Property Prediction by Correlations Based on Similarity of Molecular Structures

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

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

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

DOI 10.1002/aic.10248
Published online in Wiley InterScience (www.interscience.wiley.com).

A new approach for predicting a wide range of physical and thermodynamic properties is proposed. It involves calculation of the molecular descriptors of a target compound of unknown properties, followed by regression of this vector of molecular descriptors vs. a database of compounds with known descriptors and measured properties. The regression model, obtained for the target descriptors in terms of predictive compounds and their coefficients, is then used for prediction of properties of the target compound. The precision of the prediction can be estimated from the standard deviation of the correlation and the known precision of the property data of the predictive compounds. The proposed method was tested in predicting 31 properties of 18 compounds representing different hydrocarbon structures. The results show that the method has several unique advantages, such as the use of one structural correlation to predict all properties; estimation of the prediction error for compounds without measured data; opportunities to find alternative solutions to different problems and means to estimate their adequacy. The method can be used also for checking the consistency of measured data and data predicted by other methods. © 2004 American Institute of Chemical Engineers *AIChE J,* 50: 2481–2492, 2004
Keywords: property prediction, QSPR (molecular descriptors), molecular simulation, collinearity

Introduction

Modeling and simulation of chemical processes require, in addition to the process model, data for physical and thermodynamic properties of the various compounds, often for wide ranges of temperatures, pressures, and compositions. In many cases, experimental data for the needed properties are not available and have to be calculated with suitable quantitative structure–property relationships (QSPRs).

Correlations of acceptable accuracy can be derived from measured values of pure component constants, such as the normal boiling temperature (T_b), liquid density (d_4^{20}), and critical properties (T_c, P_c, V_c), and molecular descriptors (Poling et al., 2001). Lydersen (1955) initiated the use of functional group contributions as descriptors for estimating critical constants. Nowadays, most available methods are based on atom contributions, bond or group interaction contributions, and group contributions (Poling et al., 2001).

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

An alternative to the group contribution methods, for predicting pure component properties, is to determine—from a vast database—a combination of molecular descriptors that defines the most "significant common features" (SCFs) of the described molecules (Wakeham et al., 2002). The different molecular descriptors in the database may be groups and/or bonds and computed by procedures such as simulated molecular mechanics, quantum chemical methods, and topology of the molecules. The descriptors that are significant for the prediction of a particular constant and their weighting factors are usually found by stepwise regression techniques.

Poling et al. (2001) carried out extensive studies regarding the accuracy of the prediction of the various molecular structure-based techniques. They found that for most compounds the prediction error is less than 5%. However, for a considerable number of compounds the error of the prediction exceeds 10% and differences in the values of unknown properties predicted by different methods may amount to percentages of more than several hundred. Unfortunately, for a target compound of unmeasured pure component constants, it is impossible to assess the prediction accuracy. With no feedback on the prediction error, it is impossible to choose among the methods proposed by different authors, and to advocate the best.

The number of the compounds used at present by the industry or those of its immediate interest is estimated at around 100,000, whereas the chemical structures, which are theoretically possible and may eventually interest the industry in the future, are at least several tens of millions (Horwath, 1992). Moreover, most of the compounds that the chemical industry would like to test and eventually use as drugs, polymers, additives and so forth, possess complex structures, and this tendency will become more pronounced with the development of the methods of product design. In contrast, the number of the compounds for which measured data are available is at most several thousands and for many properties is much less. Typically, compounds chosen for experimental assessment are selected at random, so the present databases from a structural perspective are unevenly populated clusters of relatively simple compounds of medium molecular masses.

Facing the above contradictions, the user of the existing predictive correlations has to accept implicitly that they can be extrapolated toward the properties of the compounds needed, although it is obvious that such extrapolations may involve immense errors, which presently cannot be estimated.

QSPRs are built on the assumption that chemical structures are related. For instance, within homologous series, chemical structures differ by a $-CH_2$ group and their properties can be predicted with high precision from asymptotic behavior correlations (Marano and Holder, 1997a,b). It has also been suggested to use structure–structure relationships in the development of structure–property correlations for complex structures (Cholakov et al., 1999).

As a fresh first step toward overcoming the above limitations of the existing prediction techniques, we are advocating herein a novel technique. It is based on a presumption for linear dependency between the molecular descriptors of various compounds and between their pure components constants. The database described by Cholakov et al. (1999) and Wakeham et al. (2002) is used to test this presumption. The cited work used 99 molecular descriptors, presented in Appendix A, and 260 hydrocarbons. For most of the compounds, the pure component constants T_b, d_4^{20}, T_c, P_c, and V_c are available. By use of the stepwise colinearity diagnostics (SCD) algorithm of Brauner and Shacham (2000) the extent of linear dependency between the molecular descriptors of the various compounds was determined. New prediction equations were derived using the orthogonalized-variable-based stepwise regression (SROV) algorithm of Shacham and Brauner (2003).

In the next section linear relationships between properties of similar compounds will be demonstrated. After that the basic principles of the SCD and SROV algorithms will be briefly reviewed. The results of the analysis of the linear dependency between the molecular descriptors of the various compounds will be presented and prediction equations will be derived. Finally, results for the prediction of properties available from the AIChE Design Institute for Physical Properties (DIPPR) database will be given and some conclusions will be drawn.

Linear Relationship between Properties of Similar Compounds

The proposed technique is based on linear relationships between the molecular descriptors of similar compounds as well as between measured properties of similar compounds. These linear relationships are demonstrated with reference to Figures 1 and 2. In Figure 1 normalized molecular descriptors of *n*-heptane are plotted vs. those of *n*-hexane. It can be seen that the molecular descriptors are aligned along a straight line with a correlation coefficient of $R^2 = 0.9965$. Figure 2 shows a similar plot for the molecular mass and some measured properties of the two compounds. The measured properties include the critical temperature, critical volume and compressibility factor, melting point, triple point temperature and pressure, normal boiling temperature, liquid molar volume, refractive index, flash point, lower and upper flammability limits, and lower flammability limit temperature. In spite of the great variety of gas, liquid, and solid properties, these are also aligned along a straight line with a slope similar to that obtained in Figure 1 and $R^2 = 0.9994$.

This example demonstrates that there is a linear relationship between the properties of two neighboring compounds in a homological series. The SCD algorithm can be applied to identify which of the compounds in the database can be represented by a linear combination of other compounds. Such a compound can then be selected as a *target compound*, and the SROV algorithm can identify an appropriate set of *predictive compounds* that describe the entire set of molecular descriptors of the target compound. The coefficients of the *structure–*

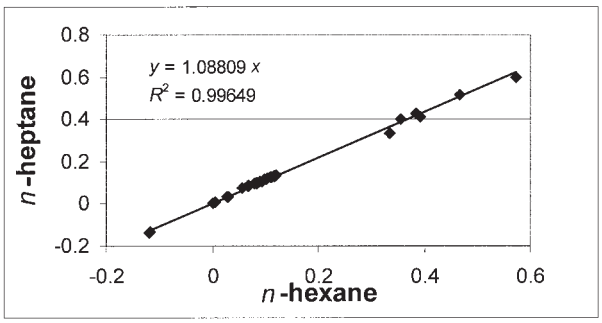

Figure 1. Plot of normalized molecular descriptors of *n*-heptane vs. those of *n*-hexane.

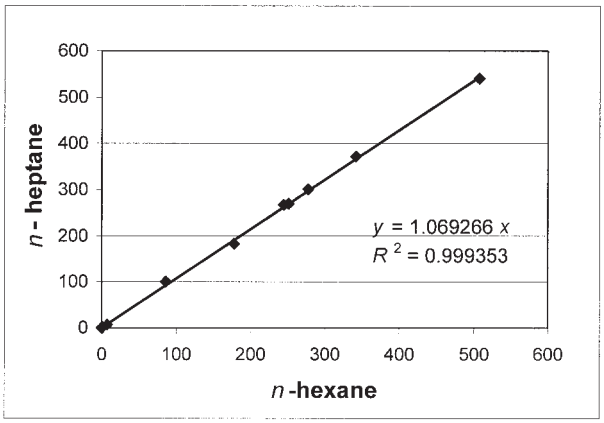

Figure 2. Plot of selected properties of *n*-heptane vs. those of *n*-hexane.

structure correlation thus obtained can then be used to predict unknown properties of the target compound.

The SCD Algorithm

The SCD algorithm is used for separating a set of variables into two subsets. The first is an orthogonal subset (base), where all the variables have orthogonal components. The second is a collinear subset, where all the variables can be represented as linear combinations of the base variables. A detailed description of this algorithm can be found in Brauner and Shacham (2000). The basic principles of the algorithm are briefly reviewed herein.

Consider a data set consisting of N-vectors of molecular descriptors (variables) $\mathbf{x}_j$ ($j = 1, 2, \ldots, n$), where the data is subject to a certain error (imprecision, noise)

$$\mathbf{x}_j = \hat{\mathbf{x}}_j + \boldsymbol{\delta}\mathbf{x}_j \tag{1}$$

where $\hat{\mathbf{x}}_j$ is an N-vector of expected values of $\mathbf{x}_j$ and $\boldsymbol{\delta}\mathbf{x}_j$ is an N-vector of stochastic terms attributed to noise. The individual elements of the noise vector $\boldsymbol{\delta}\mathbf{x}_j$ are determined, in this case, based on the round-off error of the calculated molecular descriptors data. For integer data (such as number of carbon atoms) the noise level is the computer precision and for descriptors presented with real numbers (such as those from simulated molecular mechanics; see Appendix A), the noise level is determined by the precision with which they are calculated.

The basic variables are identified by applying the Gram–Schmidt orthogonalization technique to the whole set of variables in a stepwise fashion. At each step, the variable selected to enter the basis, $\mathbf{x}_p$, is the one with the highest signal-to-noise ratio. The signal-to-noise ratio in a variable, TNR_j, is defined in terms of the corresponding Euclidean norm

$$TNR_j = \frac{\|\mathbf{x}_j\|}{\|\boldsymbol{\delta}\mathbf{x}_j\|} = \left\{\frac{\mathbf{x}_j^T}{\boldsymbol{\delta}\mathbf{x}_j^T}\frac{\mathbf{x}_j}{\boldsymbol{\delta}\mathbf{x}_j}\right\}^{1/2} \tag{2}$$

Upon selecting $\mathbf{x}_p$ at step k ($\mathbf{x}_p^k$), the subset of nonbasic variables is updated by subtracting the information that is collinear with $\mathbf{x}_p$, whereby

$$\mathbf{x}_j^{k+1} = \mathbf{x}_j^k - \mathbf{x}_p\left[\frac{(\mathbf{x}_j^l)^T(\mathbf{x}_p^k)}{(\mathbf{x}_p^k)^T(\mathbf{x}_p^k)}\right] \tag{3}$$

This process continues until the signal-to-noise ratio (TNR_j) gets close to or below the value of 1 for all the remaining nonbasic variables. The selection of this threshold value is based on our previous experience.

The SROV Algorithm

The SROV algorithm is used to identify the variables whose linear combination can adequately represent a target variable. This algorithm was detailed in Shacham and Brauner (2003). The basic principles of the algorithm are briefly reviewed.

The SROV program carries out stepwise regression to obtain a linear regression model of the form

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \cdots + \beta_m\mathbf{x}_m + \boldsymbol{\varepsilon} \tag{4}$$

where $\mathbf{y}$ is an N-vector of the molecular descriptors of the *target compound* (dependent variable); $\mathbf{x}_j$ ($j = 1, 2, \ldots m$) are N-vectors of *predictive compounds* (independent variables); β_0, $\beta_1, \ldots, \beta_m$ are the model parameters to be estimated; and $\boldsymbol{\varepsilon}$ is an N-vector of stochastic terms (measurement errors).

The SROV solves the equation $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$ using QR decomposition by decomposing $\mathbf{X}$ into the product of a matrix $\mathbf{Q}$ (of orthogonal columns) and an upper triangular matrix $\mathbf{R}$. The $\mathbf{Q}$ matrix is orthogonalized by the Gram–Schmidt method. The variables are added in the model in a stepwise fashion, as in SCD. Here, however, at each step, a variable is selected to the model on the basis of the strength of its linear correlation with the dependent variable. The strength of this correlation is measured by the vector product $YX_j = \mathbf{y}^T\mathbf{x}_j$, where $\mathbf{y}$ and $\mathbf{x}_j$ are centered and normalized to a unit length. Therefore, the value of $|YX_j|$ is in the range [0, 1]. In a case of a perfect correlation between $\mathbf{y}$ and $\mathbf{x}_j$, $|YX_j| = 1$, whereas if the two vectors are orthogonal, $YX_j = 0$. The signal-to-noise ratio in a correlation CNR_j is defined by

$$CNR_j^k = \left\{\frac{|(\mathbf{y}^k)^T\mathbf{x}_j^k|}{\sum_{i=1}^{N}(|x_{ij}^k\varepsilon_i^k| + |y_i^k\delta x_{ij}^k|)}\right\} \tag{5}$$

After the selection of $\mathbf{x}_p$ at step k ($\mathbf{x}_p^k$), the $\mathbf{Q}$ and $\mathbf{R}$ matrices are updated using the following equations

$$r_j^k = \frac{(\mathbf{x}_j^k)^T\mathbf{x}_p^k}{(\mathbf{x}_p^k)^T\mathbf{x}_p^k} \tag{6}$$

$$\mathbf{q}_j^{k+1} \equiv \mathbf{x}_j^{k+1} = \mathbf{x}_j^k - \mathbf{x}_p^k r_j^k \tag{7}$$

Note that the columns of the $\mathbf{Q}$ matrix, generated in Eq. 7, contain the updated subset of nonbasic variables. In this respect Eq. 7 is equivalent to Eq. 3. At the same time the parameter value associated with $\mathbf{x}_p$ is calculated as

$$\hat{\beta}_k = \frac{(\mathbf{y}^k)^T(\mathbf{x}_p^k)}{(\mathbf{x}_p^k)^T(\mathbf{x}_p^k)} \tag{8}$$

Table 1. Example of Subsets of Base Compounds (in Bordered Cells) and Collinear Compounds*

1-pentacosene	cyclobutane	heptadecylcyclopentane
1-hexacosene	cyclopentane	octadecylcyclopentane
1-heptacosene	cyclohexane	nonadecylcyclopentane
1-oktacosene	cycloheptane	eicosylcylcyclopentane
1-nonacosene	cyclooctane	heneicosilcylcyclopentane
1-triacontene	methylcyclohexane	docosylcylcyclopentane
1,3-butadiene	ethylcyclohexane	tricosylcylcyclopentane
c-2-butene	propylcyclohexane	tetracosylcylcyclopentane
t-2-butene	butlcyclohexane	pentacosylcylcyclopentane
i-butene	methylcyclopentane	*t*-1,3-dimethylcyclohexane
isoprene	ethylcyclopentane	cyclopentene
2,3-dimethyl-1-butene	propylcyclopentane	cyclohexene
2,3-dimethyl-2-butene	butylcyclopentane	benzene
c-2-hexene	pentylcyclopentane	toluene
t-2-hexene	hexylcyclopentane	ethylbenzene
4-methyl-1-pentene	heptylcyclopentane	propylbenzene
2,4,4-trimethyl-1-pentene	octylcycclopentane	butylbenzene
2,4,4-trimethyl-2-pentene	nonylcyclopentane	*o*-xylene
2-methyl-1-butene	decylcyclopentane	*m*-xylene
2-methyl-2-butene	undecylcyclopentane	*p*-xylene
3-methyl-1-butene	dodecylcyclopentane	1-methyl-3-ethylbenzene
2,3-dimethyl-butadiene	tridecylcyclopentane	pentylbenzene
2-methyl-2-pentene	tetradecylcyclopentane	hexylbenzene
1,5-hexadiene	pentadecylcyclopentane	heptylbenzene
cyclopropane	hexadecylcyclopentane	octylbenzene

*Part of the database of Wakeham et al. (2002).

and the **y** vector is updated to obtain the unpredicted residuals, which are orthogonal to the basic variables subset

$$\mathbf{y}^{k+1} = \mathbf{y}^k + \hat{\beta}_k \mathbf{x}_p^k \tag{9}$$

The addition of new variables to the model continues as long as CNR_j in the correlation is >1 for at least one of the remaining nonbasic variables. In an additional phase of the algorithm, the variables selected to the model are rotated to ensure that the "optimal" model has been identified, independently of the order of variables selection. Eventually, the optimal model is the one of the lowest variance and stable. It should be pointed out that because the SROV algorithm excludes variables with $TNR_j \leq 1$ from the model, potential collinearity between the basic variables is avoided.

Collinearity between Molecular Descriptors of Various Compounds

A subset of the database of Wakeham et al. (2002) was used for studying linear dependency between the structures of the various compounds. The subset contains the 208 hydrocarbons for which full sets of measured properties (T_b, d_4^{20}, T_c, P_c, and V_c) are available. Errors in the calculated values of the molecular descriptors ($\boldsymbol{\delta}\mathbf{x}_j$) were determined based on the round-off error, except for integer data (such as the total number of carbon atoms), where the word length (precision) of the computer was used (see Appendix A). The SCD program was applied to the molecular descriptors of the database. Two subsets, one of orthogonal base variables containing 68 compounds and an associated subset of 140 collinear compounds, were identified. Part of the results is presented in Table 1, where the base compounds are shown in bordered cells.

Table 2. SROV Results for Model Fitting to the Molecular Descriptors of 1-Octene

Step No.	Compound	YX_j	TNR_j	CNR_j	Variance
1	**1-nonene**	**0.99899**	2714.4	7473.1	7.615E-5
	1-heptene	0.99877	1797.1	6118.8	
	1-decene	0.99627	2263.4	6331.4	
	1-hexene	0.99432	2070.1	5791.3	
2	**1-heptene**	**0.99823**	120.3	198.16	2.72E-07
	1-hexene	0.99058	195.13	221.71	
	1-undecene	−0.98786	359.4	289.11	
	1-dodecene	−0.98602	70.429	123.92	
3	**1-hexene**	**−0.95092**	13.29	20.6	2.63E-08
	1-pentene	−0.93287	37.97	36.9	
	1-butene	−0.91151	44.9	38.2	
	propylene	−0.90183	67.54	40.05	
4	**1-decene**	**−0.44635**	18.7	7.64	2.12E-08
	1-pentadecene	−0.37967	6.54	2.96	
	1-heptadecene	−0.37934	6.16	2.83	
	1-undecene	−0.37662	7.33	4.5	

The results from the SCD analysis confirm the existence of linear dependency between the molecular structures of the various compounds as presented by their descriptors. To identify the precise linear relationship between the molecular descriptors of the various compounds, the SROV algorithm is applied to the whole database.

Deriving Linear Correlations from Molecular Descriptors Using SROV

The SROV program was used to identify linear relationships between the molecular descriptors of target compounds and the molecular descriptors of the rest of the compounds in the database. To carry out this study, the 99 molecular descriptors in the database were normalized by dividing each descriptor by the maximal absolute value of this particular descriptor over the 208 compounds.

The interactive operation of the SROV program will be demonstrated by fitting a regression model to the molecular descriptors of 1-octene (the target compound). The first four steps of the stepwise regression process are shown in Table 2. In this table, the first four compounds (out of the 207 compounds, in this case) are presented as vectors of their molecular descriptors (shown in bold) of the highest correlation with the target compound are shown, together with their YX_j, TNR_j, and CNR_j values as well as the variance of the correlation at each of the regression steps.

Inspection of the results at step 1 reveals that the highest correlation between the molecular descriptors of 1-octene is with those of 1-nonene ($YX_j = 0.99899$). The value of YX_j is slightly lower for 1-heptene ($YX_j = 0.99877$), whereas 1-decene and 1-hexene are the next compounds of choice. All the signal-to-noise ratios (TNR_j and CNR_j) are very high, indicating that the useful information included in these variables is much higher than the noise at this first step.

Upon selecting 1-nonene to the regression model (as the first predictive compound[b]), the variance of the straight-line correlation obtained is 7.615E-05. The correlation between the residual of the molecular descriptors of the target compound (1-octene) and those of the predictive compounds is only slightly reduced, but there is more than an order-of-magnitude reduction in the TNR_j and CNR_j values. Now the highest correlation is with 1-heptene ($YX_j = 0.99823$), whereas 1-hexene, 1-undecene, and 1-dodecene follow closely. After adding 1-heptene to the regression model, as the second predictive compound, the variance is reduced by two orders of magnitude to 2.72E-07, the YX_j values are slightly reduced, and there is once again an order-of-magnitude reduction in the TNR_j and CNR_j values. In the third step, 1-hexene is entered to the model, reducing the variance by one order of magnitude, to 2.63E-08. At this point there is a sharp reduction in the YX_j values and all the CNR_j values decrease to a value < 10. Adding 1-decene to the regression model results in a slight reduction of the variance, to 2.12E-08. This step brings all the CNR_j values near or below the threshold value of 1 (not shown in Table 2), and thus the use of SROV ensures that all the compounds included in the regression models contain residual (orthogonal) information above the noise level.

Table 3 shows the final form of the regression model obtained for 1-octene, along with some statistical indicators for the goodness of the fit. All those indicators (variance, R^2, 95% confidence intervals) indicate an excellent fit. The standard deviation (σ) is also shown in Table 3. Because the molecular descriptors are normalized to 1, 100σ provides a good estimate for the percentage error in the correlation. For *n*-hexane, $100\sigma = 0.0146$; thus the error in the correlation is less than 0.1%. The molecular descriptors of 1-octene can therefore be represented by the linear equation

$$\text{1-octene} = -0.22028(\text{1-hexene}) + 0.8098(\text{1-heptene}) + 0.45112(\text{1-nonene}) - 0.040664(\text{1-decene}) \quad (10)$$

Introducing (as an illustration of the above procedure) the number of carbon atoms into the right-hand side of Eq. 1 yields: $-(0.22028 \times 6) + (0.8098 \times 7) + (0.45112 \times 9) - (0.040664 \times 10) = 8.00036$. Thus the number of carbon atoms of 1-octene is represented with four decimal digits accuracy.

The excellent prediction for 1-octene by its closest neighbors in its homologous series could be expected. However, the proposed method also provides great flexibility in selecting the predictive compounds. Let us assume, for example, that 1-nonene (the compound of the highest correlation with 1-octene) cannot be used in the correlation for some reason. In such a case SROV selects an alternative regression model shown in Table 4. This regression model is of a similar accuracy. Using the coefficients shown in Table 4 and the number of carbon atoms in the predictive compounds (5, 7, 10, 12, 16, and 25, respectively), as an example, yields the value of 7.999915 for the number of carbon atoms in 1-octene. However, this model contains more predictive compounds (six), several of them being quite far from the target

Table 3. Linear Correlation for the Molecular Descriptors of 1-Octene

Compound	Coefficient	95% Confidence Interval
1-hexene	−0.22028	0.0145
1-heptene	0.8098	0.0258
1-nonene	0.45112	0.0281
1-decene	−0.04066	0.0167
Variance	2.12E-008	
R^2 (linear corr. coeff.)	1	
Standard deviation (σ)	1.46E-04	

compound in the homologous series. Thus, the SROV program identifies the most appropriate predictive compounds available, irrespective of their location in the homological series, and—as will be demonstrated further—also select compounds from different series, if necessary.

The precision of the structure–structure correlation can be improved to some extent by adding more predictive compounds to the correlation. However, this may come at the expense of larger propagated experimental error when applying it to prediction of properties. This subject will be discussed in more detail in the next section that deals with the prediction of properties.

To study further the correlation that may exist between the structures of a *target* compound and several *predictive* compounds, the SROV program was used to identify the "structure–structure" correlation equations for 18 target compounds that were used by Wakeham et al. (2002). The results of this study are summarized in Table 5. The best-fit correlation is obtained for compound 8, 1-octene (already discussed in detail above).

The prediction of the properties of members of the main homologous series, as already mentioned, does not present a practical problem. So, it is much more interesting to extend our analysis to the more complex molecular structure–structure correlations (Table 5), suggesting relationships that cannot be readily established. It may be seen from the table that for all compounds of a more complex structure the "structure–structure" correlations are of high precision, although some of the chosen predictive compounds are rather surprising. The latter fact needs further theoretical clarification in our future work.

Property Prediction Using Molecular Descriptors–Based Correlations

It is assumed that the linear correlation between the molecular descriptors of the target compound and those of the predictive compounds also apply to various structure-dependent properties of the same compounds. Thus, a property of the target compound can be predicted by introducing the known values of the same property of the predictive compounds into the linear correlation equation obtained in the previous section. For example, introducing the normal boiling temperatures of the predictive compounds into the right-hand side of Eq. 10 yields: $-(0.22028 \times 336.63) + (0.8098 \times 336.79) + (0.45112 \times 420.018) - (0.040664 \times 443.75) = 394.31$ K, compared to the measured value of 394.44 K. [Measured values are taken from the DIPPR database (http://dippr.byu.edu). The development of the database is supported by the American Institute of Chemical Engineers.] We assume that the same equation can predict other properties. To validate this assumption we have predicted 31 constant properties included in the DIPPR database for all 18 target compounds that are shown in Table 5. Detailed results of the predictions for 1-octene and *p*-diethylbenzene are shown in Tables 6 and 7, respectively.

The compound 1-octene is a representative of the compounds for which immediate neighbor compounds are used in the correlation; the error in the correlation is almost negligible and this gives the opportunity to investigate the effect of the experimental error on the precision of the prediction. Table 6 shows the value provided for the property in the DIPPR database for the target compound, the reliability assigned to these values by DIPPR, the predicted property value for 1-octene, and the corresponding relative error. It should be mentioned that some of the properties shown in the table (molecular mass, van der Waals volume and area, and standard heat of formation) are also included in the molecular descriptors database, which was used for the derivation of the correlation. However, the particular values of these properties for the target compounds are not used in the predictions, which are based only on the values for the predictive compounds.

From among the 31 properties, 10 are predicted with relative error $< 0.1\%$, 8 properties are predicted with relative error $< 1\%$, 9 properties are predicted with relative error $< 5\%$, and only in four cases is the prediction error $> 5\%$. In all cases the high prediction error can be attributed to the error in the experimental data as reported by DIPPR. To show that, the method of calculation of an estimated relative error (Δr_{est}) using the error propagation formula is described in Appendix B, where it is subsequently applied for error estimation for the triple-point vapor pressure of 1-octene. The calculated relative error (Δr_{calc}) of the predicted value for this property is 86% (Table 6), whereas the estimated prediction error as calculated in Appendix B is 84%. The contribution of the correlation error is about 1%, whereas the rest of the errors arise from errors in the data used for the triple-point pressure. An important conclusion from this error analysis is that, when predicting a particular property, the selection of the predictive compounds must be based on a balance between the attainable variance of the "structure–structure" correlation and the experimental errors in the pertinent properties of the predictive compounds.

In the particular case of 1-octene the error is also relatively high (21%) for the lower flammability limit. However, because the DIPPR data reliability is categorized as "unknown" for this property, it is impossible to determine the source of the error. Taking into consideration that the error in the prediction of the other flammability-related properties is much lower, it can be concluded that the correlation equation is valid for such properties, although the accuracy of the data must be reassessed. The results shown in Table 6 clearly indicate that the proposed technique can predict gas, liquid, and solid properties with the same high accuracy.

The compound *p*-diethylbenzene needs six predictive compounds from different homologous series for its representation. The results for 29 predicted properties are shown in Table 7 (for the two excluded properties experimental data for some of the predictive compounds are not available in DIPPR). From among the 29 properties, 14 are predicted with relative error $< 1\%$, 7 properties are predicted with relative error $< 5\%$, and in eight cases the prediction error is $>5\%$. The prediction error is higher in many cases by an order of magnitude than that for 1-octene, but most of the prediction error can be attributed to the "reliability" of the DIPPR data (compare the "reliability"

Table 4. An Alternative Linear Correlation for the Molecular Descriptors of 1-Octene

Compound	Coefficient	95% Confidence Interval
1-pentene	−0.09272	0.009694
1-heptene	0.89031	0.030171
1-decene	−0.3384	0.042598
1-dodecene	0.81852	0.040894
1-hexadecene	−0.30402	0.022498
1-pentacosene	0.026296	0.003831
Variance	1.38E-008	
R^2 (linear corr. coeff.)	1	
Standard deviation (σ)	1.176E-04	

Table 5. Structure–Structure Correlation Equations for Selected Compounds*

	No.	Target Compound	Predictive Compounds							σ	R^2
1	Name	*n*-hexane	*n*-butane	*n*-heptane	*n*-octane						
	Coefficient		0.11418	1.5482	−0.66252					2.56E-04	1
2	Name	*n*-octadecane	*n*-nonadecane	*n*-tetracosane	*n*-hexatriacontane						
	Coefficient		1.2446	−0.26027	0.015013					2.06E-04	1
3	Name	2,2,3-trimethylbutane	2,4-dimethylhexane	2,2,3-trimethylpentane	2,2,4-trimethylpentane	2,4,4-trimethylhexane	2,2,3,3-tetramethylhexane				
	Coefficient		−0.22579	1.382	0.65644	−0.59213	−0.22414			2.48E-03	0.99985
4	Name	2,2,4-trimethylpentane	2,2,3-trimethylbutane	2,2,4-trimethylhexane	3,3,5-trimethylheptane	cyclopropane	cycloheptane	cyclooctane	1,2,4-trimethyl-benzene		
	Coefficient		0.16839	1.391	−0.56081	0.071159	−0.23294	0.17135	−0.0016447	2.79E-03	0.99981
5	Name	2,2,3,3-tetramethylpentane	ethane	2,2-dimethylbutane	3,3-dimethylpentane	2,2-dimethylhexane	2,2,3,3-tetramethylhexane	2,2,3,3-tetramethylbutane			
	Coefficient		−0.017128	0.26531	0.15852	−0.40686	0.8573	0.1349		1.15E-03	0.99999
6	Name	1-butene	*n*-dodecane	propylene	1-pentene	1-hexene	1-heptene				
	Coefficient		0.0038429	0.17566	2.2859	−2.2461	0.78428			8.17E-04	0.99997
7	Name	1-heptene	1-hexene	1-octene	1-nonene						
	Coefficient		0.28017	1.1605	−0.44076					1.90E-04	1
8	Name	1-octene	1-hexene	1-heptene	1-nonene	1-decene					
	Coefficient		−0.22028	0.8098	0.45112	−0.040664				1.46E-04	1
9	Name	1-decene	1-nonene	1-dodecene	1-tetradecene	1-heptadecene	1-octadecene				
	Coefficient		0.336	1.6321	−1.3926	0.7567	−0.33238			2.38E-04	1
10	Name	1,3-butadiene	*n*-heinecosane	1-octene	2,3-dimethylbutadiene	1,5-hexadiene	2,5-dimethyl-1,5-hexadiene				
	Coefficient		0.10889	−0.44401	0.99428	1.2177	−0.99311			7.46E-03	0.99844
11	Name	2-methyl-1-butene	*n*-heptane	2,2-dimethylpropane	2,2-dimethylbutane	2,2-dimethylhexane		isobutene	2,4,4-trimethyl-1-pentene		
	Coefficient		0.43613	−0.29559	0.56864	−0.70927		0.56777	0.43312	2.39E-03	0.99975
12	Name	cyclopentane	cyclobutane	cyclohexane	methylcyclohexane	propylcyclohexane		methylcyclo-pentane	ethylcyclo-pentane		
	Coefficient		−0.0026707	1.0025	−1.4936	0.48439		2.0288	−1.0194	2.03E-03	0.99985
13	Name	methylcyclohexane	ethylcyclohexane	propylcyclohexane	methylcyclopentane	ethylcyclopentane		propylcyclopentane			
	Coefficient		3.0693	−2.0631	1.1853	−3.3756		2.1855		2.26E-03	0.99984
14	Name	trans-1,3 dimethylcyclohexane	cyclopropane	cyclobutane	trans-1,4 dimethylcyclohexane	cyclohexene		1,5-cyclooctadiene			
	Coefficient		0.035345	0.028736	0.99892	−0.10506		0.053789		2.35E-03	0.99981
15	Name	cyclohexene	cyclobutane	cycloheptane	cyclooctane	cyclopentene		1,3-cyclopentadiene			
	Coefficient		−0.27462	1.1642	−0.47293	0.097089		0.4598		2.35E-03	0.99675
16	Name	*o*-xylene	*n*-undecane	*n*-dodecane	*n*-tetradecane	*p*-xylene		*o*-ethyltoluene	*p*-ethyltoluene		
	Coefficient		−0.73863	1.0929	−0.35885	0.99493		0.99772	−0.99435	2.31E-03	0.99982
17	Name	*m*-ethyltoluene	*o*-xylene	*p*-xylene	*p*-ethyltoluene	1,2,3-trimethylbenzene		1,2,4-trimethylbenzene	isobutylbenzene		
	Coefficient		−0.90359	0.94213	0.81643	0.8865		−0.83007	0.086231	1.27E-02	0.99512
18	Name	*p*-diethylbenzene	*n*-hexatriacontane	ethylbenzene	butylbenzene	tridecylbenzene		tetradecylbenzene	mesitylene		
	Coefficient		−0.0072601	−0.72864	0.93225	−1.8129		1.6086	1.0011	1.18E-03	0.99996

*Table 7 in Wakeham et al. (2002).

Table 6. Prediction of the Properties Available in the DIPPR Database for 1-Octene

No.	Property	Unit	DIPPR Data		Prediction	
			1-Octene	Reliability	1-Octene	Rel. Error (%)
1	Molecular mass	kg/kmol	112.21264		112.22	0.005
2	Critical temperature	K	567	<0.2%	566.96	0.007
3	Critical pressure	Pa	2.68E+06	<3%	2.62E+06	2.300
4	Critical volume	m^3/kmol	0.468	<5%	0.467	0.310
5	Critical compressibility factor	unitless	0.266	None	0.258	3.049
6	Melting point	K	171.45	<1%	173.71	1.316
7	Triple-pt temperature	K	171.45	<1%	173.71	1.316
8	Triple-pt pressure	Pa	2.550E-03	<25%	4.73E-03	85.672
9	Normal boiling temperature	K	394.44	<1%	394.31	0.034
10	Liq molar volume	m^3/kmol	0.157781	<1%	0.158	0.008
11	IG heat of formation	J/kmol	−8.36E+07	<3%	−8.34E+07	0.195
12	IG Gibbs energy of formation	J/kmol	1.03E+08	<5%	1.03E+08	0.158
13	IG absolute entropy	J $kmol^{-1}$ K^{-1}	4.65E+05	<1%	4.65E+05	0.030
14	Std Heat of formation	J/kmol	−1.24E+08	<1%	−1.24E+08	0.063
15	Std Gibbs energy of formation	J/kmol	9.39E+07	<1%	9.38E+07	0.067
16	Std absolute entropy	J $kmol^{-1}$ K^{-1}	3.60E+05	<1%	3.61E+05	0.178
17	Heat fusion at melt pt	J/kmol	1.53E+07	Unknown	1.62E+07	5.702
18	Std net heat of combustion	J/kmol	−4.96E+09	<0.2%	−4.96E+09	0.004
19	Acentric factor	unitless	0.392059	None	0.382	2.671
20	Radius of gyration	m	4.46E-10	<3%	4.48E-10	0.494
21	Solubility parameter	$(J/m^3)^{0.5}$	1.55E+04	<3%	15547.86	0.309
22	Dipole moment	c m	1.13E-30	Unknown	1.15E-30	1.821
23	van der Waals volume	m^3/kmol	0.0852	<1%	0.09	0.053
24	van der Waals area	m^2	1.18E+09	<3%	1.19E+09	0.520
25	Refractive index	unitless	1.4062	<0.2%	1.41	0.002
26	Flash point	K	294	Unknown	291.23	0.941
27	Lower flammability limit	vol % in air	0.8	Unknown	0.63	21.024
28	Upper flammability limit	vol % in air	6.8	Unknown	6.54	3.785
29	Lower flamm limit temp	K	280		276.82	1.134
30	Upper flamm limit temp	K	320		319.19	0.254
31	Auto ignition temp	K	503.15	Unknown	528.41	5.020

columns in Tables 6 and 7). In cases of up to a 100% error in the available property data (as in the triple-point pressure) one cannot expect to obtain a meaningful prediction.

Table 8 shows the results of the prediction of the critical temperature, the normal boiling temperature, and the liquid molar volume for the 18 target compounds. Those properties were selected because high precision, measured values are usually available for them. The results in the table include the value of 100σ and the relative errors in the prediction of the properties for the 18 target compounds. Apparently, there is a clear connection between the values of the standard error of the "structure–structure" correlation and the precision of the predicted properties. The standard deviation of this correlation can serve as a lower bound estimate for the prediction error, but rigorous error analysis should be used to obtain accurate error estimates.

Table 9 shows the average relative error in the prediction of 29 properties for the 18 compounds. The purpose of this table is to demonstrate that the method used for property prediction of 1-octene and *p*-diethylbenzene, as shown in Tables 6 and 7, can be readily extended to all of the 18 compounds. Results related to triple-point pressure were not included because all such data in DIPPR are predicted data with poor reliability. Dipole moment was not included either because data were available for only five compounds. Note also that for some properties not all 18 target compounds were included because of lack of measured data. In such cases DIPPR included predicted values of unknown reliability. However, when the use of such values resulted in outlying prediction error for the property of some target compound (the relative prediction error is greater than five times the average error of the rest of the compounds), this prediction was excluded. It can be seen that, considering the typical data reliability figures shown in Tables 6 and 7, the prediction errors are within the experimental error range. The only exceptions are the melting point and the heat of fusion at the melting point. The cause of the excessive error in prediction of those properties is currently being investigated.

New Opportunities Provided by the Structure–Structure Correlations

In contrast to the existing methods, the new method gives its user an estimation of the errors of the prediction of the target compound with unknown properties and the opportunity for developing one's own strategy and finding one's own decisions.

For instance, the user may accept some of the errors, if they suit the purpose of the future application of the predicted properties. If measured data for some properties of the compounds suggested by the structural correlation are not available, or cannot be obtained experimentally, the user may look in several directions. Because the SROV procedure generates a cascade of potential structural correlations, it is possible to seek among them structural correlations of lower precision, but with predictive compounds that have all measured data. If such correlations could not be identified, new compounds that have data, or can be easily synthesized and their properties can be measured, may be included in the database. The SCD algorithm

Table 7. Prediction of the Properties Available in the DIPPR Database for *p*-Diethylbenzene

No.	Property	Units	DIPPR Data		Prediction	
			p-Diethylbenzene	Reliability	*p*-Diethylbenzene	Rel. Error (%)
1	Molecular weight	kg/kmol	134.21816		133.76	0.34
2	Critical temperature	K	657.9	<5%	652.42	0.83
3	Critical pressure	Pa	2.80E+06	<10%	2.76E+06	1.57
4	Critical volume	m^3/kmol	0.497	<25%	0.49	0.81
5	Critical compressibility factor	unitless	0.255	None	0.26	1.06
6	Melting point	K	230.325	<3%	220.85	4.11
7	Triple-pt temperature	K	230.325	<3%	220.84	4.12
8	Triple-pt pressure	Pa	0.21291	<100%	7.40E-01	247.6
9	Normal boiling temperature	K	456.937	<3%	454.98	0.43
10	Liq molar volume	m^3/kmol	0.156441	<1%	0.1549	0.97
11	IG Heat of Formation	J/kmol	−2.20E+07	Predicted	−3.68E+07	−67.30
12	IG Gibbs energy of formation	J/kmol	1.39E+08	<3%	1.25E+08	9.61
13	IG absolute entropy	J $kmol^{-1}$ K^{-1}	4.33E+05	<3%	4.23E+05	2.41
14	Std heat of formation	J/kmol	−7.28E+07	<25%	−8.34E+07	−14.50
15	Std Gibbs energy of formation	J/kmol	1.23E+08	<25%	1.13E+08	7.98
16	Std absolute entropy	J $kmol^{-1}$ K^{-1}	3.16E+05	<5%	3.08E+05	2.60
17	Heat fusion at melt pt	J/kmol	1.06E+07	Unknown	9.79E+06	7.60
18	Std net heat of combustion	J/kmol	−5.56E+09	<3%	−5.52E+09	−0.66
19	Acentric factor	unitless	0.402791	None	0.43	6.29
20	Radius of gyration	m	4.58E-10	<3%	4.80E-10	4.83
21	Solubility parameter	$(J/m^3)^{0.5}$	1.77E+04	<10%	17248.35	2.33
22	van der Waals volume	m^3/kmol	9.11E-02	<1%	0.09	0.62
23	van der Waals area	m^2	1.15E+09	<1%	1.15E+09	0.10
24	Refractive index	unitless	1.49245	<0.2%	1.48	0.59
25	Flash point	K	329.15	<10%	330.74	0.48
26	Lower flammability limit	vol % in air	0.8	<10%	0.79	0.68
27	Upper flammability limit	vol % in air	6.1	<25%	6.78	11.21
28	Lower flamm limit temp	K	327		325.61	0.43
29	Upper flamm limit temp	K	369		369.59	0.16

can be used in this case to test whether the potential new additions are collinear with the target compound.

If there are data for all properties, but the errors for particular properties are unacceptable, the user can look for improvement of the structural correlation and/or of the property correlations. It is possible to test the potential of structural correlations generated by SROV of lower precision, which eventually may include compounds with higher reliability of the measured properties. The user may also add more descriptors and/or compounds in the database to improve the structural correlation and, thus, the property correlations.

It should be pointed out that there are commercial software programs that can calculate thousands of descriptors. It would be better to select candidate molecules for adding to the database on the basis of the existence of measured properties but, if needed, programs for molecular design may be used for generation of structures related to that of the target.

In the proposed method, the use of more informative and a

Table 8. Accuracy of the Molecular Descriptors Based Prediction for Selected Compounds

No.	Target Compound	$\sigma \times 100$	Relative Error (%) in Prediction of		
			Critical Temperature	Normal Boiling Temperature	Liq. Molar Volume
1	*n*-hexane	0.026	0.099	0.087	0.731
2	*n*-octadecane	0.021	0.036	0.115	0.469
3	2,2,3-trimethylbutane	0.248	0.755	0.827	0.173
4	2,2,4-trimethylpentane	0.279	1.728	1.669	0.962
5	2,2,3,3-tetramethylpentane	0.115	1.800	1.404	2.392
6	1-butene	0.082	0.191	0.266	2.888
7	1-heptene	0.019	0.018	0.039	0.001
8	1-octene	0.015	0.007	0.034	0.008
9	1-decene	0.024	0.192	0.035	0.029
10	1,3-butadiene	0.746	3.188	5.62	1.906
11	2-methyl-1-butene	0.239	1.283	0.559	1.700
12	cyclopentane	0.203	0.477	0.922	1.783
13	methylcyclohexane	0.226	1.661	0.705	1.608
14	trans-1,3-dimethylcyclohexane	0.235	1.309	0.872	3.290
15	cyclohexene	0.666	1.255	0.966	0.387
16	*o*-xylene	0.231	1.083	0.975	1.183
17	*m*-ethyltoluene	1.271	0.450	0.132	0.675
18	*p*-diethylbenzene	0.128	0.834	0.429	0.972

Table 9. Average Prediction Errors for the 18 Compounds for 29 Properties

No.	Property	Data Points	Average Prediction Error (%)
1	Molecular mass	18	0.50
2	Critical temperature	18	0.91
3	Critical pressure	18	3.31
4	Critical volume	18	2.32
5	Critical compressibility factor	18	3.01
6	Melting point	16	12.98
7	Triple-pt temperature	16	12.98
9	Normal boiling temperature	18	0.59
10	Liq molar volume	18	1.18
11	IG heat of formation	15	2.29
12	IG Gibbs energy of formation	16	7.19
13	IG absolute entropy	15	1.22
14	Std heat of formation	14	3.06
15	Std Gibbs energy of formation	16	5.01
16	Std absolute entropy	17	3.61
17	Heat fusion at melt pt	14	22.74
18	Std net heat of combustion	18	0.44
19	Acentric factor	16	3.64
20	Radius of gyration	18	3.23
21	Solubility parameter	17	1.52
23	van der Waals volume	18	0.60
24	van der Waals area	18	1.07
25	Refractive index	16	0.50
26	Flash point	14	3.08
27	Lower flammability limit	16	10.88
28	Upper flammability limit	15	4.29
29	Lower flamm limit temp	12	1.04
30	Upper flamm limit temp	15	1.03
31	Auto ignition temp	11	3.91

greater number of descriptors increases the chance that the required "significant common features" will be available in the database. In contrast to the group/bond contribution methods, this does not negatively influence the statistical quality of the derived correlations. The impact of the partial correlations between certain descriptors and/or the repetition of information is crucial for the typical SCF methods, but is not an issue in the present method. In fact, in the proposed method, adding more descriptors plays a role similar to that of adding experimental data in the oversized data matrices of traditional experimental design.

The method developed in this study promises a significant breakthrough in the development of QSPRs. The present study also reveals some possible shortcomings and the main tasks for the future work on improving the new method, some of which have been outlined above.

Conclusions

A new approach for predicting a wide range of properties for pure compounds has been developed. This approach involves the calculation of the molecular descriptors of a target compound with unknown properties, followed by regression of this vector of molecular descriptors vs. a database of compounds with known molecular descriptors and measured properties. The linear regression model obtained for the molecular descriptors, in terms of predictive compounds and their coefficients, is then used for prediction of the corresponding properties of the target compound. The standard deviation of the correlation, together with the known precision of the property data of the predictive compounds, can be used to provide a good estimate on the precision of the predicted property values.

The main new advantages of the proposed method may be summarized as follows:

(1) One structural correlation is used to predict all the properties.

(2) No experimental measurements for the target compound are needed.

(3) The error in the prediction of the properties of compounds for which no experimental data are available can be estimated.

(4) If the database does not contain some of the compounds needed for property prediction, or the accuracy for a particular application is not acceptable, the standard deviation of the structural correlation and the accuracy of the property data for the predictive compounds will provide an adequate warning. The user is given several opportunities to find alternative solutions and the means to estimate the adequacy of these solutions.

(5) The proposed method can be used for checking the consistency of measured data and data predicted by other methods.

More work is required for perfecting and extending the applicability of the technique to some temperature- and pressure-dependent properties, and to additional groups of organic and inorganic compounds as outlined in this study. A simplified, more flexible version of the algorithm should be developed to enable casual users with no access to large databases of molecular descriptors to estimate unknown properties.

Literature Cited

Bondi, A., "Van der Waals Volumes and Radii," *J. Phys. Chem.*, **3**, 441 (1964).

Brauner, N., and M. Shacham, "Considering Precision of Data in Reduction of Dimensionality and PCA," *Comput. Chem. Eng.*, **24**(12), 2603 (2000).

Cholakov, G. St., W. A. Wakeham, and R. P. Stateva, "Estimation of Normal Boiling Temperature of Industrially Important Hydrocarbons from Descriptors of Molecular Structure," *Fluid Phase Equilib.*, **163**, 21 (1999).

Drapper, N. R., and H. Smith, *Applied Regression Analysis*, 3rd Edition, Wiley, New York, p. 138 (1998).

Horwath, A. L., *Molecular Design*, Elsevier, Amsterdam (1992).

Labanowski, J. K., I. Motoc, and R. A. Damkoehler, "The Physical Meaning of Topological Indexes," *Comput Chem.*, **15**, 47 (1991).

Lydersen, A. L., "Estimation of Critical Properties of Organic Compounds," Univ. Wisconsin Coll. Eng., Eng. Exp. Stn. Rep. 3, Madison, WI, April (1955).

Marano, J. J., and G. D. Holder, "General Equation for Correlating the Thermo-Physical Properties of *n*-Paraffins, *n*-Olefins, and Other Homologous Series. 1. Formalism for Developing Asymptotic Behavior Correlations," *Ind. Eng. Chem. Res.*, **36**, 1887 (1997a).

Marano, J. J., and G. D. Holder, "General Equation for Correlating the Thermo-Physical Properties of *n*-Paraffins, *n*-Olefins, and Other Homologous Series. 2. Asymptotic Behavior Correlations for PVT Properties," *Ind. Eng. Chem. Res.*, **36**, 1895 (1997b).

Poling, B. E., J. M. Prausnitz, and J. P. O'Connel, *Properties of Gases and Liquids*, 5th Edition, McGraw-Hill, New York (2001).

Reid, R. C., J. M. Prausnitz, and T. K. Sherwood, *Properties of Gases and Liquids*, 4th Edition, McGraw-Hill, New York (1987).

Shacham, M., and N. Brauner, "The SROV Program for Data Analysis and Regression Model Identification," *Comput. Chem. Eng.*, **27**(5), 701 (2003).

Somayajulu, G. R., "Estimation Procedures for Critical Constants," *J. Chem. Eng. Data*, **34**, 106 (1989).

Stein, S. E., and R. L. Brown, "Estimation of Normal Boiling Points from Group Contributions," *J. Chem. Inf. Comput. Sci.*, **34**, 581 (1994).

Wakeham, W. A., G. St. Cholakov, and R. P. Stateva, "Liquid Density and Critical Properties of Hydrocarbons Estimated from Molecular Structure," *J. Chem. Eng. Data*, **47**(3), 559 (2002).

Appendix A: Summary of the Used Descriptors of Molecular Structure and Their Assumed Precision

Descriptors and Authors of Scheme	Assumed Precision*
A. Group and Bond Contribution Descriptors	
I. Joback (Poling et al., 2001), Nos. 1–15	
Total number of carbon atoms	integer
Number of CH_3 groups	integer
Number of carbon atoms in aliphatic CH_2 groups	integer
Number of carbon atoms in aliphatic CH groups	integer
Number of carbon atoms in aliphatic C groups	integer
Total number of carbon atoms in aliphatic double bonds	integer
Number of carbon atoms in aliphatic $CH_2{=}CH_2$ groups	integer
Number of carbon atoms in aliphatic CH=CH groups	integer
Number of carbon atoms in aliphatic C=C groups	integer
Number of carbon atoms in CH_2 groups in rings	integer
Number of carbon atoms in CH groups in rings	integer
Number of carbon atoms in C groups in rings	integer
Total number of carbon atoms in double bonds in rings	integer
Number of carbon atoms in CH=CH groups in rings	integer
Number of carbon atoms in C=C groups in rings	integer
II. Ambrose (Reid et al., 1987) and Somayajulu (Somayajulu, 1989), Nos. 16–35	
Total number of carbon atoms in alkyl groups	integer
Number of carbon atoms in CH in alkyl groups	integer
Number of carbon atoms in C in alkyl groups	integer
Delta Plat Number	integer
Total number of non arene double bonds	integer
Number of arene double bonds	integer
Number of non aliphatic CH_2 groups in rings	integer
Number of CH_2 groups in fused rings	integer
Number of benzene rings	integer
Number of arene rings with single substitutions	integer
Number of arene rings with two substitutions	integer
Number of *ortho* pairs in arene rings	integer
Number of C_4H_4 groups, fused as in naphthalene	integer
Number of quadruples of carbon atoms in gauche configuration	integer
Number of carbon atoms in >C= groups in rings	integer
Number of carbon atoms in >C= groups in fused rings	integer
Number of *trans* alkene configurations	integer
Number of phenyl substitutions	integer
Number of *ortho* pairs in rings	integer
Number of *meta* pairs in rings	integer
III. Stein and Brown (Stein and Brown, 1994), Nos. 36–38	
Number of CHaa configurations (*aa* denotes an arene bond)	integer
Number of Caa configurations (*aa* denotes an arene bond)	integer
Number of Caaa configurations (*aaa* denotes an arene bond)	integer
IV. Bondi (Bondi, 1964), Nos. 39–55	
Number of cyclohexyl or cyclopentyl rings, free and *trans* condensed	integer
Number of CH_3 groups in normal alkanes	integer
Number of CH_2 groups in normal alkanes	integer
Number of CH groups in normal alkanes	integer
Number of C groups in normal alkanes	integer
Number of >C=C< groups in alkenes	integer
Number of $=CH_2$ groups in alkenes	integer
Number of >CH– groups in alkenes	integer
Number of single bonds between conjugated double bonds	integer
Number of $>C{=}CH_2$– groups in alkenes	integer
Number of >C=CH– groups in alkenes	integer
Number of >C= condensed groups in arenas	integer
Number of rings in *cis*-condensed cycloalkanes	integer
Number of =CH– groups in arene compounds	integer
Number of =C< alkyl groups in arene compounds	integer
van der Waals Volume, calculated by Bondi's method	X.xx
van der Waals Surface, calculated by Bondi's method	X.xx
B. Topological indices and molecular mass (Labanowski et al., 1991; Horwath et al., Cholakov et al., 1999), Nos. 56–84*	
Path connectivity index of order zero	X.xxx
Path connectivity index of order one	X.xxx
Path connectivity index of order two	X.xxx
Path connectivity index of order three	X.xxx
Valency path connectivity index of order zero	X.xxx
Valency path connectivity index of order one	X.xxx
Valency path connectivity index of order two	X.xxx
Valency path connectivity index of order three	X.xxx
Path connectivity index of order zero, with half of atom masses	X.xxx
Path connectivity index of order one, with half of atom masses	X.xxx
Path connectivity index of order two, with half of atom masses	X.xxx
Path connectivity index of order three, with half of atom masses	X.xxx
Wiener's index	X.xx
Information content on the realized distances	X.0
Mean information content on the realized distances	X.xx
Information content on the distribution of distances	X.xx
Mean information content on the distribution of distances	X.xxx
Average distance sum index (Balaban's index)	X.xxx
Cyclomatic number	integer
Gravitation index	X.xx
C. Energy descriptors from simulated molecular mechanics *** Nos. 85–99	
Total energy of minimized molecular model	X.xxx
Stretch energy of minimized molecular model	X.xxx
Bond energy of minimized molecular model	X.xxx
Stretch-bend energy of minimized molecular model	X.xxx
Torsion energy of minimized molecular model	X.xxx
Van der Waals energy of minimized molecular model	X.xxx
Energy of "dipole–charge" interaction of minimized molecular model	X.xxx
Electrostatic dipole moment of minimized molecular model	X.xxx
Standard heat of formation of minimized molecular model	X.xx
Strain energy of minimized molecular model	X.x
van der Waals volume	X.0
Molar volume	X.0
Total van der Waals surface	X.xxx
Saturated van der Waals surface	X.xxx
Unsaturated van der Waals surface	X.xxx

*The precision of the values of the topological indices and the descriptors from molecular mechanics was assumed on the basis of the precision of published values, and the values provided by the used molecular mechanics software, respectively.
**Calculated with integer and real bond lengths (Cholakov et al., 1999).
***PCMODEL, 5th ed., Serena Software, Bloomington, IN (1992).

Appendix B: Estimation of the Errors in the Predicted Properties

Table B1. Estimation of the Prediction Error for the Triple-Point Pressure of 1-Octene

	β	$y(x)$	Reliability (%)	$\Delta y(\Delta x)$	Σ
1-octene		0.00255	<25	0.000638	
1-hexene	−0.22028	0.000592	<25	0.000148	1.28
1-heptene	0.8098	0.00126	<25	0.000315	10.00
1-nonene	0.45112	0.0101	<10	0.00101	17.87
1-decene	−0.04066	0.0175	<25	0.004375	6.98
					36.13

If measured property values of the target compound are available, the relative error in the prediction (Δr_{calc}) can be calculated from the equation

$$\Delta r_{calc} = \frac{\tilde{y} - y}{y} \tag{B1}$$

where $\tilde{y}$ is the predicted value and y is the measured value of the property of the target compound. The relative error can also be estimated. The estimated relative error (Δr_{est}) is the sum of two components: the error attributed to the correlation (Δr_c) and the error propagated by the measurement errors in the data (Δr_m); thus $\Delta r_{est} = \Delta r_c + \Delta r_m$. Δr_c can be estimated from the standard deviation at the calculated value, which is given by (for example, Draper and Smith, 1998)

$$\Delta r_c = \left|\frac{\Delta\tilde{y}}{y}\right| = \frac{\sqrt{\mathbf{X}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_0\sigma^2}}{|y|} \tag{B2}$$

where the vector $\mathbf{X}_0$ is the vector of the measured property values of the predictive compounds, $\mathbf{X}^T\mathbf{X}$ is the normal matrix of the correlation (composed of the molecular descriptors), and σ^2 is the variance of the correlation.

Using the error propagation formula for the calculation of Δr_m, when a linear correlation (Eq. 4) is used for the prediction, yields the following equation:

$$\Delta r_m = \frac{\sum_{i=1}^{m} |\beta_i \Delta x_i|}{|y|} + \left|\frac{\tilde{y}}{y}\right|\left|\frac{\Delta y}{y}\right| \tag{B3}$$

where x_i is the value of the measured property of predictive compound i and Δx_i is the associated (absolute) error estimate.

Estimation of the prediction error for the triple-point pressure of 1-octene follows. The inverted normal matrix for this correlation, as obtained by the SROV program, is

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 2480 & -4196.7 & 2614.2 & -886.13 \\ -4196.7 & 7827.8 & -6574.9 & 2933.7 \\ 2614.2 & -6574.9 & 9273.3 & -5322.8 \\ -886.13 & 2933.7 & -5322.8 & 3284.7 \end{bmatrix} \tag{B4}$$

For 1-octene the standard deviation of the correlation is $\sigma = 0.000146$ (see Table 4), the components of the $\mathbf{X}_0$ vector are the last four values in column "$y(x)$" of Table B1, and the value of y is the first value in the same column. Introducing those numerical values into Eq. B2 yields $\Delta r_c = 1.31\%$.

In Table B1 the first term of Eq. B3 is calculated. The measured data reliability, as reported by the DIPPR, is used to estimate the errors in the measured data: Δx_i and Δy. Based on these values, the expression $\Sigma = \Sigma\,|\,\beta_i\Delta x_i\,|/|\,y\,|$ is calculated. Introducing the numerical values into Eq. B3 yields $\Delta r_m = 36.13 + (0.004735/0.00255) \times 25 = 82.55$.

Thus the total prediction error is $\Delta r_{est} = 82.55 + 1.31 \cong 84\%$.

Manuscript received Oct. 7, 2003, and revision received Jan. 24, 2004.