17th European Symposium on Computer Aided Process Engineering – ESCAPE17
V. Plesu and P.S. Agachi (Editors)

Identifying Applicability Domains for Quantitative Structure Property Relationships

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

[a]*Dept. Chem. Eng., Ben-Gurion University, Beer-Sheva, Israel,shacham@bgu.ac.il*
[b]*School of Engineering, Tel-Aviv University, Tel-Aviv, Israel,brauner@eng.tau.ac.il*
[c]*Univ. of Chemical Technology and Metallurgy, Sofia, Bulgaria, cholakov@uctm.edu*
[d]*Inst. of Chem. Eng., Bulgarian Academy of Sciences, Sofia, Bulgaria, thermod@bas.bg*

Abstract

Development of Quantitative Structure Property Relationships (QSPR) for property prediction, targeted for a particular applicability domain (AD), and definition of the AD boundaries are considered. The AD is defined in terms of the target compound (for which a property has to be predicted) belonging to a homologous series and including carbon atoms above a particular number. If the target compound satisfies these requirements simple linear QSPR, with one or two descriptors are shown to predict the property within experimental error level. The method presented can also identify the cases where lack of experimental data can prevent derivation of a reliable QSPR.

Keywords

Property prediction; QSPR; Molecular descriptors; Homologous series

1. Introduction

In recent years there is an increased interest in the development and use of Quantitative Structure-Property Relationship (QSPR) models for property prediction [1, 2]. In the traditional QSPR modeling techniques one large set of molecular descriptors and physical property data for a wide variety of compounds is used as the "training set". Because the structure-property relationship obtained is usually nonlinear, the prediction accuracy critically depends on the level of representation of the "target" compound's (for which the property has to be predicted) structural groups in the training set. If the target structure is densely represented, the prediction can be expected to be much more accurate than if the target compound is sparsely represented. Moreover, the prediction accuracy cannot be reliably assessed.

In an attempt to overcome the above limitations we have developed the "targeted" QSPR [3] method. With this method, only compounds "*similar"* to the target compound are included in the training set. The limited variability of the compounds enables developing simple, linear QSPRs for property prediction. In this paper it is demonstrated that targeted QSPR's (TQSPR's) can be developed for particular "*applicability domain*s (AD)" and can consequently predict reliably properties and errors for compounds belonging to the AD. The proposed technique will be demonstrated for the *n*-alkane homologous series, defined as an AD.

2. Applying the TQSPR method to the *n* - alkane Homologous Series

The TQSPR technique is described in detail elsewhere [3]. A brief review of the method follows. The first stage of the TQSPR involves identification of a training set (*similarity group*) structurally related to the target compound for which properties have to be predicted. For identification of the similarity group, a database of molecular descriptors, x_{ij} and property data y_{ij} for the *predictive* compounds are required, where *i* is the number of the compound and *j* is the number of the descriptor/property. The same molecular descriptors for the target (x_{tj}) and for all other compounds in the database should be available. The similarity between the *target* and potential *predictive* compounds is measured by the partial correlation coefficient, r_{ti} ($r_{ti} = \bar{\mathbf{x}}_t \bar{\mathbf{x}}_i^T$), between the vector of the molecular descriptors of the target compound, $\mathbf{x}_t$, and that of a potential predictive compound $\mathbf{x}_i$. Absolute r_{ti} values close to one ($|r_{ti}| \approx 1$) indicate high level of structural similarity. The *training set* is established by selecting the first *p* compounds of the highest $|r_{ti}|$ values for which experimental values of the desired property are available. For development of a TQSPR a linear structure-property relation is assumed of the form:

$$\mathbf{y} = \beta_0 + \beta_1 \boldsymbol{\zeta}_1 + \beta_2 \boldsymbol{\zeta}_2 \ldots \beta_m \boldsymbol{\zeta}_m + \boldsymbol{\varepsilon} \qquad (1)$$

where $\mathbf{y}$ is a *p* vector of the respective measured property values, $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2 \ldots \boldsymbol{\zeta}_m$ are *p* vectors of molecular descriptors $\beta_0, \beta_1, \beta_2 \ldots \beta_m$ are the corresponding model parameters, and $\boldsymbol{\varepsilon}$ is a *p* vector of stochastic terms (measurement errors). The selection of descriptors to the TQSPR model continues as long as the average model relative prediction error for the training set (ε_a) exceeds a prespecified error tolerance (ε_g). The so-obtained TQSPR (Eq. 1) can be subsequently employed for estimating the property value for the *target*.

Table 1. Experimental* and predicted property data taken from DIPPR database

No.	Comp. Name	Carbon Atoms	Normal Boiling Temp. (Tb, K)	Reliability %	Melting Point Temp. (Tm, K)	Reliability (%)	Critical Pressure (Mpa)	Reliability (%)	Liq. Molar Volume (M3/kmol)	Reliability (%)
1	ethane	2	***184.55***	1	***90.352***	0.2	4.872	0.2	***0.0954***	1
2	propane	3	***231.11***	1	***85.47***	0.2	**4.248**	0.2	***0.0757***	1
3	n-butane	4	***272.65***	1	134.86	0.2	**3.796**	0.2	***0.1014***	1
4	n-pentane	5	***309.22***	1	***143.42***	0.2	**3.37**	1	***0.1160***	1
5	n-hexane	6	***341.88***	1	***177.83***	0.2	**3.025**	1	***0.1314***	1
6	n-heptane	7	***371.58***	1	***182.57***	0.2	**2.74**	3	***0.1470***	1
7	n-octane	8	***398.83***	1	***216.38***	0.2	**2.49**	3	***0.1626***	1
8	n-nonane	9	***423.97***	1	***219.66***	1	**2.29**	3	***0.1789***	1
9	n-decane	10	***447.305***	1	***243.51***	1	**2.11**	3	***0.1958***	1
10	n-undecane	11	***469.078***	1	***247.571***	1	**1.95**	5	***0.2122***	1
11	n-dodecane	12	***489.473***	1	***263.568***	1	**1.82**	10	***0.2286***	1
12	n-tridecane	13	***508.616***	1	***267.76***	0.2	**1.68**	10	***0.2456***	1
13	n-tetradecane	14	***526.727***	1	279.01	0.2	**1.57**	25	***0.2613***	1
14	n-pentadecane	15	***543.835***	1	***283.072***	0.2	**1.48**	25	***0.2778***	1
15	n-hexadecane	16	***560.014***	1	***291.308***	0.2	**1.4**	25	***0.2942***	1
16	n-heptadecane	17	***575.3***	1	***295.134***	0.2	**1.34**	25	***0.3109***	1
17	n-octadecane	18	***589.86***	1	***301.31***	0.2	1.27	25	***0.3282***	1
18	n-nonadecane	19	***603.05***	1	***305.04***	0.2	1.21	25	***0.3456***	1
19	n-eicosane	20	***616.93***	1	***309.58***	0.2	1.16	25	***0.3664***	1
20	n-heneicosane	21	629.65	1	***313.35***	1	1.11	25	***0.3812***	1
21	n-docosane	22	641.75	1	***317.15***	1	1.06	25	0.3991	1
22	n-tricosane	23	653.35	1	***320.65***	1	1.02	25	0.4169	1
23	n-tetracosane	24	664.45	1	***323.75***	3	0.98	25	0.4349	1
24	n-pentacosane	25	675.05	1	***326.65***	1	0.95	25	***0.4526***	1
25	n-hexacosane	26	685.35	1	***329.25***	1	0.91	25	***0.4712***	1
26	n-heptacosane	27	695.25	3	***332.15***	1	0.883	25	***0.4882***	5
27	n-octacosane	28	704.75	1	***334.35***	1	0.85	25	***0.5077***	1
28	n-nonacosane	29	713.95	3	***336.85***	1	0.826	25	0.5238	1
29	n-triacontane	30	722.85	1	***338.65***	1	0.8	25	0.5405	1
30	n-dotriacontane	32	738.85	3	***342.35***	1	0.75	25	***0.5766***	1
31	n-pentatriacontane	35								
32	n-hexatriacontane	36	770.15	1	***349.05***	1	0.68	25	***0.6507***	1
33	n-tetracontane	40								
34	n-tetratetracontane	44								

*Experimental data is shown in bold, italic letters.

To carry out the targeted QSPR method studies, we have developed a molecular descriptor database with 1630 descriptors calculated with the Dragon program (Copyright of TALETE srl, http://www.talete.mi.it) for 259 of the hydrocarbons listed in [2]. Measured and predicted property data from the DIPPR [4] database were used in the studies. Modified versions of the stepwise regression program (SROV) of Shacham and Brauner [3] were prepared for the identification of the similarity groups and derivations of the QSPRs. The property data used, namely: normal boiling temperature T_b; melting point temperature, T_m; critical pressure, P_c; liquid molar volume, M_v; critical temp., T_c and critical volume, V_c for the compounds included in this study are shown in Table 1 (except T_c and V_c). The "reliability" assigned by DIPPR on these values (i.e. estimate of the upper error bound, %) are also given. Only experimental data were included in the training sets (shown in by bold, italic letters).

2.1. The distance between the target compound and the training set

For a target compound, which is a member of a homologous series, the highest structural similarity is with its two closest neighbors. Therefore, the maximal achievable similarity level (or distance) between the *target* and a potential training set can be well represented by the correlation coefficient value ($\left|r_{ti}\right|$) between the target and its closest neighbor in the series. The sequence of these $\left|r_{ti}\right|$ values for the *n*-alkane series is : ethane – 0.884 – propane – 0.935 – *n*-butane – 0.947 – n-pentane – 0.961 – n-hexane – 0.967 - n-heptane - 0.975 – n-octane, and reaches values over .99 for n-pentatriacontane and above. Hence, the correlation coefficient increases monotonically with increasing the carbon number. It can be therefore expected that predictions of higher accuracy are achievable for compounds of higher carbon number.

2.2. *Modeling T_b, T_m, , T_c, P_c, M_v and V_c for the n-alkane homologous series*

It is well known that properties within homologous series change asymptotically with carbon number. An example for the critical temperature, normal boiling temperature and melting temperature of the *n*-alkanes studied is presented in Fig. 1. It can be seen that, because of the asymptotic relationship, the rate of the increase of the properties becomes more moderate for higher carbon numbers, hence exhibiting a *non-linear* relation with the carbon number.
To derive the QSPR for the *n*-alkane series, *n*-hexadecane (compound which is located near the middle of the AD studied) was selected as the target compound. The compounds with experimental data available for the particular property were included in the training set, except for ethane, propane, *n*-butane and *n*-pentane, since it was determined earlier (section 2.1) that their level of similarity with the rest of the series is rather low.
Using T_b data, the targeted QSPR technique identified the descriptor *RTu* (a GETAWAY descriptor) as the *dominant descriptor* (the descriptor which has the highest correlation with the property) and the following linear QSPR was identified: T_b = -57.957+22.6315**RTu* with an average error of ε_a = 0.3% for the training set. The equation was used to calculate the T_b of all compounds included in this study. The differences between the calculated values and those reported by DIPPR (either experimental or predicted) are shown in Fig. 2. They are larger for compounds with smaller number of carbon atoms and reach 0.55 % for *n*-hexane. For ethane, propane, *n*-butane and *n*-pentane, the prediction errors are 17.7, 4.5, 1.86 and 2.1 %, respectively. Since the reliability of the T_b data is 1 % or more, the QSPR provides prediction with adequate accuracy for members of the *n*-alkane series with six or more carbon numbers. Adding more descriptors reduces the training set average percent error considerably, but does

not improve the asymptotic representation of the T_b of the compounds which do not belong to the training set and have measured data, thus does not increase the confidence in the predicted values.

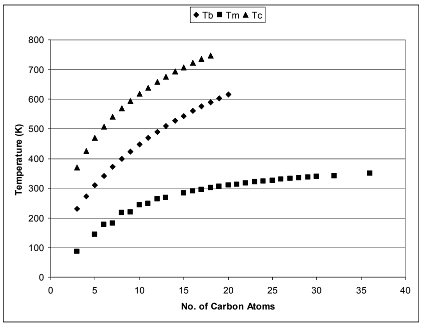

Figure 1. Variation of T_c, T_m and T_b as function of the carbon number in the *n*-alkane homologous series

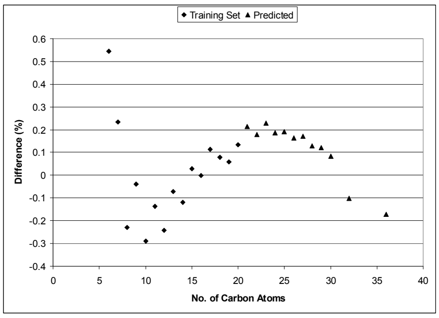

Figure 2. Difference between the QSPR boiling temperatures calculated and DIPPR values.

Using the *RTu* descriptor to represent the melting temperature data yields a QSPR with training set average percent error of 2.67. This error is too large in comparison to T_m reliability data, thus there is a need to add more descriptors. The descriptor *R2e+* (also a GETAWAY descriptor) was identified as the next one to be included in the QSPR, whereby T_m = 372.8664+0.66944**RTu*-1942.3213**R2e+*, with ε_a = 0.4 %. The T_m prediction error is < 0.5% for compounds with ten or more carbon atoms, and < 2 % for compounds with 5 to 9 carbon atoms. The prediction errors are considerably higher for ethane, propane, *n*-butane and *n*-pentane. Thus, the QSPR with the descriptors *RTu* and *R2e+* can be used with confidence for predicting melting point temperature of members of the *n*-alkane series with 10 or more carbon atoms.

The critical temperature of *n*-alkanes can be represented by: T_c = 128.677 + 21.7719 **RTu*, with ε_a = 0.29 %. This error is acceptable in view of the reliability of the T_c data (0.2 % for most of the compounds). However, the residual plot (not shown) of the difference between the experimental data and the calculated values versus the experimental data shows a curvature that is not explained by the single descriptor QSPR. Adding the RARS descriptor (a GETAWAY descriptor) to the QSPR yields the QSPR T_b = 554.2874 +12.0076**RTu*- 293.3629**RARS* with ε_a = 0.058 %. This model yields prediction errors of ≤1 % even for the light compounds; propane, *n*-butane and *n*-pentane.

The dominant descriptor selected for representing the critical pressure is the descriptor *H3p* belonging to the GETAWAY descriptors. The QSPR obtained using this descriptor is: P_c = 4.0804 -3.5271**H3p* with ε_a = 0.42 %. Considering the reliability of the P_c data that varies between 0.2 % to 25 % (Table 1), and the random error distribution indicated by the residual plot, suggest that the representation of P_c by a single descriptor QSPR is adequate. It is well within the reliability level for compounds with five or more carbon atoms.

Liquid molar volume can be represented well by a QSPR containing the *SEig* descriptor (a “geometrical” descriptor): $M_v = 0.033836 + 0.0025053 * SEig$ with $\varepsilon_a = 0.28$ %. This model represents the liquid molar volume within experimental error level for *n*-alkanes with five or more carbon atoms.

Experimental critical volume values are available only for a few members of the *n*-alkane series with 2 to 7 carbon atoms. As the above analysis has shown that the properties of the first few compounds in the homologous series correlate poorly with the remaining compounds, we should conclude that the available data is insufficient to derive a QSPR for the critical volume with reasonable confidence in the predicted values.

3. Conclusions

In order to predict properties with confidence it is essential to define an AD for the QSPR used. In this paper we demonstrated the use of similarity measures to define the AD of a QSPR for homologous series in terms of the structural similarity of a target compound and compounds with carbon atoms above a particular number. It was shown that if the target compound satisfies these requirements the QSPR can predict the property within experimental error level. The proposed method has been tested also with the 1-alkene and alkyl benzene homologous series and the same results were obtained using the same descriptor-property combinations shown here for *n*-alkanes. The presented method can also determine when the lack of experimental data can prevent derivation of a reliable QSPR.

References

1. J.C. Dearden, Environmental Toxicology and Chemistry, 22 (2003) 1696.
2. W.A. Wakeham, G.St. Cholakov and R.P.Stateva, J. Chem. Eng. Data, 47 (2002) 559.
3. N. Brauner, R.P. Stateva, G.St. Cholakov and M. Shacham, Ind. Eng. Chem. Res., 45 (2006) 8430.
4. R.L. Rowley, W.V. Wilding, J.L. Oscarson, Y. Yang, N.A. Zundel, DIPPR Data Compilation of Pure Chemical Properties, Design Institute for Physical Properties. http//dippr.byu.edu, Brigham Young University Provo Utah, 2006 .
5. M. Shacham and N. Brauner, Computers & Chem. Engng. 27(2003) 701.