European Symposium on Computer Aided Process Engineering – 15
L. Puigjaner and A. Espuña (Editors)

Identifying Equations that Represent Properties in Homologous Series using Structure-Structure Relations

Georgi St. Cholakov[a], Roumiana P. Stateva[b], Mordechai Shacham[c*] and Neima Brauner[d]

[a]Dept. Org. Synth. and Fuels, University of Chemical Technology and Metallurgy
Sofia, Bulgaria

[b]Institute of Chemical Engineering, Bulgarian Academy of Sciences,
Sofia 1113, Bulgaria

[c]Dept. Chem. Eng., Ben-Gurion University
Beer-Sheva, Israel

[d]School of Engineering, Tel-Aviv University
Tel-Aviv, Israel

Abstract

The Quantitative-Structure-Structure-Property Relationships (QS2PR) technique, which we introduced recently, is adapted to the prediction of properties of pure compounds in homologous series. The QS2PR method involves calculation of the molecular descriptors of a target compound of unknown properties, followed by regression of this vector of molecular descriptors versus a database of compounds with known descriptors and measured properties. The regression model, obtained for the target descriptor in terms of predictive compounds and their coefficients, is then used for predicting properties of the target compound.

A structure-structure relationship is derived from the number of carbon atoms and one additional, nonlinearly dependent molecular descriptor of the target compound and three predictive compounds. It is shown that such a relationship can provide predictions of satisfactory precision for many properties of the members of the homologous series. This enables users, without access to libraries of molecular descriptors, to employ the QS2PR technique for property prediction.

Keywords: property prediction**,** QSPR, molecular descriptors, homologous series

1. Introduction

Pure component properties are essential to chemical engineering calculations and computations in steady state and dynamic simulation, process and product design, environmental impact assessment, hazard and operability analysis, etc. There are compounds for which some properties cannot be measured (e. g., because the compound

* Author to who correspondence should be addressed: shacham@bgumail.bgu.ac.il

decomposes before reaching its critical properties), and extrapolated values must be used. Therefore, it is very important to develop reliable extrapolation and prediction techniques for such compounds.

Empirical correlations employing only the number of carbon atoms for predicting particular properties of compounds from some homologous series have been suggested by many authors (Marano and Holder, 1997). These correlations show asymptotic relationship and for each property a different correlation has to be derived. It can be expected that the accuracy of the prediction obtained for an unstable target compound deteriorates, since the measured data for its neighbouring compounds are less reliable than for the remote, stable compounds, in the same homologous series. The accuracy of the prediction obtained with such models depends also on the distance (in terms of the number of carbon atoms) of the predictive compounds from the target compound, and also on the relative orientation of the target compound (interpolation or extrapolation). The above assumptions have been illustrated by numerical examples (Gao et al., 2001).

Recently, Shacham et al. (2004) have demonstrated that most pure component properties of a target compound can be represented, with high precision, as linear combination of the same properties of several structurally similar predictive compounds. The structurally similar predictive compounds can be identified by a structure-structure correlation, derived from calculated molecular descriptor data without relying on any measured data. Any particular property of the target compound can then be predicted by substituting the corresponding property values of the predictive compounds into the structure-structure correlation, whereby a property-property correlation is obtained.

The use of the QS2PR (Quantitative Structure - Structure - Property Relationship) requires access to a large database of molecular descriptors**.** In this paper a variation of the QS2PR technique, which requires only a minimal amount of molecular descriptor data for members of homologous series, is presented.

2. Principles of Quantitative Structure - Structure - Property Relationship (QS2PR)

Let us assume that the vector of properties of the target compound **y** (the dependent variable) is potentially related to a set of m vectors of properties of predictive compounds (independent variables) $\mathbf{x}_1$, $\mathbf{x}_2$,…$\mathbf{x}_m$. The following partition of the **y** and **x** vectors to sub-vectors is used:

$$\mathbf{y} = \begin{Bmatrix} \mathbf{y}_c \\ \mathbf{y}_p \end{Bmatrix} \quad ; \quad \mathbf{x}_i = \begin{Bmatrix} \mathbf{x}_{cj} \\ \mathbf{x}_{pj} \end{Bmatrix} \tag{1}$$

where $\mathbf{y}_c$ is an N vector of known properties, $\mathbf{y}_p$ is a K vector of unknown properties. Both the N vector $\mathbf{x}_{cj}$ and the K vector $\mathbf{x}_{pj}$ contain known properties. Typically, the sub-vectors $\mathbf{y}_c$ and $\mathbf{x}_{cj}$ contain properties, which are directly related to the molecular structure and can be calculated with high accuracy (molecular descriptors), while the sub-vectors $\mathbf{y}_p$ and $\mathbf{x}_{pj}$ contain measured properties with various levels of experimental error. We wish to model the structure-structure relationship between $\mathbf{y}_c$ and the

independent variables $\mathbf{x}_{c1}$, $\mathbf{x}_{c2}$, … $\mathbf{x}_{cm}$ by a linear regression model, with the general form:

$$y_{ci} = \beta_1 x_{c1i} + \beta_2 x_{c2i} \ldots \beta_m x_{cmi} + \varepsilon_i \tag{2}$$

where the weighing factors $\beta_1, \beta_2 \ldots \beta_m$ are the model parameters to be estimated and ε_i represents independent normal errors of a constant variance.

The practical application of equation (2) requires preparation of a bank of potential predictive compounds as a database. The same set of molecular descriptors must be defined for all compounds included in the database, while the span of the molecular descriptors should reflect the difference between the compounds in the data-base. Having the $\mathbf{y}_c$ for a target compound defined as well, a stepwise regression procedure can be applied to the database in order to identify the most appropriate predictive compounds that should be included in the structure-structure regression model (Equation 2) and obtain the respective model parameters. Upon identifying the model parameters, the following equation can be used for predicting unknown properties of the target compound:

$$\mathbf{y}_p = \beta_1 \mathbf{x}_{p1} + \beta_2 \mathbf{x}_{p2} \ldots \beta_m \mathbf{x}_{pm} \tag{3}$$

The properties that can be predicted for the target compound include all the properties that are available for all the predictive compounds included in the structure-structure correlation.

Shacham et al. (2004) have demonstrated that the QS2PR technique predicts most constant properties with high precision for a wide range of hydrocarbons belonging to different groups (linear, branched, cyclic, etc) if a library of molecular descriptors for the potential predictive compounds and the target compound, as well as property data for all of potential predictive compounds are available.

It is going to be demonstrated hereunder, however, that when applying the QS2PR technique to members of homologous series, the structure of a target compound can be modelled by ***any*** three (sometimes even two) predictive compounds belonging to the same series. Furthermore, because of the limited structural change within homologous series the respective QS2PR requires only one molecular descriptor (in addition to the number of the carbon atoms).

3. Structure-structure relation for members of homologous series

The minimum information required for deriving a structure-structure relation for a target compound, which is a member of a homologous series, is the availability of one molecular descriptor (nonlinearly dependent of the total number of carbon atoms) for both the predictive and the target compounds. In order to find the coefficients of the structure-structure relation the following system of three linear equations is solved:

$$\begin{aligned} &\beta_1 + \beta_2 + \beta_3 = 1 \\ &\beta_1 n_{p1} + \beta_2 n_{p2} + \beta_3 n_{p3} = n_t \\ &\beta_1 x_{c1} + \beta_2 x_{c2} + \beta_3 x_{c3} = y_c \end{aligned} \tag{4}$$

where n_{p1}, n_{p2} and n_{p3} are the numbers of carbon atoms of the predictive compounds and n_c is the number of the carbon atoms of the target compound. After the coefficients β_1, β_2 and β_3 have been determined by solving equation (4), they can be introduced into equation (3) to obtain the desired properties of the target compound. The use of the proposed technique for property prediction will be demonstrated by predicting properties of ethyl-cyclopentane.

4. Prediction of Properties of Ethyl-Cyclopentane – an Example

The data used for deriving the structure-structure correlation for ethyl-cyclopentane is shown in Table 1. Three of the neighbouring compounds: methyl-, propyl- and butyl-cyclopentane were employed as predictive compounds. From amongst the molecular descriptors, which are asymptotically dependent of the number of carbon atoms, the easily calculated Wiener's index from the database of Cholakov et al. (1999) was arbitrarily selected as the additional molecular descriptor for deriving the structure-structure relationship. Introducing the numerical values of the number of carbon atoms and the Wiener's index into equation (4) yields the results: β_1 = 0.58537; β_2 = 0.2439; and β_3 = 0.17073. These coefficients can be used in the property-property correlation to predict various properties of ethyl-cyclopentane. For example, introducing the boiling point data from Table 1 and the coefficients into Eq. (3) yields an estimate for the boiling temperature of ethylcyclopentane:

0.58537*344.96+0.2439*404.11+ 0.17073*429.8 = 373.87 K

Comparison of the predicted value with the experimental data shown in Table 1 indicates a prediction error of 0.73 %.

By introducing experimental data for other properties of the predictive compounds, the respective properties for the target compound can be similarly predicted. Table 2 shows the prediction errors in predicting 28 constant properties of ethyl-cyclopentane using this structure-structure relation. More than 20 properties are predicted with error smaller than or comparable to the error assigned by published databases. When the error exceeds the assigned values, the deviation can be in most cases explained by inaccuracy or uncertainty of the in the respective database (for DIPPR data, this issue is discussed in more detail by Shacham et. al., 2004). Considering the minimum amount of descriptors needed for the derivation of the structure-structure relationship, and the fact that one correlation is used for all properties, the accuracy of the predicted values is quite satisfactory. Furthermore, the coefficients of the structural correlation derived from the particular low molecular mass compounds was used for prediction of the properties of other members of the homologous series by interpolation with the properties of their respective neighbors and errors within the above limits were obtained.

5. Prediction error as Function of the Number of Carbon Atoms

The proposed technique was utilized for predicting properties for a large number of compounds belonging to various homologous series. The following series were tested: *n*-alkanes (propane to *n*-octacosane), 1-alkenes (1-butene to 1-oktacosene), alkyl-benzenes (propylbenzene to *n*-docosylbenzene), and alkyl-cyclopentanes (ethyl-

cyclopentane to tricosyl-cylcyclopentane). In all these cases the Wiener's index taken from the database of Cholakov et al. (1999) was used to derive the structure-structure relation.

Table 1. Structure-Structure correlation data for ethyl-cyclopentane.

No.	Compound	C atoms	Wiener's index	Boiling Temp. (K) [1]
1	methylcyclopentane	6	26	344.96
2 (target)	ethylcyclopentane	7	43	376.62
3	propylcyclopentane	8	67	404.11
4	butylcyclopentane	9	99	429.8

[1]From the database of Cholakov et al. (1999)

Table 2. Percent error in predicting properties of ethyl-cyclopentane

No.	Property	Units	Reported value[1]	Prediction error (%)
1	Critical Temperature	K	569.5	1.10
2	Critical Pressure	Pa	3400000	0.57
3	Critical Volume	m^3/kmol	0.375	0.38
4	Crit Compress Factor	unitless	0.269	0.81
5	Melting Point Temp.	K	134.71	5.95
6	Triple Pt Temperature	K	134.71	5.95
7	Normal Boiling Temp.	K	376.62	0.73
8	Liq Molar Volume	m^3/kmol	0.128748	0.33
9	IG Heat of Formation	J/kmol	-1.2700E+08	0.13
10	IG Gibbs of Formation	J/kmol	4.4800E+07	0.21
11	IG Absolute Entropy	J/kmol*K	3.7800E+05	0.06
12	Std Heat of Formation	J/kmol	-1.6300E+08	0.25
13	Std Gibbs of Formation	J/kmol	3.7600E+07	0.58
14	Std Absolute Entropy	J/kmol*K	2.8000E+05	0.09
15	Heat Fusion at Melt Pt	J/kmol	6.8700E+06	22.63
16	Std Net Heat of Comb	J/kmol	-4.2800E+09	0.00
17	Acentric Factor	unitless	0.270095	2.58
18	Radius of Gyration	m	3.73E-10	0.05
19	Solubility Parameter	(J/m^3)^0.5	16300	0.46
20	van der Waals Volume	m^3/kmol	0.0704	0.02
21	van der Waals Area	m^2	8.8700E+08	0.00
22	Refractive Index	unitless	1.4173	0.17
23	Flash Point	K	269	1.00
24	Lower Flammability Limit	vol% in air	1.1	2.66
25	Upper Flammability Limit	vol% in air	6.7	11.29
26	Lower Flamm Limit Temp	K	270	0.99
27	Upper Flamm Limit Temp	K	303	0.11
28	Auto Ignition Temp	K	533.15	7.65

[1]Data from the DIPPR database.

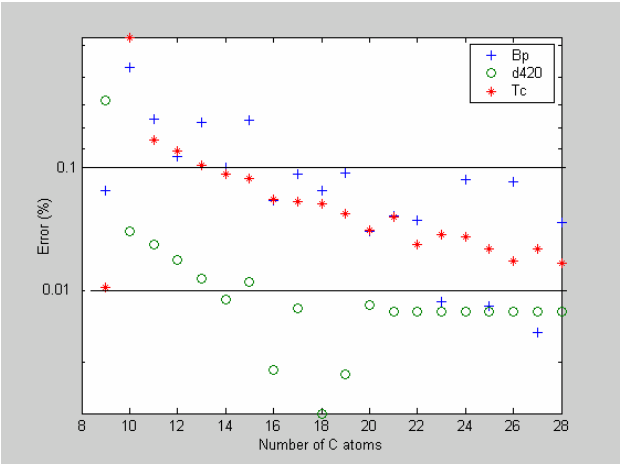

Figure 1. Prediction error (%) in predicting selected properties of alkyl-benzenes.

The properties predicted were the normal boiling temperature (T_b), the relative liquid density at 20º C (d420) and the critical temperature (T_c). Typical results are shown in Figure 1. It is important to note that the prediction error decreases with the increase of the carbon numbers. For normal boiling temperature and critical temperature the error is between 0.3 % and 0.1 % from 9 to 15 carbon atoms, and between 0.1 % to 0.01 % for molecules with more than 15 carbon atoms. The prediction error for relative liquid density is between 0.1 % to 0.01 % for C atoms between 10 and 15, and less than 0.01 % for C atoms above 15.

Conclusions

It has been demonstrated that for homologous series a single structure-structure relationship, derived using the number of carbon atoms and one additional nonlinearly dependent molecular descriptor for three predictive compounds and the target compound can provide predictions of satisfactory precision for more than 20 properties. The new technique has been tested for interpolation with the immediate neighbours on both sides of the target compound chosen as the predictive compounds. Further research is needed to determine how much the precision of the prediction may deteriorate if the distance (in terms of the number of the carbon atoms) between the predictive compounds and the target compound is increased, and/or extrapolation is used instead of interpolation. The potential benefits of increasing the number of molecular descriptors and a nonlinear structural relationship should also be investigated.

References

Cholakov, G. St., W.A.Wakeham, and R.P. Stateva, 1999, Fluid Phase Equilibria 163, 21.
Gao, W., R.L. Robinson, and K.A.M. Gasem, 2001, Fluid Phase Equilibria, 179, 43.
Marano, J. J. and G. D. Holder, 1997, Ind. Eng. Chem. Res. 36, 1887.
Shacham M. and N. Brauner, 2003, Computers Chem. Engng., 27(5), 701.
Shacham M., N. Brauner, G.St.Cholakov, and R.P. Stateva, 2004, AIChE J. 50 (10), 2481.
Wakeham, W.A, G.St.Cholakov, and R.P. Stateva, 2002, J. Chem. Eng. Data, 47 (3), 559.