DOI: 10.1002/minf.201200006

QSPR Approach to Predict Nonadditive Properties of Mixtures. Application to Bubble Point Temperatures of Binary Mixtures of Liquids

I. Oprisiu,[a, b] E. Varlamova,[c] E. Muratov,[c, d] A. Artemenko,[c] G. Marcou,[a] P. Polishchuk,[c] V. Kuz'min,[c] and A. Varnek*[a]

Abstract: This paper is devoted to the development of methodology for QSPR modeling of mixtures and its application to vapor/liquid equilibrium diagrams for bubble point temperatures of binary liquid mixtures. Two types of special mixture descriptors based on SiRMS and ISIDA approaches were developed. SiRMS-based fragment descriptors involve atoms belonging to both components of the mixture, whereas the ISIDA fragments belong only to one of these components. The models were built on the data set containing the phase diagrams for 167 mixtures represented by different combinations of 67 pure liquids. Consensus models were developed using nonlinear Support Vector Machine (SVM), Associative Neural Networks (ASNN), and Random Forest (RF) approaches. For SVM and ASNN calculations, the ISIDA fragment descriptors were used, whereas Simplex descriptors were employed in RF models. The models have been validated using three different protocols: "Points out", "Mixtures out" and "Compounds out", based on the specific rules to form training/test sets in each fold of cross-validation. A final validation of the models has been performed on an additional set of 94 mixtures represented by combinations of novel 34 compounds and modeling set chemicals with each other. The root mean squared error of predictions for new mixtures of already known liquids does not exceed 5.7 K, which outperforms COSMO-RS models. Developed QSAR methodology can be applied to the modeling of any nonadditive property of binary mixtures (antiviral activities, drug formulation, etc.)

Keywords: QSAR/QSPR · Vapor/liquid equilibrium · Bubble point curve · Mixtures prediction

1 Introduction

Vapor-liquid equilibrium (VLE) data represent one of the most important type of information required to evaluate the phase behavior of a binary liquid mixture, which is crucial for the design of separation processes.[1] Particular interest represents the dependence of bubble point or vapor pressure on the mixture composition (Figure 1). Theoretical assessment of these data could significantly reduce the costs of selection of proper agents for industrial processes.

Group contribution methods (GCM), such as UNIFAC,[2] UNIFAC-Dortmund,[3] ASOG[4] or UNIQUAC[5] are used worldwide to predict mixture behavior. UNIFAC is based on the thermodynamic equation for the activity coefficient (γ) of liquid 1 in the environment of liquid 2. To calculate γ, UNIFAC considers interactions of selected "structural groups" i ($i \in 1$) and j ($j \in 2$) accounted for the "energy parameters" A_{ij} and A_{ji}. The latter are fitted on available experimental data. An extensive table of UNIFAC group-interaction parameters was first published by Fredenslund et al.[6] in 1977. Then it has been several times revised because of the growing volume of experimental data. The latest UNIFAC update was based on the Dortmund Data Bank containing more than 39000 VLE data.

Despite its solid thermodynamics basis and excellent results of quantitative estimations of vapor-liquid equilibrium, UNIFAC has two serious drawbacks. The first one concerns the energy parameters A_{ij} and A_{ji} which cannot be assessed from the parameters of individual groups i and j but must be fitted directly on experimental VLE data for the binary

[a] *I. Oprisiu, G. Marcou, A. Varnek*
University of Strasbourg
Strasbourg, France
phone: +33.3.68.65.15.60
**e-mail: varnek@chimie.u-strasbg.fr*

[b] *I. Oprisiu*
Processium
62 Boulevard Niels Bohr, BP 2132, F 69603 Villeurbanne, France

[c] *E. Varlamova, E. Muratov, A. Artemenko, P. Polishchuk, V. Kuz'min*
A. V. Bogatsky Physical-Chemical Institute
Odessa, Ukraine

[d] *E. Muratov*
University of North Carolina
Chapel Hill, USA

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201200006

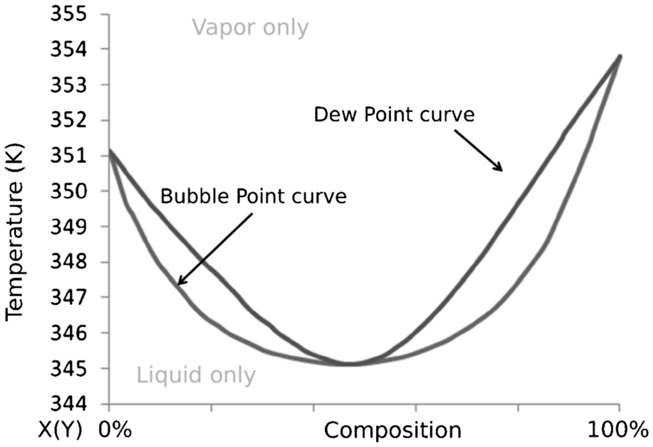


Figure 1. Vapor-liquid equilibrium curve showing the variation of equilibrium composition of the liquid mixture with the temperature at a fixed pressure. The dew-point curve represents the temperature at which the saturated vapor starts to condense whereas the bubble-point is the temperature at which the liquid starts to boil.

mixtures. Thus, the total number of the energy parameters rises as the squared number of groups. According to our estimations, for the current list of some 60 groups, less than half energy parameters were calculated. The second drawback is related to inflexible strategy of groups' selection which in most of cases makes impossible to represent a complex systems as an ensemble of groups. For instance, in recently reported UNIFAC model for ionic liquids (IL),[7] 12 new groups have been introduced and each of them represents an individual IL. Taking into account a large number of existing ILs, further development of UNIFAC models for these systems looks unrealistic. Similar conclusion could be drawn for variety of liquid binary systems involving heterocyclic molecules because each of them should be also represented as an individual group.

As an alternative of GCM, a novel method for the prediction for thermo-physical properties of fluids is used. COSMO-RS (Conductor-like Screening Model for Real Solvents)[8] approach is based on dielectric continuum models and statistical thermodynamics. The standard procedure of COSMO-RS calculations consists essentially in two steps: quantum chemical calculation of the local polarization charge density σ for each component of the mixture followed by COSMO-RS statistical calculations.[9] As a prerequisite, the 3D distribution of the polarization charge on the surface of each molecule is converted into a surface composition function (σ-profile).[8]

For the calculation of the chemical potentials (activity coefficients) using this model, analogously to GCM, the interactions between the molecules are taken into account. The surface of each molecule is subdivided into segments with equal area, and the chemical potential is derived from the interactions between the surface segments. Unlike GCM, in COSMO-RS model only a few element-based parameters must be fitted. Notice that dispersive interactions and hydrogen bonding are poorly accounted for in COSMO-RS.[10,11]

Quantitative Structure-Property Relationship (QSPR) approach could be considered as a valuable alternative to the previously described methods. A QSPR model relates a given physical property with chemical structure encoded by molecular descriptors. Although, QSPR technique is traditionally used to model individual compounds, some efforts have been recently made to model their mixtures.[12] Thus, Kravtsov et al. developed neural networks models for solvation free energies in different solvents[13] and rate constants of nucleophilic substitution reactions[14] and SN1[15] using simple concatenation of substrate and solvent descriptors. Ravindranath et al.[16] reported a "mixted" approach which integrates thermodynamics Margules, NRTL and UNIQUAC approaches with QSPR analysis. In this study, the groups interaction parameters involved in Margules, NRTL or UNIQUAC equations for activity coefficients have been modeled by QSPR. Ajmani et al.[17–19] reported QSPR models for infinite-dilution activity coefficients, excess molar volume and density of liquid binary mixtures using special mixture descriptors which were calculated as mole weighted average using the descriptor value and mole fraction of each pure component in the mixture. Recently, Katritzky et al.[20] performed QSPR modeling of normal boiling point temperature of azeotropes (T_{az}) using the CODESSA PRO program. Two different strategies have been used to prepare mixture descriptors: either simple arithmetic average of those calculated for individual molecular components 1 and 2, or weighting 1 and 2 by their molar ratios in the azeotrope. Predictive performance of the obtained in[16] linear models is rather weak: the standard deviation of about 23 K has been obtained at the fitting stage and has not even been reported for the external test set.

The most challenging problem in QSAR of mixtures is a representation of mixture by descriptors. Thus, prior to modeling, the investigators should decide which descriptors are the most suitable for the modeling of mixtures (binary liquids in the given study). Should the nonadditivity effects be included at the descriptor design level, or mixture descriptors could be simply constructed from those of individual components? Here, we examine both strategies. Another question is related to proper external validation of models for mixtures which is less obvious than in classical QSAR. Some efforts have been done by[17–19] who reported validation procedure similar to "Points Out" and "Mixture Out" strategies suggested in this work (see Section 1.2). This, however, is not sufficient to assess prediction performance for mixtures containing new compounds. Indeed, if both training and external sets include data points of the same mixture the model's performance to predict new mixture is not truly estimated. The drawbacks of conventional n-fold external cross-validation in QSAR of mixtures are discussed elsewhere.[21] Thus, new more rigorous protocol for external validation must be developed specially for QSAR modeling of mixtures. One more problem is a detection of outliers for the curves which is also different from classic QSAR.

Thus, the goal of this study is the development of the solid workflow of QSPR analysis of mixtures which includes: (i) development of two types of mixture descriptors: "non-additive" SiRMS descriptors and additive ISIDA descriptors; (ii) ensemble QSPR modeling of bubble-point temperature of mixtures of organic compounds using SVM, ASNN, and RF methods; (iii) rigorous external validation of obtained models; (iv) detection of outlier mixtures for developed QSPR models; (v) benchmarking of obtained QSPR models with COSMO-RS approach.

It should be noted that the developed QSAR methodology is not limited by particular case of phase diagrams, but it could be used to model any property of binary mixtures (antiviral activities, drug formulation, etc.)

2 Materials and Methods

2.1 Dataset Descriptions

2.1.1 Modeling Set

The dataset was compiled from Korean Data Base (KDB).[22] It consists of 67 pure liquids and 167 of their mixtures. Each mixture has been represented by several (7–57) points, thus, 167 modeling set mixtures have been described by 3185 data points. The matrix of mixtures is very sparse and consists of only 167 out of possible 2211 combinations, i.e., sparsity degree is 92.5%. One compound could be involved in different number of mixtures (from 1 to 25). The total distribution of the number of mixtures and data points per pure liquid is represented in Table 1, with an average number of 5 mixtures and 95 data points per pure compound. The bubble temperature (T_b) was expressed in Kelvin scale and has a range from 280.25 to 462.65 K for modeling set and from 315.95 to 544.26 for the external set. It creates an additional problem to predict external compounds and mixtures, because for some of them the temperature exceeds up to 80 K the maximal T_b for the modeling set.

Generally experimental measure errors reported in different publications are around 0.06 K for the bubble temperature (T_b) and 0.1% for the composition (X).[23,24] These are the measure errors made within the same laboratory; they are smaller than the errors of the same measurement made by different labs, which could be as high as 0.5 K and 1%, respectively.[25,26] In some cases due to imprecise control of atmospheric pressure and different approximation methods employed in each laboratory, the average error for T_b of pure compounds may reach 12–18 K.[27]

2.1.2 External Validation Set

The models built on the entire modeling set have been additionally validated on the external validation set of 94 new mixtures involving 66 compounds. Only 27 out 94 mixtures (632 data points) contain no new pure compounds and 67 mixtures (1386 points) contain at least one new compound. Thus, 32 external compounds are common to the modeling set, whereas other 34 are new. Four mixtures have no common compounds with the modeling set.

2.2 Strategies for Model External Validation in QSAR of Mixtures

Three different strategies of external validation were established (Figure 2): (i) "points out" – prediction of T_b for any molar ratio of the known biphasic systems, (ii) "mixtures out" – prediction of T_b for the missed data in the mixture matrix (gap-filling) formed by 67 pure liquids from the modeling set, and (iii) "compounds out" – prediction of T_b for mixtures formed by "new" pure compound(s) absent in the modeling set.

"Points out". All pure compounds were always kept in the training set and mixture data points were randomly taken to each fold of external cross-validation set. Each mixture is present both in training and test sets. Here, 2-fold external cross-validation repeated 3 times has been performed.

"Mixtures out". In each fold of external cross-validation all pure compounds were always kept in the training set but whole mixtures were randomly selected to test set. Thus, each mixture is present either in the training or in the test set, but never in both sets. Here, 5-fold external cross-validation repeated 3 times has been performed. Expected error of prediction for this models is bigger than for "points out" strategy, however, this model will not be limited by already known mixtures, but will be useful for the filling of the missed data in the mixture matrix formed from 67 training set compounds. Because this matrix is very sparse, 2211–167 = 2044 new mixtures could be predicted.

"Compounds out". Pure liquid and all its mixtures were simultaneously taken to an external fold. Thus each mixture in the external set contains at least one compound which is absent in the training set. The difference with classical CV algorithm is that the folds were not created randomly, but supervised in order to keep the number of both pure liquids and the mixtures amongst the folds more or less constant. The supervision is needed because one pure liquid, for instance bromobenzene, can participate in only one mixture, while another – carbon tetrachloride – can create 25 mixtures and the classical random algorithm is unable to consider such situation during external folds creation. Moreover, despite the supervised process of folds creation, we are aware that some folds could be predicted badly because they are still anisotropic and sufficient lack of information in the training set could be observed for some external folds. That was the reason of triple repetition of 10-fold external CV. It is necessary to note that, because we are dealing with binary mixtures, every mixture will be taken to the external set twice, except the case when both

Table 1. Distribution of number of mixtures and data points per pure liquid.

Compound	Mixtures	Points	Compound	Mixtures	Points
Carbontetrachloride	25	444	*m*-Xylene	2	59
Methanol	16	313	Water	3	59
Benzene	16	310	Dibromomethane	4	54
Ethanol	12	279	2,4-Dimethylpentane	3	53
n-Butanol	13	260	1,1-Dichloroethane	2	44
Cyclohexane	12	250	Benzotrifluoride	3	43
Acetonitrile	10	242	Methylcyclopentane	3	43
1-Bromopropane	10	226	Cyclohexene	2	40
Toluene	11	221	1-Hexene	2	39
Ethylacetate	9	217	1-Chlorobutane	3	35
Methylacetate	10	213	2,2,5-Trimethylhexane	2	35
n-Propanol	12	207	1-Octene	2	34
n-Octane	8	159	*n*-Butylacetate	1	29
n-Heptane	9	148	*n*-Butylformate	1	28
Hexafluorobenzene	5	139	*n*-Butyl-*n*-butyrate	1	28
sec-Butanol	8	136	*n*-Butylpropionate	1	28
1,2-Dichloroethane	8	132	*p*-Cresol	1	23
Chloroform	6	132	*n*-Decane	1	22
Vinylacetate	6	117	Cumene	1	19
Isopropanol	6	114	Phenol	1	19
n-Hexane	6	114	2-Methylpentane	1	18
Acetone	7	107	3-Methylpentane	1	18
Ethylbenzene	6	105	Cyclopentane	1	18
2,3-Dimethylbutane	5	102	2,2-Dimethylbutane	1	18
tert-Butanol	6	102	1,1,2-Trichlorotrifluoroethane	1	17
Chlorobenzene	6	87	1,2-Dichlorotetrafluoroethane	1	17
p-Xylene	4	83	1,1,1-Trichloroethane	1	16
Isobutanol	6	75	Dibutylether	1	14
2,2,4-Trimethylpentane	4	72	*n*-Propylacetate	1	14
o-Xylene	5	72	Bromobenzene	1	13
Methylcyclohexane	4	69	Ethyleneglycol	1	12
Methylmethacrylate	3	69	*n*-Propylbenzene	1	10
Tetrachloroethylene	4	64	Acrylonitrile	1	7
Trichloroethylene	4	64			

compounds are belonging to the same external fold. It simulates the addition of novel component to existing matrix of mixtures. This is the most rigorous way of external validation of QSAR models for mixtures. Although the error of prediction for this strategy is expected to be the biggest, the models passed the validation will be able to predict T_b for mixtures created by a new pure compound beyond the modeling set.

2.3 Mixture Descriptors

2.3.1 Simplex Representation of Molecular Structure (SiRMS)

In the frameworks of Simplex representation of molecular structure (SiRMS)[28–30] any molecule can be represented as an ensemble of different 2D tetratomic fragments of fixed composition, structure, chirality and symmetry simplexes. A number of identical simplexes in a molecule is a descriptor value of that simplex. The connectivity of atoms in simplex, atom type and bond nature (single, double, triple, or aromatic) have been considered at the 2D level (Figure 3). Bounded and unbounded 2D simplexes were used. Not only atom type, but some other physical-chemical characteristics of atoms, i.e., partial charge, lipophilicity, refraction, and atom's ability for being a donor/acceptor in hydrogen-bond formation were used for atom labeling in simplexes (Figure 3). For these atom characteristics the binning procedure has been used to transform real values (charge, lipophilicity, and refraction) to four categories corresponding to their (i) partial charge $A \leq -0.05 < B \leq 0 < C \leq 0.05 < D$, (ii) lipophilicity $A \leq -0.5 < B \leq 0 < C \leq 0.5 < D$, and (iii) refraction $A \leq 1.5 < B \leq 3 < C \leq 8 < D$. Three characteristics of atom H-bond formation ability were specified A (acceptor of hydrogen in H-bond), D (donor of hydrogen in H-bond), and I (indifferent atom).

Bounded simplexes describe only single components of the mixture (Compounds 1 or 2), whereas unbounded simplexes can describe both the constituent parts and the mixture as a whole (Figure 4). With this purpose it is necessary to indicate whether the parts of unbounded simplexes are belonging to the same molecule or to different ones. A special mark is used during descriptors generation to distinguish such simplexes. Descriptors of constituent parts (Compounds 1 and 2) are weighted according to their

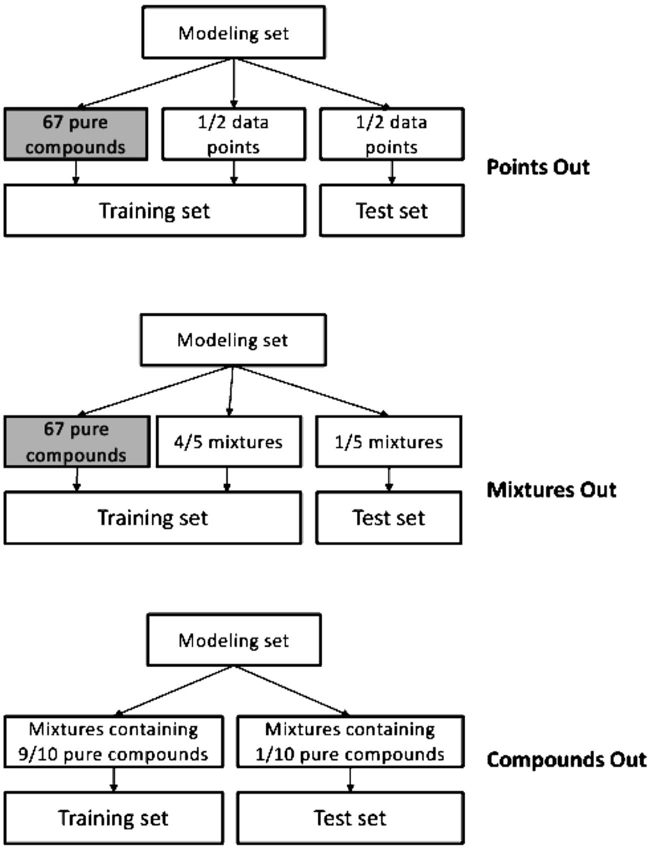


Figure 2. Different strategies used for models validation.

molar fraction and summarized, and mixture descriptors are multiplied on doubled minimal weight according to Equation 1. If both mixtures and pure liquids are considered, descriptors of pure liquid have the weight equal to 1.

$$D = \begin{cases} x_1 D_1 + x_2 D_2 \\ 2x_1 D_{1+2} \end{cases}, \tag{1}$$

where D is the descriptor value, x_1 and x_2 are molar fractions of Components 1 and 2 ($x_1 < x_2$ and $x_1 + x_2 = 1$), D_1, D_2, and D_{1+2} are descriptor values for individual Compounds 1 and 2, and for their mixture, respectively.

2.3.2 ISIDA Fragment Descriptors

ISIDA fragment descriptors[31] were used in combination with SVM and ASNN machine learning methods. Two different types of molecular subgraphs are considered (Figure 5): "sequences" (**I**) and "augmented atoms" (**II**).

The sequences correspond to consecutive set of atoms linked by chemical bonds, where either atom types (C, N, O, …) or bond types (single, double, …) or both of them are considered explicitly. In the following, we specify the number of atoms of a given sequence. Thus, **I**(**AB**, n_{min}–n_{max}) refers to all sequences containing from n_{min} to n_{max} atoms connected by bonds of specified type. For **I**(**A**, n_{min}–n_{max}), the definition is similar, but bond types are omitted. Only shortest paths from one atom to the other are used, as shown in Figure 5.

An "augmented atom" represents a selected atom within its nearest environment including both neighboring atoms and bonds (**AB**), or atoms only (**A**), or bonds only (**B**). The neighborhood is described by concatenating all sequences starting at a given atom and of a given length. For instance **II**(**AB**, n_{min}–n_{max}) refers to fragments concatenating sequences of length from n_{min} to n_{max} around the selected atom. In this work, $n_{min} \geq 2$ and $n_{max} \leq 10$ were used for the sequence and $n_{min} \geq 2$ and $n_{max} \leq 4$ for augmented atoms. In QSPR models each fragment is considered as an individual descriptor, whereas its occurrence in the given molecule is the descriptor value.

Mixture descriptors have been obtained by combining ISIDA descriptors for each component (Equation 2, Figure 6).

$$D = \begin{cases} x_1 D_1 + x_2 D_2 \\ |x_1 D_1 - x_2 D_2| \end{cases} \tag{2}$$

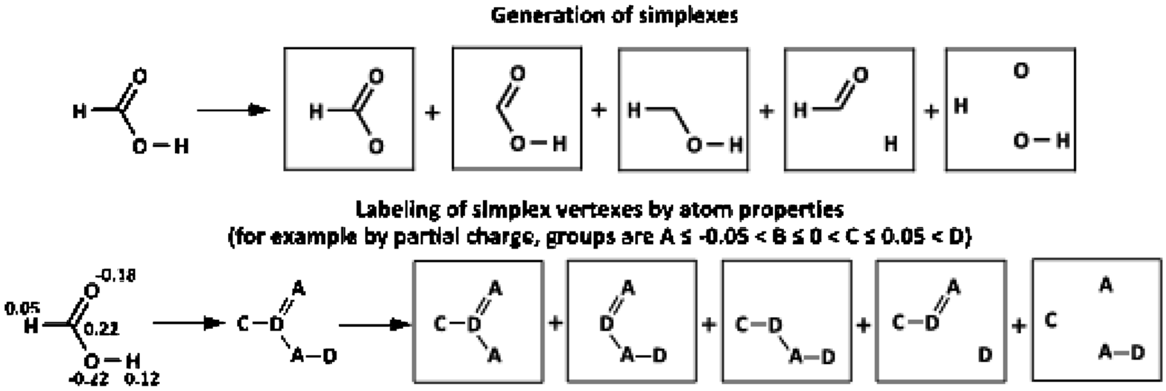

Generation of simplexes
Labeling of simplex vertexes by atom properties
(for example by partial charge, groups are A ≤ -0.05 < B ≤ 0 < C ≤ 0.05 < D)
-0.18
0.05
0.22
-0.42
0.12

Figure 3. Simplex representation of molecular structure (SiRMS).

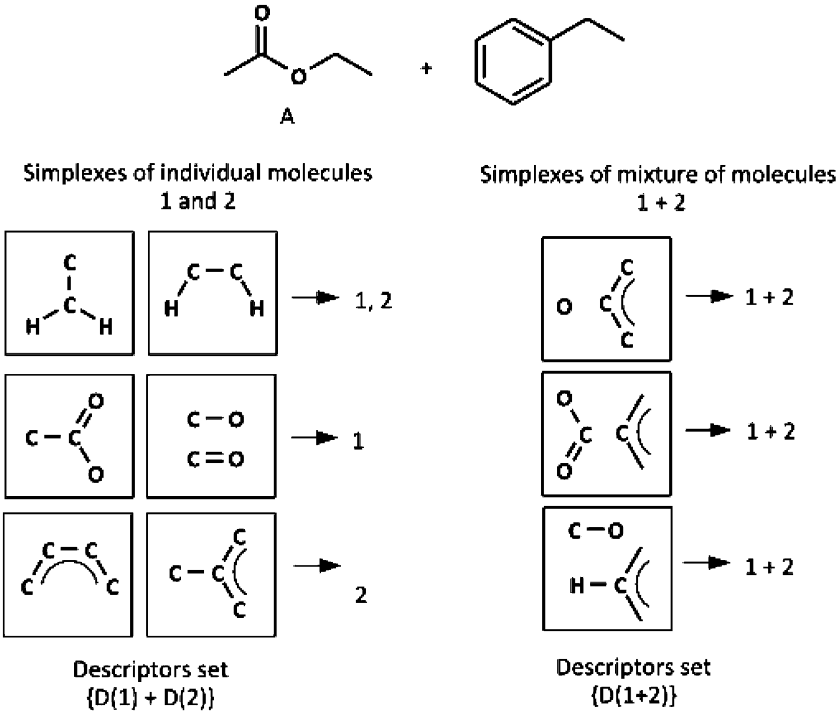
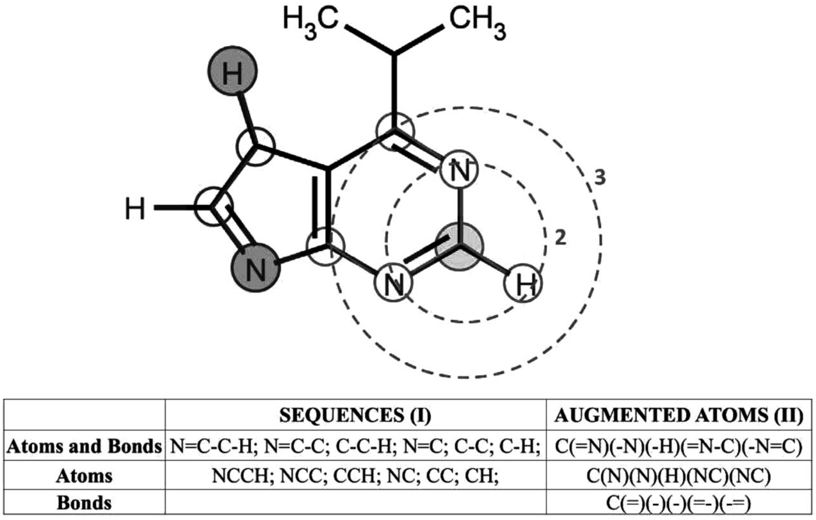


Figure 4. Simplex descriptors for a binary mixture.



	SEQUENCES (I)	AUGMENTED ATOMS (II)
Atoms and Bonds	N=C-C-H; N=C-C; C-C-H; N=C; C-C; C-H;	C(=N)(-N)(-H)(=N-C)(-N=C)
Atoms	NCCH; NCC; CCH; NC; CC; CH;	C(N)(N)(H)(NC)(NC)
Bonds		C(=)(-)(-)(=-)(-=)

Figure 5. ISIDA Fragmentation. Two classes of substructural fragments: atom/bond sequences and augmented atoms. From top to bottom: the sequences (I) correspond to the I (AB, 2–4) and I (A, 2–4) types involving the shortest paths between each pair of atoms. Augmented atoms (II) correspond to the II (AB, 2–3), II (A, 2–3) and II (B, 2–3) types.

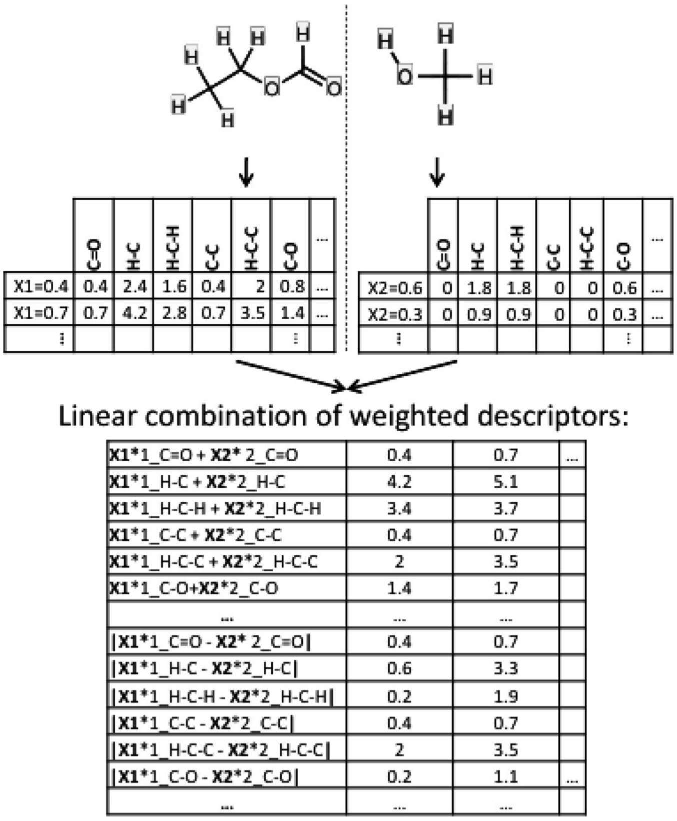
where D_1 and D_2 are descriptor values for individual Compounds 1 and 2, x_1 and x_2 are molar fractions of components 1 and 2.

2.4 Machine-Learning Methods

Three following statistical approaches were used: Random Forest (RF) in combination with Simplex descriptors, when ISIDA fragment descriptors were employed in SVM and ASNN methods.

2.4.1 Random Forest

RF models were obtained according to the original RF algorithm described in[32] and realized in.[33] RF is an ensemble of single decision trees. This ensemble produces a corre-



	C=O	H-C	H-C-H	C-C	H-C-C	C-O	...
X1=0.4	0.4	2.4	1.6	0.4	2	0.8	...
X1=0.7	0.7	4.2	2.8	0.7	3.5	1.4	...
⋮						⋮	

	C=O	H-C	H-C-H	C-C	H-C-C	C-O	...
X2=0.6	0	1.8	1.8	0	0	0.6	...
X2=0.3	0	0.9	0.9	0	0	0.3	...
⋮						⋮	

Linear combination of weighted descriptors:

X1*1_C=O + **X2*** 2_C=O	0.4	0.7	...
X1*1_H-C + **X2***2_H-C	4.2	5.1	
X1*1_H-C-H + **X2***2_H-C-H	3.4	3.7	
X1*1_C-C + **X2***2_C-C	0.4	0.7	
X1*1_H-C-C + **X2***2_H-C-C	2	3.5	
X1*1_C-O+**X2***2_C-O	1.4	1.7	
...	...	...	
|**X1***1_C=O - **X2*** 2_C=O|	0.4	0.7	
|**X1***1_H-C - **X2***2_H-C|	0.6	3.3	
|**X1***1_H-C-H - **X2***2_H-C-H|	0.2	1.9	
|**X1***1_C-C - **X2***2_C-C|	0.4	0.7	
|**X1***1_H-C-C - **X2***2_H-C-C|	2	3.5	
|**X1***1_C-O - **X2***2_C-O|	0.2	1.1	...
...	...	...	

Figure 6. ISIDA mixture descriptors. X1 and X2 are the molar ratios of the first and the second component in the mixture, respectively. For each component of the liquid, the numbers correspond to the product of the fragment's occurrence and molar ratio X.

sponding number of outputs. Outputs of all trees are aggregated to obtain one final prediction as an average of the individual tree predictions. Each tree has been grown as follows: (i) A bootstrap sample, which will be a training set for the current tree, is produced from the whole training set of *N* compounds. Compounds which are not in the current tree training set are placed in an out-of-bag (OOB) set (~*N*/3 molecules). (ii) The best split among the m randomly selected parameters from the initial set of M descriptors is chosen in each node by CART algorithm.[34] The value of m is just one tuning parameter for which RF models are sensitive. (iii) Each tree is grown to the largest possible extent without any pruning. Performance of the models has been assessed on OOB sets, which values are similar to those obtained in 5-fold external cross-validation procedure.[35] The model selection has been performed according to R^2_{OOB} values. Each individual RF model involved 457–508 simplex descriptors.

Associative Neural Network (ASNN)

An associative neural network (ASNN) is a combination of an ensemble of 100 feed-forward neural networks and the KNN technique. Three layers architecture of neural network has been used. The number of neurons in the input layer corresponds to the number of descriptors, whereas the output layer consists of one neuron (modeled property). The number of neurons in the hidden layer is equal to 5, as recommended by Tetko et al.[36] ASNN uses correlation between ensemble responses as a measure of distance among the analyzed cases for the nearest neighbor technique. This method corrects a bias of a global model for a considered data case by analyzing the biases of its nearest neighbors determined in the space of calculated models. Early stopping technique has been used to avoid an overfit of the models.[37] The ASNN 1.0 program provided by Dr. Igor V. Tetko has been used in this work.

Support Vector Machines (SVM)

In SVM a nonlinear function is learned by linear fitting in the feature space which is a nonlinear mapping of the initial (input) space in which the fitting problem was expressed.[38] The libSVM, a C library was used to generate SVM models. The calculations have been performed with the RBF kernel. The γ and ε parameters have been optimized in grid calculations.

Using different initial pools of ISIDA descriptors corresponding to different fragmentation types, several tens of SVM and ASNN models have been trained. Only those models with determination coefficient $R^2 > 0.8$ at cross-validation have been retained and used for the predictions for SVM and ASNN consensus models.

Once the ensemble of models issued from a given machine-learning method is obtained, a consensus model can be calculated either by simple averaging of predictions or by using Stacking technique[39] which develops a linear regression using predictions made by each model as independent variables.

2.4.2 Applicability Domain

QSPR models are obtained on a training set, which, no matter how large, may never represent a significant sample of the entire chemical space. Applicability domain (AD) of a model defines a region of the chemical space that can be adequately covered by training set compounds. Statistical models can deliver reliable predictions for the compounds belonging to this region. Presumably, the models cannot be trustfully used outside of their AD.

Minimum spanning tree AD approach[28] was used for RF models. Minimum spanning tree has been built in the space of decision trees predictions for the given RF model using Kruskal's algorithm.[40] Then average distance (d_{av}) and its root-mean-square deviation (σ) among all tree edges have been calculated. Substantially, such distance is the characteristic of average density of molecules distribution in the considered space. If any of external set molecules has been situated on the distance bigger than $d_{av}+$

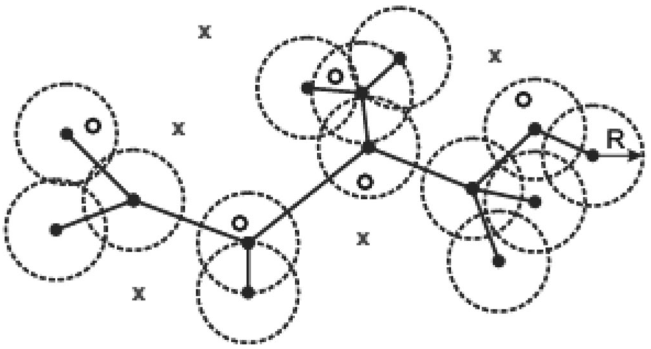


Figure 7. Scheme of Local AD approach based on minimum spanning tree. ● – training set molecules, ○ – test set compounds within AD, × – test set compounds outside AD, radius of each sphere is $R = d_{av} + 3\sigma$.

3σ from the nearest training set point, it means that this external set molecule is situated outside AD (Figure 7).

Fragment control[41] AD approach has been used for SVM and ASNN models. Any molecules containing the fragments which do not occur in compounds of the training set are considered to be outside AD in this method.

2.5 Statistical Characteristics Used

Determination Coefficient (R^2) and Root Mean Squared Error (*RMSE*) for the external set were used to estimate the predictivity of the obtained models.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_{\text{pred},i} - y_{\text{exp},i}\right)^2}{\sum_{i=1}^{n} \left(y_{\text{exp},i} - \overline{y_{\text{exp}}}\right)^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(y_{\text{pred},i} - y_{\text{exp},i}\right)^2} \tag{4}$$

where n is the number of compounds in the external set (folds); y_{exp} is the experimental value of T_b and y_{pred} is the predicted value of T_b.

2.6 Outliers Analysis

Outliers analysis for modeled curves has been performed using z^2 parameter:

$$z^2 = \frac{\sum \left(Y_{\text{exp}} - Y_{\text{pred}}\right)^2}{\frac{1}{n-1}\sum \left(Y_{\text{exp}} - \bar{Y}_{\text{exp}}\right)^2}, \tag{5}$$

where Y_{exp} represents the experimental temperature, $\bar{Y}_{\text{exp}}$ is the average of the experimental values of T_b, and Y_{pred} is T_b predicted value.

The z^2 values follow a Chi-square law[42] with n degrees of freedom, where n is the number of points in the mixture. If z^2 for a given curve is higher than the threshold value at 99% of confidence the corresponding mixture is considered an outlier.

2.7 Benchmarking

Predictive performance of developed QSPR models has been compared with the ones of COSMO-RS model.

All COSMO-RS calculations of this work were performed using the COSMOtherm program (version C21_0108) using precomputed σ-profiles[43] for pure compounds. Since σ-profiles for some molecules were not reported in,[43] the calculations have been performed for 166 out of 167 mixtures of the modeling set and for 89 out of 94 mixtures of the test set.

3 Results and Discussion

3.1 External Cross-Validation

Results of n-fold external CV are shown at Table 2. As expected, the error of prediction increased in the order "points out" < "mixtures out" < "compounds out". In "points out" strategy data points of the same mixture are simultaneously present in both training and external set. Therefore, high accuracy of predictions ($RMSE = 1.1$ K) is not surprising. In "mixtures out" strategy the data points for a given mixture are present either in training set or in test set, but not simultaneously in both. Therefore, *RMSE* of 5.2 K is bigger than for "points out" strategy, but it remains on the acceptable level. The "compound out" is the most strict validation strategy ($RMSE = 7.0$ K) because a given pure compound and all its mixtures were simultaneously placed to external set during external cross-validation. Stacking consensus model significantly outperformed all other models for "mixtures out" and "compounds out" strategies (Table 2).

Typical examples of predicted bubble point curves are given on Figure 8. In most cases the calculations correctly reproduce the curves behavior: azeotrop predicted as azeotrop (Figure 8a) and zeotrop predicted as zeotrop (Figure 8b). However, in some cases, the "mixtures out" and "compounds out" models fail (Figure 8c). Generally, except of two mixtures (metanol-1-bromopropan and benzotrifluoride-toluene), the "points out" curves very well reproduce the experimental ones, see Figure SM1 in Supporting Information. In the "mixture out" calculations, the experimental trend has not been reproduced only for 15% of studied mixtures. On the other hand, in most of the "failed" predictions, the difference between experimental and calculated T_b does not exceed 5 K, which is within the error bar of the model. The "compounds out" predictions led to bad predictions in about half of all curves (Figure SM1) which corresponds to bad statistical parameters given in Table 2.

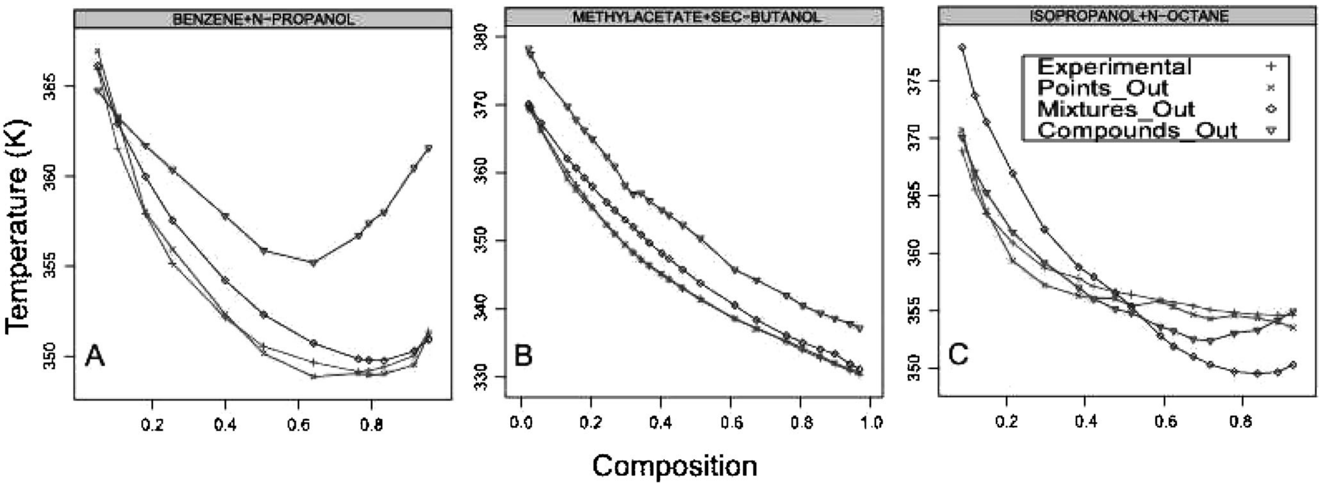


Figure 8. Typical examples of bubble point curves predictions in 5-CV for the modeling set: azeotrop predicted as azeotrop (A), zeotrop predicted as zeotrop (B) and zeotrop predicted as azeotrop in "mixture out" and "compounds out" strategies (C).

Table 2. Results of external validation of developed QSPR models.

		"Points Out"					"Mixtures Out"					"Compounds Out"				
		SVM	ASNN	RF	Av.[a]	Stack.[b]	SVM	ASNN	RF	Av.[a]	Stack.[b]	SVM	ASNN	RF	Av.[a]	Stack.[b]
Modeling set	R^2	0.98	0.99	0.98	0.99	0.99	0.86	0.92	0.90	0.93	0.95	0.73	0.89	0.79	0.85	0.90
	RMSE (K)	3.5	1.2	3.2	2.3	1.1	8.5	6.2	6.9	5.7	5.2	11.6	7.3	10.3	8.7	7.0
External set[c]	R^2						0.72	0.90	0.81	0.85	0.88	0.23[d]	0.42[d]	0.48[e]	0.39	0.42
	RMSE (K)						8.8	5.2	7.2	6.6	5.9	24.3[d]	21.0[d]	18.5[e]	22.0	21.4

[a], [b] Av.: average; Stack.: stacking. Prediction for each mixture have been calculated as a simple average of the results of SVM, ASNN and RF or using stacking approach, respectively. [c] "Mixture out" and "Compounds out" predictions were made for 27 and 67 mixtures, respectively. [d] Diethylamine + chloroform mixture was found outside of AD. [e] Formic acid + *n*,*n*-dimethylformamide, *p*-xylene + *n*,*n*-dimethylformamide and *n*-dodecane + 1-hexadecene mixtures were found outside of AD.

3.2 External Validation

The resulting models were built on the entire modeling set and then were applied to the prediction of external set compounds. The difference in range of bubble point temperatures for compounds of modeling and external sets is up to 80 K. This significantly complicates the accurate prediction of external set compounds. Calculation on 27 mixtures of external set containing no new components were considered as "mixture out" predictions, whereas those for 67 mixtures containing at least one new compound were considered as "compound out". Performance of the "mixtures out" predictions of resulting consensus model (R^2_{ext} = 0.88; *RMSE* = 5.7 K) was close to 5-fold external cross-validation results (R^2_{ext} = 0.95; *RMSE* = 5.2 K). On the other hand, "compounds out" predictions resulted in worse statistical parameters (R^2_{ext} = 0.44, *RMSE* = 21.0 K) than external cross-validation ones (R^2_{ext} = 0.90; *RMSE* = 7.0 K). Thus, the developed models are able to fill the gaps in the matrix of mixtures formed by 67 individual compounds of modeling set, i.e., to predict T_b for 2044 missing mixtures with reasonable accuracy. But both considered methodologies (ISIDA and SiRMS) still need some tuning to be able to predict T_b of mixtures containing at least one new compound.

At the same time, results of external cross-validation given in Table 2 and Table 3 show that consensus predictions are comparable or better than predictions obtained within one machine-learning method. The way of consensus model development is also important: stacking systematically shows better results than simple averaging.

3.3 Models Applicability Domain

An applicability domain (AD) analysis has been performed for the modeling set with "mixtures out" and "compounds out" strategies. 1,2-Dichlorotetrafluoroethane + 1,1,2-trichlorotrifluoroethane, acetonitrile + methylmethacrylate, acrylonitrile + acetonitrile and vinylacetate + methylmethacrylate were out of AD of RF models during external cross-validation. For the external set no compounds were outside AD for "mixtures out" strategy. Formic acid + *n*,*n*-dimethylformamide, *p*-xylene + *n*,*n*-dimethylformamide and *n*-dodecane + 1-hexadecene were out of AD of RF models for "compounds out" case, while ASNN and SVM models designate only diethylamine + chloroform mixture to be out of AD.

The use of AD for RF models does not lead to a significant improvement of the model quality with the exception of external validation for “compounds out” strategy. In this case AD usage increases R_{ext}^2 value from 0.36 to 0.48 and decreases *RMSE* from 23.2 K to 18.5 K. The AD methodology applied for ASNN and SVM did not improve predictive performance of the models.

3.4 Outliers Detection

As expected, the number of outliers increases in the order: “points out” < “mixture out” < “compounds out”. In “mixture out”, 14 mixtures have been classified as outliers for SVM, ASNN, and RF models: 2,3-dimethylbutane + acetone, 2,3-dimethylbutane + methanol, 2,4-dimethylpentane + benzene, acetonitrile + 1-bromopropane, benzene + cyclohexane, benzene + hexafluorobenzene, ethanol + 2,2,4-trimethylpentane, ethylacetate + benzene, methanol + 1-bromopropane, methanol + 1,1-dichloroethane, methylacetate + 1-hexene, methylmethacrylate + toluene, *n*-butanol + *n*-octane, *n*-propanol + *n*-octane.

Bad predictions for these mixtures could be explained either by the noise in experimental data used for model development or by specific physical phenomena observed only for few mixtures. For example, the data for isopropanol + 1-bromopropane mixture were erroneously attributed to the methanol + 1-bromopropane mixture.[44] Another example concerns the benzene + hexafluorobenzene mixture that forms both positive and negative azeotrope which is the only case regarding to 167 mixtures in the modeling set.[45]

3.5 Benchmarking Results

Besides “1,2-dichlorotetrafluoroethane + 1,1,2-trichlorotrifluoroethane” (for which σ-profile was not reported in[43]), the COSMOtherm program predicts all others 166 mixtures of the modeling set. Results are considerably less good ($R^2 = 0.75$, $RMSE = 11.0$ K) than QSPR stacking predictions for 5-fold external cross-validation using either “mixture out” or “compounds out” strategy (Table 2). As expected, prediction performance of the COSMO-RS model for 89 mixtures of external set ($R^2 = 0.80$, $RMSE = 11.3$ K) is similar to that for the 166 training set compounds. QSPR stacking model applied to this set led to worse statistics parameters ($R^2 =$ 0.52, $RMSE = 17.6$ K). In addition, we split the external set onto “mixture out” (27 mixtures) and “compound out” (62 mixtures) test sets, predictions for which were performed separately. For the “mixture out” set the accuracy of predictions for the QSPR stacking and COSMO-RS models were found similar, but for the “compound out” set the COSMO-RS model outperforms our results (Table 3). Predicted and

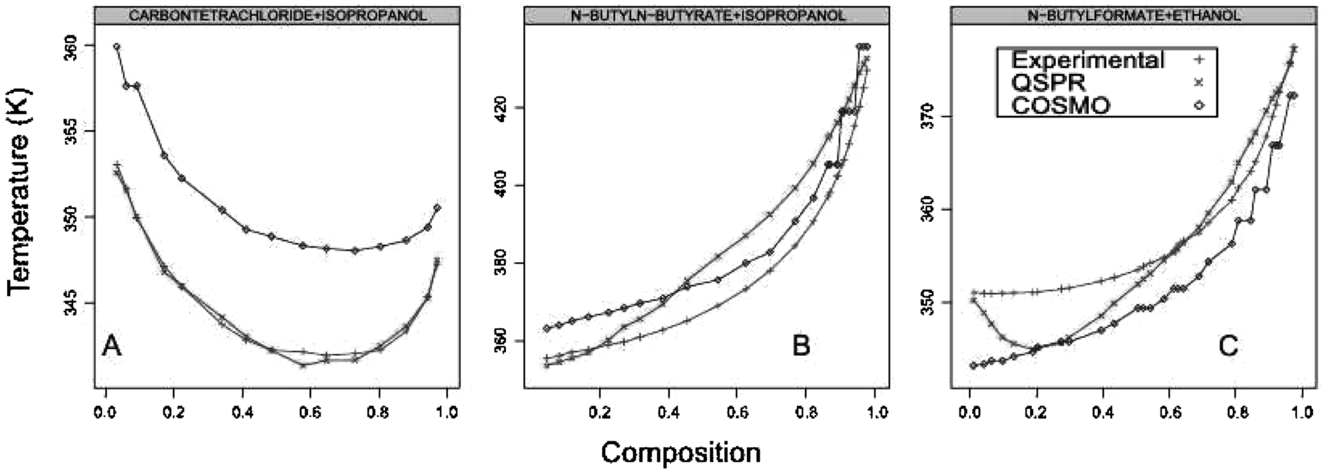


Figure 9. Typical examples of bubble point curves predicted for external “mixture out” test set of 27 compounds: azeotrop predicted as azeotrop (A), zeotrop predicted as zeotrop (B) and zeotrop predicted by QSPR as azeotrop (C).

Table 3. Statistical parameters obtained for mixtures of external test set.

	Entire set of 89 compounds[a]		“Mixture out” set of 27 compounds		“Compound out” set of 62 compounds	
	R^2	*RMSE* (K)	R^2	*RMSE* (K)	R^2	*RMSE* (K)
Stacking	0.52	17.6	0.88	5.7	0.44	21.2
COSMO-RS	0.80	11.3	0.84	6.6	0.78	13.0

[a] Only 89 out of 94 external set compounds have been predicted with COSMO-RS model.

experimental curves for the "mixture out" and "compound out" test sets are given on Figures SM2 and SM3, respectively. As shown in Figures 9a and 9b, in most cases, both QSPR and COSMO-RS models correctly reproduce the curves behavior, although some erroneous predictions have also been observed (Figure 9c).

4 Conclusions

In this paper we described some general aspects of QSPR methodology to model the compound mixtures with nonadditive response. This includes (i) design of additive or nonadditive "mixture" descriptors and (ii) cross-validation of "mixture" models which differs from that of QSPR models for individual compounds.

The developed workflow for QSPR analysis of mixtures was tested on boiling point temperatures of binary mixtures of organic compounds. Special descriptors of mixtures were developed on the basis of ISIDA (additive) and SiRMS (nonadditive) approaches. They were successfully used to predict bubble point temperatures of binary liquid mixtures from the initial data matrix. Thus, using experimental data on 167 mixtures formed by 67 individual liquids, we are able to predict with the reasonable accuracy the bubble point temperatures for $67\times(67–1)/2–167=2044$ "missing" mixtures, i.e., to fill remaining 92.5% of the sparse data matrix. On the other hand, both SiRMS and ISIDA approaches need an additional tuning to be able to predict the mixtures containing new components absent in the modeling set.

As follows from comparing the results obtained with two different types of descriptors, nonadditivity effects taken into account at the descriptor design level do not affect the prediction performance of the models. This opens an opportunity to apply a huge variety of molecular descriptors in the modeling of nonadditive properties of mixtures.

An important achievement of this work is the establishment of the strategies of external cross-validation for QSPR of mixtures that should be different and more rigorous than the ones used in classic QSPR analysis. Developed strategies allow one to obtain realistic estimation of the model predictivity in the most rigorous way and can simulate both (i) gap-filling of initial matrix of mixtures and (ii) addition of new component (liquid) absent during the modeling.

It has been shown that consensus predictions are usually better than predictions obtained within one machine-learning method. Moreover, the way of consensus model development is critical to overall prediction performance. Thus, the stacking approach leads to better results compared to simple arithmetic averaging of the values predicted by individual models. Benchmarking studies show that QSPR models represent a good alternative to the COSMO-RS approach, especially for mixtures the individual components of which were present in the modeling set.

Developed QSAR/QSPR methodology can obviously be used for the modeling of any property of binary mixtures (antiviral activities, drug formulation, etc).

Acknowledgements

VK, PP, EM, and EV thank the *ARCUS "Alsace-Russia/Ukraine"* and the *SupraChem Projects* for support of their stay at the *University of Strasbourg*. EM acknowledges the support from *NIH* (Grant GM66940). IO thanks the *Processium company* for the PhD fellowship.

References

[1] J. Gmehling, R. Bölts, *J. Chem. Eng. Data* **1996**, *41*, 202–209.

[2] J. Gmehling, P. Rasmussen, A. Fredenslund, *Chem.-Ing.-Tech.* **1980**, *52*, 724.

[3] U. Weidlich, J. Gmehling, *Indust. Eng. Chem. Res.* **1987**, *26*, 1372–1381.

[4] K. Tochigi, D. Tiegs, J. Gmehling, K. Kojima, *J. Chem. Eng. Jpn.* **1990**, *23*, 453–463.

[5] T. F. Anderson, J. M. Prausnitz, *Ind. Eng. Chem. Proc. Dd.* **1978**, *17*, 552–561.

[6] A. Fredenslund, J. Gmehling, M. L. Michelsen, P. Rasmussen, J. M. Prausnitz, *Ind. Eng. Chem. Proc. Dd.* **1977**, *16*, 450–462.

[7] Z. Lei, J. Zhang, Q. Li, B. Chen, *Indust. Eng. Chem. Res.* **2009**, *48*, 2697–2704.

[8] A. Klamt, *J. Phys. Chem.-Us.* **1995**, *99*, 2224–2235.

[9] F. Eckert, *COSMOtherm User's Manual*, C2.1 Release 01.08, **2010**.

[10] T. Mu, J. Rarey, J. Gmehling, *Indust. Eng. Chem. Res.* **2007**, *46*, 6612–6629.

[11] A. Klamt, F. Eckert, W. Arlt, *Ann. Rev. Chem. Biomol. Eng.* **2010**, *1*, 101–122.

[12] E. N. Muratov, E. V. Varlamova, A. G. Artemenko, T. Khristova, V. E. Kuz'min, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, *Future Med. Chem.* **2011**, *3*, 15–27.

[13] A. Kravtsov, P. Karpov, I. Baskin, V. Palyulin, N. Zefirov, *Dokl. Chem.* **2007**, *414*, 128–131.

[14] A. Kravtsov, P. Karpov, I. Baskin, V. Palyulin, N. Zefirov, *Dokl. Chem.* **2011**, *440*, 299–301.

[15] A. Kravtsov, P. Karpov, I. Baskin, V. Palyulin, N. Zefirov, *Dokl. Chem.* **2011**, *441*, 314–317.

[16] D. Ravindranath, B. J. Neely, R. L. Robinson, K. A.M. Gasem, *Fluid Phase Equilibr.* **2007**, *257*, 53–62.

[17] S. Ajmani, S. C. Rogers, M. H. Barley, D. J. Livingstone, *J. Chem. Inf. Model.* **2006**, *46*, 2043–2055.

[18] S. Ajmani, S. C. Rogers, M. H. Barley, A. N. Burgess, D. J. Livingstone, *QSAR Comb. Sci.* **2008**, *27*, 1346–1361.

[19] S. Ajmani, S. C. Rogers, M. H. Barley, A. N. Burgess, D. J. Livingstone, *Mol. Inf.* **2010**, *29*, 645–653.

[20] A. R. Katritzky, I. B. Stoyanova-Slavova, K. Tämm, T. Tamm, M. Karelson, *J. Phys. Chem.* A **2011**, *115*, 3475–3479.

[21] E. N. Muratov, E. V. Varlamova, A. G. Artemenko, P. G. Polishchuk, V. E. Kuz'min, *Mol. Inf.* **2012**, *31*, 202–221.

[22] J. W. Kang, K. P. Yoo, H. Y. Kim, H. Lee, D. R. Yang, C. S. Lee, *Int. J. Thermophys.* **2001**, *22*, 487–494.

[23] N. Alpert, P. J. Elvinq, *Indust. Eng. Chem.* **1951**, *43*, 1174–1177.

[24] K. J. Miller, H.-S. Huang, *J. Chem. Eng. Data* **1972**, *17*, 77–78.

[25] T. Hiaki, K. Yamato, K. Kojima, *J. Chem. Eng. Data* **1992**, *37*, 203–206.

[26] C. E. Kirby, M. Van Winkle, *J. Chem. Eng. Data* **1970**, *15*, 177–182.

[27] I. V. Tetko, *The Prediction of Physicochemical Properties in Computational Toxicology*, Wiley, New York, **2006**, p. 240–275.

[28] E. N. Muratov, A. G. Artemenko, E. V. Varlamova, P. G. Polischuk, V. P. Lozitsky, A. S. Fedchuk, R. L. Lozitska, T. L. Gridina, L. S. Koroleva, V. N. Sil'nikov, A. S. Galabov, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, V. E. Kuz'min, *Future Med. Chem.* **2010**, *2*, 1205–1226.

[29] V. E. Kuz'min, A. G. Artemenko, E. N. Muratov, *J. Comput. Aided Mol. Des.* **2008**, *22*, 403–421.

[30] V. E. Kuz'min, A. G. Artemenko, E. N. Muratov, I. L. Volineckaya, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, *J. Med. Chem.* **2007**, *50*, 4205–4213.

[31] Available from: http://infochim.u-strasbg.fr/.

[32] L. Breiman, *Mach. Learn.* **2001**, *45*, 5–32.

[33] P. G. Polishchuk, E. N. Muratov, A. G. Artemenko, O. G. Kolumbin, N. N. Muratov, V. E. Kuz'min, *J. Chem. Inf. Model.* **2009**, *49*, 2481–2488.

[34] L. Breiman, *The Wadsworth Statistics/Probability Series*, Wadsworth International Group, Belmont, CA, **1984**, p. 358.

[35] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–58.

[36] I. V. Tetko, *Neural Process Lett.* **2002**, *16*, 187–199.

[37] I. V. Tetko, D. J. Livingstone, A. I. Luik, *J. Chem. Inf. Comp. Sci.* **1995**, *35*, 826–833.

[38] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, *IEEE Trans. Neural Net*w. **2001**, *12*, 181–201.

[39] A. K. Seewald, in *Proc. 19th Int. Conf. Mach. Learn.*, Morgan Kaufmann, Waltham, MA, **2002**, pp. 554–561.

[40] J. B. Kruskal, *Proc. Am. Math. Soc.* **1956**, *7*, 48–50.

[41] V. P. Solov'ev, I. Oprisiu, G. Marcou and A. Varnek, *Indust. Eng. Chem. Res.* **2011**, *50 (24)*, 14162–14167.

[42] E. B. Wilson, M. M. Hilferty, *Proc. Natl. Acad. Sci. USA* **1931**, *17*, 684–688.

[43] E. Mullins, R. Oldland, Y. A. Liu, S. Wang, S. I. Sandler, C.-C. Chen, M. Zwolak, K. C. Seavey, *Indust. Eng. Chem. Res.* **2006**, *45*, 4389–4415.

[44] J. Wisniak, A. Tamir, *J. Chem. Eng. Data* **1985**, *30*, 339–344.

[45] A. Chinikamala, *J. Chem. Eng. Data* **1973**, *18*, 322–325.

Received: January 16, 2012
Accepted: April 23, 2012
Published online: July 6, 2012