Estimation of Properties of Homologous Series with Targeted Quantitative Structure−Property Relationships

Georgi St. Cholakov,[†] Roumiana P. Stateva,[‡] Neima Brauner,[§] and Mordechai Shacham*[,||]

Department of Organic Synthesis and Fuels, University of Chemical Technology and Metallurgy, Sofia 1756, Bulgaria, Institute of Chemical Engineering, Bulgarian Academy of Sciences, Sofia 1113, Bulgaria, School of Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel, and Department of Chemical Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

The ability of the targeted quantitative structure−property relationships (TQSPR) method to predict properties for groups of congeneric compounds was tested with T_c and p_c data for five homologous series: *n*-alkanes, 1-alkenes, 1-alkanols, *n*-alkylbenzenes, and *n*-alkanoic acids. Training sets were identified from a database of 326 hydrocarbon and oxygen compounds with different structures, described with 1664 descriptors, or from the respective series only. It has been established that the TQSPR method can identify descriptors collinear with the property studied and develop linear equations for the series from measured data. In most cases, the respective collinear descriptors could be identified with the controls imbedded in the TQSPR program. Comparison with presently available methods shows that TQSPR achieves deviations from measured data in most cases within the average experimental uncertainties, like the best ABC methods, but it needs smaller amounts of measured data and provides higher statistical confidence in long-range prediction. The method has been tested with only five homologous series, but the existence of descriptors collinear with properties found in the present work is relevant to all homologous series. When applied to simple molecules, TQSPR can also provide insight into the way compounds are selected by structural similarity and outline eventual inefficiencies in this selection.

Introduction

Prediction of properties of pure compounds by quantitative structure−property relationships (QSPRs) is a valuable tool in chemistry and chemical engineering, environmental engineering, and environmental impact assessment, hazard, and operability analysis, etc.[1] Probably one of the main reasons for its relatively wide application is the success in drug design of the quantitative structure−activity relationships (QSARs), which employ a similar methodology. Other reasons are the lack or high price of experimental data, the impossibility to determine experimentally property values of some compounds, the increasing use of pure component data for prediction of properties of their mixtures, the possibility for molecular design, and upgrading to simulation of properties of materials at the macro scale, etc.[2]

Except for those obtained by "ab initio" quantum−chemical methods, which require huge computational time, QSPRs are typically empirical correlations derived from a limited amount of available structural information and experimental property data. Although sometimes implied otherwise, independent studies[3−6] advise that QSPR predictions, especially when extrapolating outside of the available data, can have significant errors. That is why it is recommended to define a "model applicability domain",[7] outlined in the space of the selected descriptors. This follows from the fact that no matter how robust a QSPR might be it cannot be expected to predict for all possible compounds because it is derived from a limited number of compounds with particular chemical structures. Moreover, even within the applicability domain, the uncertainties of the predictions depend on the presentation of the structure of the given molecule in the databases used; i.e., if the chemical structure of a compound is well represented in the training set, the prediction is expected to be much more accurate than when its structure is sparsely represented.[8] Basak et al.[9] also point out that QSPR models work best for "congeneric" molecules, i.e., belonging to a narrow class defined by structural analogy or similarity of action for a particular application, and suggest tailoring of the training sets by structural similarity. Therefore, any QSPR model has an applicability domain, defined by the compounds employed in its development, and the training set should contain compounds structurally related between themselves and representative for the rest of the compounds in the model application domain.

The targeted QSPR (TQSPR) method (described in detail by Brauner et al.[10]) is designed to answer the above requirements. It defines structural similarity between *potential predictive compounds* and *a target compound* as measured by the partial correlation coefficient between the vector of the molecular descriptors of the target compound and that of a potential predictive compound.[11] Absolute values of the partial correlation close to one indicate high correlation between the vectors of the target and a predictive compound, and thus a high level of similarity between the molecular structures of the target compound and the predictive compound. The number of the compounds structurally similar to the target, which form the applicability domain of the model, can be selected by choosing an acceptable lowest value of the correlation coefficients of the similarity group. The *training set* from which the model is developed usually includes a selected number of compounds

* Corresponding author. E-mail: shacham@bgu.ac.il. Fax: +972-8-64-72916.
[†] University of Chemical Technology and Metallurgy.
[‡] Bulgarian Academy of Sciences.
[§] Tel-Aviv University.
[||] Ben-Gurion University of the Negev.

10.1021/je800272x CCC: $40.75
Published on Web 09/30/2008

with the highest values of the correlation coefficient, but the training set compounds may be also chosen, applying other criteria. The *prediction (validation) set* used for *validation* of the developed model may include all applicability domain compounds, except those in the training set, or a selected number of them. *Cross-validation* can be performed in the usual manner, by exchanging compounds between the training and prediction sets.[12]

Methods designed for prediction of properties of homologous series have a specific importance in the use of structure−property relationships for property estimation. Their predictions are important for high molecular mass compounds like polymers,[13] synthetic fuels, and lubricants, etc. However, because the structural variations within a homologous series are eminimal, and thus the structural relations between the respective members of the series are most comprehensive, predictions for homologous series provide also a good opportunity for understanding how QSPRs work in general, in terms of selection of compounds for the model, extrapolation, etc.

Properties of homologous series are usually predicted with asymptotic behavior correlations (ABCs), employing the well-known relationship between properties and number of repeating units in the molecules of their members, based on the lattice-fluid theory.[14] The parameters of ABC models are typically obtained by regressing most of the available experimental data. The key parameter value of the property at an infinite number of repeating units, however, is obtained by extrapolation, and because of that, there are significant differences between values, predicted by different authors for the same homologous series.[4,14]

The TQSPR method up to now has been developed as a tool for direct prediction of a property of a single compound, the target. The aim of the present work is to investigate the capabilities of the method to predict properties of groups of compounds, i.e., members of homologous series. In our previous work,[5] we have shown that it is possible to derive a linear QSPR for a property of a homologous series if a descriptor collinear with the property can be identified. In this work, we study the options for identification of collinear descriptors with TQSPR and provide linear relationships for critical properties of five widely used homologous series.

Methodology

Typical QSPR Technique and the TQSPR Method. For a given property, a QSPR can be mathematically represented by the following equation

$$y_t = f(\mathbf{x}_s, \mathbf{x}_p, \boldsymbol{\beta}) \tag{1}$$

where y_t is the property (e.g., boiling temperature, melting temperature, toxicity, etc.) to be predicted; $\mathbf{x}_s$ is a vector of descriptors selected from a huge databank, which represent numerically the molecular structures of the compounds in the used database (including also the target compound); $\mathbf{x}_p$ is a vector of other known properties, if the QSPR uses more widely available physical property data (such as normal boiling temperatures) to predict the y_t property (e.g., critical temperature); and $\boldsymbol{\beta}$ is the vector of the QSPR parameters.

To derive the QSPR, the available data (descriptors and experimental property values) are divided into a *training set* and a *validation set*. Model validation is typically carried out using only one validation set. *Cross-validation techniques* use alternatively defined training and validation sets. Multiple linear or nonlinear regression, partial least-squares techniques, etc. are applied to the training set, in order to select the "significant common features"[15] molecular descriptors and the property values to be included in the right-hand side of eq 1 and to calculate the model parameter values. A recent review of the traditional QSPR technique has been published by Godavarthy et al.[16] The *targeted QSPR (TQSPR) technique* is described hereunder, only in principle. A detailed description is given elsewhere.[10]

The first stage of the method differs significantly from the typical QSPR methodology. It involves identification of a *similarity group* (typically of around 50 compounds) structurally related to the compound for which properties have to be predicted (the *target* compound). For identification of the similarity group, a much wider database of molecular descriptors, x_{ij}, where i is the number of the compound and j is the number of the descriptor/property, is used.

The similarity between potential predictive compounds and the target compound is measured by the partial correlation coefficient, r_{ti}, between the vector of the molecular descriptors of the target compound, $\mathbf{x}_t$, and that of a potential predictive compound $\mathbf{x}_i$. The partial correlation coefficient in this work is defined as $r_{ti} = \bar{\mathbf{x}}_t \bar{\mathbf{x}}_i^T$, where $\bar{\mathbf{x}}_t$ and $\bar{\mathbf{x}}_i$ are row vectors, centered (by subtracting the mean) and normalized to a unit length (after dividing by the Euclidean norm of the vector). Different methods for adding the predictive compounds to the similarity group, all related to cluster algorithms, and different methods for scaling the descriptors were compared.[17] It was found that while the effects of the various similarity measures and scaling methods on the average accuracy of prediction of the data measured were of minor importance they somewhat changed the particular compounds selected to the training sets.

The *training set* is established by selecting the first n compounds, with the highest $|r_{ti}|$ values, for which experimental values of the desired property are available. To represent the level of structural relationship between the similarity group and the target compound with a single number, we used the geometric average correlation coefficient (GACC) for the training set, defined by Shacham et al.[8]

$$\mathrm{GACC} = \sqrt{|r_{t1} r_{tn}|} \tag{2}$$

For the TQSPR model development for a particular property of the target compound, a linear structure−property relation is assumed of the form

$$\mathbf{y} = \beta_0 + \beta_1 \boldsymbol{\zeta}_1 + \beta_2 \boldsymbol{\zeta}_2 \ldots \beta_m \boldsymbol{\zeta}_m + \boldsymbol{\varepsilon} \tag{3}$$

where $\mathbf{y}$ is an n vector of the respective property (known, measured) values; $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2 \ldots$ $\boldsymbol{\zeta}_m$ are n vectors of m predictive molecular descriptors (to be identified via a stepwise regression algorithm); β_0, β_1, $\beta_2 \ldots$ β_m are the corresponding model parameters to be estimated; and $\boldsymbol{\varepsilon}$ is an n vector of stochastic terms (due to measurement uncertainties). Note that in the above TQSPR model, only descriptors (and not property values) are included on the right hand side of eq 3.

The second stage of the TQSPR method is similar to the typical QSPR technique. The stepwise regression program SROV[18] is used for the selection of the independent variables. In each step, it includes in the model one molecular descriptor that reduces the prediction error most strongly. The descriptors are selected to the model in a stepwise manner 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 $\rho_{yj} = \bar{\mathbf{y}} \bar{\boldsymbol{\zeta}}_j^T$, where $\bar{\mathbf{y}}$ and $\bar{\boldsymbol{\zeta}}_j$ are row vectors, centered (by subtracting the mean) and normalized to a unit length. Values

close to one indicate high correlation between molecular descriptor and the property.

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 that overfitting is avoided. Additionally, the SROV program provides a procedure for rotation of descriptors, so that eventually a better combination of descriptors might be found. The final model is validated with a selection of (or with all) compounds in the similarity group which are not members of the training set.

The brief description above reveals that the TQSPR method, as compared to the traditional QSPR techniques, selects the sets of compounds from which QSPRs are developed. The quality of the established structural relationship is limited only by the eventual inefficiency of the description of the molecular structure with the available descriptors. However, the TQSPR method at present has been developed mainly toward prediction of the properties of a single compound, the target, and its ability to predict for groups of compounds, as typical QSPRs do, needs further elaboration.

The TQSPR method provides several opportunities for prediction of properties of members of homologous series, which are explored consecutively in this work:

- Selecting a suitable target compound from the homologous series, defining a similarity group and a training set, and predicting the property for the members of the homologous series, that were not used for the model derivation, with the thus derived model. The similarity group may be selected from all database compounds (which have various structures), or the homologous series might be a priori defined as the similarity group;

- Targeting separately each of the members of a homologous series one by one and thus providing estimated values of a property for all compounds. The similarity group may be selected as above;

- Identifying a descriptor which for a homologous series is collinear with the estimated property, thus being able to derive a linear relationship for prediction of the property values from the values of this descriptor.

Sources of Data, Databases, and Software. Experimental values of the critical temperatures, T_c, and pressures, p_c, of five homologous series with the general formula $H(CH_2)_nR$, where R is the following end groups, H (normal alkanes), C_2H_3 (1-alkenes), OH (1-alkanols), C_6H_5 (*n*-alkylbenzenes), and COOH (*n*-alkanoic acids), have been used to test the different options for prediction with the TQSPR method. While we believe that the results we have obtained are relevant to most homologous series, which differ only by their end groups, the homologous series chosen have been used for comparison of the abilities of several QSPR methods recently published.[19]

The experimental data used for the critical properties studied and references to the literature sources are presented in Appendix A, Table A1 to Table A10, respectively (Appendix A can be found in the Supporting Information, SI). We would like to stress that for the derivation of our models for the homologous series *only* the experimental data, given in the tables of Appendix A, were used.

In the tables of Appendix A, we present the uncertainties of all experimental data in percentage form. This was needed to be able to compare them with predictions from different QSPR methods, which were published in percentage form.[19] When the uncertainties were given in the original publication only in percentage form, we used them directly. When the uncertainties in the respective experimental work were given only in absolute values of measurement, we recalculated them in percentage form. When the uncertainties of the experimental data were given in the original publication in both ways (i.e., in the text in percentage form and in the respective tables in absolute values of measurement), we assumed that the tabulated data are the representative ones and the uncertainties in the text were rounded numbers. So, we calculated the uncertainties in percentage form from the experimental uncertainties given in the original tables in absolute values of measurement.

The tables in Appendix A contain also average and maximum uncertainties of the experimental data for the carbon atom range in each of the series, which has been used for comparison.

For the critical temperatures, we compare the average (AAPE) and maximum (MAPE) deviations of the predicted values of the different methods from the experimental values by

$$\mathrm{AAPE} = \left(\frac{1}{N}\right)\left(\sum_{j=1}^{N} \frac{|T_j^{\mathrm{exp}} - T_j^{\mathrm{calc}}|}{T_j^{\mathrm{exp}}}\right) \tag{4}$$

where T_j^{exp} and T_j^{calc} are the experimental and calculated values of the critical temperature and N is the number of data points.

$$\mathrm{MAPE} = \left|\frac{T_j^{\mathrm{exp}} - T_j^{\mathrm{calc}}}{T_j^{\mathrm{exp}}}\right|_{\mathrm{max}} \tag{5}$$

For the critical pressures, the temperature values in eqs 4 and 5 are replaced by the respective pressure values. We would like to note here that for consistency we keep the abbreviations used by Nikitin et al.[19] but use the formulas and interpretation their team has adopted recently.[20]

For the prediction by targeting each of the members of the homologous series, similarity groups of 50 compounds and training sets of 10 compounds were identified from a database of 326 compounds of different structure (the 260 hydrocarbons, which we have listed in a previous work[15] plus 66 oxygen-containing compounds, namely, the members of the studied series and a selection of polyvalent alcohols). Like in our previous studies, we have used for the derivation of all our models *only* the available experimental data from our database.

The chemical structures of all molecules have been characterized by a total of 1664 descriptors, calculated with the Dragon, version 5.4. software[21] (DRAGON, TALETE srl, http://www.talete.mi.it) from minimized molecular models. The molecular geometries were optimized using the CNDO (Complete Neglect of Differential Overlap) semiempirical method implemented in the HyperChem package (Version 7.01, Hypercube Inc.). Excluding the constant and near constant descriptors led to a selection of 1280 descriptors, with which we performed our work.

The software program used for the selection of the similarity group and for deriving the QSPRs is the one we developed in the MATLAB^R (ref 22) environment for the TQSPR method[17] on the basis of the SROV program.[18]

Results and Discussion

Prediction of the Critical Temperature of n-Alkanes with Compounds from the Whole Database and from the Series Only. The TQSPR method in its original form develops equations for prediction of the properties of single compounds (targets) from training sets, selected from a group of structurally similar compounds (similarity group). These equations can also predict satisfactorily for most of the members of the similarity

Table 1. Prediction of T_c by Targeting Each of the *n*-Alkanes

		training set from all 326 compounds			training set from the *n*-alkane series		
target[a]	T_c/K, exp	dominant (selected) descriptor	T_c/K predicted	deviation/%	dominant (selected) descriptor	T_c/K predicted	deviation/%
C2	305.32	HTm	285.76	6.41	CIC0 (HVcpx)	299.5 (307.7)	1.91 (0.78)
C3	369.83	TPC	374.21	1.18	HVcpx (CIC0)	363.8 (370.6)	1.63 (0.21)
n-C4	425.12	EPS1	428.87	0.88	CIC0	424.9	0.06
n-C5	469.7	IVDM	473.2	0.74	CIC0	469.6	0.02
n-C6	507.6	DP01	510.8	0.63	CIC0	507.6	0.00
n-C7	540.2	IVDM	540.2	0.00	CIC0	540.1	0.02
n-C8	568.7	IVDM	568.8	0.02	CIC0	568.9	0.03
n-C9	594.6	DP01	594.7	0.01	CIC1	594.3	0.05
n-C10	617.7	DP01	617.8	0.02	piPC01	618.0	0.05
n-C11	639	DP01	638.7	0.05	BELe7	637.5	0.23
n-C12	658	R4p	659.2	0.18	R4p	659.2	0.18
n-C13	675	H4p	676.8	0.27	R4p	676.1	0.16
n-C14	693	H3m	696.0	0.43	H3m	696.0	0.43
n-C15	708	RARS	708.1	0.01	*R3m*	707.5	0.07
n-C16	723	R4p	721.2	0.25	R4p	721.2	0.25
n-C17	736	HATSp	736.5	0.07	HATSp	736.5	0.07
n-C18	747	R1m	748.2	0.16	R1m	748.2	0.16
n-C19	755	*R3m*	755.6	0.08	*R3m*	755.6	0.08
n-C20	768	HATS4p	764.6	0.45	SIC2	766.8	0.15
n-C21	778	Mor21v	776.8	0.15	Mor21v	776.8	0.15
n-C22	786	HATS5v	782.7	0.42	HATS5v	782.7	0.42
n-C23	790	Mor21v	793.2	0.40	Mor21v	793.2	0.40
n-C24	800	HATS6u	801.7	0.21	HATS6u	801.7	0.21
n-C25	-	R7m	808.1	-	R7m	808.1	-
n-C26	816	Espm02u	814.3	0.21	ESpm02u	814.3	0.21
n-C27	-	R7m	816.3	-	R7m	816.3	-
n-C28	824	R6v	829.9	0.72	H8e	829.6	0.68
n-C29	-	HATS5v (REIG)	**832.7** (835.0)	-	HATS5v (REIG)	**832.7** (835.00)	-
n-C30	843	Mor26p (REIG)	**831.7** (839.5)	1.34 (0.41)	Mor26p (REIG)	**831.7** (839.5)	1.34 (0.41)
n-C32	-	HATS5v	849.0	-	HATS5v	849.0	-
n-C35	-	R7m (REIG)	**873.6** (865.5)	-	R7m (REIG)	**873.6** (865.5)	-
n-C36	872	RTm	875	0.34	JGI1 (DP01)	**863.9** (874.3)	0.93 (0.26)
n-C40	-	RTm	892	-	HATS5v	890.0	-
n-C44	-	RTm	909	-	HATS5v	906.3	-
n-C60	-	RTm (Espm04x)	**1425.0** (955)	-	R7m (DP01)	**1389.0** (950.3)	-
		(C3 to C36), AAPE/%		0.35 (0.32)			0.31 (0.19)
		(C3 to C36), MAPE/%		1.34 (0.41)			1.63 (0.41)

[a] Short notation; i.e., C2 is ethane, C3 is propane, *n*-C4 is butane, etc.

group as well. For homologous series, the similarity group may be selected from all available compounds of different structures in the database, or the particular series may be considered naturally to be the similarity group.

In this part of our work, each member of the homologous series was consecutively chosen as a target. When the whole database of 326 compounds was used, similarity groups of 50 and training sets of 10 compounds were selected for each targeted *n*-alkane. When the homologous series was assumed to be the natural similarity group for all *n*-alkanes, the 10-member training sets for each member were selected from the *n*-alkane series. In both cases only single descriptor equations were developed from the training sets.

Table 1 compares the selected descriptors and predictions of the models for the T_c of *n*-alkanes obtained with the training sets selected from the whole database and from the homologous series only. The descriptors are classified as "dominant", i.e., the first chosen by the program as most correlated with the property, and "selected", descriptors found to provide a better prediction when the deviation from measured values provided by the dominant descriptors was considered unsatisfactory. The predicted value was judged unsatisfactory in cases when the deviation was greater than the experimental uncertainty shown in Table A1 (SI, Appendix A), or there were systematic deviations from the experimental data. Cases of systematic deviations are marked in bold in Table 1. The selected descriptors were chosen by a "trial and error" procedure, testing first the dominant descriptors selected when the neighboring compounds were targeted and the descriptors identified by the TQSPR algorithm to be among the ten most correlated with T_c. Typically, the differences in the correlation coefficients of the first ten descriptors, suggested by the algorithm, are within the range of the third digit.

It is seen from Table 1 that the AAPE and MAPE, obtained with the dominant descriptors, are within the average experimental uncertainties (Table A1 in Appendix A, SI), except for the first two *n*-alkanes and four compounds in the end of the series, for which systematic deviation is also observed. In both cases, the equations with the "selected" descriptors fix the problems. However, to identify unreasonable predictions like those given in bold, they should always be checked against the known measured values for the series, and a descriptor different from the dominant should be selected if necessary. It is interesting to note also that the differences between the models developed from the whole database, and those from the series only, do not seem to be as significant as originally expected. The systematic deviations at the end of the series seem to be the result of the lack of experimental data for closer members rather than the incorrect selection of compounds for the training sets (Table 9). It is seen also that the descriptors selected for a given target vary when the training sets selected vary significantly (Table 1 and Table 9).

The TQSPR method provides predicted property values also for the members of the similarity group not included in the

Table 2. Predictions of T_c within the Similarity Group of 50 Compounds with Decane as the Target[a]

DBNo.	compound name	C atoms	T_c/K, exp.	T_c/K, pred.	Deviation %
5	hexane	6	507.6	507.3	0.07
6	heptane	7	540.2	540.1	0.02
7	*octane*	*8*	*568.7*	*569.1*	*0.07*
8	*nonane*	*9*	*594.6*	*594.7*	*0.01*
9	decane-target	10	617.7	594.7	0.01
10	*undecane*	*11*	*639*	*638.8*	*0.03*
11	*dodecane*	*12*	*658*	*658.1*	*0.01*
12	*tridecane*	*13*	*675*	*675.8*	*0.12*
13	*tetradecane*	*14*	*693*	*692.2*	*0.11*
14	*pentadecane*	*15*	*708*	*707.6*	*0.06*
15	*hexadecane*	*16*	*723*	*722.0*	*0.13*
16	heptadecane	17	736	735.6	0.06
17	*octadecane*	*18*	*747*	*748.3*	*0.18*
18	nonadecane	19	755	760.4	0.71
19	eicosane	20	768	771.9	0.51
20	heneicosane	21	778	782.8	0.62
21	docosane	22	786	793.2	0.92
22	tricosane	23	790	803.1	1.66
23	tetracosane	24	800	812.7	1.59
51	2-methylheptane	8	559.7	559.4	0.06
53	4-methylheptane	8	561.7	554.2	1.34
57	2,5-dimethylhexane	8	550	549.0	0.18
75	2-methyloctane	9	582.87	586.30	0.59
91	1-heptene	7	537.4	537.7	0.06
92	1-octene	8	567	566.9	0.01
93	1-nonene	9	593.1	593.0	0.02
94	*1-decene*	*10*	*616.6*	*616.3*	*0.04*
95	1-undecene	11	-	637.5	-
96	1-dodecene	12	658	656.9	0.16
97	1-tridecene	13	673	674.8	0.26
98	1-tetradecene	14	691	691.3	0.05
99	1-pentadecene	15	705	706.8	0.26
100	1-hexadecene	16	718	721.3	0.47
101	1-heptadecene	17	734	734.9	0.12
102	1-octadecene	18	748	747.7	0.03
103	1-nonadecene	19	755	759.8	0.64
151	butylcyclohexane	10	653.1	588.1	9.95
155	butylcyclopentane	9	-	570.2	-
197	pentylbenzene	11	675	608.8	9.81
234	p-diethylbenzene	10	657.9	576.9	12.32
252	4-methyloctane	9	-	580.8	-

[a] Deviations from the experimental values if available. The training set compounds are shown in italic. In the TQSPR method, target compounds are not used in training sets.

training sets, some of which may be structurally different from the particular homologous series. Table 2 shows such predictions with the model developed for the target *n*-decane from its training set, when selected from all 326 compounds. The average errors of the predictions outside the training set, except for the structurally most distant compounds (three of the last five in Table 2), are within the average experimental uncertainties for the respective series (Appendix A, SI).

Comparison of T_c and p_c Predictions by Consecutive Targeting of Each Series Member with Those of Other Methods. Critical temperatures and pressures were predicted by consecutive targeting of each member of the *n*-alkane, 1-alkene, 1-alkanol, *n*-alkylbenzene, and *n*-alkanoic acid homologous series. In Tables 3 and 4, the obtained results are compared with two well-known QSPR methods: the group contribution method of Constantinou and Gani[23] and the group/bond contribution method of Marero and Gani.[24] The AAPE and MAPE deviations for these QSPR methods were taken from Nikitin et al.[19] Tables 3 and 4 contain the deviations of the models, developed in this work only from the respective homologous series, while the two group contribution methods are developed from a much wider database and are expected to be applicable not only to homologous series, but also to any structure. As demonstrated above, targeting only homologous series does not improve significantly the TQSPR predictions for homologous series, so the comparison is correct. Table 3 and Table 4 show that the deviations of the TQSPR method compare favorably with the deviations of the two well-recognized QSPR methods.

Linear Equations for T_c and p_c, of Each Series. Comparison with ABC Methods. Our previous work[5] demonstrated that it was possible to develop linear QSPRs for prediction of properties in homologous series, if a descriptor collinear with a given property could be identified. There the similarity group was defined as the whole homologous series, and a collinear descriptor was found with all available experimental data, even complemented by predicted data when long-range extrapolation was needed; but, the experimental data were insufficient. The parameters of the traditional QSPR model were then determined from experimental data only.

In the present work, we use the TQSPR method to identify descriptors collinear with a given property with a possible minimum of experimental data *only*, employing the descriptor−property correlation coefficients (ρ_{yj}) and the CNR values, provided by the TQSPR program, as a guidance.

The following tables and figures present the linear equations, obtained in this manner. The coefficients of the equations were determined by using a smallest possible amount of measured data, which would allow for adequate prediction of the rest of the available experimental data. The respective data used can be identified from the references of the tables given in the SI, Appendix A (given in the Supporting Information).

It should be possible in principle to develop linear equations from only two measured points. Table 5 shows the parameters of equations, obtained with a minimum of experimental data, and with only two measured points. Details for all models are presented in the respective tables in Appendix B (given in the Supporting Information).

Figures 1 to 8 present the experimental and the predicted values obtained with the linear equations. Except for two cases (for the T_c of *n*-alkanes and for the p_c of *n*-alkanoic acids), which require more than two measured data, all equations presented on the figures are obtained from two measured data for the lower members of the series only. These figures, and the detailed data presented in Appendix B of the Supporting Information, clearly demonstrate that even with only two measured points the studied properties can be predicted with deviations comparable to those typical for contemporary ABC methods and the linear TQSPR equations derived from more experimental data. For certain homologous series which might have only several measured data, the procedure for the identification of collinear descriptors might still require the use of data predicted with a reliable ABC method.[5] The equations obtained in such cases must be reported with an exact description of the data, predicted from another correlation, and why and how they have been used. We would like again to point out, however, that for all series studied in the present work the measured data have been sufficient, and they have been developed from experimental data only.

Table 6 and Table 7 compare the predictions of all linear equations, developed with TQSPR, with those of the recent ABC equations, proposed by Nikitin et al.[19] These tables show similar average deviations of the compared methods, within the experimental uncertainties of the measured properties. Comparison with our previous work,[5] where linear equations for homologous series were developed with traditional QSPR methodology, outlines clear advantages of using the TQSPR method suggested in this work, which achieves deviations close

Table 3. Experimental Uncertainties and Average (AAPE) and Maximum (MAPE) Deviations from Experimental Values for Predictions of T_c by TQSPR and by Other QSPR Methods[a]

		uncertainty/%		this work, consecutive training sets from series (whole database)		Constantinou and Gani (Nikitin et al.[19])		Marero and Gani (Nikitin et al.[19])	
series	range of C atoms	avg.	max	AAPE/%	MAPE/%	AAPE/%	MAPE/%	AAPE/%	MAPE/%
n-alkanes	3 to 36	0.91	1.50	0.19 (0.32)	0.41 (0.41)	1.05	5.76	0.92	3.11
1-alkenes	3 to 18	0.49	1.04	0.19	0.87	0.79	1.87	0.39	1.38
n-alkylbenzenes	3 to 13	0.69	1.06	0.31	0.71	1.24	2.90	0.41	0.96
1-alkanols	3 to 22	0.51	1.06	0.16	0.68	1.44	2.77	1.18	2.78
n-alkanoic acids	3 to 21	0.69	1.05	0.43	1.20	0.73	1.77	5.97	10.00

[a] The number of significant digits is given as in Nikitin et al.[19]

Table 4. Experimental Uncertainties and Average (AAPE) and Maximum (MAPE) Deviations from the Experimental Values for Predictions of p_c by TQSPR and by Other QSPR Methods[a]

		uncertainty/%		this work, consecutive training sets from series		Constantinou and Gani (Nikitin et al.[19])		Marero and Gani (Nikitin et al.[19])	
series	range of C atoms	avg.	max.	AAPE/%	MAPE/%	AAPE/%	MAPE/%	AAPE/%	MAPE/%
n-alkanes	3 to 36	8.80	22.99	1.68	7.63	3.65	14.13	14.29	84.28
1-alkenes	3 to 18	2.55	10.36	1.68	3.24	3.61	7.85	4.73	11.84
n-alkylbenzenes	3 to 13	2.53	3.25	1.28	3.05	2.41	11.81	2.02	3.74
1-alkanols	3 to 22	1.99	3.13	1.27	5.45	5.51	9.57	2.02	3.74
n-alkanoic acids	3 to 21	4.62	23.81	1.48	4.26	5.03	10.98	4.24	7.93

[a] The number of significant digits is given as in Nikitin et al.[19]

Table 5. Linear Equations for Prediction of Properties from Selected Data[a]

series	equation[b]	used data, figure[c]
n-alkanes	T_c/K = 154.4557 + 141.0302 IVDM (5.1)	*n*-C3 to *n*-C12
	p_c/MPa = −2.5244 + 20.4267 BIC1 (5.2)	*n*-C3 to *n*-C14
	p_c/MPa = −2.6027 + 20.7597 BIC1 (5.3)	*n*-C3 and *n*-C14, Figure 1
1-*n*-alkenes	T_c/K = 195.178+ 111.440 DP01 (5.4)	*n*-C3 to *n*-C7
	T_c/K = 195.421 + 111.466 DP01 (5.5)	*n*-C3 and *n*-C7, Figure 2
	p_c/MPa = −32.0022 + 31.9809 PCR (5.6)	*n*-C3 to *n*-C7; *n*-C16
	p_c/MPa = −32.5258 + 32.4242 PCR (5.7)	*n*-C3 and *n*-C16, Figure 3
n-alkylbenzenes	T_c/K = −26.6100 + 259.648 piPC01 (5.8)	*n*-C3 to *n*-C7
	T_c/K = −28.2629 + 259.888 piPC01 (5.9)	*n*-C3 and *n*-C7, Figure 4
	p_c/MPa = −0.12651 + 4.9908 ARR (5.10)	*n*-C0 to *n*-C4
	p_c/MPa = −0.1175 + 5.0125 ARR (5.11)	*n*-C0 and *n*-C4, Figure 5
n-1-alkanols	T_c/K = 334.329 + 15.5792 RTu (5.12)	*n*-C3 to *n*-C12
	T_c/K = 340.994 + 15.4458 RTu (5.13)	*n*-C3 and *n*-C12, Figure 6
	p_c/MPa = −1.97054 + 14.4979 VEZ2 (5.14)	*n*-C3 to *n*-C7
	p_c/MPa = −2.0084 + 14.5586 VEZ2 (5.15)	*n*-C3 and *n*-C7, Figure 7
n-alkanoic acids	T_c/K = 588.957 + 16.6285 DP07 (5.16)	*n*-C2 to *n*-C6
	T_c/K = 586.197 + 17.2789 DP07 (5.17)	*n*-C2 and *n*-C6, Figure 8
	p_c/MPa = 5.3283 − 7.4108 H2v (5.18)	*n*-C4 to *n*-C10

[a] Details are given in the Supporting Information (Appendix B). [b] The equations are numbered consecutively; the first number corresponds to the number of this Table. [c] Short notation; the number of the Figure on which the results are displayed is also shown.

to the experimental uncertainties for the respective properties using a smaller amount of measured data and no predicted data. The advantages of the linear equations developed with TQSPR over traditional ABCs again is the use of less experimental data and the higher level of confidence in linear long-range extrapolation.

The comparison of the different methods, presented above, is more likely to be of interest to readers that are "data experts" and/or develop their own QSPRs. For practicing chemical engineers looking for experimental and/or predicted data, the primary choice would rather be among the commercial databases. That is why we have also compared our predictions and the predictions of Nikitin et al.[19] to predicted data, recommended by DIPPR.[25] In order to do that, we have assumed that the "reliabilities" (in percentage form) assigned to each predicted value by DIPPR correspond to the deviations from the experimental values, used in this work. Then we have calculated AAPE and MAPE for the respective DIPPR data, as shown above. The comparison showed that for all homologous series the values predicted by Nikitin et al.[19] and in this work were significantly closer to the measured values than the predicted values, recommended by DIPPR.

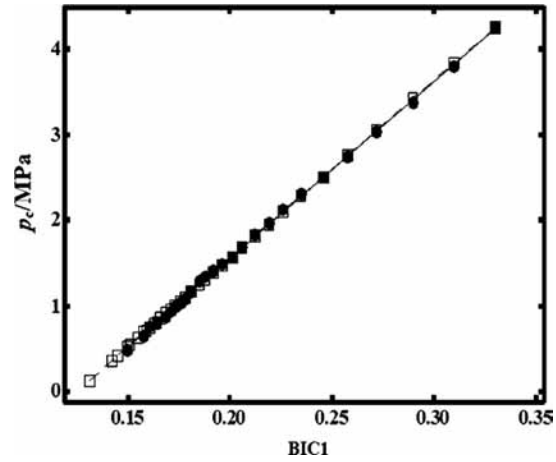

Figure 1. Linear prediction of the p_c of *n*-alkanes with eq 5.3, Table 5, using only two of the measured data. ●, experimental values of p_c; □, predicted values of p_c; —, line of approximated experimental values; - - -, line of approximated predicted values. BIC1 is the collinear descriptor.

Identification of Descriptor−Property Collinearity. As seen above, the identification of descriptors collinear with properties

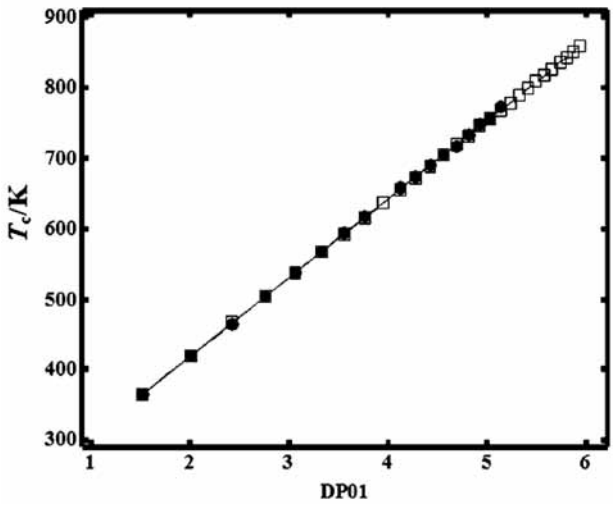

Figure 2. Linear prediction of the T_c of *n*-1-alkenes with eq 5.5, Table 5, using only two of the measured data. ●, experimental values of T_c; □, predicted values of T_c; —, line approximating experimental values; - - -, line approximating predicted values. DP01 is the collinear descriptor.

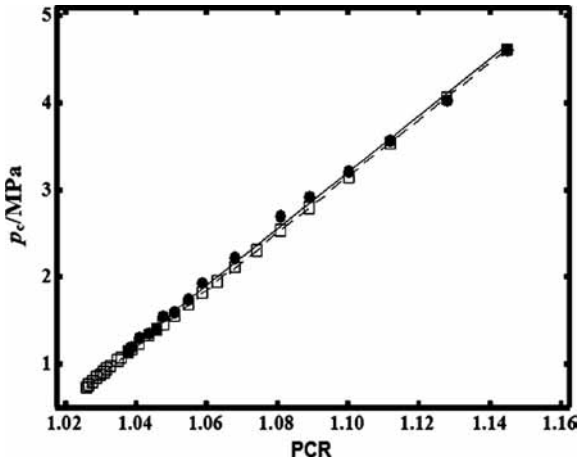

Figure 3. Linear prediction of the p_c of *n*-1-alkenes with eq 5.7, Table 5, using only two of the measured data. ●, experimental values of p_c; □, predicted values of p_c; —, line approximating experimental values; - - -, line approximating predicted values. PCR is the collinear descriptor.

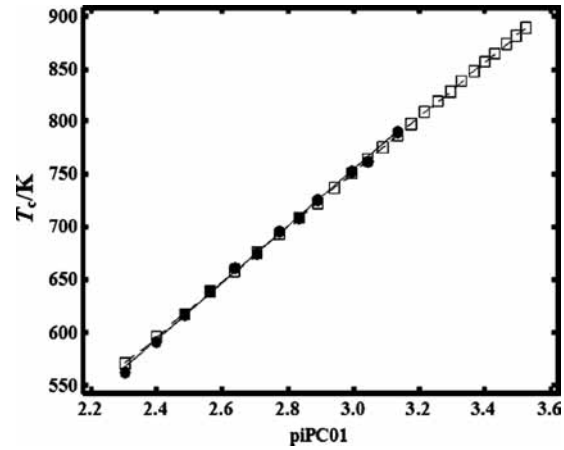

Figure 4. Linear prediction of the T_c of *n*-alkylbenzenes with eq 5.9, Table 5, using only two of the measured data. ●, experimental values of T_c; □, predicted values of T_c; —, line of approximated experimental values; - - -, line of approximated predicted values. PiPC01 is the collinear descriptor.

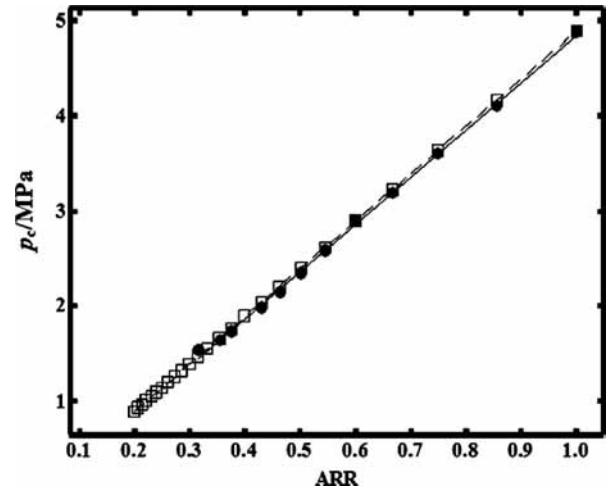

Figure 5. Linear prediction of the p_c of *n*-alkylbenzenes with eq 5.11, Table 5, using only two of the measured data. -●-, experimental values of p_c; - -□- -, predicted values of p_c; —, line of approximated experimental values; - - -, line of approximated predicted values. ARR is the collinear descriptor.

studied is an important prerequisite of the proposed method. Table 8 summarizes the descriptors found in this work to be collinear with T_c or p_c of the respective homologous series, together with the compounds used for their identification by the values of the property−descriptor correlation coefficients (ρ_{yj}) and/or the CNR values. It includes also the definition of the descriptors, as given in the User Manual of the Dragon 5.4. software.[21]

It is seen from Table 8 that only in one case (p_c, 1-*n*-alkenes) the descriptor collinear with the respective property could not be identified from experimental data for a relatively small group of lower members of the respective homologous series by its ρ_{yj} and/or CNR value. Typically, the TQSPR program will find such a descriptor by interactive generation of candidate descriptors and linear equations and selecting different target compounds and/or amounts of measured data, until the user is satisfied with the prediction of all experimental values.

The present work raises a question which is fundamental for the QSPRs methodology; namely: Does the existence of descriptors collinear with a given property imply that they represent the "most significant common features" of the chemical structure important for the particular property? The answer to this question is not straightforward. Presently we cannot provide a definite answer because it requires systematic studies, which go beyond the aim of this work. Table 8 shows that the ten collinear descriptors belong to six groups. Both descriptors for the *n*-alkanes are information indices. The descriptors for T_c of 1-*n*-alkenes and 1-*n*-alkanoic acids are Randic molecular profiles. Walk and path count indices have been chosen for the p_c of 1-*n*-alkenes and the T_c of *n*-alkylbenzenes, and GETAWAY descriptors - for the T_c of *n*-1-alkanols and p_c of 1-*n*-alkanoic acids. The remaining two collinear descriptors belong to two different groups.

It should be pointed out, however, that our work indicated that in addition to the collinear descriptors identified above there might exist other collinear descriptors, which we have not sought, as explained previously. Thus, the professional answer to the posted question would require first analysis of the collinear descriptors, suggested by the algorithm with the view to establish one or more most suitable for the property. Then, the relationships between collinear descriptors and respective properties, and the factors which affect them, should be studied systematically. We trust that the present findings might serve as an introduction to such studies.

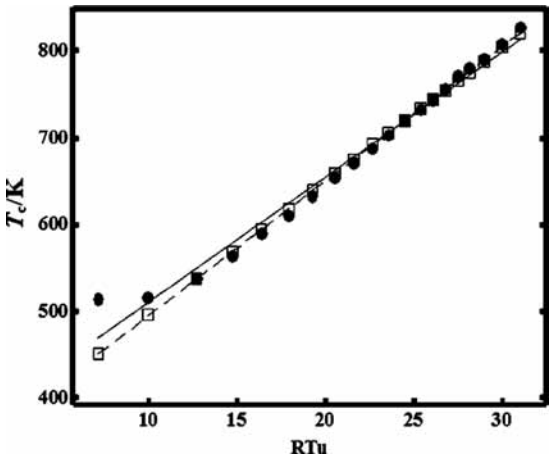

Figure 6. Linear prediction of the T_c of *n*-1-alkanols with eq 5.13, Table 5, using only two of the measured data. ●, experimental values of T_c; □, predicted values of T_c; —, line of approximated experimental values; - - -, line of approximated predicted values. RTu is the collinear descriptor.

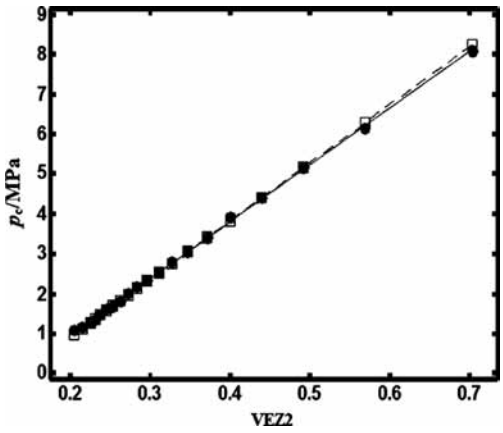

Figure 7. Linear prediction of the p_c of *n*-1-alkanols with eq 5.15, Table 5, using only two of the measured data. ●, experimental values of p_c; □, predicted values of p_c; —, line of approximated experimental values; - - -, line of approximated predicted values. VEZ2 is the collinear descriptor.

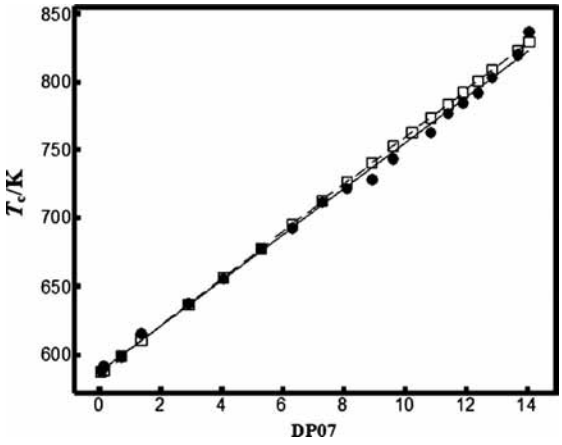

Figure 8. Linear prediction of the T_c of *n*-1-alkanols with eq 5.17, Table 5, using only two of the measured data. ●, experimental values of T_c; □, predicted values of T_c; —, line of approximated experimental values; - - -, line of approximated predicted values. DP07 is the collinear descriptor.

Structural Similarity and Selection of Compounds for the Training Sets. In the course of the work presented above, it was found that many of the training sets contain compounds which, from the point of view of common knowledge of chemical structure, seem to be less related to the target than others. This observation holds both for the similarity groups of 50 and training sets of 10 compounds selected from the whole database of 326 compounds, for each targeted *n*-alkane (Table 9) and for the 10-member training sets determined only from homologous series (Table 10). The two tables show the order into which the members of the training set have been selected only for some of the particular targets, but the observations hereunder are based on all selections.

It can be seen from both tables that the problem is more pronounced for the low carbon number members of the series. For example, in the training set for ethane, identified from the whole database (Table 9), propane was selected first, and butane was the ninth. No other member of the *n*-alkane series was identified as structurally similar to ethane. The *n*-alkane most structurally related to propane was identified as pentane, whereas butane was selected last; neither ethane nor any other *n*-alkane was included in the propane training set. Instead, the propane training set includes alcohols, *iso*-alkenes, cycloalkanes, etc., most of them with up to four heavy atoms. The number of *n*-alkanes selected in the training sets increases with the increase of the number of carbon atoms in the target, and the nonmembers gradually move to the last places in the training sets.

The GACC values shown in the two tables provide a good single number indication of the level of similarity between the target compound and the training set. They are lowest for the low carbon number compounds (up to approximately 8C atoms), highest for the compounds in the middle of the homologous series (C9 to C30), and somewhat reduced toward the upper end of the homologous series. For *n*-C12, *n*-C14, and *n*-C16 to *n*-C26 from the whole database (Table 9), only members of the homologous series are selected. Then, single nonmember compounds start to appear again occasionally in the last places of the training sets, but without a profound effect on the GACC. The values of the latter increase or decrease significantly only when the appearance of nonmembers is combined with the limitation of the ability for balanced interpolation, i.e., when the number of compounds with lower and higher carbon numbers is limited and the distance (in terms of n_C) between the predictive compounds (above *n*-C30) increases. When the training sets are selected solely from the *n*-alkanes (Table 10), only the latter effect is observed.

The order of selection of the compounds for the training sets from the series in most cases seems correlated with the known alternation of the structures with odd and even numbers of carbon atoms, which influences properties related to crystals, e.g., melting temperatures. Indeed, for all targets with more and for some with less than nine carbon atoms, the first chosen compound is always with two more carbon atoms than the target. Moreover, we have demonstrated[5] that the TQSPR method can exploit successfully these subtle structural features. The deviations for compounds with lower number of carbon atoms, however, indicate that there is a strong influence of factors in addition to chemical structure on the selection of training sets. For instance, it has been demonstrated[26] that certain descriptors (e.g., topological indices) cannot be defined and/or have the same values for the first members of homologous series; some descriptors do not distinguish different heavy atoms, etc. In our work, we have observed that the values of the descriptors depend also on the extent of minimization of the molecular structure. One example of values inconsistent with the general trend is given in bold in Table B10 of the SI, but we have identified more cases. There might be yet other nonidentified factors

Table 6. Experimental Uncertainty and Average and Maximum Deviations from the Experimental Values for the Prediction of T_c by Linear Equations[a], Compared to a Recent Asymptotic Correlation

	compared range	uncertainty/%		this work, training sets from series		Nikitin et al.[19]	
series	of C atoms	avg.	max.	AAPE/%	MAPE/%	AAPE/%	MAPE/%
n-alkanes	3 to 36	0.91	1.50	0.47 (eq 5.1)	1.15 (eq 5.1)	0.35	1.34
1-alkenes	3 to 18	0.49	1.04	0.21 (eq 5.4) 0.19 (eq 5.5)	0.47 (eq 5.4) 0.47 (eq 5.5)	0.21	0.66
n-alkylbenzenes	3 to 13	0.69	1.06	0.19 (eq 5.8) 0.21 (eq 5.9)	0.33 (eq 5.8) 0.44 (eq 5.9)	0.59	1.18
1-alkanols	3 to 22	0.51	1.06	0.48 (eq 5.12) 0.53 (eq 5.13)	1.09 (eq 5.12) 1.31 (eq 5.13)	0.43	1.21
n-alkanoic acids	3 to 21	0.69	1.05	0.46 (eq 5.16) 0.68 (eq 5.17)	1.59 (eq 5.16) 1.73 (eq 5.17)	0.72	1.37

[a] The experimental uncertainties and the deviations from the measured values are presented in detail in the respective tables of the Supporting Information (Appendices A and B). The number of significant digits is given as in Nikitin et al.[19]

Table 7. Experimental Uncertainty and Average and Maximum Deviations from the Experimental Values for the Prediction of p_c by Linear Equations[a], Compared to a Recent Asymptotic Correlation

	compared range	uncertainty/%		this work, training sets from series		Nikitin et al.[19]	
series	of C atoms	avg.	max.	AAPE/%	MAPE/%	AAPE/%	MAPE/%
n-alkanes	3 to 36	8.70	22.99	2.52 (eq 5.2) 1.76 (eq 5.3)	14.41 (eq 5.2) 8.26 (eq 5.3)	2.32	18.01
1-alkenes	3 to 18	2.55	10.36	2.22 (eq 5.6) 2.98 (eq 5.7)	4.38 (eq 5.6) 6.22 (eq 5.7)	1.45	2.94
n-alkylbenzenes	3 to 13	2.53	3.25	1.43 (eq 5.10) 1.84 (eq 5.11)	5.84 (eq 5.10) 4.54 (eq 5.11)	2.17	9.38
1-alkanols	3 to 22	1.99	3.13	1.28 (eq 5.14) 1.24 (eq 5.15)	6.54 (eq 5.14) 8.41 (eq 5.15)	1.14	2.56
n-alkanoic acids	3 to 21	4.62	23.81	2.99 (eq 5.18)	7.38 (eq 5.18)	4.90	9.76

[a] The experimental uncertainties and the deviations from the measured values are presented in detail in the respective tables of the Supporting Information (Appendices A and B). The number of significant digits is given as in Nikitin et al.[19]

Table 8. Descriptors Collinear with the Critical Properties of the Homologous Series Studied

descriptor[a]	collinear with	Dragon definition of descriptor
IVDM (C_4 to C_{20}; 1st CNR)	T_c, *n*-alkanes	Information index. Mean information content on the vertex degree magnitude.
BIC1 (C_4 to C_{16}; 1st ρ_{yj})	p_c, *n*-alkanes	Information index. Bonding information content (neighbor symmetry of 1-order).
DP01 (C_4 to C_{12}; within first four of equal ρ_{yj})	T_c, 1-*n*-alkenes	Randic molecular profile indices. DP01 is molecular profile 01.
PCR (−)	p_c, 1-*n*-alkenes	Walk and path count index. The ratio of multiple path count over path count.
piPC01 (C_0 to C_{10}; ρ_{yj}; 4 CNR)	T_c, *n*-alkylbenzenes	Walk and path count index. Molecular multiple path count of order 01 (sum of conventional bond orders.)
ARR (C_0 to C_{10}; 1st ρ_{yj})	p_c, *n*-alkylbenzenes	Constitutional descriptor. Aromatic ratio.
RTu (C_4 to C_{14}; 1st ρ_{yj})	T_c, *n*-1-alkanols	GETAWAY descriptor. R total index/unweighted.
VEZ2 (C_4 to C_{14}; within first 4 of equal ρ_{yj})	p_c, *n*-1-alkanols	eigenvalue based index. Average eigenvector coef. sum from *Z* weighted distance (Barysz) matrix.
DP07 (C_4 to C_{14}; 1st ρ_{yj})	T_c, *n*-alkanoic acids	Randic molecular profile. Molecular profile 07.
H2v (C_4 to C_{14}; 1st CNR)	p_c, *n*-alkanoic acids	GETAWAY descriptor. H autocorrelation of lag 2, weighted by atomic van der Waals volumes

[a] The data in brackets are, respectively: the compounds needed to identify the collinear descriptor; the place of this descriptor in the ρ_{yj} and/or the CNR selection; (−) means "not chosen" by either criterion, in this case the descriptor has been found by interactive tries. The target compound is always the first member of the series, compared in Table 6 and Table 7. The similarity group does not contain compounds not included in the comparison.

Table 9. Order of Selection of the *n*-Alkane Training Sets from the Whole Database

	target compound[a]					
order	C2	C3	*n*-C11	*n*-C12	*n*-C36	*n*-C40
1st	**C3**[b]	propylene	***n*-C13**	***n*-C14**	***n*-C40**	***n*-C44**
2nd	methanol	isobutene	***n*-C9**	***n*-C10**	***n*-C35**	***n*-C36**
3rd	propylene	isobutane	***n*-C12**	***n*-C11**	***n*-C32**	***n*-C35**
4th	ethanol	***n*-C5**	***n*-C10**	***n*-C13**	***n*-C30**	***n*-C32**
5th	cyclobutane	*c*-2-butene	***n*-C15**	***n*-C16**	***n*-C44**	***n*-C30**
6th	*c*-2-butene	*t*-2-butene	***n*-C14**	***n*-C15**	***n*-C29**	***n*-C29**
7th	*t*-2-butene	cyclobutane	1-*n*-C11-ene	***n*-C18**	***n*-C28**	***n*-C28**
8th	cyclopropane	1,3-butadiene	***n*-C17**	***n*-C9**	***n*-C27**	***n*-C27**
9th	***n*-C4**	1-butene	1-*n*-C10-ene	***n*-C17**	***n*-C26**	**n-C26**
10th	1,2-ethanediol	***n*-C4**	1-*n*-C12-ene	***n*-C20**	1-*n*-C30-ene	1-*n*-C30-ene
GACC[c]	0.796	0.909	0.978	0.977	0.989	0.986

[a] Short notation; i.e., C2 is ethane. [b] The chosen members of the *n*-alkane series are shown in bold. [c] GACC is the Geometric Average Correlation Coefficient (eq 2).

as well. Thus, it is important to identify and study systematically all factors and their influence on descriptor selection, predictive ability, meaning of collinearity, etc.

It is important to point out also that the results presented in Table 9 and Table 10 are subject to variation when a different descriptor normalization (e.g., the Euclidean norm instead of infinite norm) and/or different property and descriptor databases are used. In a previous work, we analyzed different normalization techniques for prediction of properties of more than 100 compounds.[17] As already stated above, while differences were detected in the training sets, the prediction errors were not significantly influenced.

Table 10. Order of Selection of the Training Sets Only from the *n*-Alkane Homologous Series

	target compound[a]							
order	**C2**	**C3**	***n*-C4**	***n*-C5**	***n*-C11**	***n*-C12**	***n*-C36**	***n*-C40**
1st	*n*-C3	*n*-C5	*n*-C5	*n*-C6	*n*-C13	*n*-C14	*n*-C40	*n*-C44
2nd	*n*-C4	*n*-C4	*n*-C6	*n*-C7	*n*-C12	*n*-C13	*n*-C35	*n*-C36
3rd	*n*-C5	*n*-C2	*n*-C3	*n*-C4	*n*-C10	*n*-C10	*n*-C32	*n*-C35
4th	*n*-C6	*n*-C6	*n*-C7	*n*-C8	*n*-C9	*n*-C11	*n*-C30	*n*-C32
5th	*n*-C8	*n*-C7	*n*-C8	*n*-C9	*n*-C15	*n*-C16	*n*-C44	*n*-C30
6th	*n*-C7	*n*-C8	*n*-C10	*n*-C3	*n*-C14	*n*-C15	*n*-C29	*n*-C29
7th	*n*-C10	*n*-C9	*n*-C9	*n*-C10	*n*-C17	*n*-C18	*n*-C28	*n*-C28
8th	*n*-C9	*n*-C10	*n*-C12	*n*-C11	*n*-C16	*n*-C17	*n*-C27	*n*-C27
9th	*n*-C12	*n*-C11	*n*-C11	*n*-C12	*n*-C19	*n*-C9	*n*-C26	*n*-C26
10th	*n*-C11	*n*-C12	*n*-C14	*n*-C13	*n*-C18	*n*-C20	*n*-C25	*n*-C25
GACC[b]	0.698	0.788	0.837	0.896	0.970	0.977	0.989	0.983

[a] Short notation; i.e., *n*-C4 is *n*-butane. [b] GACC is the Geometric Average Correlation Coefficient (eq 2).

Labanowski et al.[27] have shown that many of the popular topological indices are highly correlated with the van der Waals molecular surface area or the van der Waals volume. It was suggested that the available measured T_c and p_c of homologous series depend linearly on the natural logarithm of the surface area of minimized molecular models.[28] The same authors also pointed out the relationship between their models and the parameters of the van der Waals equation of state. The practical use of the surface area itself as a descriptor is limited by the fact that its values are significantly different when calculated by different programs, but its importance has been proved in many later studies, which have selected descriptors related to it in general QSPR models.[16]

The TQSPR method, when applied to congeneric compounds (homologous series in the present work), provides new opportunities for studying systematically the relationships between descriptors and properties, comparing descriptors and evaluating their degeneracy, comparing methods for identification of structural similarity between molecular structures.

Conclusions

The ability of the TQSPR method to predict properties for members of homologous series has been tested with experimental T_c and p_c data from a database of 326 hydrocarbon- and oxygen-containing compounds of different structures, described with 1664 descriptors, and with five of the homologous series contained in the database, having a general formula $H(CH_2)_nR$, where R is the following end groups: H (normal alkanes), C_2H_3 (1-alkenes), OH (1-alkanols), C_6H_5 (*n*-alkylbenzenes), and COOH (*n*-alkanoic acids).

The TQSPR method can be used for development of linear equations for homologous series with descriptors collinear with the studied property. Only in one case out of ten, the respective collinear descriptors could not be identified with the controls imbedded in the TQSPR program (ρ_{yj} and CNR). The comparison with presently available methods showed that while achieving precision in most cases within average experimental uncertainties like the best ABC methods, the TQSPR method needs smaller amounts of measured data and provides higher statistical confidence in long-range prediction.

The TQSPR method has been tested with only five homologous series, but the findings of the present work are relevant to all homologous series because, by definition, the structure of the members of a given homologous series is different only in terms of the number of carbon atoms.

The results obtained reveal that the TQSPR method, when applied to simple molecules, can provide insight into the way compounds are selected for training sets by structural similarity. They outline the existence of certain inefficiencies in this selection, even when members of homologous series for which a significant amount of measured data are available are considered.

Our general approach in this and previous work has been to test novel methods with the simplest structures and structural variation. The present work showed that this approach provides advantages from a methodological point of view over the tendency in the literature published to develop QSPRs from as large a database and structural variation as possible because it allows for better understanding of how the novel methods, and QSPRs in general, work.

The TQSPR method explicitly exploits the relationships between chemical structures, their descriptors, and properties and thus allows for systematic future studies of the causes for inefficiencies and the eventual physical meaning of QSPRs.

Supporting Information Available:

Appendices A and B, containing ten tables each. This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Poling, B. E.; Prausnitz, J. M.; O′Connel, J. P. *Properties of Gases and Liquids*, 5th ed.; McGraw-Hill: New York, 2001.

(2) Westmoreland, P. R.; Kollman, P. A.; Chaka, A. M.; Cummings, P. T.; Morokuma, K.; Neurock, M.; Stechel, E. B.; Vashishta, P. *WTEC Panel Report on Applications of Molecular and Materials Modeling*; NIST: USA, 2002.

(3) 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.

(4) 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.

(5) Brauner, N.; Cholakov, G. St.; Kahrs, O.; Stateva, R. P.; Shacham, M. Linear QSPRs for Predicting Pure Compound Properties in Homologous Series. *AIChE J.* **2008**, *54*, 978–990.

(6) Sola, D.; Ferri, A.; Banchero, M.; Manna, L.; Sicardi, S. QSPR Prediction of N-boiling Point and Critical Properties of Organic Compounds and Comparison with a Group-Contribution Method. *Fluid Phase Equilib.* **2008**, *263* (1), 33–42.

(7) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Combin. Sci.* **2003**, *22*, 69–77.

(8) Shacham, M.; Kahrs, O.; Cholakov, G. St.; Stateva, R. P.; Marquardt, W.; Brauner, N. The Role of the Dominant Descriptor in Targeted Quantitative Structure Property Relationships. *Chem. Eng. Sci.* **2007**, *62*, 6222–6233.

(9) Basak, S. C.; Gute, B. D.; Mills, D.; Hawkins, D. M. 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.

(10) Brauner, N.; Stateva, R. P.; Cholakov, G. St.; Shacham, M. A. Structurally "Targeted" QSPR Method for Property Prediction. *Ind. Eng. Chem. Res.* **2006**, *45*, 8430–8437.

(11) Shacham, M.; Brauner, N.; Cholakov, G. St.; Stateva, R. P. Property Prediction by Correlations Based on Similarity of Molecular Structures. *AIChE J.* **2004**, *50*, 2481–2492.

(12) Martinez W. L. ; Martinez, A. R., *Computational Statistics Handbook with MATLAB*; Chapman & Hall/CRC: London, 2002.

(13) Chickos, J. S. Hypothetical Thermodynamic Properties: The Boiling and Critical Temperatures of Polyethylene and Polytetrafluoroethylene. *J. Chem. Eng. Data* **2004**, *49*, 518–526.

(14) Gao, W., Jr.; Gasem, K. A. M. Improved Correlations for Heavyn-Paraffin Physical Properties. *Fluid Phase Equilib.* **2001**, *179*, 207–216.

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

(16) Godavarthy, S. S., Jr.; Gasem, K. A. M. Improved Structure-Property Relationship Models for Prediction of Critical Properties. *Fluid Phase Equilib.* **2008**, *264*, 122–136.

(17) Kahrs, O.; Brauner, N.; Cholakov, G. St.; Stateva, R. P.; Marquardt, W.; Shacham, M. Analysis and Refinement of the Targeted QSPR Method. *Comput. Chem. Eng.* **2008**, *32*, 1397–1410.
(18) Shacham, M.; Brauner, N. The SROV Program for Data Analysis and Regression Model Identification. *Comput. Chem. Eng.* **2003**, *27*, 701–714.
(19) Nikitin, E. D.; Pavlov, P. A.; Bogatishcheva, N. S. Critical Properties of Long-Chain Substances from the Hypothesis of Functional Self-Similarity. *Fluid Phase Equilib.* **2005**, *235*, 18–23.
(20) Nikitin, E. D.; Popov, A. P.; Simakina, V. A. Vapor Liquid Critical Properties of Some Tetraalkoxysilanes. *J. Chem. Eng. Data* **2008**, *53*, 1371–1374.
(21) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *DRAGON User Manual*; Talete srl: Milano, Italy, 2006.
(22) MATLAB^{R}, ver. 7.04. (R14); copyrighted by The MathWorks Inc., 2005.
(23) Constantinou, L.; Gani, R. A New Group-Contribution Method for the Estimation of Properties of Pure Compounds. *AIChE J.* **1994**, *40*, 1697–1710.
(24) Marrero, J.; Gani, R. Group-Contribution Based Estimation of Pure Component Properties. *Fluid Phase Equilib.* **2001**, *183−184*, 183–208.
(25) Rowley, R. L.; Wilding, W. V.; Oscarson, J. L.; Yang, Y.; Zundel, N. A. *DIPPR Data Compilation of Pure Chemical Properties*; Design Institute for Physical Properties. Brigham Young University: Provo Utah, 2006 (http://dippr.byu.edu).
(26) Horwath, A. L. *Molecular Design*; Elsevier: Amsterdam, 1992.
(27) Labanowski, J. K.; Motoc, I.; Damkoehler, R. A. The Physical Meaning of Tological Indexes. *Comput. Chem.* **1991**, *15*, 47–53.
(28) Mebane, R. C.; Williams, C. D.; Rybolt, T. R. Correlations of Critical properties with Computed Molecular Surface Areas for 11 Homologous Series of 118 Organic Compounds. *Fluid Phase Equilib.* **1996**, *124*, 111–122.

Received for review April 18, 2008. Accepted August 25, 2008.

JE800272X