

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/265254076>

'Quasi-Mixture' Descriptors for QSPR Analysis of Molecular Macroscopic Properties. The Critical Properties of Organic Compounds

ARTICLE · OCTOBER 2014

DOI: 10.1002/minf.201400036

READS

26

4 AUTHORS, INCLUDING:



Olena Mokshyna

National Academy of Sciences of Ukraine

9 PUBLICATIONS 5 CITATIONS

SEE PROFILE



Pavel G Polishchuk

Palacký University of Olomouc

27 PUBLICATIONS 259 CITATIONS

SEE PROFILE

'Quasi-Mixture' Descriptors for QSPR Analysis of Molecular Macroscopic Properties. The Critical Properties of Organic Compounds

E. Mokshyna,^{*,[a]} V. I. Nedostup,^[a] P. G. Polishchuk,^[a] and V. E. Kuzmin^[a]

Abstract: Rational approach towards the QSAR/QSPR modeling requires the descriptors to be computationally efficient, yet physically and chemically meaningful. On the basis of existing simplex representation of molecular structure (SiRMS) the novel 'quasi-mixture' descriptors were developed in order to accomplish the goal of characterization molecules on 2D level (i.e. without explicit generation of 3D structure and exhaustive conformational search) with account for potential intermolecular interactions. The critical properties of organic compounds were chosen as target properties for the estimation of descriptors' efficacy be-

cause of their well-known physical nature, rigorously estimated experimental errors and large quantity of experimental data. Among described properties are critical temperature, pressure and volume. Obtained models have high statistical characteristics, therefore showing the efficacy of suggested 'quasi-mixture' approach. Moreover, 'quasi-mixture' approach, as a branch of the SiRMS, allows to interpret results in terms of simple basic molecular properties. The obtained picture of influences corresponds to the accepted theoretical views.

Keywords: Quasi-mixture descriptors • Macroscopic properties • Organic compounds

1 Introduction

Since its creation, the idea of QSAR/QSPR analysis has gained great popularity. The problem of evaluation for novel descriptors is one of the most important, as there is a great variety of possible descriptors of molecular structure, but there is no single methodology to validate them in terms of physical meaning and precision. The different types of descriptors are highlighting various aspects of molecular structure, i.e. electronic or topological.^[1]

Formally, any structural characteristics can be used for QSAR/QSPR models development. However, rational approach towards QSAR/QSPR modelling requires selection of descriptors that are related to investigated property. To create interpretable models there is a need for the descriptors with clear physico-chemical meaning, that reflect the interaction features of target molecules with functional target (For every property (function) of the molecule the interaction between it and the functional target (i.e., receptor, enzyme, catalyst, other molecule) is obligatory condition to display this property.). According to that, for macroscopic properties investigation of the descriptors that describe intermolecular interactions is needed.

As the experimental data are the only reference point that chemoinformatics has, the question of a great importance is on what kind of experimental data we need to evaluate efficacy of descriptors. Historically, the main object of QSAR analysis were biological data. However, according to large experimental errors, descriptors' efficacy cannot be precisely estimated with it. The big advantage of thermody-

namic properties is that they possess clear physical and theoretical meaning, i.e. they can be described by high-level fundamental physical equations. And, what is even more important, the nature of these properties is well-known, that allows to validate models in terms of their physicochemical meaning.

Thermodynamic properties, particularly critical parameters, can be strictly defined in terms of interactions. It is important that QSPR analysis of thermodynamic properties allows to estimate quality of applied approaches as this thermodynamic properties are measured with high (acceptable error rate is less than 5%) and well-known experimental accuracy.

There are range of existing QSPR works, dedicated to the problem of description and prediction of critical properties. Some developed approaches were applicable to critical

[a] E. Mokshyna, V. I. Nedostup, P. G. Polishchuk, V. E. Kuzmin
Physico-Chemical Institute NAS of Ukraine
86 Lustdorfs'ka doroga, 65081 Odesa, Ukraine
*e-mail: mokshinaelena@ukr.net

Supporting information for this article is available on the WWW under <http://dx.doi.org/10.1002/minf.201400036>.

© 2014 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

temperatures for a particular class of compounds, i.e. alkylbenzenes.^[2] The dataset of critical temperatures for 298 compounds was analysed by Jurs et al.^[3] and later in work of Katritzky et al.^[4] Of particular interest are the latest research by Kazakov^[5] on critical temperatures and ratios of critical temperature to critical pressure, which employs the largest set of experimental data.

However, several drawbacks of the previous works should be pointed out. Almost all the authors apply laborious methodology of model development including conformational search and use of semi-empirical quantum methods. The validation procedures are not always rigorously performed, and/or descriptors are not unambiguously interpreted.

Also almost all of the models are aimed to predict critical temperature only, whereas critical pressure is subject of less research, and there are no articles dedicated to modelling of critical volume.

Therefore, the main goals of the current work were the development of a descriptor system, able to account for potential intermolecular interactions in pure substances, and estimation of its efficacy towards building of adequate QSPR models for all critical properties of various organic compounds. On the basis of existing mixture simplex descriptors,^[6] the new approach to the representation of pure compounds was developed, called 'quasi-mixture' descriptors.

Moreover, despite the existence of the huge collections of data on critical properties, there are still "blind spots" in knowledge about critical properties of complex organic compounds.

The goals of the current work are therefore formulated in two ways:

- The efficacy estimation and modification of the existing 2D simplex descriptors towards development of the new 'quasi-mixture' descriptors;
- Development of the simple 2D QSPR models based on the simplex representation of molecular structure,^[7] which are able to predict critical properties and correspond to the existing theoretical knowledge, without launching complex and time-consuming 3D workflows.

2 Materials and Methods

2.1 Experimental Data Collection and Curation

The experimental data were taken from the comprehensive handbook.^[8] Then wrong or incomplete data were cured using NIST Webbook database.^[9] Among considered compounds were those of various classes, such as saturated and unsaturated hydrocarbons, aromatic hydrocarbons and their derivatives, heterocyclic compounds, alcohols, ethers, esters, various halogenated compounds, etc. Number of considered compounds was not equal for different proper-

ties. The experimental T_c values were available for 407 compounds, P_c – for 382 compounds, V_c – for 309 compounds. Collected data cover the large range of values (for critical temperature from about 100 K to about 900 K, critical pressure from about 10 to about 90 bar, critical volume from about 100 to about 1000 cm³/mol), see Supporting Information 1 for more details).

Collected data were thoroughly checked for errors, i.e. wrong structures, misspelled names of compounds, wrong endpoint values. Duplicates were deleted using TheorChem in-house software, afterwards all structures were standardized using ChemAxon Standardizer plugin.^[10]

The data (SMILES structures and property values) are available in Supporting Information 2.

2.2 Structure Description

The main idea of SiRMS is based on the rigorous mathematical concept of 11 basic tetraatomic fragments with fixed composition and connectivity (Table 1). Di-, tri- or pentaatomic fragments are used as well.

Table 1. Basic types of simplexes.

Basic type	Fragment	Example
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		



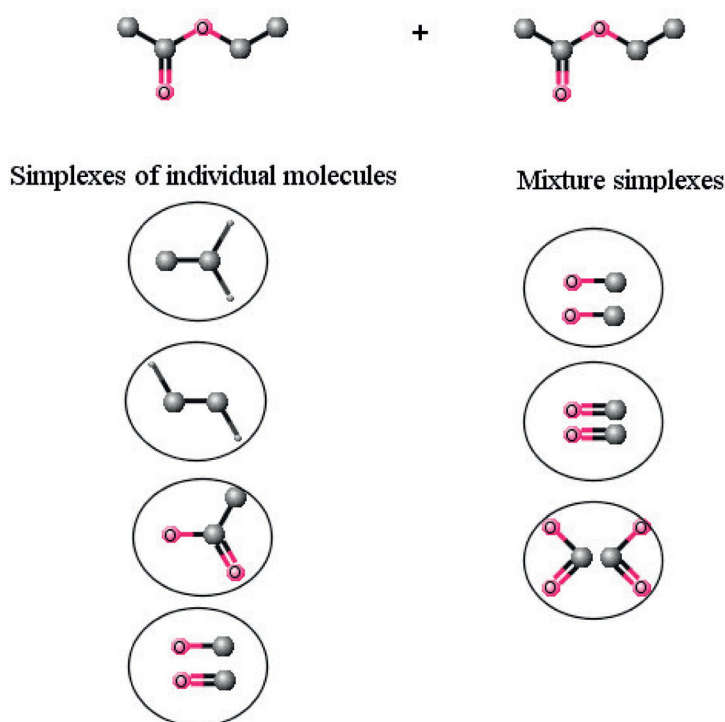


Figure 2. Simplex descriptors of pure compound in 'quasi-mixture' approach.

chemical" thermodynamic methodologies.^[16] Within the framework of the "quasi-mixture" approach, the pure compound is depicted as mixture of two molecules. Due to the character of the mixture descriptors it does not lead to simple duplication of simplexes, but leads to appearance of new unique types of mixture simplexes (Figure 2). All descriptors were calculated using in-house software TheorChem.

2.3 Modelling Workflow

The workflow of QSPR modelling was the following:

- Calculation of the quasi-mixture SiRMS descriptors
- Preprocessing of the calculated descriptors, which included removing of variables with low variability and cross-correlations with cut-off level more than 0.9
- Development of the Random Forest models
- Models validation with the 3×5-fold external cross-validation procedure
- Applicability domain estimation

Table 2. Performance characteristics and their 95%-confidence intervals for the consensus models' predictions (for the 3 times 5-fold external cross-validation).

		T_c	V_c	P_c
Single molecule	R^2	0.87 ± 0.04	0.96 ± 0.01	0.88 ± 0.03
	RMSE	40.1 ± 0.7 (K)	28.0 ± 0.10 (cm ³ /mol)	4.45 ± 0.07 (bar)
Quasi-mixture	R^2	0.89 ± 0.02	0.96 ± 0.01	0.90 ± 0.02
	RMSE	37.2 ± 0.7 (K)	24.0 ± 0.12 (cm ³ /mol)	4.17 ± 0.05 (bar)

The Random Forest method is a non-linear method of the machine learning, based on use of the ensembles of decision trees.^[17] The main advantages of this method are speed, robustness and its own measure of predictive ability of the models – so called out-of-bag correlation coefficient (R^2_{oob}). For model development in-house implementation of RandomForest method was used.^[18]

Random forests can be used to rank the importance of variables in a regression or classification problem. To measure the importance of the j -th feature after training, the values of the j -th feature are permuted among the training data and the error is computed again on these perturbed data. The importance score for the j -th feature is computed by averaging the difference in the error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences.

Features which produce large values for this score are ranked as more important than features which produce small values.

The 3 times 5-fold external stratified cross-validation^[7] is quite useful for obtaining the objective predictive charac-

teristics of the models. During this procedure data are preliminary ranged by value of the property investigated, divided into 5 parts or folds, and then each 20% of compounds in each fold are placed into the test set. The procedure is repeated so the each compound is in the test set once. The 5-fold external cross-validation procedure is repeated three times with reshuffling. All the resulting models were combined in a final consensus model, which can be further used for prediction of various compounds' critical properties.

The variable importance was calculated following the procedure, developed by Breiman.^[17] The first step in measuring the variable importance in a dataset is to fit a random forest to the data. During the fitting process the out-of-bag error for each data point is recorded and averaged over the forest (errors on an independent test set can be substituted if bagging is not used during training). To measure the importance of the j -th feature after training, the values of the j -th feature are permuted among the training data and the out-of-bag error is again computed on this perturbed data set. The importance score for the j -th feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences. Features which produce large values for this score are ranked as more important than features which produce small values.

3 Results and Discussion

The series of models were built using the workflow described above. For the sake of comparison there were two sets of models: with traditional SiRMS descriptors (further "single molecule" models) and with "quasi-mixture" approach (correspondingly "quasi-mixture" models).

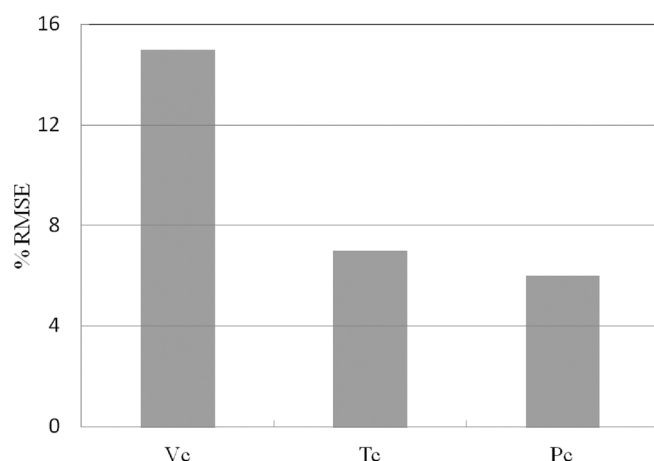


Figure 3. Percentage increase in 'quasi-mixture' models prediction accuracy relative to 'single molecule' models.

The characteristics of the models' performance are presented in Table 2. The calculation of the 95% confidence intervals for statistical parameters was performed across the models for different folds (15 values altogether). Despite the fact, that correlation coefficients of the obtained

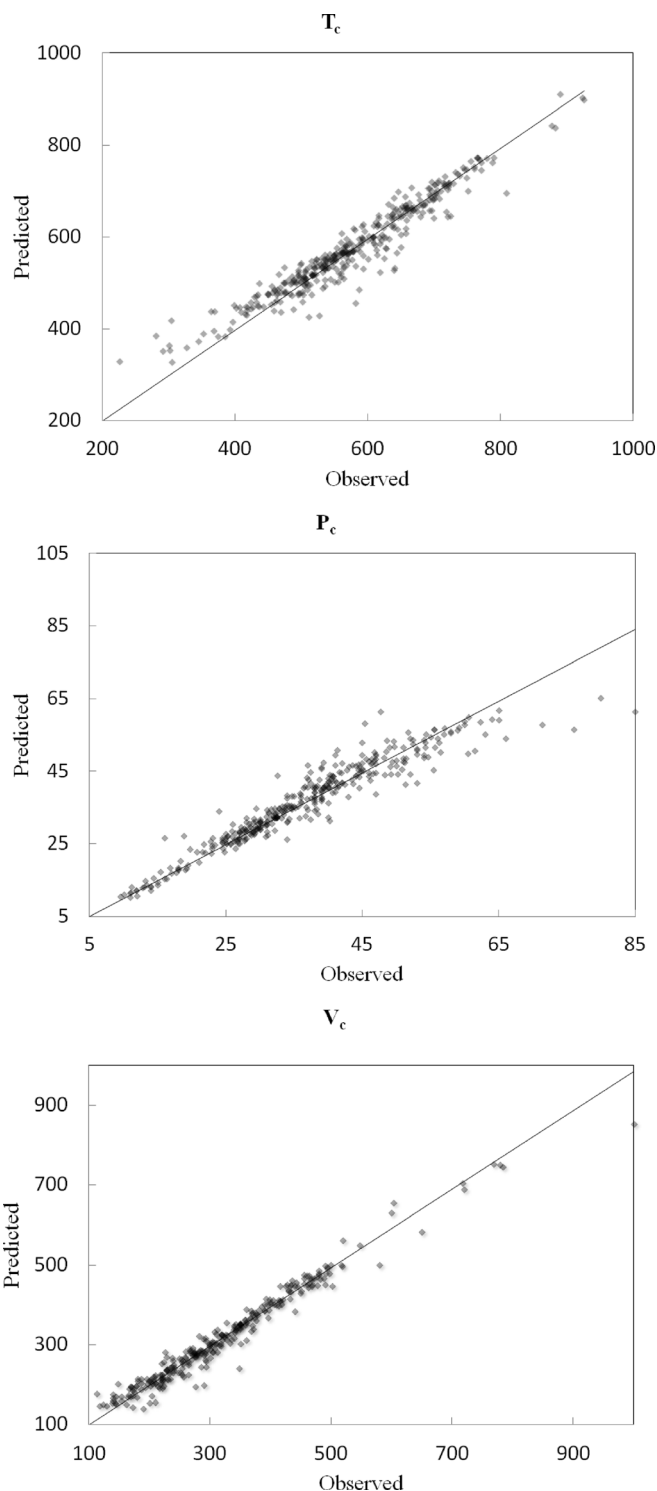
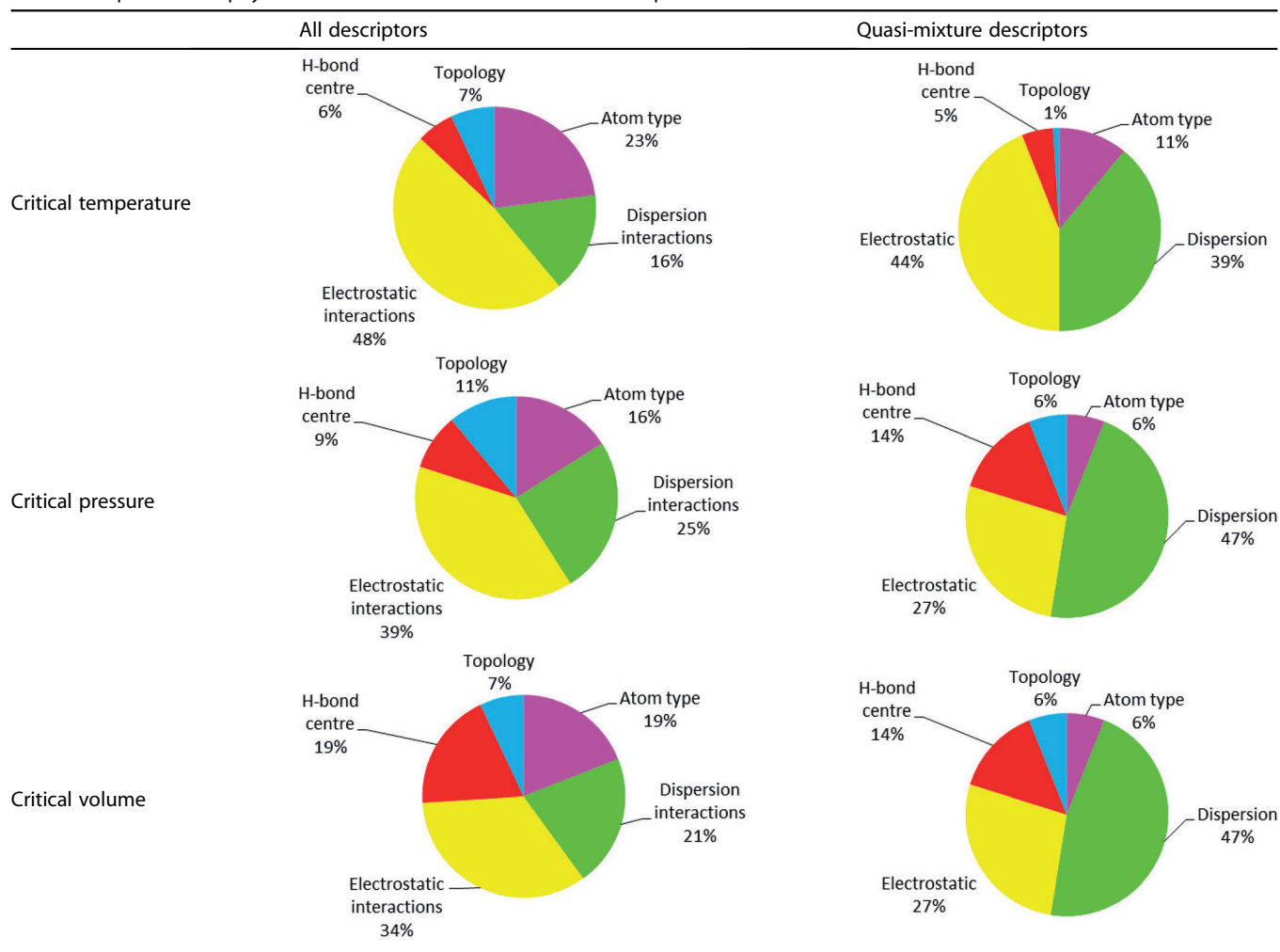


Figure 4. Scatterplots 'observed vs. predicted' for 'quasi-mixture' models.

Table 3. Importance of physicochemical molecular characteristics for 'quasi-mixture' models.

models are close and differ insignificantly, the difference in RMSE is statistically significant within the calculated 95 % confidence interval. To demonstrate the difference between models the decrease in *RMSE* (%) was shown on Figure 3. The best increase in performance is regarded for V_c – about 15%, then for T_c – about 7% and P_c – 6%. The largest improvement in V_c models can be explained by the additive nature of the property.

The largest deviations are observed near the margins of the scatterplots (Figure 4). The larger prediction errors for the smaller values of critical properties are due to the complexity of description of quite simple molecules as methane because of the lack of unique simplex features that characterize them. On the other hand, the bigger prediction errors for large values of critical properties are provided by the increase in experimental error for more complex molecules, especially their oxygen containing derivatives. For example, experimental error for methane and ethane are only 0.3 K, while for nonadecane and eicosane this error is 8 K. For oxygen containing compounds like ethanol, acetalde-

hyde and acetic acid experimental errors correspondingly are 7 K, 2 K and 30 K.^[9]

One of the biggest advantages of usage of property-labeled descriptors is possibility to interpret models in terms of basic molecular properties. In Table 3 the importance of different molecular features for the critical properties in the 'single'- and 'quasi-mixture' models correspondingly are shown. Calculated rate of "quasi-mixture" descriptors for all three properties (T_c , P_c , V_c) were approximately 25%.

For the 'quasi-mixture' models the most important are electrostatic and dispersion interactions, which play important role for the critical properties. In this case by dispersion interactions we mean sum of importances for all descriptors that are weighted by constants of Lennard-Jones potential 6–12 and by electrostatic interactions – sum of importances for refractivity, partial charges and electronegativity, which correspond to dipole-dipole, dipole-induced dipole interactions mainly. The role of topology is relatively small for all the critical properties investigated. Quite remarkable is the fact that atom's ability to form hydrogen

Table 4. Examples of the most important unconnected “quasi-mixture” simplexes. EI: electrostatic interactions, DI: dispersion interactions, AT: atom type, D/A: donor/acceptor of hydrogen bond; C.ar: aromatic carbon; C.3: sp³ carbon.

Tc			Pc			Vc		
Fragment	Type	Example	Fragment	Type	Example	Fragment	Type	Example
	EI	-0.020 -0.045		EI	-0.135 -0.045		EI	0.260 -0.020 -0.045
	EI	-0.031 -0.009 -0.170		DI	-0.020 0.044		DI	0.105 0.060 0.105
	AT	0.041 H		EI	-0.214 0.250 0.010 H		AT	0.247 H
	DI	0.050 0.105		AT	0.105 H		DI	0.247 0.105
	AT	0.105 C.3 H		D/A	D A I		D/A	I A I

bond is more important for critical pressure – 19%, in comparison with 6% and 9% for critical temperature and critical volume correspondingly. The clear explanation of this fact is complicated due to the close relations between all the three properties and requires further investigation.

Comparison of variable importance plots for all descriptors and ‘quasi-mixture’ descriptors only (Table 3) shows that the importance of the dispersion interactions is higher. In case of critical pressure the importance of hydrogen bonding is significantly higher in case of ‘quasi-mixture’ descriptors. The importance of ‘quasi-mixture’ hydrogen bond can be seen as an importance of intermolecular hydrogen bonding, which responds to formation of associates, and therefore influences most of all the critical pressure. The

topology importance for ‘quasi-mixture’ descriptors is much smaller. The importance of simplexes labeled by atom types is reduced, especially in case of critical pressure and volume.

Detailed analysis of descriptors’ importance for ‘single’-models can be found in Supporting Information 3.

Among different fragments the majority are simplex tetraatomic fragment. The most important simplexes were chosen and results are presented in Table 4. Certainly, this shouldn’t be interpreted as four-center interaction, but as importance of pair atomic interactions for atoms in certain environment. It can be seen that the most important simplexes are those of type 2 and 3. General trends of property importance slightly changed. For example, for critical

volume and critical pressure one of the most important fragments are type 2 simplexes weighted by donor/acceptor of hydrogen bond, whilst for critical temperature those are absent in the most important simplexes. For all three properties the most important simplexes are those of type 2 that describe electrostatic interactions, followed by type 2 for dispersion interactions.

4 Conclusions

In current work, new “quasi-mixture” descriptors were developed. It was shown that they can meaningfully describe complex thermodynamic properties using unique features of intermolecular interactions. The obtained “quasi-mixture” QSPR models showed statistically significant increase in performance in comparison with ordinary (“single molecule”) models. Despite the close values of determination coefficients for prediction, difference in RMSE is statistically significant.

Generally, the most influential are descriptors that characterize electrostatic and dispersion interactions, and ability to form hydrogen bond. Obtained interpretation is consistent with the accepted thermodynamic theoretical dependencies. In general, taking into account lower prediction error and clearer physico-chemical interpretation of descriptors, the “quasi-mixture” approach shows perspectives for further use and development.

Moreover, obtained 2D-SiRMS models allow predicting critical properties with the acceptable accuracy for various classes of organic compounds. Therefore properly selected 2D descriptors are able to accurately describe critical properties, but also they have a great advantage of simplicity and quick computational workflow without exploiting, i.e. procedures of conformational search and/or quantum chemical optimization. On the other side, fundamental physico-chemical properties of atoms are included in 2D descriptor's weighting so they obtain clear physical meaning

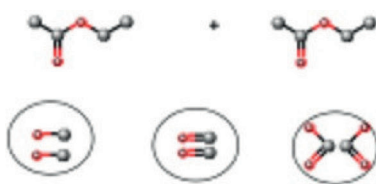
References

- [1] V. Consonni, R. Todeschini, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, **2000**.
- [2] S. Yang, W. Lu, N. Chen, Q. Hu, *J. Mol. Struct.: THEOCHEM* **2005**, 719, 119–127.
- [3] B. E. Turner, C. L. Costello, P. C. Jurs, *J. Chem. Inf. Model.* **1998**, 38, 639–645.
- [4] A. R. Katritzky, U. Maran, V. S. Lobanov, M. Karelson, *J. Chem. Inf. Model.* **2000**, 40, 1–18.
- [5] A. Kazakov, C. D. Muzny, V. Diky, R. D. Chirico, M. Frenkel, *Fluid Phase Equilib.* **2010**, 298, 131–142.
- [6] I. Oprisiu, E. Varlamova, E. Muratov, A. Artemenko, G. Marcou, P. Polishchuk, V. Kuz'min, A. Varnek, *Mol. Inf.* **2012**, 31, 491–502.
- [7] T. Puzyn, J. Leszczynski, M. T. D. Cronin, in *Challenges and Advances in Computational Chemistry and Physics*, Springer, Heidelberg, **2010**, pp. xiv, 423 p.
- [8] R. C. Reid, J. M. Prausnitz, B. E. Poling, *The Properties of Gases and Liquids*, McGraw-Hill, New York, **1987**.
- [9] M. Frenkel, in *NIST Chemistry WebBook, NIST Standard Reference Database Number 69* (Ed: P. J. Lindstrom, W. G. Mallard), National Institute of Standards and Technology, Gaithersburg MD, 20899, **2013**.
- [10] *Standardizer*, ChemAxon: www.chemaxon.com **2013**.
- [11] *Tripos*, http://www.tripos.com/mol2/atom_types.html.
- [12] W. L. Jolly, W. B. Perry, *J. Am. Chem. Soc.* **1973**, 95, 5442–5450.
- [13] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, W. M. Skiff, *J. Am. Chem. Soc.* **1992**, 114, 10024–10035.
- [14] A. G. Artemenko, E. N. Muratov, V. E. Kuz'min, N. A. Kovdienko, A. I. Hromov, V. A. Makarov, O. B. Riabova, P. Wutzler, M. Schmidtke, *J. Antimicrob. Chemother.* **2007**, 60, 68–77.
- [15] E. N. Muratov, E. V. Varlamova, A. G. Artemenko, P. G. Polishchuk, V. E. Kuz'min, *Mol. Inf.* **2012**, 31, 202–221.
- [16] B. E. Poling, J. M. Prausnitz, J. P. O'Connell, *The properties of gases and liquids*, McGraw-Hill, **2001**.
- [17] L. Breiman, *Machine Learning* **2001**, 45, 5–32.
- [18] a) 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; b) *Random Forest PCI*, <http://qsar4u.com/pages/rf.php>.

Received: March 17, 2014


Accepted: June 11, 2014

Published online: ■ ■ ■, 0000



E. Mokshyna, V. I. Nedostup,
P. G. Polishchuk, V. E. Kuzmin*

■■■ – ■■■

‘Quasi-Mixture’ Descriptors for QSPR 
**Analysis of Molecular Macroscopic
Properties. The Critical Properties of
Organic Compounds**