CHEMICAL ENGINEERING TRANSACTIONS
VOL. 31, 2013
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at: www.aidic.it/cet
Guest Editors: Eddy De Rademaeker, Bruno Fabiano, Simberto Senni Buratti

ISBN 978-88-95608-22-8; **ISSN** 1974-9791
DOI: 10.3303/CET1331155

Prediction of Physico-Chemical Properties for REACH Based on QSPR Models

Guillaume Fayet[a], Patricia Rotureau[a*], Vinca Prana[a,b], Carlo Adamo[b,c]

[a]INERIS, Parc Technologique Alata, BP2, 60550 Verneuil-en-Halatte, France
[b]Laboratoire d'Electrochimie, Chimie des Interfaces et Modélisation pour l'Energie, CNRS UMR-7575, Chimie ParisTech, 11 rue P. et M. Curie, 75231 Paris, France
[c]Institut Universitaire de France, 103 Boulevard Saint Michel, F-75005 Paris, France.

Quantitative Structure Property Relationship models have been developed for the prediction of flash points of two families of organic compounds selected in the PREDIMOL French Project: amines and organic peroxides. If the model dedicated to amines respected all OECD validation principles with excellent performances in predictivity, the one dedicated to organic peroxides was not validated on an external validation set, due to the low number of available data, but already presented high performances in fitting and robustness. This work highlighted the need of gathering experimental data, as in progress in the PREDIMOL project, to achieve validated reliable models that could be used in a regulatory framework, like REACH. Such models are expected to be submitted to the European Joint Research Comity (JRC) and to existing tools (like the OECD ECHA QSAR Toolbox) to be available for use by industrials and regulatory instances.

1. Introduction

The new EU regulation REACH requires the evaluation of the physico-chemical properties of a large number of existing substances (143 000 pre-registered substances in 2008) before 2018 in order to allow their use. Taking into account the number of substances and properties, the timing, the economic costs, the feasibility at the R&D level and the risks for the operator, in particular for the characterization of the hazardous physico-chemical properties (explosibility, flammability), the experimental measurement of all the data reveals not realistic. Thus, the development of alternative predictive methods for the evaluation of the properties of substances was recommended in the framework of REACH.
In this context, the French PREDIMOL (molecular modeling prediction of physico-chemical properties of products) project (2012) funded by ANR (National Research Agency) has started in November 2010 for 3 years. This project is conducted by INERIS in partnership with several public and private partners. Its objective is to demonstrate that molecular modeling, notably through use of QSPR (Quantitative Structure-Property Relationships) models, is a credible alternative approach to experimental characterization to access, in a reliable and fast manner, to the whole range of physico-chemical properties of substances required by EU-REACH's regulation (annexes VII and IX) as well as for the industry in terms of property-screening method. QSPR methods allow predicting properties from the molecular structures of chemicals. The project is focused on the prediction of physico-chemical properties related to particular families of compounds, like amines and organic peroxides.

In this paper, we describe in a first part existing QSPR models applicable to predict properties relevant to these substances, debating onto their performances and limits. An inventory of existing experimental data from literature was also established in this part. Furthermore, as predictivity of QSPR models highly depends on the database suitability in terms of number of data and uncertainties of measurement, experimental databases were also consolidated in a consistent way in this project for these families of compounds. In the second part of this paper, new QSPR models developed for the prediction of the flash point property of amines and organic peroxides are presented.

Please cite this article as: Fayet G., Rotureau P., Prana V. and Adamo C., 2013, Prediction of physico-chemical properties for reach based on qspr models, Chemical Engineering Transactions, 31, 925-930 DOI: 10.3303/CET1331155

2. Inventory of existing QSPR models and experimental data

2.1 Inventory of existing QSPR models

Even if a large number of QSPR models have been developed for the prediction of physico-chemical properties required by REACH, only few references have been dedicated to models specifically developed for amines or organic peroxides. Considering amines, Koziol (2001) developed neural networks for the prediction of the boiling point of 190 compounds. This model based on 46 constitutional descriptors demonstrated a good predictive power with a standard deviation of 6°C (and R^2=0.98) on an external validation set.

Two references have been found about QSPR models focused on organic peroxides. Firstly, Romanelli et al. (2001) have proposed several models for the prediction of the density based on 14 molecules with a R^2 reaching 0.999 but without any external validation. A more recent QSPR study presented (MLR-multilinear regressions and PLS-partial least square) models developed by Lu et al. (2011) based on 16 compounds for the onset temperature and the heat of decomposition. Even if PLS models presented interesting performances in terms of fitting and robustness (e.g. R^2=0.957 and Q^2= 0.859 for heat of decomposition), once again, they were not validated on an external validation set.

Other more global models (dedicated to organic compounds) are applicable for amines and organic peroxides, but their reliability for these particular families of compounds have not been robustly evidenced since they only used few representatives (typically less than 10 in large datasets of thousands of molecules).

2.2 Inventorying experimental data

In order to develop QSPR models, an inventory of available experimental data was made. Concerning amines, different sources of data were investigated like DIPPR, NIST, Handbook of Chemistry and Physics (Haynes, 2011). In particular, the CarAtex database (2012), available online, gathers properties related to process safety (auto-ignition temperature, lower and upper flammability limits, flash points) for about 80 amines.

Considering organic peroxides, a particular attention focused on the *Datatop* (2005), a database developed by TNO collecting until 40 explosive properties for 116 compounds at different concentrations and diluents. Indeed, for stability reason, organic peroxides are rarely pure compounds. These data allow classifying organic peroxides (represented by a large variety of structures including hydroperoxydes, peroxyesters and peroxydicarbonates for example) according to the Transport of Dangerous Goods regulation (UN, 2011). In this database, reliability of data is not guaranteed and all data have to be regarded as indicative. Moreover, the source is unknown for many data. Nevertheless, it allowed observing that some property values vary for the same substance with concentration in organic peroxides.

In this context, a robust database is under consolidation in the framework of the PREDIMOL project with new data obtained in homogenous experimental conditions for 30 organic compounds. The targeted properties are heats and temperatures of decomposition, impact sensitivities, densities and flash points that will allow developing new QSPR models.

3. Development of QSPR models for the prediction of flash points

In this study, QSPR models to predict the flash point (FP) of amines and organic peroxides were developed as data were available for these both families. This property is a key measure of the flammability hazard of liquids. It is defined as the lowest temperature, corrected to a barometric pressure of 101.3 kPa, at which the vapour/air mixture above the liquid can be ignited. Substances with low flash points present higher flammability than those with higher flash points. All experimental data were obtained in closed cup apparatus.

3.1 Principle of the QSPR method

The QSPR method is based on the principle that molecules with similar structures have similar properties. The chemical structure is represented at molecular level by a series of descriptors that can be mathematically connected to experimental properties by a QSPR model. So, such model will have the following form:

$$\text{Property} = f(\text{descriptors}) \quad (1)$$

A large number of descriptors (constitutional, topological, geometrical and quantum chemical) can be calculated to describe the structure of molecules (Karelson, 2000). Many statistical tools can be also used to develop QSPR models (multi-linear regression, neural network, etc...).

In this paper, multilinear regressions were developed using the Best Multi Linear Regression (BMLR) approach (Karelson, 2000) as implemented in Codessa software (2002). This stepwise approach started with constructing two-parameter MLR models based on non-intercorrelated descriptors (with R^2 between descriptors lower than 0.1) and then it built higher rank models by adding new non-intercorrelated descriptors (i.e. with R^2 lower than 0.6 with each of the previous ones). By this way, the method guaranteed that two intercorrelated descriptors were not selected in the same model. The algorithm gave, at each rank (i.e. for each number of descriptors), the model presenting the highest correlation with the studied property. The final model was chosen as the best compromise between correlation refinement and number of descriptors.

Within the context of REACH, the development of QSPR models is encouraged providing that they respect the 5 driving principles for the validation of QSPR models drawn up by OECD (2007):

1. A defined endpoint (including experimental protocol);
2. An unambiguous algorithm;
3. A defined domain of applicability;
4. Appropriate measures of goodness-of-fit, robustness and predictive power;
5. A mechanistic interpretation, when it's possible.

The fourth OECD principle requires suitable measures of performances. To measure the goodness-of-fit of a model, the determination coefficient R^2 is calculated between predicted and experimental values. For robustness, leave-one-out (LOO) and leave-many-out (LMO) cross-validation were performed (Gramatica, 2007). Y-scrambling (Lindgren et al., 1996; Rücker et al., 2007) was processed to prevent from chance correlation. Moreover the predictive power was evaluated on an external validation set on a series of coefficients: R^2_{ext} (characterizing the correlation between predicted and experimental values in the validation set) and coefficients Q^2_{F1} (Tropsha et al., 2003), Q^2_{F2} (Schüürman et al., 2008), Q^2_{F3} (Consonni et al., 2009) and CCC (Lin, 1989; Lin, 1992).

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{TR})^2} \tag{2}$$

With $\hat{y}_i$ the predicted value of the property, y_i the experimental value of the property, $\bar{y}_{TR}$ the mean experimental value in the training set and n_{ext} the number of molecules in the validation set.

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{EXT})^2} \tag{3}$$

With $\bar{y}_{EXT}$ the mean experimental value in the validation set.

$$Q^2_{F3} = 1 - \frac{\left[\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2\right] / n_{EXT}}{\left[\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2\right] / n_{TR}} \tag{4}$$

With n_{TR} the number of molecules in the training set.

$$CCC = \frac{2\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{x} - \bar{y})^2} \tag{5}$$

In this last formula, x and y represent the experimental and predicted values, respectively.

The applicability domain (Netzeva et al. 2005; Jaworska et al., 2005) required by the third OECD principle was determined based on the descriptors included in the model. The Euclidean distance method available in Ambit discovery software (Jeliazkova and Jaworska, 2007) was used with a 95% threshold, i.e. the domain was calculated to contain 95% of the molecules of the training set. Then, the performances inside

the applicability domain were also calculated based on the sole molecules of the validation set that belong to this domain using coefficients previously presented.

3.2 Amines

A first QSPR model was derived to predict the flash point of a series of 68 amines, including both aromatic and aliphatic compounds and also ethanolamines which were extracted from the CarAtex database (2012) of INRS. For each of these compounds, 165 descriptors were calculated in Codessa software (2002) based on geometric structures optimized at the AM1 level using Gaussian09 (2009).
To allow an external validation of models, the data set was then divided into a training set, containing two thirds of the molecules of the data set and a validation set constituted by the remaining molecules. To ensure that the validation set well represented the domain of property of the model (defined by the property values in the training set), molecules were classified by increasing order of flash points and one molecule out of four was selected (2nd, 6th, etc.) to build up the validation set. Moreover, the chemical structures of the molecules in both sets were analyzed and no bias was evidenced.
Then, the BMLR approach was applied to the 51 molecules of the training set and the three-parameter equation 6 was found as the best compromise between the correlation performance and the number of descriptors.

$$FP(°C) = 337.96 - 735.48\ n_H + 4715.5\ HDCA2 + 0.46\ PPSA1 \quad (6)$$

where n_H is the relative number of H atoms, HDCA2 is the area-weighted surface charge of hydrogen-donor atoms and PPSA1 is the partial positive surface area. It is worth noting that HDCA2 is related to the hydrogen bonding ability of the molecule which can be related also to its volatility as already highlighted by Katritzky (2001).
This model is well correlated with a coefficient of determination R^2=0.956. It is also robust as evaluated by LOO and LMO cross validations (Q^2_{LOO}=0.947, Q^2_{10CV}=0.946 and Q^2_{5CV}=0.949). The Y-scrambling approach was used to check that it did not issue from chance correlation. Indeed, low R^2 were exhibited for the models obtained from randomized data.

The model was applied to the 17 molecules of the validation set as shown in Figure 2. It demonstrated high predictive power based on the various coefficients dedicated to the external validation of QSPR models (R^2_{EXT} = 0.905, Q^2_{F1} = 0.897, Q^2_{F2} = 0.897, Q^2_{F3} = 0.966, CCC = 0.949).

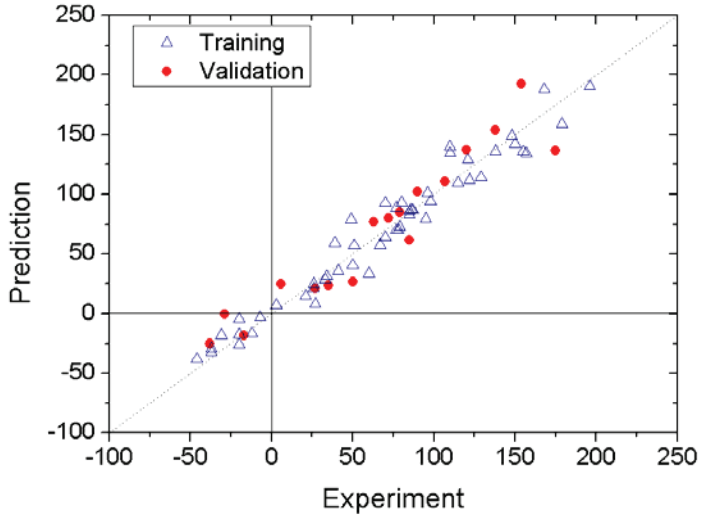

Figure 2: Calculated vs. experimental values of flash point (in °C) for amines

Finally, the applicability domain (AD) was computed and two molecules of the validation set were evidenced as out of the determined AD. So, the external validation of the model was computed again taking into account the AD. The external validation coefficients reached higher values (R^2_{EXT} = 0.905, Q^2_{F1} = 0.907, Q^2_{F2} = 0.902, Q^2_{F3} = 0.971, CCC = 0.946).
In a future step of the project, other descriptors, methods and level of theory will be investigated to achieve potentially more accurate models.

3.3 Organic peroxides

Another QSPR model was developed for 23 organic peroxides including diverse structures like hydroperoxides and peroxyesters, based on experimental results obtained in the framework of the PREDIMOL project. In this section, the DFT (density functional theory) at PBE0/6-31+G(d,p) level was used to optimize the structures of organic peroxides with the Gaussian09 (2009) software from which quantum chemical descriptors were calculated. From more than 350 descriptors and using the BMLR method, a 5-parameter model was found:

$$FP(°C) = -6127.1 + 3.46\ n_H + 2.51\ WNSA2 + 5300.2\ FPSA3 - 3.55\ PNSA3 + 1534.7\ V_{avg,C} \quad (7)$$

Where n_H is the number of H atoms, WNSA2 is the surface-weighted negative charged surface area, FPSA3 is the fractional atomic charge-weighted positive surface area, PNSA3 is atomic charge-weighted negatively charged surface area and $V_{avg,C}$ is the average valency of a C atom.

The model was characterized by a good correlation (R^2=0.921) as shown in Figure 3 and robustness (Q^2_{LOO}= 0.866, Q^2_{10CV}= 0.868 and Q^2_{5CV}=0.888). The Y-scrambling method also validated the model because of low values of R^2 for the models obtained after randomisation. However, due to the small number of data available for organic peroxides, no external validation was performed in this study. A data acquisition phase is in progress to overcome this point. It must be also pointed out that the concept of flash point as to reflect the flammability hazard of organic peroxides is limited due to potential decomposition issues, in comparison with conventional flammable liquids.

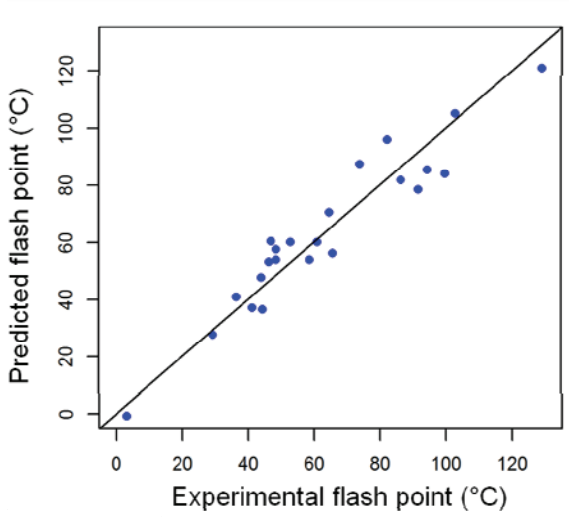

Figure 3: Experimental vs. predicted flash point for organic peroxides

4. Conclusion and perspectives

Browsing existing models and experimental data performed in the PREDIMOL project highlighted the need to consolidate robust databases, notably for organic peroxides, in order to develop accurate QSPR models for the prediction of the hazardous physico-chemical properties required by REACH. Two QSPR models were developed to predict the flash point of amines and organic peroxides respectively. The first one respects all the OECD principles and presents excellent performances in fitting, robustness and predictivity. The second one dedicated to organic peroxides was not validated on an external validation set, notably due to the low number of available data but it already presents good performance in fitting and robustness.

Further investigation will be done in this project to obtain additional data for hazardous properties of physico-chemical nature that will allow the development of new QSPR models. These models are expected to be available to industrials and regulatory instances in order that predictive data could be used for registration. For this reason, accurate QSPR models will be submitted to the European Joint Research Comity (JRC) and to existing tools (like the OECD ECHA QSAR Toolbox).

Acknowledgments

The PREDIMOL project (ANR-10-CDII-007) is financed by ANR and the French ministry in charge of environment.

References

CarAtex Database, 2012, INRS, <http://www.inrs.fr/accueil/produits/bdd/caratex.html> accessed 20.04.2012.

Codessa software, 2002, University of Florida.

Consonni V., Ballabio D., Todeschini R., 2009, Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 49, 1669-1678.

Datatop, 2005, TNO Defence, Security and Safety; Energetic Materials Research Group, Rijswijk, The Netherlands.

Haynes W. M., 2011, CRC Handbook of Chemistry and Physics, CRC Press, Colorado.

Gaussian 09, Revision B.01, 2009, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J., Gaussian, Inc., Wallingford CT.

Gramatica P., 2007, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26, 694-701.

Jeliazkova N., Jaworska J., 2007, Ambit Discovery, version 1.20.

Jaworska J, Nikolova-Jeliazkova N, Aldenberg T., 2005, QSAR applicability domain estimation by projection of the training set descriptor space: a review, Altern. Lab. Anim. 33, 445-459.

Karelson M., 2000, Molecular Descriptors in QSAR/QSPR, Wiley, New York.

Katritzky A.R., Petrukhin R., Jain R., Karelson M., 2001, QSPR analysis of flash points, J. Chem. Inf. Comput. Sci. 41, 1521-1530.

Koziol J, 2001, Neural network modelling of physical properties of chemical compounds, Int. J. Quantum Chem. 84, 17-26.

Lin L. I., 1989, A Concordance Correlation Coefficient to Evaluate Reproducibility, Biometrics 45, 255-268.

Lin L. I., 1992, Assay Validation Using the Concordance Correlation Coefficient, Biometrics 48, 599-604.

Lindgren F., Hansen B., Karcher W., 1996, Model validation by permutation tests: applications to variable selection, J. Chemometr. 10, 521-532.

Lu Y., Ng D., Mannan M. S., 2011, Prediction of the Reactivity Hazards for Organic Peroxides Using the QSPR Approach, Ind. Eng. Chem. Res. 50, 3, 1515-1522.

Netzeva T.I., Worth A., Aldenberg T., Benigni R., Cronin M.T.D., Gramatica P., Jaworska J. S., Kahn S., Klopman G., Marchant C. A., Myatt G., Nikolova-Jeliazkova N., Patlewicz G. Y., Perkins R., Roberts D. W., Schultz T. W., Stanton D. T., van de Sandt J. J.M., Tong W., Veith G., Yang C., 2005, Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52, Altern. Lab. Anim. 33, 155-173.

OECD, Organisation for Economic Co-operation and Development, 2007, Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models, OECD, Paris.

PREDIMOL project, <www.ineris.fr/predimol/> accessed 24.07.2012.

Romanelli G.P., Cafferata L.R.F., Castro E.A., 2001, Ameliorate QSPR study of alkyl Hydroperoxides, Russ. J. Gen. Chem. 71, 257-260.

Rücker C., Rücker G., Meringer M., 2007, y-Randomization and Its Variants in QSPR/QSAR, J. Chem. Inf. Moldel. 47, 2345-2357.

Schüürmann G., Ebert R., Chen J., Wang B., Kühne R., 2008, External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean, J. Chem. Inf. Model. 48, 2140-2145.

UN, Recommendations on the transport of dangerous goods: Manual of tests and criteria, ST/SG/AC.10/Rev.5 fifth revised edition, United Nations, Geneva/New-York, 2011.

Tropsha A., Gramatica P., Gombar V. K., 2003, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22, 69-77.