HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Marie Curie Initial Training Network

Environmental Chemoinformatics (ECO)

Final project report

30.06.2012

Modeling of non-additive mixtures properties using On-line CHEmical database and Modeling Environment (OCHEM)

Duration of Short Term fellowship:

4.5 months

Early stage researcher:

Miss OPRISIU IOANA

Project supervisor:

Dr. TETKO IGOR

Research Institution:

Helmholtz Zentrum München

1 SUMMARY

The Online Chemical Modeling Environment (OCHEM, http://ochem.eu)[1] is a web-based platform that provides tools for automation of typical steps necessary to create a predictive QSAR/QSPR model. The platform consists of two major subsystems: the database of experimental measurements and the modeling framework. So far, OCHEM was limited to processing of individual compounds. The aim of this work was to implement new features into OCHEM that allow to store and model properties of binary non-additive mixtures.

The study was developed in two directions: Firstly, new features were implemented in OCHEM that allow reading and uploading data for mixtures, creating special descriptors for mixtures and validating models. Secondly, models to predict non-additive properties for binary mixtures, where developed. For this, data for binary mixtures was collected from different sources[2, 3, 4]. Next, data were prepared in a specific format to be uploaded into OCHEM and ultimately, models were developed, using different descriptors type and machine learning implemented in OCHEM.

2 INRODUCTION

Generally QSPR (Quantitative Structure Property Relationship) models are limited to predict properties of pure compounds. Nevertheless, in the last years several studies to develop QSPR models to predict non-additive properties (density[2], infinite dilution activity coefficient[5], bubble temperature[3] azeotropic behavior[4], or excess molar volume[6]) of mixtures were carried out.

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. Another question is related to proper external validation of models for mixtures, which is less obvious than in classical QSAR.

2.1 Descriptors

Mixtures' descriptors developed in this study were constructed as suggested in the previous work [3, 7], based on the descriptors of individual components of the mixture. Figure 1 sows two different descriptor types developed according to the modeled property.

(i) When the value of the mixture's property doesn't depend on the concentration of its components, mixtures' descriptors are obtained by simple (unweighted) average or sum and absolute difference of the descriptor values corresponding to the individual constituents of the mixture. These mixtures' descriptors have been used for the prediction of azeotropic behavior of binary mixtures in previous study[4] and also in our work. It should be noted that to each mixture corresponds one property value.

(ii) When the value of the mixture's property changes with the concentration of its components, mixtures' descriptors are calculated as mole weighted sum or weighted sum and weighted absolute difference, using the descriptor value and mole fraction of each pure component in the mixture. These mixtures' descriptors have been used to predict density and bubble point of binary mixtures in previous studies[2, 3] and also in our work. In should be noted that in this case each mixture is defined by several points (several property values) corresponding to different concentration values of its components.

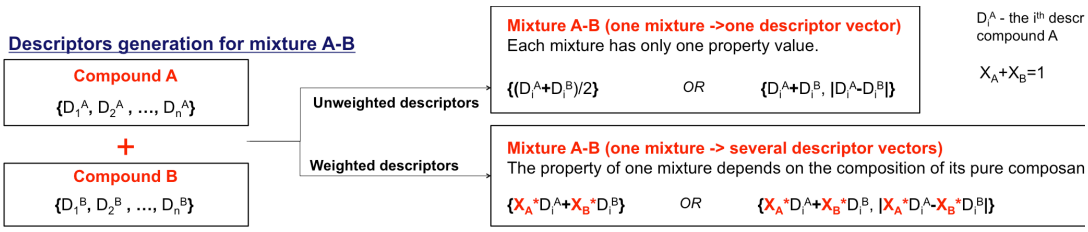

Figure 1 Methodology used for obtaining descriptors for mixtures.

2.2 Validation protocol

As in classical QSAR modeling, rigorous external validation is a needed and it is also important for QSAR modeling of mixtures. However, the conventional external cross-validation procedure, when the points are randomly placed in the external set is insufficient because it leads to an overestimation of the predictive performances of the developed models, especially when mixtures of the same compounds with different ratios are present several times in the dataset. Indeed, if both training and external sets include data points corresponding to the same mixture, the model's true predictive performance will not be estimated properly.

A rigorous protocol for external validation was developed specially for QSAR modeling of mixtures and presented in [3] and involves three different strategies:

- “Points out”: data points are randomly placed in each fold of the external cross-validation set. Every mixture is present simultaneously in both training and external sets.

- “Mixtures out”: All data points corresponding to mixtures composed of the same constituents but in different ratios are simultaneously removed and placed in the same external fold. Thus, every mixture is present either in the training or external set, but never in both sets.

- “Compounds out”: Pure compounds and their mixtures are simultaneously placed in the same external fold. Thus, every mixture in the external set contains at least one compound that is absent in the training set.

“Points out” strategy is applicable only if several mixtures with the same components with different ratios are present in the modeling set. This method reflects the capability of models to predict only existing mixtures with novel composition and, therefore, its usefulness is rather low. Therefore, only “Mixtures Out” and “Compounds Out” strategies were used in our study.

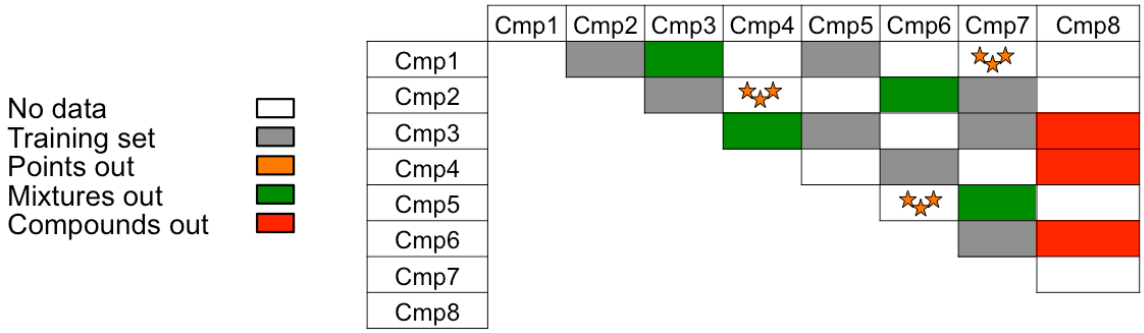

Figure 2 Protocol for external validation of mixtures’ properties models

3 METHOD: INTEGRATION WITH ONLINE CHEMICAL MODELING ENVIRONMENT

3.1 Data format

Henceforth, data for mixtures can be stored in OCHEM database and for that, an excel file should be provided which contains the needed information for mixtures. Each data point is represented by a row in the file, which must contain: the structure of the first compound (eg. SMILES) in the mixture, its molar fraction, the id number corresponding to the identification of the pure compound in OCHEM database and the

value of the property of mixture data point. It should be noted that all pure compounds contained in the mixtures should be present independently in the database.

The first compound in the binary mixture is always the one with the highest molar fraction value than the second one. Therefore, the values of molar fraction range between 0.5 and 1. It should be noted that for binary mixtures, the sum of molar fractions of its two components equals always 1. Therefore the value of the molar fraction of the second compound can be easily obtained when the molar fraction of the first compound is known.

3.2 Descriptors and validation protocol

As it was mentioned in section 2.1, special descriptors for mixtures were constructed based on the descriptors computed for single compounds constituting the mixture. The user can choose between four different descriptor types obtained by simple averaging of the descriptors, sum and absolute difference of the descriptors, weighted sum of the descriptors or weighted sum and weighted absolute difference of the descriptors.

Moreover, two different validation types are implemented "mixtures out" and "compounds out". The default validation type is "compounds out". Figure 3 represent a screenshot of the OCHEM web interface showing the new features.

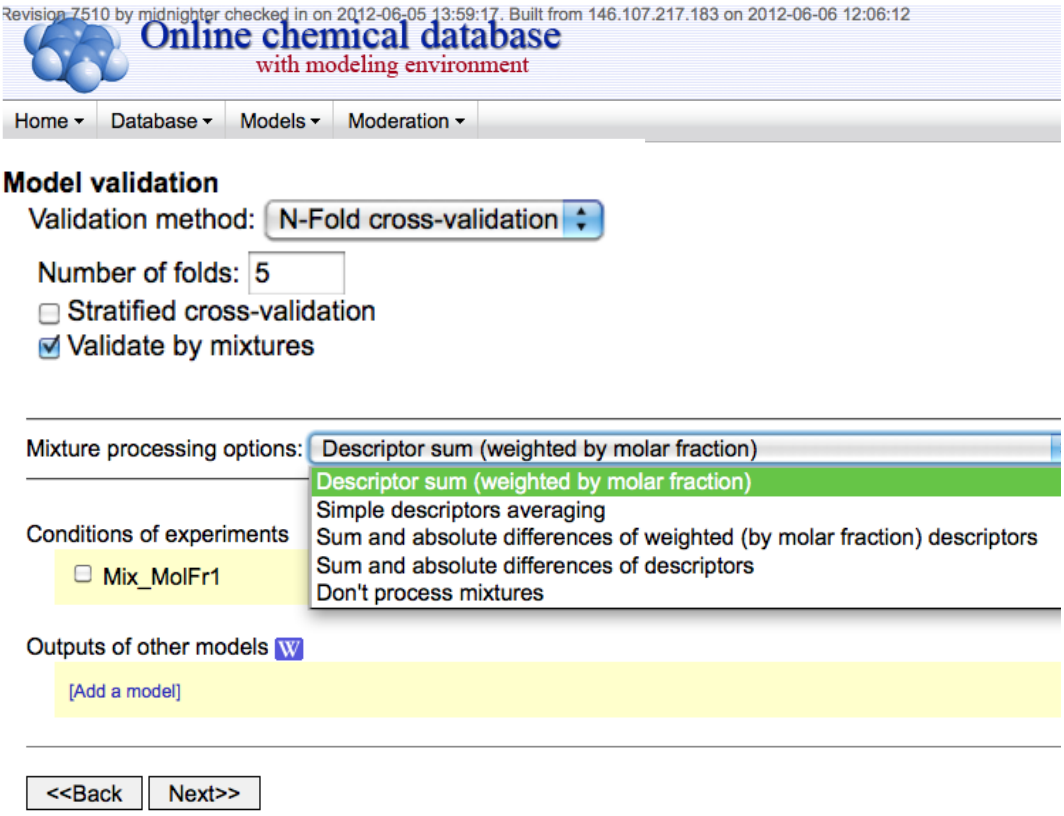

Figure 3 Screenshot of the new OCHEM features

4 RESULTS: USE CASE

QSPR models were developed both for qualitative endpoints (azeotrope/zeotrope) and for quantitative ones (density and bubble points) using different learning methods (Asociative Neural Network[8] and Random Forest[9]) and special descriptors for mixtures based on different sets of descriptors (Chemaxon[10] and ISIDA substructural fragments[11]) and the particular validation methods presented before. In all cases the prediction performance reached similar or better accuracy as reported in previous studies[3, 4].

4.1 Density

Ajmani et al. studied binary mixtures' density in [2] where the QSPR methodology was applied to 4679 data points of experimental measured density of binary liquid mixtures compiled from the literature, corresponding to 271 binary mixtures. QSPR models were developed to predict the deviation of the experimental mixture density from the "ideal" mixture density calculated by combining the densities of the single components according to their ratio in the mixture. Two different ways of training/test set creation were used: QMD-1 and QMD-2, which corresponds to "Points Out" and respectively "Mixtures Out" validation strategies (see section 2.2).

All the models developed were predictive ($Q2>0.75$), however, their predictability was limited by the missing points in the constitution of the given mixture (QMD-1) or missing mixtures of pure compounds in the modeling set (QMD-2) that limits the application of developed models. In addition, 8 duplicate mixtures were found in the data set, which definitely contributed to the higher statistical results of the models reported in the previous study. Moreover, the authors used only once 39 out of 271 mixtures (for QMD-2 strategy) for external validation. 10-fold cross validation, a more suitable method, is reported. Unfortunately the lack of details suggests that cross validation was used on data points, which is the equivalent of "Points Out" cross-validation.

For the reasons aforementioned, we used our own cross-validation strategy and we modeled directly the density instead of "delta" values.

After eliminating duplicates, the training set contained 3858 data points and 672 for the test set as suggested by Adjami.

Models to predict density of binary mixtures were developed using ASNN machine learning and Chemaxon descriptors combined by simple weighted sum.

Good results were obtained in all cases with a determination coefficient Q^2>0.98 and RMSE<20.1 Kg/cm3 (see Table 1).

Figure 4 shows the predicted values of density versus the experimental ones for "Mixtures Out" and "Compounds out" cross-validation strategies (blue points) and also for the test set (green points).

Table 1 Performances of the developed models to predict density, for external cross-validation (CV) and external test set.

	CV-Mixtures Out (3858 data points)	**CV-Compounds Out (3858 data points)**	**Test set (672 data points)**
Q^2	0.99	0.98	0.99
RMSE	16.14	20.09	14.98
MAE	10.04	14.32	10.07

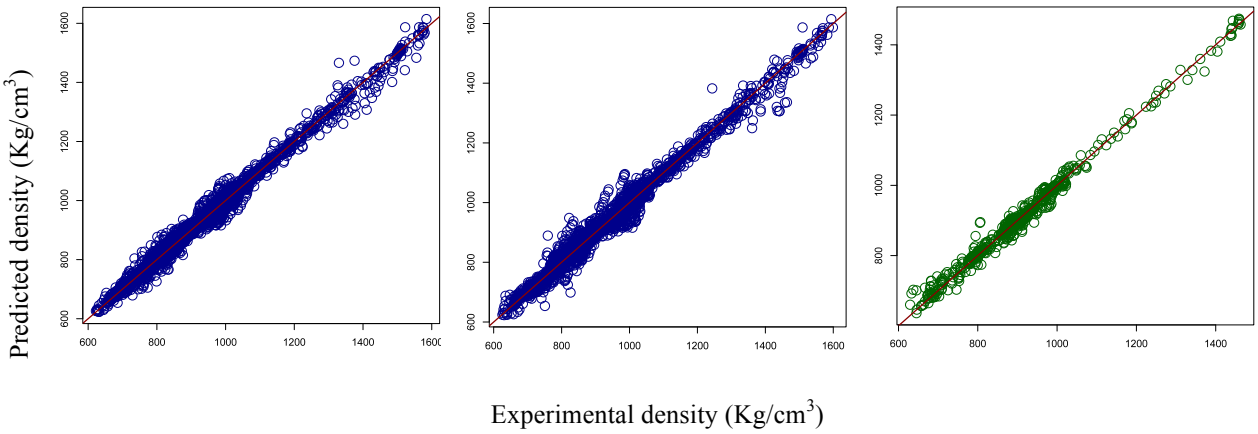

Figure 4 Predicted densities versus the experimental ones for "Mixtures Out" and "Compounds out" strategies (blue points) and for the test set (green points).

4.2 Bubble Point Temperatures

Vapor-liquid equilibrium is a condition where a liquid and its vapor are in equilibrium with each other, a condition where the rate of evaporation equals the rate of condensation. VLE curve shows 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.

Bubble point temperatures of 167 mixtures containing 3232 data points were modeled by Oprisiu et al. [3]. An external test set of 94 mixtures containing almost 2000 data points was also used. Consensus models were developed using three different machines learning: Support Vector Machine (SVM), Associative Neural Networks

(ASNN), and Random Forest (RF). For SVM and ASNN calculations, the ISIDA[11] fragment descriptors were used, whereas Simplex[12] descriptors were employed in RF models.

The same data was modeled with OCHEM using ASNN machine learning and Chemaxon[10] and ISIDA descriptors. Table 2 shows that the performances of the developed models are similar to those from the previous work. This shows the usefulness of our tool.

Table 2 Performances of the models obtained with OCHEM compared with those from previous work. [1]82 data points out of AD. [2]205 data points out of AD. [3]Results obtained in previous work[3].

		OCHEM_Mix[1]	OCHEM_CmpOut[2]	MixOut[3]	CmpOut[3]
Training set	Q^2	0.95	0.91	0.95	0.9
	RMSE	4.86	6.95	5.2	7.0
Test set	Q^2	0.89	0.48	0.88	0.4
	RMSE	5.6	20.5	5.9	21.4

4.3 Azeotropic behavior

Azeotropic data are most important for the design of distillation processes. Their theoretical assessment could significantly reduce the costs of selection of proper agents for industrial processes. An azeotrope is a liquid mixture, which boils at constant temperature keeping his composition fixed. When an azeotrope is boiled, at a certain composition, the resulting vapor has the same ratio of constituents as the liquid phase with which it is in equilibrium.

Classification models were developed using 400 mixtures (200 azeotropes/200 zeotropes) and validated on a data set of 95 mixtures containing only pure compounds already included in the training set. Only one 5-fold cross validation was randomly done on mixtures data set. Each mixture present in the test set may contain only known pure compounds or one or two new compounds.

The same data sets were used in OCHEM with the difference that also pure compounds of the mixtures were considered as zeotropes. Thus 465 mixtures where used as training set and 95 for the test set. „Mixtures out“ and „Compounds out“ cross-validations were considered.

Table 3 shows the classification results obtained with OCHEM, using Random Forest machine learning of Weka[13] and Chemaxon and ISIDA descriptors combined by simple averaging. The performances of the obtained models are similar or even better than those from the previous work.

Table 3 Classification results using OCHEM tool and the comparison with the results obtained in the previous work. * 465 mixtures were used with OCHEM.

	5-Fold Cross Validation (400 mixtures*)			Test set (95 mixtures)	
	OCHEM-Mix Out	OCHEM-Comp Out	**Oprisiu[4]**	**OCHEM**	**Oprisiu[4]**
BA	0.83	0.80	0.82	0.81	0.82
Recall (0)	0.85	0.81	0.78	0.76	0.73
Recall (1)	0.76	0.78	0.85	0.86	0.91

5 CONCLUSION

New features were implemented in OCHEM, http://ochem.eu, which was extended to offer free web tools to store and model binary mixtures. The theoretical part of the modeling approach was based on my PhD work carried out in the Laboratory of Chemoinformatics in Strasbourg [3, 4].

Particular descriptors for mixtures were used, based on classical (one compound) descriptors derived from the mixture's components. Moreover, specific cross-validation protocols for mixtures were implemented: “mixtures out“ and "compounds out”. In order to validate our developments, qualitative and quantitative models were developed with good predictive performances, using different learning methods and different sets of descriptors.

The main purpose of this work was to contribute publicly available tools for modelling of mixtures of chemical compounds on the web. The availability of such tools will stimulate developments in this area of research. Secondly, while it was a logical to derive descriptors for mixtures based on fragmental descriptors, the extension of this methodology to arbitrary descriptors was not obvious. We have shown that the same methodology can be successfully used with very different types of descriptors and different machine learning methods thus improving results compared to the previous studies.

Moreover, the developed approach can be used to model any other non-additive properties of binary mixtures, such as toxicity, viscosity or antiviral activity. This approach can be also extended to multiple compounds mixtures containing more than two compounds.

6 ACKNOWLEDGEMENTS

This study was partially supported with FP7 MC INT project "Enviromental Chemoinformatics", grant agreement number 238701. I would like to thank members of eADMET http://www.eadmet.com team, in particular Mr. Sergey Novotarskyi, for their help, development and programming of code required for my work.

7 ORAL COMUNICATION

Poster presented at the 3rd Summer School on Chemoinformatics in Strasbourg, France (June 25-29 , 2012):

I. OPRISIU, S. Novotarskyi, I. TETKO, G. MARCOU, A. VARNEK, "*Modeling of non-additive mixtures properties using On-line CHEmical database and Modeling Environment (OCHEM)*"

8 REFERENCES

[1] Sushko, I.;Novotarskyi, S.; et al. (**2011**), Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information, *J Comput Aid Mol Des*, 25(6), 533-554.

[2] Ajmani, S.;Rogers, S. C.; et al. (**2006**), Application of QSPR to mixtures, *J Chem Inf Model*, 46(5), 2043-2055.

[3] Oprisiu, I.;Varlamova, E.; et al. (**2012**), QSPR Approach to Predict Nonadditive Properties of Mixtures. Application to Bubble Point Temperatures of Binary Mixtures of Liquids, *Mol. Inf.*, in press.

[4] Oprisiu, I. (**2012**), Modélisation QSPR de mélanges binaires non-additifs. Application au comportement azéotropique., Strasbourg, Université de Strasbourg, Ph. D.**,** 192.

[5] Ajmani, S.;Rogers, S. C.; et al. (**2008**), Characterization of Mixtures Part 1: Prediction of Infinite-Dilution Activity Coefficients Using Neural Network-Based QSPR Models, *Qsar Comb Sci*, 27(11-12), 1346-1361.

[6] Ajmani, S.;Rogers, S. C.; et al. (**2010**), Characterization of Mixtures. Part 2: QSPR Models for Prediction of Excess Molar Volume and Liquid Density Using Neural Networks, *Mol. Inf.*, 29(8-9), 645-653.

[7] Muratov, E. N.;Varlamova, E. V.; et al. (**2012**), Existing and Developing Approaches for QSAR Analysis of Mixtures, *Mol. Inf.*, 31(3-4), 202-221.

[8] Tetko, I. V. (**2002**), Associative neural network., *Neural Processing Letters*, 16(2), 187-199.

[9] Breiman, L. (**2001**), Random forests, *Mach Learn*, 45(1), 5-32.

[10] Weber, L. (**2008**), JChem Base - ChemAxon, *Chem World-Uk*, 5(10), 65-66.

[11] Varnek, A.;Fourches, D.; et al. (**2008**), ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors, *Curr Comput-Aid Drug*, 4(3), 191-198.

[12] Kuz'min, V. E.;Artemenko, A. G.; et al. (**2008**), Hierarchical QSAR technology based on the Simplex representation of molecular structure, *J Comput Aided Mol Des*, 22(6-7), 403-421.

[13] Frank, E.;Hall, M.; et al. (**2004**), Data mining in bioinformatics using Weka., *Bioinformatics*, 20(15), 2479-2481.