; Bengio, Y. Collaborative Filtering on a Family of Biological Targets. *J. Chem. Inf. Model.* **2006**, *46* (2), 626–635.
- (69) Acuña, G.; Pinto, E., Development of a Matlab Toolbox for the Design of Grey-Box Neural Models. *Int. J. Comp. Commun. Control* **2006**, *1* (2), 7–14.

CI600493X