16th European Symposium on Computer Aided Process Engineering
and 9th International Symposium on Process Systems Engineering
W. Marquardt, C. Pantelides (Editors)

A "Targeted" QSPR for Prediction of Properties

Neima Brauner[a], Roumiana P. Stateva[b], G. St. Cholakov[c] and M. Shacham[d]

[a]*School of Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel*
[b]*Institute of Chem. Eng., Bulgarian Academy of Sciences, Sofia 1113, Bulgaria*
[c]*Dept. of Organic Synthesis and Fuels, Univ. Chem. Technol., Sofia 1756, Bulgaria*
[d]*Dept. of Chem. Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel*

Abstract

In order to improve the reliability of the Quantitative Structure-Property Relationships (QSPR) for property prediction, a "targeted" QSPR (TQSPR) method is developed, from a training set, which contains only compounds structurally similar to the target compound. Structural similarity is measured by the partial correlation coefficients between the vectors of the molecular descriptors of the target compound and those of the predictive compounds. The available properties of the compounds in the training set are then used in the usual manner for predicting the properties of the target and the rest of the compounds of unknown properties in the set. Preliminary results show that the targeted QSPR method yields predictions within the experimental error level for compounds well represented in the database and fairly accurate estimates for complex compounds that are sparsely represented. The cut-off value of the partial correlation coefficient provides an indication of the expected prediction error.

Keywords: *Quantitative structure-property relationship; QSPR, QS2PR; Property prediction; Process design;*.

1. Introduction

Modeling and simulation of chemical processes require, in addition to the process model, correlations of physical and thermodynamic properties of the various compounds, often for wide ranges of temperatures, pressures and compositions. Pure component properties are needed to derive the correlations. However, often those properties cannot be measured, or the measurements are expensive and/or unreliable.
In the recent years there has been increased interest in the development and use of Quantitative Structure-Property Relationship (QSPR) models [1-7]. The QSPR models are being extensively used for predicting a variety of pure component properties pertaining to chemistry and chemical engineering, environmental engineering and environmental impact assessment, hazard and operability analysis, etc. In the present work we will concentrate on the "most significant common features" QSPR methods, as defined in [1] which we shall call for short QSPRs henceforward. The above QSPRs can be schematically represented by the following equation:

$$y_p = f(x_{s1}, x_{s2}, \ldots x_{sk}; x_{p1}, x_{p2} \ldots x_{pm}; \beta_0, \beta_1 \ldots \beta_n) \qquad (1)$$

where x_{s1}, x_{s2},… x_{sk} are the molecular structure descriptors of a particular pure compound, x_{p1}, x_{p2},… x_{pm} are measurable properties of the same compound (such as boiling temperature, melting temperature, toxicity, etc.), β_0, β_1,… β_n are the QSPR parameters and y_p is the target property (to be predicted) of the same compound.

To derive the QSPR, the available data is divided into a "training set" and an "evaluation set". Using the "training set", multiple linear or nonlinear regression, and partial least squares techniques are employed to select the molecular descriptors and/or properties to be included in the RHS of Eq. (1), and to calculate the model parameter values. Model validation is carried out using the "evaluation set".

A limitation of the traditional QSPR approach is that if the molecular structure of the target compound belongs to a group that is well represented in the "training set", the prediction can be expected to be much more accurate than if the target compound belongs to a group which is sparsely represented [e.g. 8]. The structure-property relationships are usually nonlinear, therefore, extrapolation toward a target compound of unmeasured pure component constants can be rather risky and at present the prediction accuracy cannot be assessed. Recently Shacham et al.[9, 10] and Brauner et al. [11] presented a different approach: the Quantitative Structure - Structure Property relationship (QS2PR). This technique enables the derivation of linear property-property correlations based on a structure-structure relationship and provides an estimate of the prediction error. However it can be envisioned that in some cases it will be difficult to apply the QS2PR technique because of the lack of enough predictive compounds for which reliable measured property values exist.

In an attempt to overcome the limitations of both the QSPR and QS2PR techniques we have developed a quantitative measure of similarity between molecules and a new "targeted QSPR" (TQSPR) technique, which are described in the next section.

2. The Targeted-QSPR method

The TQSPR method attempts to tailor a QSPR to an unknown (*target property*) of a particular compound (*target compound*). For its effective use a database of molecular descriptors, x_{ij} and physical properties y_{ij} for the predictive compounds, where i is the number of the compound and j is the number of the descriptor/property, is required. Molecular descriptors for the target compound (x_{tj}) should also be available. The same set of descriptors is defined for all compounds in the database, and the span of molecular descriptors should reflect the difference between any two compounds in the database. In principle, the database should be as large as possible, as adding more molecular descriptors and more compounds to the database can increase its predictive capability.

At the first stage of the targeted QSPR method, a similarity group (cluster, training set) for the target compound is established. The similarity group includes the predictive compounds, identified as structurally similar to the target compound by the partial correlation coefficient, r_{ti}, between the vector of the molecular descriptors of the target compound, $\mathbf{x}_t$, and that of a potential predictive compound $\mathbf{x}_i$, i.e., $r_{ti} = \mathbf{x}_t\ \mathbf{x}_i^{\mathrm{T}}$, where $\mathbf{x}_t$ and $\mathbf{x}_i$ are row vectors, centered and normalized to a unit length. Absolute r_{ti} values close to one ($|r_{ti}| \approx 1$) indicate high correlation between vectors $\mathbf{x}_t$ and $\mathbf{x}_i$ (high level of similarity) between the molecular structures of the target compound and the predictive compound i. The similarity group includes the first p compounds with highest $|r_{ti}|$ values. Another option is to form the similarity group only with compounds for which the $|r_{ti}|$ values exceed a prescribed threshold value.

To tailor a QSPR for a property of the target compound (applicable for all members of the similarity group) only members of the group for which data for the particular property are available are considered (N compounds). In view of the limited variability

of the property values within the similarity group, a linear structure-property relation is assumed of the form:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 \ldots \beta_m \mathbf{x}_m \qquad (2)$$

where **y** is an *N* vector of the target property values, *N* is the number of compounds included in the similarity group, $\mathbf{x}_1$, $\mathbf{x}_2$ **…** $\mathbf{x}_m$ are *N* vectors of predictive molecular descriptors (to be identified via a stepwise regression algorithm), and $\beta_0, \beta_1, \beta_2 \ldots \beta_m$ are the corresponding model parameters to be estimated. The signal-to-noise ratio in the partial correlation coefficient (*CNRj*) is used as a criterion for determining the number of the molecular descriptors that should be included in the model (*m*). The calculation of *CNRj* requires specification of error levels for the molecular descriptor data. The error (noise) in the molecular descriptors is assumed to be of the order of the round-off error of the calculated values. For integer data (no. of carbon atoms, for example) the noise level is the computer precision. Addition of new descriptors to the model can continue as long as the *CNRj* is greater than one for, at least, one of the descriptors which are not yet included. Detailed description of this stopping criterion can be found in Shacham and Brauner[9-11]. It should be noted that if necessary, nonlinear functions of molecular descriptors may also be considered in the RHS of Eq. (2).

As in a typical most "significant common features" method [1], a stepwise regression program is used to determine which molecular descriptors should be included in the QSPR to best represent the measured property data of the similarity group and to calculate the QSPR parameter values. The QSPR so obtained can be subsequently used for calculating the estimated value of the corresponding property values for the target compound and for other (*p*-*N*) compounds in the group that do not have measured data, i.e. using the equation:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} \ldots \beta_m x_{tm} \qquad (3)$$

where y_t is the estimated property value of the target compound and x_{t1}, x_{t2}, **…** x_{tm} are the corresponding molecular descriptors values of the target compound.

The targeted QSPR method ensures that the most pertinent information available in the data base (as measured values and molecular descriptors) is used for prediction of each property of the structurally similar compounds.

2. Application of the Targeted QSPR method for Property Prediction

For practical study of the targeted QSPR method, we used the molecular descriptor and property database of Cholakov *et al.* [2] and Wakeham *et al*. [1]. The database contains 260 hydrocarbons, the molecular structure of which is represented by 99 molecular descriptors, and values for five physical properties.

The properties included in the database are the normal boiling temperature (*NBT*), relative liquid density at 20º C (d_4^{20}), critical temperature (T_c), critical pressure (P_c) and critical volume (V_c). The list of the hydrocarbons in the database, the sources and quality of the property data are given in the corresponding references [1, 2].

In general, the molecular descriptors include the molar mass along with carbon atom descriptors, descriptors from simulated molecular mechanics (total energy, bond stretch energy, etc.) and some of the most popular topological indices, calculated with unit

bond lengths and with the bond lengths of the minimized molecular model, obtained by molecular mechanics. A complete list of all molecular descriptors in the database can be found elsewhere [10]. The 99 molecular descriptors in the data base were normalized dividing each descriptor by its maximal absolute value over the 260 database compounds. The stepwise regression program SROV [9] was used for identification of the similarity group, by sorting the compounds in descending order according to their $|r_{ti}|$ values. The first p = 50 compounds were included in the similarity group. This number was arbitrarily set. The SROV program was also used for deriving the structure-property relation (Eq. 3).
In the two examples hereunder the practical application of the targeted QSPR method is illustrated.

2.1. Example 1. Prediction of the Properties of n-tetradecane

The compound *n*-tetradecane is a representative of compounds for which accurate experimental data is available for most physical properties, it is densely represented in the database (meaning that there are many similar compounds included) and its properties can be predicted fairly well with existing QSPRs and homologous series techniques.
The results of the similarity group selection are displayed in Figure 1. It can be seen that the database contains a large number of compounds with high level of similarity to *n*-tetradecane ($|r_{ti}|$ between 0.93195 and 0.99968). The highest correlations are with the immediate neighbors of the target compound in the homologous series, *n*-pentadecane and *n*-tridecane. The lowest $|r_{ti}|$ is with 1-nonacosene.
The similarity group was used to derive QSPRs for the NBT, d_4^{20}, T_c, P_c and V_c for compounds structurally related to *n*-tetradecane in the form of Eq. (2). Those QSPRs were subsequently used for predicting the properties using Eq. (3). A summary of the QSPRs for the various properties derived for the similarity group of *n*-tetradecane is shown in Table 1. It can be seen that the QSPRs for the various properties include different molecular descriptors. The linear correlation coefficient R^2 values (>0.999 in all the cases) indicate an excellent fit. Only three descriptors were enough for R^2>0.999 for prediction of P_c, while for prediction of the other properties four predictors were needed.
In Table 1 the property prediction errors obtained with the "targeted" QSPR are compared with experimental errors assigned by DIPPR and with the corresponding prediction errors obtained in previous works [1, 2, 10-11] by applying the QSPR and QS2PR methods to the same data.
In general the "targeted" QSPR advocated in this work predicts the properties of *n*-tetradecane better than the traditional QSPRs and with precision comparable to that of the QS2PR [10-11] method (Table 1). However, the errors of both the QS2PR and the "targeted" QSPR (except for T_c) are well within the experimental errors assigned by DIPPR for the target, and hence, when its structure is well represented in the data base, either method can be used.

2.2. Example 2. Prediction of Unmeasured Properties of Members of the Similarity Group of n-tetradecane

For three members belonging to the similarity group of *n*-tetradecane, namely 2,5-dimethyldecane, 2,5-dimethyldodecane and 4-methyloctane, there are no experimental

values for the critical properties and the relative liquid density (except for 4-methyloctane). The unknown properties of those compounds can be predicted using the same targeted QSPR that was derived for *n*-tetradecane.
In Table 3 the property values obtained with the TQSPR are compared with measured values (whenever available) and with predictions obtained with the QSPR method of Wakeham *et al*. [1]. The largest differences between measured and predicted values for 4-methyloctane are: for *NBT* - 0.4 %; for d_{20}^{4} - 0.36 %, for T_c - 1.6 %, for P_c - 1.6 % and for V_c - 3.6 %, all within experimental error.

3. Conclusions

The partial correlation coefficient between vectors of molecular descriptors has been found to be an efficient and convenient measure for identifying structurally similar compounds and creating a training set of structurally similar compounds for traditional QSPR techniques.
The preliminary results obtained with the new targeted QSPR method show that it yields predictions within the experimental error level for compounds that are well represented in the database, and fairly accurate, reliable estimates for complex compounds which are sparsely represented. The cut-off value of the partial correlation coefficient provides an indication for the expected prediction error. Thus, the new method can complement the QS2PR and the traditional QSPR technique for prediction of properties of compounds which are sparsely represented in the molecular descriptor – property database.
More research is required in order to determine the relationships between the prediction reliability, the threshold value used for the partial correlation coefficient, the number of compounds included in the similarity group and the accuracy of their property data, and the improvement that might be eventually achieved by inclusion of nonlinear terms in the QSPR model.
Another important avenue for future research is the potential for application of the partial correlation coefficient between the vectors of molecular descriptors in computer aided design of molecules, structurally related to a compound with well established useful properties.

Bibliography

1. Wakeham, W.A.; Cholakov, G.St. and Stateva, R.P. J. Chem. Eng. Data. 47(2002) 559.
2. Cholakov, G.St.; Wakeham, W.A. and Stateva, R.P. Fluid Phase Equil. 163 (1999) 21.
3. Lydersen, A.L., Univ. Wisconsin Coll. Eng., Eng. Exp. Stn. Rep. 3, Madison, Wis. (1955).
4. Daubert, T. E. J., Chem. Eng. Data, 41(1996) 942.
5. Boethling, R.S.; Mackay D., eds, Handbook of Property Estimation Methods for Chem., Lewis, Boca Raton, FL, USA (2000).
6. Poling, B.E.; Prausnitz, J. M. and O'Connel, J. P., Properties of Gases and Liquids, 5th Ed., McGraw-Hill, New York (2001).
7. Dearden J. C., Environmental Toxicology and Chemistry, 22 (2003) 1696.
8. Yan, X.; Dong, Q. and Hong, X., J. Chem. Eng. Data, 48 (2003) 380.
9. Shacham, M. and Brauner, N. Comp. Chem. Engng. 27 (2003) 701.
10. Shacham, M.; Brauner, N.; Cholakov, G.St. and Stateva R.P. AIChE J. 50 (2004) 2481.
11. Brauner, N.; Shacham, M.; Cholakov, G.St. Stateva, R.P. Chem. Eng. Sci. 60 (2005) 5458.

Table 1. **Summary of structure-property correlations for various properties of *n*-tetradecane**

Property	Descriptors	R^2	Prediction error, %			
			Experiment (DIPPR)	Targeted QSPR	QSPR*	QS2PR**
NBT	x_3,x_{84},x_{85},x_{86}	0.99988	< 1	<0.01	1.92	0.05
d_4^{20}	x_3,x_{42}, x_{88}, x_{95}	0.99932	< 1	0.09	0.40	0.04
T_c	x_{59}, x_{88}, x_{92},x_{95}	0.99956	< 0.2	0.29	0.42	0.06
P_c	x_{65}, x_{77}, x_{85}	0.99946	< 10	1.46	0.07	0.70
V_c	x_{72}, x_{86}, x_{95}, x_{98}	0.99987	< 10	0.547	1.04	0.10

* 8 descriptors[1, 2], ** 4 descriptors[10].

Table 2. Prediction of properties of members of the *n*-tetradecane similarity group

Properties	2,5-dimethyldecane		2,5-dimethyldodecane		4-methyloctane	
	Targeted QSPR	Published* (QSPR**)	Targeted QSPR	Published* (QSPR**)	Targeted QSPR	Published* (QSPR**)
NBT	470.33	471.25	506.76	506.75	417.24	415.60
d_4^{20}	0.7502	(0.7502)	0.7630	(0.7646)	0.7225	0.7199
T_c	647.8	(642.2)	683.6	(672.5)	588.1	(589.0)
P_c	1.900	(1.878)	1.659	(1.633)	2.337	(2.355)
V_c	709	(728)	843	(854)	533	(553)

* Published values (without brackets), ** Predicted (inside brackets), 8 descriptors[1],[2]

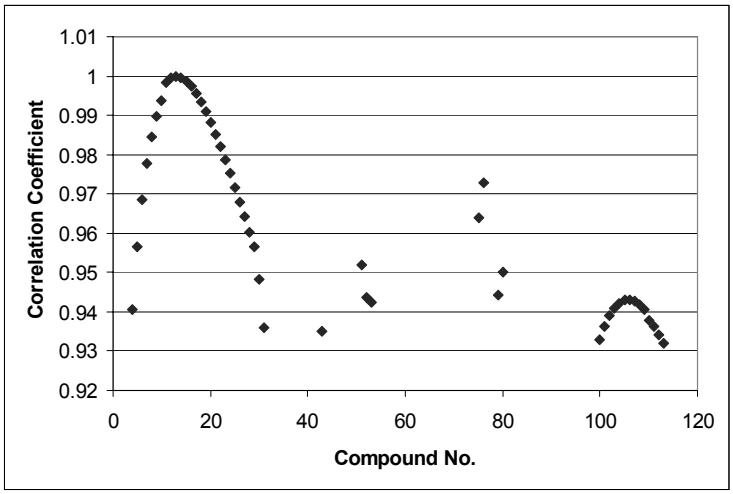

Figure 1. Partial correlation coefficients in the group of compounds similar to *n*-tetradecane.