A UNIFIED CORRELATION FOR PREDICTION OF PURE COMPONENT PROPERTIES BASED ON SIMILARITY OF MOLECULAR DESCRIPTORS OF VARIOUS COMPOUNDS

Mordechai Shacham[1*], Neima Brauner[2], George St. Cholakov[3], and Roumiana P. Stateva[4]

[1]Dept. Chem. Eng., Ben-Gurion University, Beer-Sheva, Israel, e-mail shacham@bgumail.bgu.ac.il
[2]School of Engineering, Tel-Aviv University, Tel-Aviv, Israel
[3]University of Chemical Technology and Metallurgy, Sofia, Bulgaria
[4]Inst. of Chem. Eng., Bulgarian Academy of Sciences, Sofia, Bulgaria

Abstract

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 versus a database of compounds with known descriptors and measured properties. The regression model, obtained for the descriptors of a target compound in terms of those of predictive compounds and their weighting factors, is then used for prediction of properties of the target compound. The precision of the prediction can be estimated based on the standard deviation of the correlation and the known precision of the property data of the predictive compounds. The use of the proposed technique is demonstrated by using regression models of various precision and complexity to predict properties of *n*-tetradecane.

Keywords

Prediction of properties, Molecular descriptors, Collinearity, Linear regression model

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. The number of the compounds used at present by the industry, or being of its immediate interest, is estimated at around 100 000, while 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). 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. In cases where experimental data for the needed properties are not available, they have to be estimated by using suitable quantitative structure – property relationships (QSPRs). Correlations of acceptable accuracy can be derived between measured values of pure component constants, such as the normal boiling temperature (T_b), liquid density (d_4^{20}), critical properties (T_c, P_c, V_c), etc., 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 (Constantinou and Gani, 1994, Poling et al., 2001). An alternative to the group contribution methods is to find, out of a huge database, a combination of molecular descriptors, which defines the most "significant common features" (SCF) of the described molecules (Wakeham et al., 2002). The different molecular descriptors in the database may be

* To whom all correspondence should be addressed

computed by simulated molecular mechanics, quantum chemical methods, the topology of the molecules, etc. The descriptors that are significant for the prediction of a particular constant and their weighting factors are 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 %. Moreover, differences in the values of unknown properties predicted by different methods may amount to more than several hundred percents. Unfortunately, for a target compound of unmeasured pure component constants, it is impossible to assess the prediction accuracy. With no feed-back on the prediction error, it is impossible to choose among the methods proposed by different authors, and to advocate the best.

As a fresh first step towards overcoming the limitations of the existing prediction techniques, we are advocating hereunder a novel technique. It is based on a presumption for a 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. It contains 99 molecular descriptors for 260 compounds. The molecular descriptors include molecular mass, carbon atom descriptors and descriptors obtained from simulated molecular mechanics such as total energy, stretch energy and standard heat of formation. Using the SROV algorithm of Brauner and Shacham (2003), a linear *structure-structure* correlation of the molecular descriptors of a *target compound* versus the molecular descriptors of several *predictive compounds* is developed. In the prediction stage the same correlation is used as property-property correlation to predict the properties of the target compound using the measured properties of the predictive compounds. The constant properties that available in the DIPPR database[1] are used in the prediction stage. In the following, the derivation of the structure-structure correlation and the use of the proposed technique will be briefly reviewed.

Derivation of the Structure-Structure Correlation

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

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

where $\mathbf{y}_c$ is an N vector of known properties, $\mathbf{y}_p$ is a K vector of unknown properties. Both the N vector $\mathbf{x}_{cj}$ and the K vector $\mathbf{x}_{pj}$ contain known properties. Typically, the sub-vectors $\mathbf{y}_c$ and $\mathbf{x}_{cj}$ contain properties, which are directly related to the molecular structure and can be calculated with high accuracy (molecular descriptors), while the sub-vectors $\mathbf{y}_p$ and $\mathbf{x}_{pj}$ contain measured properties with various levels of experimental error. We wish to model the structure-structure relationship between $\mathbf{y}_c$ and the independent variables $\mathbf{x}_{c1}$, $\mathbf{x}_{c2}$, … $\mathbf{x}_{cm}$ by a linear regression model, with the general form:

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

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

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

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

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

The SROV algorithm of Shacham and Brauner (2003) is used for selection of the predictive compounds that should be included in the structure-structure regression model and for calculation of the model parameters. The SROV algorithm solves Eq. (1) using QR decomposition, by decomposing $\mathbf{X}_c$ into the product of a matrix **Q** (of orthogonal columns) and an upper triangular matrix **R**. The basic variables (those included in the regression model) are identified by applying the Gram-Schmidt orthogonalization technique to the whole set of variables in a stepwise fashion. At each step, an independent variable, $\mathbf{x}_p$ is selected to enter 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 $YXj = \mathbf{y}^T \mathbf{x}_j$, where **y** and $\mathbf{x}_j$ are centered and normalized to a unit length. Therefore, the value of $|YXj|$ is in the range [0,1]. In a case of a perfect correlation between **y** and $\mathbf{x}_j$, $|YXj| = 1$, while if the two vectors are orthogonal, $YXj = 0$. The YXj is often denoted as the partial correlation coefficient.

Upon the selection of $\mathbf{x}_p$ at step k, the **Q** and **R** matrices are updated. The update is carried out for all the columns associated with non-basic variables, whereby the columns of the **Q** matrix contain the updated subset of non-basic variables (not yet included in the regression model), which are orthogonal to the subset of the basic variables. At the same time, the parameter value associated with $\mathbf{x}_p$ is calculated and the **y** vector is updated to obtain

[1] The development of the DIPPR database (http://dippr.byu.edu) is supported by the AIChE organization

the unpredicted residuals, which are orthogonal to the basic variables subset. The signal-to-noise ratio in the correlation is used as a criterion for stopping addition of new variables to the model. Only those variables, which are associated with signal-to-noise ratio greater than one, can be selected to the model.

Upon obtaining a regression model, the SROV algorithm proceeds to additional phases, where 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 that of the lowest standard deviation and all its parameters are significantly different from zero.

Prediction of the Properties of *n*-tetradecane

The SROV program was used to identify a linear relationship between the molecular descriptors of *n*-tetradecane (the target compound) and the molecular descriptors of the rest of the compounds in the database. To carry out this study, the 99 molecular descriptors in the data-base were normalized by dividing each descriptor by its maximal absolute value over the 260 compounds in the data-base.

At the first step, SROV identified *n*-pentadecane as having the highest correlation with the target compound *n*-tetradecane (*YXj* = 0.99967). The value of *YXj* is slightly lower for n-tridecane (*YXj* = 0.99964) and *n*-dodecane (*YXj* = 0.99842). The high correlation (collinearity) that exist between the normalized molecular descriptors of *n*-tetradecane and those of *n*-pentadecane is further demonstrated in Fig. 1, where the relation between the normalized molecular descriptors of the two compounds is depicted. Evidently, the descriptors of these two compound align on a straight line with a zero intercept, a slope of 0.9578 and a linear correlation coefficient of: R^2 = 0.9991. The standard deviation of this correlation as calculated by SROV is σ = 0.00486. Thus, the first structure-structure correlation identified for *n*-tetradecane is $\mathbf{y}_c = 0.9578\ \mathbf{x}_{c1}$, where $\mathbf{x}_{c1}$ is the vector of normalized molecular descriptors of *n*-pentadecane.

After selecting *n*-pentadecane to the model, *n*-dodecane is identified as having highest correlation with the target compound, with *YXj* =0.9981. Adding *n*-dodecane to the model yields the correlation $\mathbf{y}_c = 0.65806\ \mathbf{x}_{c1} + 0.34426\ \mathbf{x}_{c2}$, where $\mathbf{x}_{c2}$ is the vector of normalized molecular descriptors of *n*-dodecane. The standard deviation of this correlation is σ = 0.0003, thus smaller by more than an order of magnitude than that obtained with one predictive compound.

Upon adding *n*-dodecane to the model, the *YXj* values drop considerably. Propane is identified as having highest correlation with the target compound, with *YXj* = -0.47654. After adding propane to the model, SROV continues to generate structure-structure models with increasing number of predictive compounds and decreasing standard deviation values. The model of the minimal standard deviation ($\sigma = 8.18\times10^{-5}$) that is generated contains six predictive compounds. The details of three structure-structure models are shown in Table 1. The models shown include model No.1 with one predictive compound, model No. 2 with two predictive compounds and model No. 3, which is the minimal standard deviation model, with six predictive compounds.

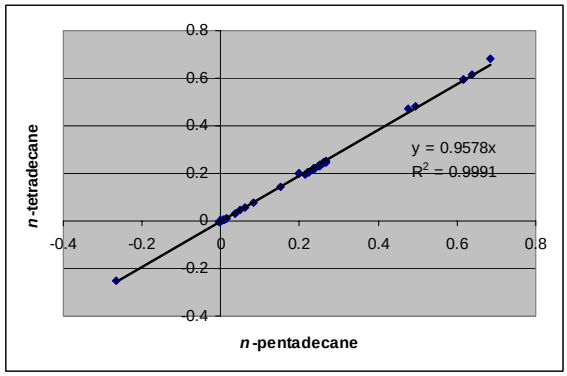

Figure 1. Plot of normalized molecular descriptors of n-tetradecane versus those of n-pentadecane

Table 1. Structure-structure correlations for the target compound n-tetradecane

	Coefficients of Models No.		
Compound	1	2	3
n-dodecane	-	0.3443	-
n-tridecane	-	-	0.50976
n-pentadecane	0.95779	0.65806	0.45603
n-heneicosane	-	-	0.04587
n-docosane	-	-	0.57114
n-tricosane	-	-	-0.9856
n-tetracosane	-	-	0.4031
Standard dev.	4.86E-03	3.01E-04	8.18E-05

The structure-structure models can be used for prediction of properties of *n*-tetradecane by introducing the coefficients and the vector of measured properties of the predictive compounds into Eq. (2). We have applied the correlations obtained to predict 29 measured properties, which are available for most of the predictive compounds and for the target compound in the DIPPR database. Six of those properties are listed in Table 2. The properties include solid (such as melting point) liquid (such as normal boiling point) and gas phase (such us critical temperature) properties. Predicted values for the target compound were calculated using the three structure-structure correlations shown in Table 1. The corresponding relative errors in the prediction are also shown in Table 2. In analyzing the results of the predictions for the various properties the precision of the experimental data should also be considered. In the DIPPR database, the "reliability" of the measured value is given as an indicator for the upper limit on the experimental error. In general, the properties can be classified into "high precision" (for which high precision data is available for all the predictive compounds) "medium" and "low" precision. Typical examples for high precision properties are the normal boiling point, liquid molar volume and

critical temperature (reliability in DIPPR for *n*-pentadecane is <1%), medium precision properties include the critical pressure and volume (reliability for *n*-dodecane is <10%) and a typical low precision property is the triple point pressure (reliability for *n*-tricosane is <50%).

Model No. 1 of one predictive compound cannot predict the properties within the experimental error level. The prediction error is greater than 1% even for the "high precision" properties (see Table 2). Model No. 2 of two predictive compounds predicts all but one property (heat of fusion at melting point) within the experimental error level, in most cases the error even lower. Low prediction errors compared to the reliability assigned by DIPPR may indicate that the precision of the experimental data is actually higher than the assigned reliability. Model No. 3 of six predictive compounds yields in many cases more accurate prediction than model No. 2, but there are almost the same number of properties for which the prediction is less accurate. This shows that the higher precision of the structure-structure correlation may not necessarily improve the property prediction as a result of larger experimental errors in the predictive compounds that added to the model.

Conclusions

The new approach for predicting a wide range of properties for pure compounds has been demonstrated. In addition to the example presented here, the new technique has been tested for predicting properties of many hydrocarbons of different homological series, in particular the compounds included in Table 7 of Cholakov *et al*. (1999) and in Table 7 of Wakeham *et al*. (2002). The results of those tests (which could not be included here due to space limitations) confirm the following conclusions:

- The proposed technique predicts most of the properties for the compounds tested within experimental error level with structure-structure correlations containing between two to eight predictive compounds. It should be emphasized that one structure-structure correlation is used to predict all the properties of a target compound.
- The structure-structure correlation can be used to check consistency of the experimental data available for the target and predictive compounds and to verify the accuracy of the error bounds specified for those data.
- Several alternative structure-structure correlations can be generated for the same target compound enabling the use of different correlations for different properties according to the experimental data available for the predictive compounds.
- The error in prediction of the properties of a target compound for which no experimental data is available can be estimated using the standard deviation of the structure-structure correlation and the error level estimates of the predictive compounds.

More work is required for investigating the source of unusually large errors associated with some properties, which apparently cannot be explained based on the experimental errors alone. Further research will be carried out to extend the applicability of the method to some temperature and pressure dependent properties and to additional groups of organic and inorganic compounds.

References

Cholakov, G. St, Wakeham, W. A, and Stateva, R. P.(1999), "Estimation of Normal Boiling Temperature of Industrially Important Hydrocarbons from Descriptors of Molecular Structure", *Fluid. Phase. Equlib*. **163**, 21-42 .

Constantinou L, Gani R. (1994), New Group-Contribution Method For Estimating Properties Of Pure Compounds, *AIChE J* **40**(10): 1697-1710.

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

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

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

Shacham, M. and Brauner, N. (2003). The SROV Program for Data Analysis and Regression Model Identification, *Comp. Chem. Engng*. **27**(5), 701-714 .

Wakeham, W. A, Cholakov, G. St, and Stateva, R. P.(2002). Liquid Density and Critical Properties of Hydrocarbons Estimated from Molecular Structure, J. Chem. Eng. Data, 47(3), 559-570 .

.

Table 2. Prediction of properties of n-tetradecane with various regression models

			Prediction Error (%) with model No.		
Property	Units	Value	1	2	3
Critical Temperature	K	693	2.15	0.08	0.19
Critical Pressure	Pa	1.57E+06	9.71	1.94	0.47
Critical Volume	m^3/kmol	0.83	2.59	0.18	0.18
Normal Boiling Point	K	526.727	1.11	0.07	0.03
Liq Molar Volume	m^3/kmol	0.261271	1.83	0.09	0.01
Melting Point	K	279.01	2.83	0.72	1.23