Linear QSPRs for Predicting Pure Compound Properties in Homologous Series

Neima Brauner
School of Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel

Georgi St. Cholakov
Dept. of Organic Synthesis and Fuels, University of Chemical Technology and Metallurgy, Sofia 1756, Bulgaria

Olaf Kahrs
Process Systems Engineering, RWTH Aachen University, Aachen 52064, Germany

Roumiana P. Stateva
Institute of Chemical Engineering, Bulgarian Academy of Sciences, Sofia 1113, Bulgaria

Mordechai Shacham
Dept. of Chemical Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

DOI 10.1002/aic.11424
Published online February 8, 2008 in Wiley InterScience (www.interscience.wiley.com).

Linear QSPRs, containing 1 through 4 descriptors, are developed for predicting the normal boiling temperature, melting point temperature, and critical properties for the n-alkane, 1-alkene, n-alkylbenzene, 1-alcohol, and alkanoic monocarboxylic acid homologous series. It has been shown that property values for which experimental data are available can be predicted within experimental error level (with very few and very small exceptions), irrespective of whether interpolation or extrapolation is involved. Property values for which predicted literature data are available can be matched within the reported "reliability" level, even when extrapolation is carried out from very small training sets containing experimental data. Thus, the linear QSPRs developed represent well the nonlinear variation of the particular property with the carbon number, and increase the confidence in the values predicted when extrapolation is involved. © 2008 American Institute of Chemical Engineers *AIChE J,* 54: 978–990, 2008
Keywords: property prediction, QSPR, molecular descriptors, homologous series

Introduction

Experimental property data are available only for a small fraction of the pure compounds pertaining to chemistry and chemical engineering, environmental engineering and environmental impact assessment, hazard and operability analysis, and others. Therefore, methods for reliable prediction of property data are needed.

One class of modern property prediction methods are based on the identification of quantitative structure–property relationships (QSPRs) between a compound and its physical properties. Various classes of descriptors (topological, graph theoretical, topochemical, quantum chemical, structural, physicochemical, electrotopological, etc.) have been proposed

Correspondence concerning this article should be addressed to M. Shacham at shacham@bgumail.bgu.ac.il.

to adequately represent molecular structures numerically. Group and bond contribution, most significant common features and neural network models have been developed as QSPRs. A recent review of the field was published by Dearden.[1]

In the present work, we adopt the methodology of the "most significant common features" QSPR methods, as defined in Ref. 2, which we shall call for short QSPRs henceforward.

QSPRs are typically targeted at estimating a given property for all possible compounds. For a given property, a QSPR can be schematically represented by the following equation:

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

where y_p is the property (e.g., boiling temperature, melting temperature, toxicity) to be predicted; x_{s1}, $x_{s2}, \ldots x_{sk}$ are descriptors representing numerically the molecular structure of the compounds in the database; x_{p1}, $x_{p2}, \ldots x_{pm}$ are known property data of the compounds in the database; and β_0, $\beta_1, \ldots \beta_n$ are the QSPR regression parameters.

To derive the QSPR, the available data (descriptors and property values) is divided into a *training set* and a *validation set*. Multiple linear or nonlinear regression, and partial least squares techniques are applied to the training set, to select the significant common feature molecular descriptors and the property values to be included in the right hand side of Eq. 1, and to calculate the model parameter values. Model validation is typically carried out using only one validation set, though cross-validation techniques, which use alternatively defined training and validation sets, are also available. Applications of QSPRs for prediction of physical and thermodynamic properties are reported by many authors.[2–6]

Independent studies of QSPR prediction errors (e.g. Refs. 1, 7, and 8 show that if the chemical structure of the *target compound* (the compound for which a property has to be predicted) is well represented in the training set, the prediction can be expected to be much more accurate than when its structure is sparsely represented. Since no measures for the level of representation of a particular group in a training set are available, it is difficult to estimate the prediction error. Therefore, Tropsha et al.[9] recommended defining a "model applicability domain" and Basak et al.[10] advocated the use of "tailored" training sets, meaning that the compounds in the training set are not selected randomly, but they should be structurally related to the target compound.

The quantitative structure–structure property relationship (QS2PR) method of Shacham et al.[11] and Brauner et al.[12] exploits the available structural similarity between the target compound and potential predictive compounds. If successful, a small set of structurally related *predictive compounds* can be found, so that the vector of molecular descriptors of the target compound can be represented by a linear combination of the vectors of molecular descriptors of the predictive compounds. This structure–structure relation can be then used as a property–property relation, to predict all properties of the target for which reliable experimental data are available for the predictive compounds. However, it has been observed that lack of reliable experimental data for structurally similar predictive compounds and summation of experimental errors, especially when extrapolating toward a distant target,[13] tend to considerably reduce the accuracy of the predicted property values when using the QS2PR method.

In order to better utilize the experimental data available for the predictive compounds, Brauner et al.[14] developed the "targeted" QSPR (TQSPR) method. This method is based on the identification of a set of compounds structurally similar to the target compound (i.e., the similarity group) and of the training and validation set, which are subsets of the similarity group. The QSPR parameters are estimated from the training set in the usual manner. Structural similarity is measured by the partial correlation coefficients between the vectors of the molecular descriptors of the target compound and each potential predictive compound. The results obtained[14] showed that the TQSPR method yields predictions within the experimental reliability for compounds well represented in the database and fairly reliable estimates for complex compounds that are sparsely represented.

Examination of databases that contain experimental property data reveals that such data are usually available for compounds with low carbon numbers and not for high carbon number compounds. When high carbon number compounds cannot be included in the training set, the application of the QSPRs for the prediction of their properties involves extrapolation from low carbon number compounds.

For homologous series, the extrapolation is usually carried out by an "asymptotic behavior correlation" (ABC).[15–17] These correlations are simple and need only the number of carbon atoms as an independent variable. For interpolation and close range extrapolation, their prediction results are of high precision and may be considered "pseudoexperimental," as discussed in Ref. 2. However, in ABCs one of the key constants—the property value at an infinitely high carbon number—varies in different publications, and cannot be validated experimentally. As a result, the long-range extrapolated values can be considered only as "reasonable."[15]

The prediction of melting point temperatures represents a special challenge because of the many factors affecting the melting point.[18] Many authors report prediction errors for melting points, which are significantly higher than for similar properties, such as normal boiling point, or critical temperature.[1,7,12,14,18]

The objective of this paper is to increase the level of confidence in the predicted values in long-range extrapolation, and to bring the prediction error down to the experimental reliability level for all properties considered in this study, including the melting point temperature.

Methodology

To carry out the studies described in this paper, we developed a molecular descriptor database for homologous series of hydrocarbons and oxygen containing organic compounds (Table 1). The Dragon program (version 5.4, DRAGON is copyrighted by TALETE srl, http://www.talete.mi.it)[19] was used to calculate 1664 descriptors for the compounds in the database from minimized energy molecular models. The molecular geometries were optimized using the complete neglect of differential overlap semiempirical method implemented in the HyperChem package (Version 7.01, Hyperchem is copyrighted by Hypercube Inc.). The number of descriptors was

Table 1. List of Compounds Included in the Study

C Atoms	Name	C Atoms	Name	C Atoms	Name
n-Alkanes		1-Alkenes (continued)		Aliphatic-alcohols	
2	Ethane	12	1-Dodecene	1	Methanol
3	Propane	13	1-Tridecene	2	Ethanol
4	*n*-Butane	14	1-Tetradecene	3	1-Propanol
5	*n*-Pentane	15	1-Pentadecene	4	1-Butanol
6	*n*-Hexane	16	1-Hexadecene	5	1-Pentanol
7	*n*-Heptane	17	1-Heptadecene	6	1-Hexanol
8	*n*-Octane	18	1-Octadecene	7	1-Heptanol
9	*n*-Nonane	19	1-Nonadecene	8	1-Octanol
10	*n*-Decane	20	1-Eicosene	9	1-Nonanol
11	*n*-Undecane	21	1-Heneicosene	10	1-Decanol
12	*n*-Dodecane	22	1-Docosene	11	1-Undecanol
13	*n*-Tridecane	23	1-Tricosene	12	1-Dodecanol
14	*n*-Tetradecane	24	1-Tetracosene	13	1-Tridecanol
15	*n*-Pentadecane	25	1-Pentacosene	14	1-Tetradecanol
16	*n*-Hexadecane	26	1-Hexacosene	15	1-Pentadecanol
17	*n*-Heptadecane	27	1-Heptacosene	16	1-Hexadecanol
18	*n*-Octadecane	28	1-Octacosene	17	1-Heptadecanol
19	*n*-Nonadecane	29	1-Nonacosene	18	1-Octadecanol
20	*n*-Eicosane	30	1-Triacontene	20	1-Eicosanol
21	*n*-Heneicosane	Alkyl-benzenes		22	1-Docosanol
22	*n*-Docosane	8	Ethylbenzene	Alkanoic monocarboxylic acids	
23	*n*-Tricosane	9	Propylbenzene	1	Methanoic acid, formic acid
24	*n*-Tetracosane	10	Butylbenzene	2	Ethanoic acid, acetic acid
25	*n*-Pentacosane	11	Pentylbenzene	3	Propanoic acid
26	*n*-Hexacosane	12	Hexylbenzene	4	Butanoic acid, butyric acid
27	*n*-Heptacosane	13	Heptylbenzene	5	Pentanoic acid, valeric acid
28	*n*-Octacosane	14	Octylbenzene	6	Hexanoic acid, caproic acid
29	*n*-Nonacosane	15	*n*-Nonylbenzene	7	Heptanoic acid, enanthic acid
30	*n*-Triacontane	16	*n*-Decylbenzene	8	Octanoic acid, caprilic acid
32	*n*-Dotriacontane	17	*n*-Undecylbenzene	9	Nonanoic acid, pelargonic acid
35	*n*-Pentatriacontane	18	*n*-Dodecylbenzene	10	Decanoic acid, capric acid
36	*n*-Hexatriacontane	19	*n*-Tridecylbenzene	11	Undecanoic acid
40	*n*-Tetracontane	20	*n*-Tetradecylbenzene	12	Dodecanoic acid
44	*n*-Tetratetracontane	21	*n*-Pentadecylbenzene	13	Tridecanoic acid
1-Alkenes		22	*n*-Hexadecylbenzene	14	Tetradecanoic acid
4	1-Butene	23	*n*-Heptadecylbenzene	15	Pentadecanoic acid
5	1-Pentene	24	*n*-Octadecylbenzene	16	Hexadecanoic acid
6	1-Hexene	25	*n*-Nonadecylbenzene	17	Heptadecanoic acid
7	1-Heptene	26	*n*-Eicosylbenzene	18	Octadecanoic acid, stearic acid
8	1-Octene	27	*n*-Heneicosylbenzene	20	Eicosanoic acid, arachidic acid
9	1-Nonene	28	*n*-Docosylbenzene		
10	1-Decene	29	*n*-Tricosylbenzene		
11	1-Undecene	30	*n*-Tetracosylbenzene		

reduced to 1280 by removing the descriptors that had the same value for all compounds. Property data (measured and predicted) were taken from several databases.[2,20–22] A modified version of the stepwise regression program (SROV)[23] was used for the identification of the most appropriate QSPRs. The following properties were predicted: normal boiling temperature (T_b), melting point temperature (T_m), critical temperature (T_c), critical pressure (P_c), and critical volume (V_c).

Property-descriptor relationship for members of homologous series

In a previous work[24] we have shown that for homologous series it is possible to find *dominant* descriptors, which are highly collinear with particular properties that have to be predicted. In such cases, a single descriptor linear QSPR, providing a high level of confidence in long-range extrapolation, can be developed.

The two most common forms of the relationships between properties of compounds belonging to homological series and molecular descriptors will be demonstrated with reference to normal boiling temperature (T_b) and normal melting temperature (T_m) of 19 compounds from the alkanoic monocarboxylic acid series. The T_b values of these compounds are shown as function of the number of carbon atoms (n_C) in Figure 1. Experimental T_b values are available[20] only for compounds with carbon numbers between 1 and 12. It can be observed in Figure 1 that the T_b values increase monotonically and smoothly with increasing n_C value. However, there is a nonlinear (asymptotic) relationship between n_C and T_b, as is the case with most of the presently used descriptors.

Measured T_m values are also available for all the compounds included in this study. In contrast to the normal boiling temperature, the change of T_m as function of n_C is very irregular (Figure 2). The general trend for small number of C atoms is a decrease of T_m with increasing n_C. Starting with $n_C = 6$, the trend is reversed. Moreover, there are consider-

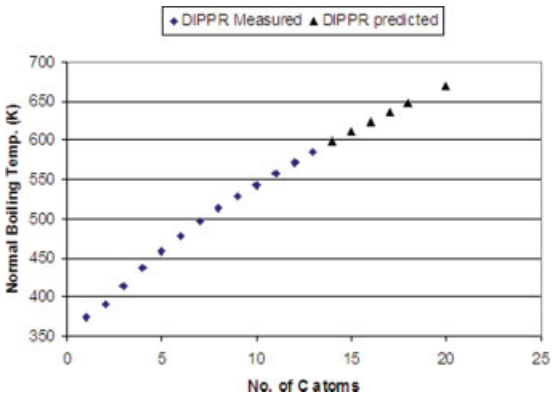

Figure 1. Normal boiling temperature (T_b) of alkanoic, monocarboxylic acids (data from Ref. 20).

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

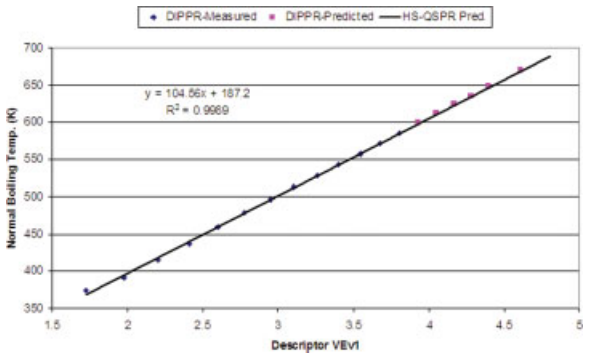

Figure 3. Normal boiling temperature T_b plotted vs. the descriptor *VEv1* for alkanoic, monocarboxylic acids.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

able oscillations of the T_m values between the closest neighbors. Thus, in this case, a simple correlation between T_m and n_C does not exist even though the compounds belong to the same homologous series.

In the database we have used in this study, there are 1280 molecular descriptors for each compound. Some of these descriptors are highly correlated (collinear) with the normal boiling temperature. Figure 3 demonstrates the collinearity between T_b and the descriptor *VEv1* (an eigenvalue-based descriptor: eigenvector coefficient sum from van der Waals weighted distance matrix). Further information on the molecular descriptors used in this study can be found in Ref. 25. The relationship between T_b and the descriptor can be represented by a linear model, $T_b = 104.56\ VEv1 + 187.7$, with a correlation coefficient value of $R^2 = 0.9989$.

However, none of the descriptors in the database can represent satisfactorily the complex behavior of T_m for the alkanoic, monocarboxylic acid series. The descriptor *EEig06x* (an edge adjacency index: eigenvalue 06 from edge adjacency matrix weighted by edge degrees) has the highest correlation with T_m. The plot of T_m vs. *EEig06x* is shown in Figure 4. A linear model, $T_m = 31.496\ EEig06x + 270.54$, represents fairly well the general trend of the change of the property values ($R^2 = 0.9307$); however, the error in representing the individual T_m values is in some cases very high. More descriptors have to be added to the QSPR to improve the accuracy of the representation. This example demonstrates that in spite of the complex behavior of the normal melting temperature, it is possible to find descriptors that represent the general trend, provided that the compounds included in the training set belong to the same homologous series. Including in the training set additional compounds that do not belong to the same series would increase the variety of the structural differences between the members of the set, and the identification of descriptors that represent the general trend of the property variation would be then much more difficult.

This is further demonstrated in Figure 5, where 35 branched alkanes containing between 5 and 11 carbon atoms are included. The number of branches varies between one (2-methylbutane) to four (2,2,3,3-tetramethylhexane). The additional structural differences include the location of the branch and the number of carbon atoms in the branch (one or two). The descriptor *X3A* (a connectivity index; average connectivity index chi-3) has the highest correlation with T_m.

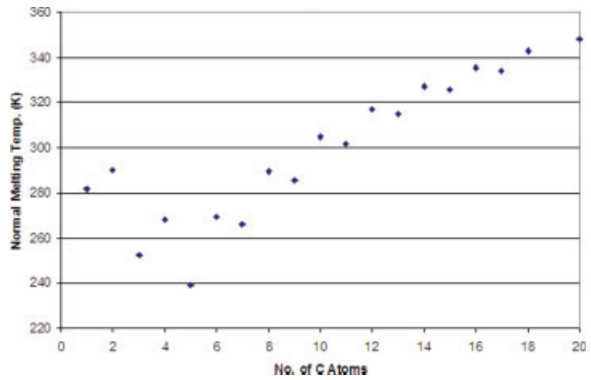

Figure 2. Normal melting temperature (T_m) of alkanoic, monocarboxylic acids (data from Ref. 20).

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

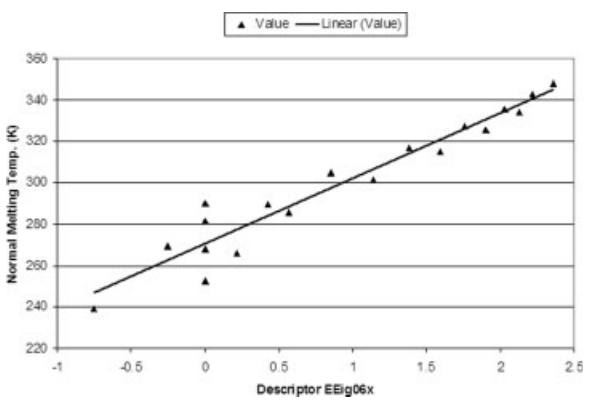

Figure 4. Normal melting temperature T_m plotted vs. the descriptor *EEig06x* for alkanoic, monocarboxylic acids.

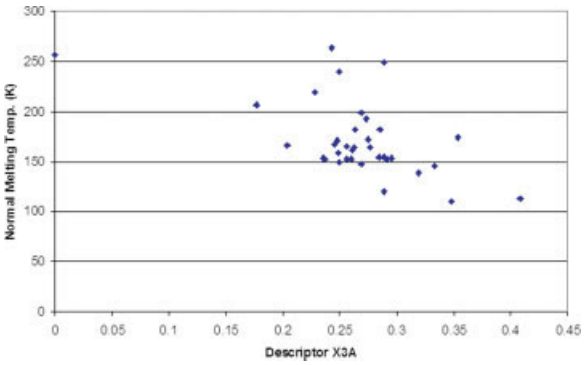

Figure 5. Normal melting temperature T_m plotted vs. descriptor *X3A* for branched alkanes.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

However, in this case, the level of correlation between T_m and the descriptor *X3A* is very low ($R^2 = 0.33$).

The LHS-QSPR method

The linear QSPR method for homologous series (LHS-QSPR) employs the earlier-illustrated advantage of collinearity between certain descriptors and properties. In order to achieve that with the available data, two strategies have been adopted.

In both strategies, the data is divided into a training set and a prediction set. If the extrapolation is not long-ranged (e.g, for compounds differing by less than 10 carbon atoms), only experimental data are employed in both sets.

When long-range extrapolation is needed (or the available data are very few), pseudoexperimental data predicted by a reliable ABC can be added. In this case, however, choosing the right ABC is very important, because it has been demonstrated in Refs. 16 and 17 that, between the different ABCs published, significant differences can be found. Data predicted by QSPRs developed for compounds with wider structural variation should not be used, because their errors might be much higher.[8,26] Since the correct identification of the collinear descriptor(s) is very important, all data from both sets (experimental and predicted) are used first in order to find the collinear descriptors. To this aim, we used the data reported in Ref. 20. However, for the calculation of the QSPR parameter values, the training set is restricted to include only the compounds for which measured property values are available.

For development of the QSPR for a particular property of the homologous series, a linear structure–property relation is assumed of the form:

$$\mathbf{y} = \beta_0 + \beta_1\boldsymbol{\zeta}_1 + \beta_2\boldsymbol{\zeta}_2 \ldots \beta_m\boldsymbol{\zeta}_m + \boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{y}$ is a p-dimensional vector of the respective property (known, measured) values (p is the number of compounds included in the training set), $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \ldots, \boldsymbol{\zeta}_m$ are p-dimensional vectors of predictive molecular descriptors, $\beta_0, \beta_1, \beta_2 \ldots \beta_m$ are the corresponding model parameters to be estimated, and $\boldsymbol{\varepsilon}$ is a p-dimensional vector of stochastic terms (due to measurement errors). Studies involving the TQSPR method[24,27] have shown that using 10 *similar* compounds (for which experimental data are available) is sufficient as a training set.

The descriptors enter the model according to the value of the partial correlation coefficient, $|\rho_{yj}|$ between the vector of the property values $\mathbf{y}$, and that of a potential predictive descriptor $\boldsymbol{\zeta}_j$. The partial correlation coefficient is defined as $\rho_{yj} = \overline{\mathbf{y}}\overline{\boldsymbol{\zeta}}_j^T$, where $\overline{\mathbf{y}}$ and $\overline{\boldsymbol{\zeta}}_j$ are row vectors, centered (by subtracting the mean) and normalized to a unit length. Values of $|\rho_{yj}|$ close to one indicate high correlation between molecular descriptor j and the vector of properties $\mathbf{y}$.

The training set average percent error, defined as

$$\varepsilon_a = \frac{\sum_{i=1}^{p} 100\left|y_i - (\beta_0 + \beta_1\zeta_{1,i} + \beta_2\zeta_{2,i} \ldots \beta_m\zeta_{m,i})\right|/y_i}{p}, \tag{3}$$

can be used for estimating the expected prediction error. Addition of new descriptors to the model may continue as long as the calculated average error is greater than the prespecified error tolerance ($\varepsilon_a > \varepsilon_g$). The SROV[23] method is used, which selects in each step one molecular descriptor that reduces the prediction error most strongly. Two criteria for measuring the signal-to-noise ratio in the j-th candidate descriptor (*TNRj*) and for the partial correlation of the j-th candidate descriptor with the prediction residual (*CNRj*) ensure that the selected descriptors contain valuable information and ensure that overfitting is avoided. A detailed description of the *TNR* and *CNR* criteria and further algorithmic details can be found in Ref. 23.

The property value for a *trial* compound (from the validation set, or any compound in the homologous series that do not have measured property value) is estimated by

$$\tilde{y}_t = \beta_0 + \beta_1\zeta_{t1} + \beta_2\zeta_{t2} \ldots \beta_m\zeta_{tm}, \tag{4}$$

where $\tilde{y}_t$ is the estimated (unknown) property value of the respective compound and $\zeta_{t1}, \zeta_{t2}, \ldots, \zeta_{tm}$ are its corresponding molecular descriptors values.

Prediction of properties of the 1-alcohol series with LHS-QSPRs

The special aspects of the LHS-QSPR method will be demonstrated by predicting T_b, T_c, and T_m for the aliphatic 1-alcohol homologous series.

Normal Boiling Temperature T_b. The T_b data and some additional information regarding the members of the 1-alcohol series, which are included in this study, are shown in Table 2. Alcohols containing between one (methanol) and 22 (1-docosanol) carbon atoms are included. For alcohols up to 1-dodecanol experimental T_b data are available in the DIPPR database.[20] For the compounds 1-tridecanol through 1-eicosanol only predicted data are available (in the DIPPR database).

The DIPPR database[20] provides, in addition to the property values, the reliability of these values. The "reliability" provides an estimate of the upper error bound (%) in a property value. The reliability of most of the experimental T_b values is 1% (except 1-octanol: 0.2%, and 1-dodecanol 3%), and for most of the predicted values is 3% (except 1-tridecanol: 1% and 1-octadecanol 5%).

Table 2. Normal Boiling Temperature Data for the Aliphatic Alcohol Series

Comp. No.	Component Name	No. of C Atoms	T_b* (K) Measured	T_b* (K) Predicted	Reliability (%)
261	Methanol	1	337.85	–	1
262	Ethanol	2	351.44	–	1
263	1-Propanol	3	370.35	–	1
264	1-Butanol	4	390.81	–	1
265	1-Pentanol	5	410.95	–	1
266	1-Hexanol	6	430.55	–	1
267	1-Heptanol	7	449.45	–	1
268	1-Octanol	8	468.35	–	0.2
269	1-Nonanol	9	486.25	–	1
270	1-Decanol	10	504.07	–	1
271	1-Undecanol	11	518.15	–	1
272	1-Dodecanol	12	536.95	–	3
273	1-Tridecanol	13	–	553.6	1
274	1-Tetradecanol	14	–	568.8	3
275	1-Pentadecanol	15	–	583.4	3
276	1-Hexadecanol	16	–	597.23	3
277	1-Heptadecanol	17	–	610.5	3
278	1-Octadecanol	18	–	624	5
279	1-Eicosanol	20	–	645.5	3
280	1-Docosanol	22	–		–

*Data from Ref. 20.

To demonstrate the relevance of the procedure adopted for the descriptor selection, we first show the results obtained when only compounds, for which experimental T_b values are available, are used for the descriptor selection. In this case, the compounds included in the training set range from methanol to 1-dodecanol, except the trial compound (1-decanol, left out for validation). The validation set includes 1-decanol (experimental T_b available) and the range from 1-tridecanol to 1-eicosanol (DIPPR predicted T_b values are used). For this training set, the SROV program has sorted the molecular descriptors with ascending order of their $|\rho_{yj}|$ values. The 10 descriptors with the highest values are shown in Table 3. Many of them are highly collinear with the training set T_b values. For the first descriptor in the list (the descriptor *Mor03u* from the "3D-MoRSE descriptor" group) $|\rho_{yj}| = 0.9998$, while for the ninth descriptor (the descriptor *RDF015v* from the "radial distribution function (RDF)" group) $|\rho_{yj}| = 0.9995$. Most of the descriptors (7 out of 10) in the list are from these two groups. However, there are also descriptors from the connectivity indices, GETAWAY and molecular properties groups. The differences between the $|\rho_{yj}|$ values of the various descriptors shown in Table 3 are however very small, suggesting that each of them can be selected for the QSPR.

Table 3. The 10 Descriptors with the Highest Correlation with T_b for the 1-Alcohol Homologous Series

Descriptor Name	Descriptor Category	ρ_{yj}
Mor03u	3D-MoRSE descriptors	−0.9998
RDF015m	RDF descriptors	0.99976
Mor03e	3D-MoRSE descriptors	−0.99969
RDF015u	RDF descriptors	0.99969
Mor03p	3D-MoRSE descriptors	−0.99969
RDCHI	connectivity indices	0.99959
The	GETAWAY descriptor	0.99952
ALOGP	Molecular properties	0.99951
RDF015v	RDF descriptors	0.9995
Mor03v	3D-MoRSE descriptors	−0.99949

Selecting the descriptor with the highest $|\rho_{yj}|$ value to enter the regression model first yields the following QSPR: $T_b = 336.9205 - 14.7865\ Mor03u$. This single descriptor model already satisfies the signal-to-noise ratio criterion used by the SROV program, thus no more descriptors need to be added to the model. The training set residual plot of this model is shown in Figure 6. The residuals are distributed randomly and the maximal error is ~2 K. For the trial compound, 1-decanol (which is not a member of the training set), the error is also 2 K. Thus, it can be concluded that this QSPR represents the normal boiling temperature for the whole training set (and for the trial compound) well below the 1% reliability. Applying this model to all the alcohols included in Table 1 yields the predicted values shown in the second column of Table 4. The percent errors (based on the experimental and predicted values reported in Ref. 20) are shown in the third column. Amongst the training set, the error percent is the highest for ethanol, 0.45%, below the 1% reliability limit;

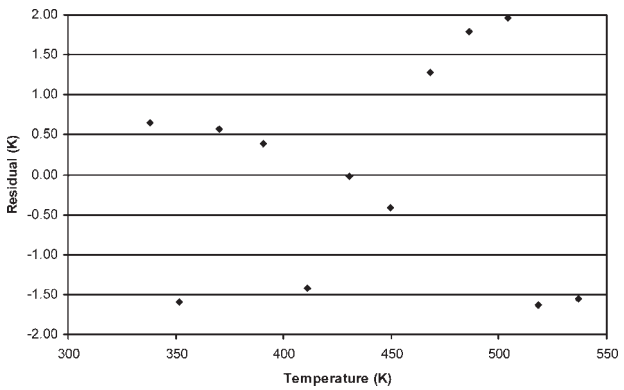

Figure 6. Residual plot for the model $T_b = 336.9205 - 14.7865\ Mor03u$ for the training set of the 1-alcohol homologous series.

Table 4. QSPRs for Predicting the Normal Boiling Temperature of 1-Alcohols

	T_b = 336.9205 − 14.7865 *Mor03u*			T_b = 309.6267+105.4689 *H3v* + 7.2727 *HTe*			
Component Name	LHS-QSPR Prediction	% Error*	Descriptor *Mor03u*	LHS-QSPR Prediction	% Error*	Descriptor *H3v*	Descriptor *HTe*
Methanol	337.20	0.19	−0.019	338.6664	0.242	0	3.993
Ethanol	353.04	0.45	−1.09	350.6268	0.231	0.031	5.188
1-Propanol	369.78	0.15	−2.222	371.3071	0.258	0.066	7.524
1-Butanol	390.42	0.10	−3.618	390.4097	0.102	0.123	9.324
1-Pentanol	412.38	0.35	−5.103	409.5377	0.344	0.187	11.026
1-Hexanol	430.56	0.00	−6.333	430.8477	0.069	0.257	12.941
1-Heptanol	449.86	0.09	−7.638	449.983	0.119	0.315	14.731
1-Octanol	467.07	0.27	−8.802	468.369	0.004	0.363	16.563
1-Nonanol	484.46	0.37	−9.978	485.9987	0.052	0.409	18.32
1-Decanol	502.12	0.39	−11.172	503.37	0.139	0.448	20.143
1-Undecanol	519.79	0.32	−12.367	519.7268	0.304	0.482	21.899
1-Dodecanol	538.50	0.29	−13.633	536.3271	0.116	0.513	23.732
1-Tridecanol	548.65	0.89	−14.319	550.3486	0.587	0.503	25.805
1-Tetradecanol	576.09	1.28	−16.175	567.8767	0.162	0.567	27.287
1-Pentadecanol	594.90	1.97	−17.447	582.8515	0.094	0.587	29.056
1-Hexadecanol	613.40	2.71	−18.698	598.1462	0.153	0.607	30.869
1-Heptadecanol	632.00	3.52	−19.956	612.8736	0.389	0.625	32.633
1-Octadecanol	645.56	3.46	−20.873	625.19	0.191	0.632	34.225
1-Eicosanol	686.50	6.35	−23.642	656.5721	1.715	0.67	37.989
1-Docosanol	723.17	–	−26.122	685.1577	–	0.693	41.586

*Relative error is calculated based on the difference between the DIPPR values and the LHS-QSPR predictions.

however, the percent error increases with increasing n_C, and it exceeds the 3% reliability assigned by DIPPR already for 1-heptadecanol (predicted data). Thus, extrapolating with this QSPR from the training set to higher n_C yields results that are inconsistent with predictions reported in the DIPPR database.[20]

To improve the extrapolation ability of the QSPR, an extended training set that includes both the experimental and the predicted T_b values is used to identify the descriptors included in the QSPR. Using the extended training set the descriptor *HTe* (the seventh descriptor in Table 3) obtains the highest $|\rho_{yj}|$ value. Subsequently, it is used in the QSPR, together with the additional descriptor *H3v* (a "GETAWAY" descriptor) to yield the QSPR: T_b = 309.6267 ± 105.4689 *H3v* + 7.2727 *HTe*. Note that the parameters of the QSPR are calculated with a training set that includes only the compounds for which measured property values are available.

The residual plot of this model is shown in Figure 7. The residuals are distributed randomly and the maximal error is ~1.7 K. The predicted temperature values obtained using this QSPR are shown in the fourth column of Table 4, while the percent error is shown in the fifth column. Except for 1-eicosanol, for all compounds, the errors are below 1%. For 1-eicosanol, the difference between the QSPR prediction and the DIPPR predicted value is 1.75%, below the 3% reliability assigned by DIPPR to their value. It can be concluded that the QSPR containing the *H3v* and *HTe* descriptors yields satisfactory predictions for the members of the training set and also when extrapolating to higher n_C values.

Several techniques for predicting T_b for various compounds, including alcohols, are compared in Ref. 28. Their comparison (like most of the comparisons published in the literature) is based on the average error in the temperature prediction and the number of cases where the prediction error is higher than the threshold value. However, measures for the prediction accuracy and reliability must also consider the experimental error (reliability) of the measured property values, as predicted values cannot be more accurate than the data used for their prediction and evaluation. In fact, all predicted values that are within the experimental error bounds (as obtained via the above LHS-QSPR) should be considered of equivalent reliability. Physical considerations must also be used in assessing the reliability of the predicted values, as overestimation of the experimental error may conceal well-established physical phenomena (e.g., oscillatory behavior of T_m between even and odd carbon numbers caused by different crystal packing).

Critical Temperature T_c. The T_c data for the members of the 1-alcohol series, which are included in this study, are shown in Table 5. In the DIPPR database,[20] experimental T_c, data are available for alcohols up to 1-dodecanol, and for the compounds 1-tridecanol through 1-eicosanol, only predicted data are available. Experimental T_c data for the compounds

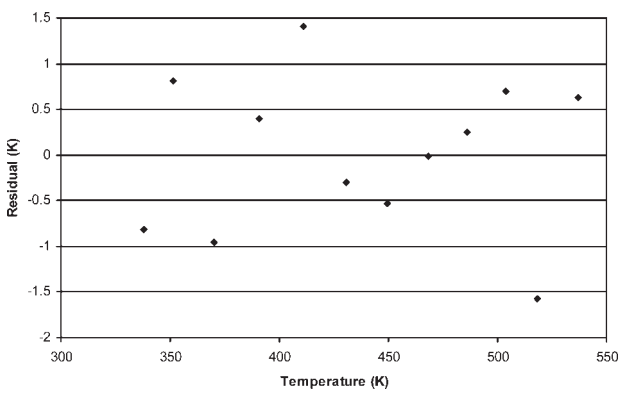

Figure 7. Residual plot for the model, T_b = 309.6267 + 105.4689 *H3v* + 7.2727 *Hte*, for the training set of the 1-alcohol homologous series.

Table 5. Reference Data and QSPR for Predicting the Critical Temperature of 1-Alcohols

	T_c* (K)			$T_c = 514.1452 + 10.7194\ DP07 + 208.2999\ H4p$			
Component Name	Measured	Predicted	Reliability* (%)	LHS-QSPR Prediction	% Error	Descriptor *DP07*	Descriptor *H4p*
Methanol	512.50	–	0.2	514.17	0.33	0.002	0
Ethanol	514.00	–	0.2	514.70	0.14	0.052	0
1-Propanol	536.80	–	0.2	534.75	0.38	0.679	0.064
1-Butanol	563.00	–	0.2	563.59	0.10	2.028	0.133
1-Pentanol	588.10	–	0.2	587.11	0.17	3.523	0.169
1-Hexanol	610.30	–	0.2	610.37	0.01	4.935	0.208
1-Heptanol	632.60	–	0.2	632.58	0.00	6.055	0.257
1-Octanol	652.50	–	0.2	650.46	0.31	7.023	0.293
1-Nonanol	670.70	–	0.2	669.59	0.17	7.894	0.34
1-Decanol	687.30	–	0.2	688.34	0.15	8.672	0.39
1-Undecanol	703.60	–	3	704.93	0.19	9.384	0.433
1-Dodecanol	719.40	–	1	720.87	0.20	10.035	0.476
1-Tridecanol	(732.00)	734.00	3 (1)	729.28	0.64 (0.37)	10.451	0.495
1-Tetradecanol	(743.00)	747.00	5 (1)	748.52	0.20 (0.74)	11.196	0.549
1-Pentadecanol	(757.00)	759.00	5 (1)	761.01	0.26 (0.52)	11.72	0.582
1-Hexadecanol	(770.00)	770.00	5 (1)	772.50	0.32 (0.32)	12.209	0.612
1-Heptadecanol	(780.00)	780.00	5 (1)	783.07	0.39 (0.39)	12.671	0.639
1-Octadecanol	(790.00)	790.00	3 (1)	791.81	0.23 (0.23)	12.981	0.665
1-Eicosanol	(808.00)	809.00	3 (1)	810.77	0.22 (0.34)	13.914	0.708
1-Docosanol	(827.00)	–	(1)	826.34	(0.08)	14.647	0.745

*Data either from Ref. 20 or from Ref. 21 (shown inside parentheses). The data from Ref. 21 are not used in the training set.

1-tridecanol through 1-docosanol are listed in Ref. 21. These data are shown inside parentheses in Table 5. The reliability of most of the experimental T_c values (as reported by DIPPR) is 0.1% (except for 1-undecanol, 3% and for 1-dodecanol 1%) and for predicted values it is either 3% or 5%. The reliability of the experimental data (reported in Ref. 21 in absolute values) corresponds to relative errors of about 1%.

The sequence of the compounds from methanol through 1-dodecanol (excluding a trial compound 1-decanol) was used as the training set. Selecting the descriptor with the highest $|\rho_{yj}|$ value to enter the regression model first, yields the following QSPR: $T_c = 513.7078 - 20.5474\ DP07$ (*DP07* is a "Randic molecular profile" descriptor). The percent error (based on the values reported by DIPPR[20]) of the predicted values obtained by this QSPR for the whole data set are shown in Figure 8. The error is below 2% for all the compounds. This is satisfactory for the predicted values (for which the reliability is either 3% or 5%), but is too high for the members of the training set. Adding one more descriptor to the QSPR yields the model: $T_c = 514.1452 + 10.7194\ DP07 + 208.2999 H4p$ (*H4p* is a GETAWAY descriptor). The values of these two descriptors, the T_c predicted by this model, and the percent error values are shown in Table 5. The addition of the *H4p* descriptor reduced the percent error values below the reliability limit (for both Refs. 20 and 21), except for three members of the training set: methanol, 1-propanol, and 1-octanol. For these compounds, the error percent exceeds slightly the reliability limit of 0.2%. These prediction errors can be further reduced by adding more descriptors; however, considering that the predicted values are well within the reliability limits, further complication of the QSPR may not be justified.

Melting Temperature T_m. The T_m data for the members of the 1-alcohol series included in this study are shown in Table 6. Experimental T_m data are available in the DIPPR database[20] for all the compounds, except for 1-docosanol, for which the melting temperature has been taken from the NIST database.[29] The reliability assigned by DIPPR for most of the T_m values is 1% (except for 1-octanol, 1-nonanol, and 1-undecanol for which the reliability is 3%, for 1-octadecanol it is 0.2%, and for 1-docosanol the reliability is not available).

Figure 9 shows the experimental T_m values plotted vs. the number of carbon atoms in the molecule (n_C). The first three compounds in the homologous series (methanol, ethanol, and 1-propanol) exhibit anomalous variation of T_m, with a decreasing trend with the n_C, rather than an increase, as the rest of the compounds in the series.

To validate the extrapolating capabilities of the QSPRs to be developed for this property, the same training set that was used for T_b and T_c is used also here, even though in this case, experimental values are available for all of the

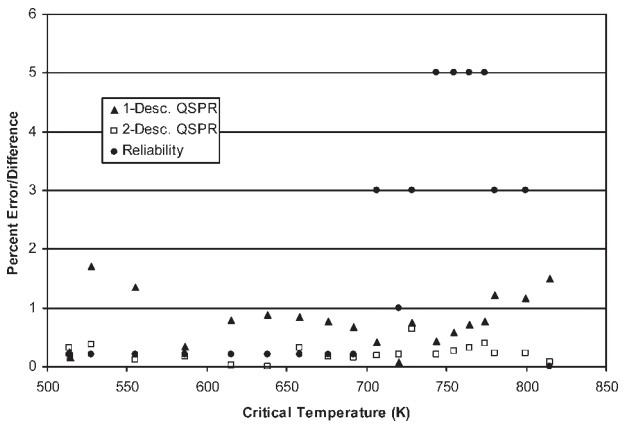

Figure 8. Prediction errors and reliabilities for critical temperatures of the compounds in the 1-alcohol homologous series.

Table 6. Reference Data and QSPR for Predicting the Melting Temperature of 1-Alcohols

		T_m = 257.226 − 45.6894 *IC1* − 8.7167 *EEig06d* − 30.9892 *BELm3* + 131.5317 *BELv5*						
Component Name	T_m* (K) Measured	Reliability* (%)	LHS-QSPR Prediction	% Error	Descriptor *IC1*	Descriptor *EEig06d5*	Descriptor *BELm3*	Descriptor *BElv5*
Methanol	175.47	1	175.35	0.07	1.792	0	0	0
Ethanol	159.05	1	159.24	0.12	1.88	0	0.39	0
1-Propanol	146.95	1	146.95	0.00	1.947	0	0.688	0
1-Butanol	183.85	1	183.84	0.01	1.872	0	0.924	0.31
1-Pentanol	195.56	1	195.49	0.04	1.792	0	1.12	0.417
1-Hexanol	228.55	1	227.71	0.37	1.721	−1.774	1.274	0.556
1-Heptanol	239.15	1	242.06	1.21	1.659	−1.345	1.393	0.7
1-Octanol	257.65	3	255.57	0.81	1.607	−0.901	1.486	0.836
1-Nonanol	268.15	3	267.98	0.06	1.561	−0.503	1.559	0.958
1-Decanol	280.05	1	279.08	0.35	1.522	−0.163	1.617	1.065
1-Undecanol	288.45	3	288.96	0.18	1.487	0.123	1.664	1.158
1-Dodecanol	296.95	1	297.60	0.22	1.457	0.361	1.702	1.238
1-Tridecanol	303.75	1	305.19	0.48	1.43	0.559	1.734	1.307
1-Tetradecanol	310.65	1	311.94	0.41	1.405	0.726	1.761	1.367
1-Pentadecanol	317.05	1	317.87	0.26	1.383	0.867	1.783	1.419
1-Hexadecanol	322.35	1	323.08	0.23	1.363	0.986	1.802	1.464
1-Heptadecanol	327.05	1	327.64	0.18	1.345	1.089	1.818	1.503
1-Octadecanol	331.05	0.2	331.65	0.18	1.329	1.176	1.832	1.537
1-Eicosanol	338.55	1	338.69	0.04	1.299	1.319	1.855	1.595
1-Docosanol	343.00†	–	344.25	0.36	1.274	1.427	1.873	1.64

*Data from Ref. 20.
†Data from Ref. 29.

compounds. It was found that a QSPR of two descriptors, T_m = 144.3501 + 120.1861 *BELv5* + 114.1821 *Qmean*, already represents correctly the shape of the melting temperature curve shown in Figure 9. Using this QSPR, however, yields prediction errors that exceed the reliability for eight compounds, the largest error being for ethanol (3.86%). With a three-descriptor QSPR, the number of excessive errors is reduced to two (see Table 7).

There are several four-descriptor models for which there is only one case of excessive error. From among these, we have selected the one which does not include 3D-descriptors (calculated from the three-dimensional structure of the molecule). Excluding 3D descriptors is preferable as the values of such descriptors may be sensitive to the level of convergence of the energy minimization programs, which are used to determine the 3D structure. The results obtained with the selected four-descriptor QSPR are shown in Table 6. The descriptors associated with correlation are as follows: *IC1*, an information content index, from the information indices group; *EEig06d*, from the edge adjacency indices category; *BELm3* and *Belv5*, from the Burden eigenvalue category. All the descriptors included are 2D. With this model, there is only one excessive error value (1.2% for 1-heptanol).

To check whether removing the first three compounds in the series can yield a simpler QSPR (which can be adequate for extrapolation to higher carbon numbers), we have carried out the descriptor selection after removing methanol, ethanol, and 1-propanol from the training set. Such exclusions are often done when developing ABCs.

The two-descriptor QSPR obtained is T_m = 369.8751 − 244.6928 *BIC2* + 16.3417 *BEHm7* (*BIC2* belongs to the information indices category and *BEHm7* to the Burden eigenvalues category). This QSPR yields satisfactory predictions as long as it is not used for compounds with less than four carbon atoms. There is only one case of excessive error in the region of validity (1.3% error for 1-heptanol); however for methanol, the error is larger than 100%.

The predictions of the melting temperatures with the LHS-QSPR method are compared with predictions obtained by other prediction programs (Table 8): *MPBVP* from Syracuse Research Corporation, *ChemOffice* from CambridgeSoft and *ProPred* from the Technical University of Denmark. The prediction results obtained with these programs are reported in Ref. 1. Additional information regarding these programs can be found in the same reference. Comparison is done for four compounds from the 1-alcohol series. The maximal

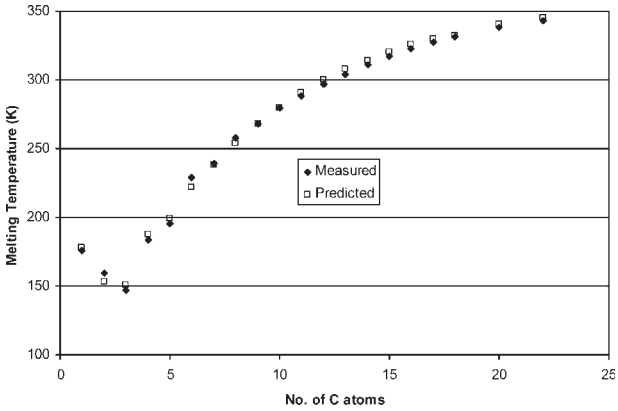

Figure 9. Melting temperatures of 1-alcohols vs. number of carbon atoms, experimental and 2-descriptor QSPR predicted values.

Table 7. Various QSPRs for Representing the Melting Temperature of 1-Alcohols

Description	QSPR	No. of Cases of Excessive Error
Two descriptors	T_c = 144.3501+120.1861 *BELv5*+114.1821 *Qmean*	8
Three descriptors	T_c = −296.357+124.043 *BELv5* + 326.6768 *Qmean* +280.2383*BLI*	2
Four descriptors	T_m = 257.226 − 45.6894 *IC1* − 8.7167 *EEig06d* − 30.9892 *BELm3* + 131.5317 *BELv5*	
Two descriptors, three compounds excluded	T_c = 369.8751−244.6928 *BIC2* +16.3417*BEHm7*	1

error for the LHS-QSPR method is 2.6 K (for 1-octanol), while the maximal errors for the other methods are −53.8 K for MPBPVP, −57 K for ProPed and −47.4 K for ChemOffice. Thus, the use of the LHS-QSPR method reduces the uncertainty associated with the prediction of T_m by more than one order of magnitude.

Application of the LHS-QSPRs for long-range extrapolation

Cases where experimental property values can be accurately represented by a linear model containing a single descriptor (such as the case for T_b, see Figure 3) suggest higher reliability of the predicted values even when extrapolation is involved. The use of single descriptor LHS-QSPRs for long-range extrapolation will be demonstrated by prediction of critical properties for compounds in various homologous series.

Critical Temperature T_c for the 1-Alkene Homologous Series. *The* T_c data for the members of the 1-alkene series that are included in this study are shown in Table 9. This series includes 27 compounds with n_C ranging from 4 to 30. In the DIPPR database experimental data are available for six of the compounds, while predicted values are available for 12 additional compounds. Wakeham et al.[2] list experimental T_c data for 10 additional compounds. These data are shown inside parentheses in Table 9.

The descriptor *DP02* from the Randic molecular profiles group (molecular profile no. 2) was selected based on the T_c data (DIPPR experimental and predicted) available for 17 compounds in the training set. For the calculation of the QSPR parameters, a training set containing only the six compounds with the DIPPR experimental data was used, resulting in the LHS-QSPR: T_c = 258.8382 + 73.1259 *DP02*. The DIPPR predicted and the Wakeham et al.[2] experimental data were used as *validation set*. As shown in Table 9, the one-descriptor QSPR yields T_c values that match the DIPPR predicted values and the experimental values of Ref. 2 within the reliability limits, even though a long-range extrapolation is involved.

Critical Pressure P_c for the Alkyl-Benzene Homologous Series. *The* P_c data for the members of the alkyl-benzene series that are included in this study are shown in Table 10. This series contains 23 compounds with n_C values ranging 8 through 30 (six of those in the benzene ring). DIPPR experimental data are available for the first 3 compounds of the series and predicted values are available for 14 additional compounds. Nikitin et al.[22] provide seven additional experimental values. Using the *BLI* descriptor (a topological descriptor, Kier benzene-likeliness index) and the three DIPPR experimental P_c values yields P_c = 14.0174 − 9.3452 *BLI*. This LHS-QSPR is used to predict the P_c values for the 12 compounds for which only DIPPR predicted values are available (shown in Figure 10) and the seven experimental values of Ref. 22. In all cases, the LHS-QSPR predictions are very close to the DIPPR reported values, and they are inside the bounds that were established using the reliability of the DIPPR data reported. There are four compounds for which the reported reliability limit of the Nikitin et al.[22] data is slightly exceeded (see Table 10). Thus, long-range extrapolation from only three data points yields satisfactory results using the LHS-QSPR method.

Critical Volume V_c for the Monocarboxylic Acid Series. This series includes 19 compounds with n_C range of 1–20. Experimental V_c data are available for seven compounds (ethanoic acid through octanoic acid). Predicted values are reported by DIPPR for the remaining 12 compounds. Using the *Mor21p* descriptor (3D-MoRSE-signal 21/weighted by atomic polarizabilities descriptor) and the seven experimental V_c values yields a LHS-QSPR, V_c = 0.13386 − 0.51211 *Mor21p* (Figure 11). In all cases tested, the LHS-QSPR predictions are very close to the DIPPR reported predictions, and they are inside the bounds that were set according to the reported reliability of the DIPPR data.

Table 8. Comparison of Normal Melting Temperature Prediction Methods for 1-Alcohols

	T_m (K)	LHS-QSPR		MPBPVP*		ProPed*		ChemOffice*	
Compound	Exp.	T_m (K)	Error (K)	T_m (K)	Error (K)	T_m (K)	Error (K)	T_m (K)	Error (K)
Methanol	175.47	175.50	0.0	172.15	3.3	184.05	−8.6	161.45	14.0
Ethanol	159.05	159.07	0.0	209.55	−50.5	194.35	−35.3	172.65	−13.6
1-Propanol	146.95	148.25	−1.3	198.15	−51.2	203.95	−57.0	183.95	−37.0
1-Octanol	257.65	255.08	2.6	258.45	−0.8	244.35	13.3	240.25	17.4

*Data from Ref. 1. In the reference the temperature is reported in °C.

Table 9. Reference Data and QSPR for Predicting the Critical Temperature of 1-Alkenes Using the Model: $T_c = 258.8382 + 73.1259\ DP02$

Component Name	T_c* (K) Experimental	T_c* (K) Predicted	Reliability* (%)	HS-QSPR Prediction	% Error	Descriptor DP02
1-Butene	419.5	–	0.2	418.9	0.14	2.189
1-Pentene	464.8	–	0.2	465.3	0.12	2.824
1-Hexene	504	–	0.2	504.0	0.006	3.353
1-Heptene	537.4	–	0.2	537.8	0.077	3.815
1-Octene	(567)	566.9	1 (0.14)	567.0	0.016 (0.001)	4.214
1-Nonene	593.1	–	1	593.0	0.013	4.57
1-Decene	616.6	–	1	616.3	0.052	4.888
1-Undecene	–	637.8	1	637.3	0.072	5.176
1-Dodecene	(658.0)	657.1	3 (1)	656.6	0.081 (0.22)	5.439
1-Tridecene	(673.0)	674.8	3 (1)	674.3	0.079 (0.19)	5.681
1-Tetradecene	(691.0)	691	3 (1)	690.6	0.062 (0.062)	5.904
1-Pentadecene	(705.0)	708	3 (1)	705.8	0.31 (0.11)	6.112
1-Hexadecene	(718.0)	722	3 (1)	720.0	0.27 (0.28)	6.307
1-Heptadecene	(734.0)	736	3 (1)	733.4	0.35 (0.78)	6.49
1-Oktadecene	(748.0)	748	3 (1)	746.0	0.27 (0.27)	6.662
1-Nonadecene	(755.0)	760	3 (1)	757.9	0.27 (0.39)	6.825
1-Eicosene	(772.0)	771	3 (1.3)	769.3	0.23 (0.35)	6.98
1-Heneicosene	–	–	–	780.0	–	7.127
1-Docosene	–	–	–	790.2	–	7.267
1-Tricosene	–	–	–	800.0	–	7.401
1-Tetracosene	–	–	–	809.4	–	7.529
1-Pentacosene	–	–	–	818.3	–	7.651
1-Hexacosene	–	–	–	827.0	–	7.769
1-Heptacosene	–	–	–	835.3	–	7.883
1-Oktacosene	–	–	–	843.3	–	7.992
1-Nonacosene	–	–	–	851.0	–	8.098
1-Triacontene	–	–	–	858.5	–	8.2

*Data either from Ref. 20 or recommended experimental data from Ref. 2 (shown inside parentheses). The data from Ref. 2 are not used in the training set.

Conclusions

The LHS-QSPR method was applied for predicting T_b, T_c, P_c, V_c, and T_m for the *n*-alkane, 1-alkene, alkyl benzene, aliphatic alcohol, and alkanoic monocarboxylic acid homologous series. The results indicate that property values for which experimental data are available can be predicted within experimental error level, irrespective of whether interpolation or extrapolation to high carbon numbers (n_C) is involved.

Table 10. Reference Data and Results for Predicting Critical Pressure of Alkyl-Benzenes Using the Model: $P_c = 14.0174 - 9.3452\ BLI$

Component Name	P_c* (MPa) Experimental	P_c* (MPa) Predicted	Reliability* (%)	LHS-QSPR Prediction	% Error	Descriptor BLI
Ethylbenzene	3.609	–	1	3.6068	0.061	1.114
Propylbenzene	3.2	–	1	3.205	0.155	1.157
Butylbenzene	2.89	–	3	2.8872	0.096	1.191
Pentilbenzene	(2.58)	2.604	10 (3.1)	2.6256	0.83 (1.77)	1.219
Hexilbenzene	(2.35)	2.38	10 (3.0)	2.4013	0.89 (2.18)	1.243
Heptilbenzene	(2.14)	2.18	10 (2.8)	2.2144	1.58 (3.47)	1.263
Oktylbenzene	(1.98)	2.02	10 (3.0)	2.0555	1.758 (3.81)	1.28
n-Nonylbenzene	–	1.895	10	1.9247	1.57	1.294
n-Decylbenzene	(1.72)	1.77	5 (2.9)	1.8032	1.88 (4.84)	1.307
n-Undecylbenzene	(1.64)	1.672	10 (3.0)	1.7004	1.7 (3.68)	1.318
n-Dodecylbenzene	–	1.56	10	1.5976	2.41	1.329
n-Tridecylbenzene	(1.54)	1.48	10 (3.2)	1.5135	2.26 (1.72)	1.338
n-Tetradecylbenzene	–	1.4	10	1.4387	2.77	1.346
n-Pentadecylbenzene	–	1.33	10	1.3733	3.26	1.353
n-Hexadecylbenzene	–	1.27	10	1.3079	2.98	1.36
n-Heptadecylbenzene	–	1.21	10	1.2518	3.46	1.366
n-Octadecylbenzene	–	1.16	10	1.2051	3.89	1.371
n-Nonadecylbenzene	–	–	–	1.149	–	1.377
n-Eicosylbenzene	–	–	–	1.1116	–	1.381
n-Heneicosylbenzene	–	–	–	1.0649	–	1.386
n-Docosylbenzene	–	–	–	1.0275	–	1.39
n-Tricosylbenzene	–	–	–	0.99015	–	1.394
n-Tetracosylbenzene	–	–	–	0.96212	–	1.397

*Data either from Ref. 20 or from Ref. 22 (shown inside parentheses). The data from Ref. 22 are not used in the training set.

These results show that the linear combination of molecular descriptors included in a QSPR represent well the nonlinear behavior of the particular property in the homologous series in the range of n_C tested. Thus, the use of such QSPR increases considerably the confidence in the values predicted when extrapolation is involved. Property values for which predicted data are available can be matched within the reliability level, even when extrapolation is carried out from an experimental training set as small as with only three compounds. From theoretical considerations it seems, at first glance, unreasonable to use QSPR models for extrapolation. However, the presented LHS-QSPRs can be reliably extrapolated to some extent because of two reasons.

First, it is very unlikely that a linear QSPR model, such as shown in Figure 3, exists just by chance. It can be reasonably assumed that such a linear relationship can be extrapolated towards larger compounds in the homologous series. Table 4, however, showed that the prediction errors may increase with increasing extrapolation range.

Second, long-range extrapolation with one-descriptor LHS-QSPRs is possible because all measured and predicted compounds are used as an extended training set to select one molecular descriptor vector with highest correlation with the property vector. This procedure is equivalent to identifying the best linear one-descriptor QSPR for the extended training set. After the best molecular descriptor is selected using all information available, only measured property data are used to estimate the LHS-QSPR parameters. Thus, these LHS-QSPR model structures incorporate knowledge about predicted property values of the validation set and, thereby, exhibit high prediction accuracy therein.

The use of LHS-QSPRs enabled reducing the maximal error in predicting the melting temperatures of the compounds included in this study below 3 K. This value is smaller by at least one order of magnitude than the deviations of predicted values reported in the literature.[1] It should, however, be noted that the methods tested by Dearden[1] are not tailored to property prediction in a homologous series, but are more generally applicable.

Other (ABC) methods for predicting properties (T_b, T_m, and critical properties) of compounds belonging to homologous series were suggested in the literature (see for example

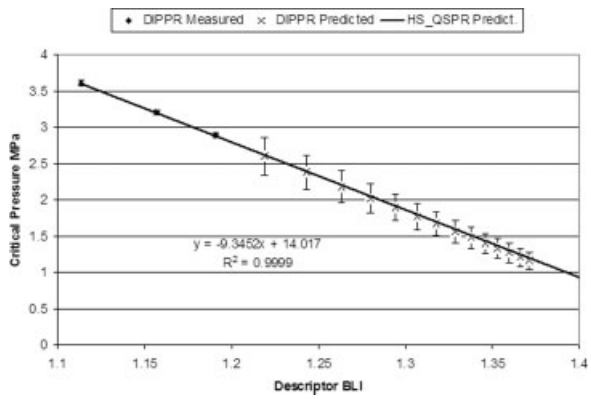

Figure 10. Critical pressure P_c plotted vs. the descriptor *BLI* for alkyl-benzenes.

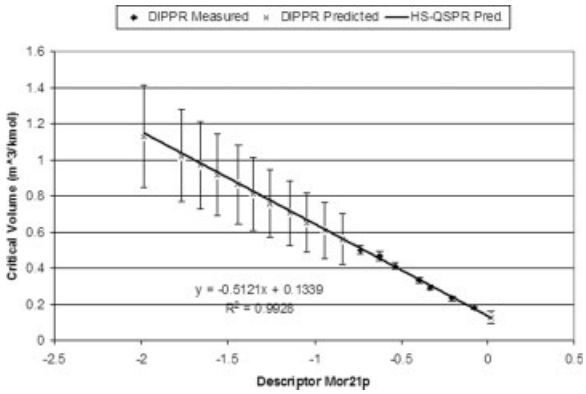

Figure 11. Critical volume V_c plotted vs. the descriptor *Mor21p* for monocarboxylic acids.

Refs. 15 and 21). In the range of carbon number where experimental data are available, the current LHS-QSPR method and other methods that predict property values within the experimental error bounds can be considered of equivalent reliability as long as interpolation is involved. The principal difference between the LHS-QSPR and the ABC methods is that the former attempts to identify a descriptor that is collinear with the predicted property, while the later attempts to represent the nonlinear variation of the property with n_C by a nonlinear function (using nonlinear regression), which may contain up to five parameters (Ref. 15). When long-range extrapolation is involved, the selection of the preferred prediction method should be based on the availability of the information required for its application and the level of confidence the method offers in predicted values. In such cases, the use of single descriptor linear QSPRs when possible, provides higher level of confidence.

Literature Cited

1. Dearden JC. Quantitative structure–property relationships for prediction of boiling point, vapor pressure, and melting point. *Environ Toxicol Chem.* 2003;22:1696–1709.
2. Wakeham WA, Cholakov GSt, Stateva RP. Liquid density and critical properties of hydrocarbons estimated from molecular structure. *J Chem Eng Data.* 2002;47:559–570.
3. Cholakov GSt, Wakeham WA, Stateva RP. Estimation of normal boiling temperature of industrially important hydrocarbons from descriptors of molecular structure. *Fluid Phase Equilibr.* 1999;163:21–42.
4. Marrero J, Gani R. Group-contribution-based estimation of octanol/water partition coefficient and aqueous solubility. *Ind Eng Chem Res.* 2002;41:6623–6633.
5. Dyekjaer JD, Jonsdottir SO. QSPR models based on molecular mechanics and quantum chemical calculations, Part 2: Thermodynamic properties of alkanes, alcohols, polyols and ethers. *Ind Eng Chem Res.* 2003;42:4241–4259.
6. Li Q, Chen X, Hu Z. Quantitative structure–property relationship studies for estimating boiling points of alcohols using calculated molecular descriptors with radial basis-function neural networks. *Chemom Intell Lab Syst.* 2004;72:93–100.
7. Poling BE, Prausnitz JM, O'Connel JP. *Properties of Gases and Liquids*, 5th ed. New York: McGraw-Hill, 2001.
8. Yan X, Dong Q, Hong X. Reliability analysis of group-contribution methods in predicting critical temperatures of organic compounds. *J Chem Eng Data.* 2003;48:374–380.
9. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci.* 2003:22:69–77.

10. Basak SC, Gute BD, Mills D, Hawkins DM. Quantitative molecular similarity methods in the property/toxicity estimation of chemicals: a comparison of arbitrary versus tailored similarity spaces. *J Mol Struct Theochem*. 2003;622:127–145.
11. Shacham M, Brauner N, Cholakov GSt, Stateva RP. Property prediction by correlations based on similarity of molecular structures. *AIChE J*. 2004;50:2481–2492.
12. Brauner N, Shacham M, Cholakov Gst, Stateva RP. Property prediction by similarity of molecular structures—practical application and consistency analysis. *Chem Eng Sci*. 2005;60:5458–5471.
13. Cholakov GSt, Stateva RP, Shacham M, Brauner N. Identifying equations that represent properties in homologous series using structure–structure relations. *AIChE J*. 2007;53(1):150–159.
14. Brauner N, Stateva RP, Cholakov GSt, Shacham M. A structurally "targeted" QSPR method for property prediction. *Ind Eng Chem Res*. 2006;45:8430–8437.
15. Marano JJ, Holder GD. General equation for correlating the thermophysical properties of *n*-paraffins, *n*-olefins and other homologous series, Part 2: Asymptotic behavior correlations for PVT properties. *Ind Eng Chem Res*. 1997;36:1895–1907.
16. Gao W, Robinson RL, Gasem KAM. Improved correlations for heavy *n*-paraffin physical properties. *Fluid Phase Equilibr*. 2001; 179:207–216.
17. Nikitin ED, Popov AP, Bogatishcheva NS. Critical properties of long-chain substances from the hypothesis of functional self-similarity. *Fluid Phase Equilib*. 2005;235:18–23.
18. Godavarthy SS, Robinson RL, Gasen KAM. An improved structure–property model for predicting melting point temperatures. *Ind Eng Chem Res*. 2006;45:5117–5126.
19. Todeschini R, Consonni V, Mauri A, Pavan M. *DRAGON User Manual*. Milano, Italy: Talete srl, 2006.
20. Rowley RL, Wilding WV, Oscarson JL, Yang Y, Zundel NA. *DIPPR Data Compilation of Pure Chemical Properties Design Institute for Physical Properties*. Provo, Utah: Brigham Young University, 2006. Available at http//dippr.byu.edu.
21. Nikitin ED, Pavlov PA, Popov AP. Critical temperatures and pressures of 1-alkanols with 13 to 22 carbon atoms. *Fluid Phase Equilibr*. 1998;149:223–232.
22. Nikitin ED, Polov AP, Bogatishcheva NS, Yatluk YG. Vapor–liquid critical properties of *n*-alkylbenzenes from toluene to 1-phenyltridecane. *J Chem Eng Data*. 2002;47:1012–1016.
23. Shacham M, Brauner N. The SROV program for data analysis and regression model identification. *Comput Chem Eng*. 2003;27:701–714.
24. Shacham M, Kahrs O, Cholakov GSt, Stateva RP, Marquardt W, Brauner N. The role of the dominant descriptor in targeted quantitative structure property relationships. *Chem Eng Sci*. 2007;62:6222–6233.
25. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim, 2000.
26. Nannoolal Y, Rarey J, Ramjugernath D. Estimation of pure component properties, Part 2: Estimation of critical property data by group contribution. *Fluid Phase Equilib*. 2007;252:1–27.
27. Kahrs O, Brauner N, Cholakov GSt, Stateva RP, Marquardt W, Shacham M. Analysis and refinement of the targeted QSPR method. *Comput Chem Eng*. doi: 10.1016/j.compchemeng.2007.06.006 (2007).
28. Cordes W, Rarey J. A new method for estimation of normal boiling point of non-electrolyte organic compounds. *Fluid Phase Equilibr*. 2002;201:409–433.
29. Chickos JS, Acree WE, Jr., Liebman JF, Students of Chem 202 (Introduction to the Literature of Chemistry), University of Missouri–St. Louis, "Heat of Fusion data." In: Linstrom PJ, Mallard WG, editors. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. Gaithersburg, MD: NIST, June 2005; 20899. Available at http://webbook.nist.gov.

Manuscript received Jun. 9, 2007, and revision received Dec. 2, 2007.