

Adjustable QSPRs for Prediction of Properties of Long-Chain Substances

Inga Paster and 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

DOI 10.1002/aic.12279

Published online May 12, 2010 in Wiley Online Library (wileyonlinelibrary.com).

The development of quantitative structure property relationships (QSPRs) with good extrapolation capabilities for high carbon number (n_C) substances in homologous series is considered. Based on the available experimental data, molecular descriptors collinear with a particular property are identified. Among these, the ones whose behavior at the limit $n_C \rightarrow \infty$ is similar to the properties behavior, are eventually used for prediction. A linear QSPR in terms of the selected descriptor with an optional additional correction term which exponentially decays with n_C can be developed. The confidence level in the property prediction can be adjusted to the quantity, precision, and reliability of the available data. The proposed method has been tested by developing QSPRs for predicting T_C and P_C for several homologous series and T_m for the n -alkane series. In all cases, the QSPRs developed represent the available experimental data satisfactorily and converge to theoretically accepted values for $n_C \rightarrow \infty$. © 2010 American Institute of Chemical Engineers AIChE J, 57: 423–433, 2011

Keywords: property prediction, QSPR, molecular descriptors, homologous series

Introduction

Physical property data are extensively used in chemical process design, environmental impact assessment, hazard and operability analysis, and additional applications. Critical properties and the acentric factor values are needed, for example, as parameters in equation of state models which are widely used for phase equilibrium calculations. However, measured property values are available only for a small fraction of the chemicals used in the industry, as reactants, products, or side products. Long-chain substances pose special challenges, as their critical constants cannot be measured because of thermal instability. Nikitin et al.¹ noted that critical temperatures (T_C) and pressures (P_C) of n -alkanes have been measured only up to hexatriacontane ($C_{36}H_{74}$) and for 1-alkanols up to 1-docosanol ($C_{22}H_{45}OH$). The critical constants of the heavier members of the homologous series can only be predicted.

Current methods used to predict physical and thermodynamic properties can be classified into group contribution

methods (GC^2), asymptotic behavior correlations ($ABCs^{3-6}$) and various quantitative structure property relationships ($QSPRs^{7-10}$). All of these methods use available experimental data for low carbon number (n_C) compounds to obtain either “group contribution” values or regression model parameter values. The so-obtained group contributions or regression models are used for the prediction of properties of long-chain members of homologous series by extrapolation. The $ABCs$ are nonlinear correlations in terms of n_C , which use in addition to the experimental property data also an estimation of the property value at the limit $n_C \rightarrow \infty$, y^∞ . Kontogeorgis and Tassios¹¹ compared several GC methods and methods that converge to a finite y^∞ value, for predicting T_C , and P_C , of heavy alkanes. They concluded that only methods that converge to finite y^∞ values yield reliable predictions for T_C and P_C of heavy alkanes.

The correlation equations included in the review of Kontogeorgis and Tassios¹¹ can be used for one particular property and many of them are specific to a particular homologous series. Gasem et al.³ introduced a generalized ABC correlation, applicable to several different properties, where y^∞ is included as one of the fitted parameters. Marano and Holder⁴

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

argued that some properties (such as molar mass and molar volume) change linearly with n_C and as such they do not converge to a finite value when $n_C \rightarrow \infty$. They derived a general equation which is applicable to both type of behaviors (finite or infinite y^∞).

The prediction of the y^∞ is based on the Lattice-Fluid models of Kurata and Isida¹² and Sanchez and Lacombe¹³ or the Flory Cell Model.¹⁴ Detailed discussion of these theories and of y^∞ values for critical properties, internal energy and entropy, vapor pressure, boiling point, heat of vaporization, and melting point can be found in Marano and Holder.⁵

It has been realized⁴ that the limiting property values for various homologous series (e.g., n -paraffins, 1-alcohols, and n -alkanoic acids) must be the same as the effect of the particular functional groups (e.g., $-\text{CH}_3$, $-\text{COOH}$, and $-\text{CO}$) diminishes with increasing n_C , where the role of the $-\text{CH}_2-$ chain becomes the dominant one.

The existing correlations for predicting properties of long-chain molecules rely mostly on expression of a change of the properties as nonlinear functions of n_C (and in some cases as function of molecular mass as well). It is generally accepted that extrapolation using nonlinear functions is very risky and unreliable. Cholakov et al.¹⁵ have shown that critical temperature and critical pressure can be represented as linear functions of some molecular descriptors in the regions where experimental data are available, for several homologous series. Linear relationships are more reliable for extrapolation than nonlinear ones. However, Cholakov et al.¹⁵ did not consider the behavior of the selected descriptors at $n_C \rightarrow \infty$, which is also very important when attempting long-range extrapolation to high n_C values.

The objective of this study is to develop QSPRs for homologous series using descriptors that are collinear with the predicted properties in the region where experimental data are available but, in addition, their limiting behavior enables matching to acceptable y^∞ values.

In the Methodology section, the methods for selection of the molecular descriptors and the experimental physical property values that are included in the study, are discussed. In the following section, an introductory example is presented, where one of the descriptors used by Cholakov et al.¹⁵ for predicting the critical temperature is analyzed for its limiting behavior. The general methodology for selecting the appropriate descriptors for representation of properties of long chain molecules is presented in the Basic principles section, and demonstrated for critical temperature of n -alkanes. The descriptor used for n -alkanes is employed for developing linear QSPR's for prediction of critical temperatures of five additional homologous series. Development of linear QSPR's for predicting P_C for six homologous series is discussed in the section entitled Developing a QSPR for the prediction of P_C . A QSPR with an exponential adjustment term for predicting the normal melting temperature of n -alkanes is introduced and tested in the following section. Finally, some conclusions are drawn.

Methodology

To carry out the studies described in this article, we developed a molecular descriptor database for the n -alkane, 1-alkene, 1-alcohol, n -aliphatic acid, aldehyde, and alkyl-ben-

zenes homologous series. Molecular structures of the various compounds for up to $n_C = 330$ were drawn using the HyperChem package (Version 7.01, Hyperchem is copyrighted by Hypercube). No geometry optimization of the structures was carried out. The Dragon program (version 5.5, DRAGON is copyrighted by TALETE srl, <http://www.taletе.mi.it>, Todeschini et al.¹⁶) was used to calculate the descriptors. The user must provide, to this aim, a file describing the structure of the molecule, using one of the standard formats [e.g., Sybil© Mol (*.mol), HyperChem (*.hin), and SMILES notation (*.smi) files]. Using the provided structure file, the Dragon program automatically calculates (in a batch mode) the molecular descriptors. The limit for molecular size in Dragon 5.5 is 1000 atoms per molecule. This limit dictated the maximal n_C (=330) for the molecules used in this study. As for the long-chain homologs the 3D minimized structures were not available, the 3D descriptors were excluded from the data base.

The criteria established by Paster et al.¹⁷ for selecting descriptors which can be appropriate for representing particular properties were used for initial screening of the descriptors. From the remaining descriptors only the ones that can be characterized as having trend of "linear or nearly linear increase (or decrease)" or "nonlinear monotonic increase (or decrease) with decreasing slope" were retained. A total of 377 descriptors fulfilling these criteria were identified and used in the study.

As for physical property data, only measured values which follow the general trend of the homologous series for the particular property were used. For this aim, the experimental data reported by different authors were plotted vs. n_C , and data points that could be considered as outliers were removed. In many cases, the selected property data coincide with the recommended values of the DIPPR¹⁸ and the NIST¹⁹ databases, however data from other sources were also used as necessary. Estimates of the experimental uncertainties were also considered when deciding on the data used for derivation of the QSPRs. The uncertainties provided in the DIPPR database were used (rather than using the uncertainty estimates provided by individual authors), as the DIPPR uncertainties are based on comparison of property values obtained by several researchers.

Predicting T_C For Members of the n -Alkane Homologous Series—An Introductory Example

In Table 1, the experimental T_C values of members of the n -alkane homologous series are shown. The values up to $n_C = 4$ were reported by Ambrose and Tsonopoulos²⁰ and the values between $5 \leq n_C \leq 36$ were reported by Nikitin et al.²¹ The uncertainties of the T_C values (as determined by the DIPPR staff and published in the DIPPR database¹⁸) are also shown in Table 1. The uncertainties for low n_C compounds (up to $n_C = 17$) are <0.2% and they increase up to < 3% for compounds of higher n_C . Plotting T_C vs. n_C (see Figure 1) shows nonlinear change of the critical temperature with the carbon number for the n -alkane series, thus extrapolating to higher carbon numbers can be very risky.

It is possible to identify several molecular descriptors which are collinear with the T_C data shown in Table 1. One such descriptor is the *IVDM* (mean-information content, also

Table 1. Data for Modeling Critical Temperature of *n*-Alkanes with the *IVDM* Descriptor

n_C	Critical Temperature (K, Experimental)			Descriptor <i>IVDM</i>	Predicted T_C	
	Value	Source*	Uncertainty [†]		Value (K)	Uncertainty (%)
1	190.5641	A	<0.2%	Undefined		
2	305.321	A	<0.2%	1.00	299.12	2.03
3	369.831	A	<0.2%	1.50	369.36	0.13
4	425.121	A	<0.2%	1.92	428.11	-0.70
5	469.7	B	<0.2%	2.25	474.71	-1.07
6	507.6	B	<0.2%	2.52	512.91	-1.05
7	540.2	B	<0.2%	2.75	545.17	-0.92
8	568.7	B	<0.2%	2.95	573.07	-0.77
9	594.6	B	<0.2%	3.13	597.62	-0.51
10	617.7	B	<0.2%	3.28	619.54	-0.30
11	639	B	<0.2%	3.42	639.33	-0.05
12	658	B	<0.2%	3.55	657.37	0.10
13	675	B	<0.2%	3.67	673.94	0.16
14	693	B	<0.2%	3.78	689.26	0.54
15	708	B	<0.2%	3.88	703.50	0.64
16	723	B	<0.2%	3.97	716.82	0.86
17	736	B	<0.2%	4.06	729.31	0.91
18	747	B	<1%	4.15	741.08	0.79
19	755	B	<1%	4.23	752.20	0.37
20	768	B	<1%	4.30	762.75	0.68
21	778	B	<3%	4.37	772.77	0.67
22	786	B	<3%	4.44	782.33	0.47
23	790	B	<3%	4.50	791.45	-0.18
24	800	B	<3%	4.57	800.18	-0.02
26	816	B	<3%	4.68	816.59	-0.07
28	824	B	<3%	4.79	831.77	-0.94
30	843	B	<3%	4.89	845.89	-0.34
36	872	B	<3%	5.16	883.17	-1.28

*Source: A. Ambrose and Tsonopoulos²⁰; B. Nikitin et al.²¹

[†]Source: DIPPR database.¹⁸

called Shannon's entropy, Shannon and Weaver,²² on the vertex degree magnitude, which is a measure of the molecular complexity) descriptor. This descriptor was selected by Cholakov et al.¹⁵ for representation of T_C for the *n*-alkane series in a linear QSPR. This descriptor belongs to the "information indices" and is based on the partition of vertices according to the vertex degree magnitude, the vertex degree of an atom being the number of connected non-H atoms. This index was proposed as a measure of molecular complexity by Raychaudhury et al.²³ It is defined as

$$IVDM = - \sum_{a=1}^A \frac{\delta_a}{A_V} \cdot \log_2 \frac{\delta_a}{A_V}, \quad (1)$$

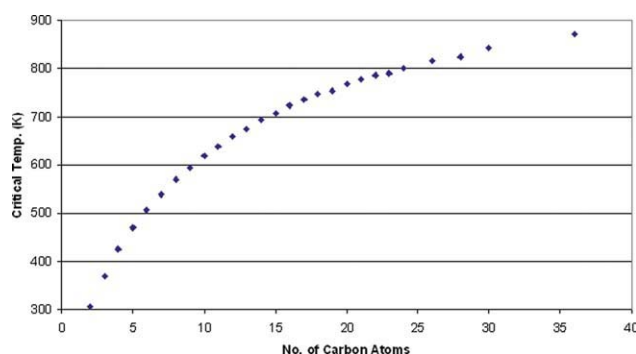


Figure 1. Plot of the critical temperatures of *n*-alkanes vs. the number of carbon atoms.

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

where δ_a is the vertex degree of the *a*-th atom, *A* is the atom number, and A_V is total adjacency index. For the *n*-alkane homologous series, the *IVDM* descriptor can be expressed as a function of n_C (see Appendix A)

$$IVDM = \left[1 + \frac{\ln(n_C - 1)}{\ln 2} - \frac{n_C}{n_C - 1} + \frac{2}{n_C - 1} \right]. \quad (2)$$

Using Eq. 2, the *IVDM* descriptor values can be calculated for all the compounds for which experimental T_C values are available (except for methane, $n_C = 1$, for which the descriptor is undefined). The calculated values of *IVDM* were entered into Table 1. Plotting T_C vs. *IVDM* (Figure 2) yields a linear relationship (QSPR): $T_{C,pred} = 158.65 +$

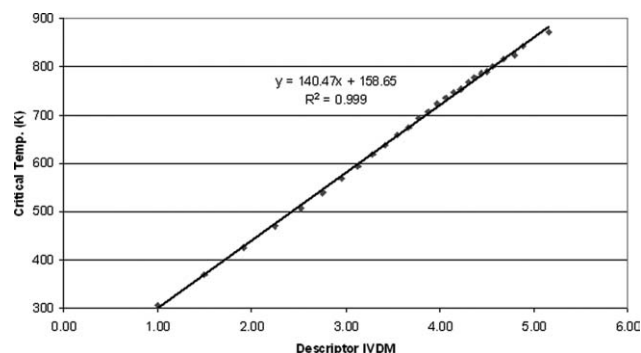


Figure 2. Plot of the critical temperatures of *n*-alkanes vs. the molecular descriptor *IVDM*.

Table 2. Descriptors Collinear with T_C for the n -Alkane Group in the Range of $5 \leq n_C \leq 36$

Descriptor*	R^2	Variance (K)	β_0	β_1
<i>IVDE</i>	0.999	12.638	1045.911 ± 5.002	-591.805 ± 8.271
<i>IVDM</i>	0.998	23.734	154.839 ± 10.733	141.416 ± 2.710
<i>ATS1e</i>	0.998	27.450	139.117 ± 11.863	208.200 ± 4.291
<i>ATS2e</i>	0.999	14.785	201.404 ± 7.779	190.919 ± 2.886
<i>ESpm01r</i>	0.998	27.450	139.117 ± 11.863	208.200 ± 4.291
<i>HVcpx</i>	0.997	42.440	172.749 ± 13.913	152.282 ± 3.905
<i>IDDM</i>	0.998	27.957	143.480 ± 11.883	144.460 ± 3.005
<i>IDE</i>	0.999	17.233	215.834 ± 8.169	136.745 ± 2.232
<i>IDM</i>	0.998	22.435	248.084 ± 8.737	71.006 ± 1.323
<i>piPC01</i>	0.998	27.450	139.117 ± 11.863	208.200 ± 4.291
<i>piPC02</i>	0.999	14.785	201.404 ± 7.779	190.919 ± 2.886

*Definitions of the descriptors are available in Todeschini et al.¹⁶ and Todeschini and Consonni.²⁵

$140.47 \times IVDM$, with a correlation coefficient of $R^2 = 0.999$. This linear equation was used to predict the $T_{C,pred}$ values and the associated prediction uncertainties: $(T_C - T_{C,pred}) \times 100/T_C$. The predicted critical temperature and the prediction uncertainty values are shown in Table 1. The prediction uncertainty is $<1\%$ for all but four compounds, and it is higher than the experimental uncertainty in the region where high precision experimental data are available. In any case, such level of precision can be considered very high for a QSPR with only one descriptor (two adjustable parameters).

Equation 2 and the QSPR developed can be used for extrapolating to high carbon number compounds. For example, for $n_C = 100$ the value $T_{C,pred} = 1091.29$ K is obtained. This value is within the range of values obtained by Kontogeorgis and Tassios,¹¹ who compared several methods for predicting T_C for long-chain alkanes. An additional test for the appropriateness of the selected descriptor is its value at the limit of $n_C \rightarrow \infty$. The critical temperature is commonly categorized as a property which approaches a finite value for large carbon numbers. At the $n_C \rightarrow \infty$ limit, addition of more carbon units to the chain has no effect on T_C . As there is a linear relationship between the descriptor *IVDM* and T_C , a constant value of T_C at the limit requires constant value of *IVDM*, as well. However at the limit $n_C \rightarrow \infty$, *IVDM* = ∞ . Because of that, the *IVDM* descriptor cannot be used for extrapolation of T_C to high n_C values, in spite of its ability to represent well the available experimental data.

Developing QSPRs for Prediction of Properties at High n_C Values—Basic Principles

Following the example in the previous section, the first step of the development of the QSPR is the identification of molecular descriptors which are highly correlated with the property values of compounds included in the training set, for which experimental values of the desired property are available. Selection of the compounds to the training set must take into account the fact that the functional groups (such as $-\text{CH}_3$, $-\text{COOH}$, $-\text{CO}$, etc.) effects are dominant for low n_C members of the homologous series. These effects diminish with the increase of n_C , where the $-\text{CH}_2-$ chain becomes the dominant one.

To identify the descriptors collinear with the T_C values of n -alkanes a training set, which includes the compounds in the range of $n_C = 5$ through $n_C = 36$ (Table 1) were used. The results regarding the descriptors that are of the highest correlation with T_C for the training set, are shown in Table 2. In this table, the correlation coefficient (R^2) between the descriptor ζ and T_C , the coefficients β_0 and β_1 of the linear relationship

$$T_C = \beta_0 + \beta_1 \zeta, \quad (3)$$

and the variance of Eq. 3 are shown. Observe that the correlation coefficients in all cases are in the range of $0.997 \leq R^2 \leq 0.999$ and the confidence limits on the parameter values β_0 and β_1 are narrow, thus all of the descriptors shown represent the data well.

Additional information regarding the quality of the fit can be obtained from the residual plots. In Figure 3 the residuals: $(T_C - T_{C,pred}) \times 100/T_C$ are shown for the case when the *IVDE* descriptor and Eq. 3 (with parameter values of $\beta_0 = 1045.911$ and $\beta_1 = -591.805$) are used for prediction. The maximal residual value (uncertainty) is $\sim 1\%$. While the residual distribution is nonrandom, its trend can be partially explained by some increase in the residual values in the range $500 \text{ K} \leq T_C \leq 550 \text{ K}$ (caused possibly by the influence of the $-\text{CH}_3$ end groups for low carbon number compounds), and some increase in the region when $T_C > 850 \text{ K}$ (high carbon number compounds where the experimental uncertainty is also higher, see Table 1).

The next step is to identify those descriptors that converge to a finite limit at $n_C \rightarrow \infty$, and their QSPR yields a T_C^∞ value that is consistent with the values that have been suggested previously.⁶ Let us consider, in this respect, the *IVDE* descriptor. The symbol *IVDE* corresponds to the mean information content of the vertex degree equality. This descriptor belongs to the “information indices,” and is based on the partition of vertices according to vertex degree equality, the vertex degree of an atom being the number of connected non-H atom.²³ It is defined as

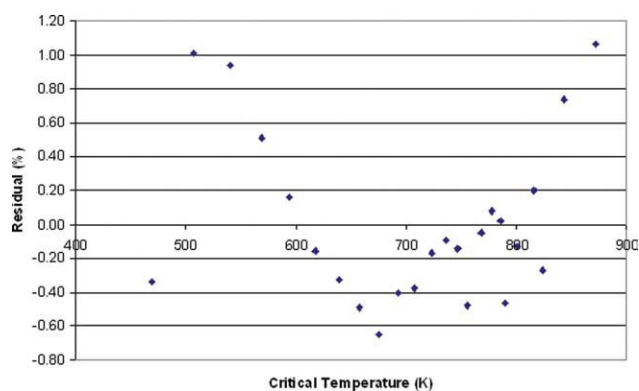


Figure 3. Residual plot of the critical temperatures predicted by the *IVDE* descriptor for n -alkanes.

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

Table 3. Critical Temperature Data for Five Homologous Series

n_C	1-Alkenes		1-Alcohols		<i>n</i> -Aliphatic Acids		Aldehydes		<i>n</i> -Alkyl Benzenes	
	T_C (K)	Source*	T_C (K)	Source*	T_C (K)	Source*	T_C (K)	Source*	T_C (K)	Source*
1			512.5	A	588	A				
2	282.34	A	514	A	591.95	A				
3	364.85	A	536.8	A	600.81	A	504.4	A		
4	419.5	A	563.05	D	615.7	A	537.2	A		
5	464.8	A	588.1	A	639.16	A	566.1	A		
6	504	A	610.3	E	660.2	A	591	A	562.05	A
7	537.3	B	632.6	E	677.3	A	616.8	A	591.75	A
8	567	B	652.5	E	694.26	A	638.9	A	617.15	A
9	594	B	670.7	E	710.7	A	659	H	638.35	A
10	617	B	687.3	E	722.1	A	674	H	660.5	A
11	–	B	703.6	E	731.26	A				
12	658	B	719.4	E	743	A				
13	673	C	732	F	–					
14	691	C	743	F	763	A				
15	705	C	757	F	777	G				
16	718	C	770	F	785	A				
17	734	C	780	F	792	A				
18	747	C	790	F	803	A				
19	755	C			–					
20	772	C			820	A				
21					–					
22					837	G				

*Sources: A. DIPPR database recommended value¹⁸; B. Tsionopoulos and Ambrose²⁷; C. Nikitin and Popov²⁸; D. Ambrose and Walton²⁹; E. Gude and Teja³⁰; F. Nikitin et al.³¹; G. Nikitin et al.¹; H. Tsionopoulos et al.³²

$$IVDE = - \sum_{g=1}^G \frac{F}{A} \log_2 \frac{F}{A}, \quad (4)$$

where F is the vertex degree count, A is the atom number, and G is maximal vertex degree value. For the n -alkane homologous series the $IVDE$ descriptor can be expressed as function of n_C

$$IVDE = - \left[\frac{2}{n_C} \log_2 \frac{2}{n_C} + \frac{n_C - 2}{n_C} \log_2 \left(\frac{n_C - 2}{n_C} \right) \right], \quad (5)$$

whereby

$$\lim_{n_C \rightarrow \infty} [IVDE] \rightarrow 0.$$

Thus, $IVDE$ converges to a constant, zero value, where Eq. 3 yields $T_C^\infty = 1045.911 \pm 5.003$ K. In the literature several T_C^∞ values are suggested in the 900–1357 K range.^{11,6} The value obtained using the $IVDE$ descriptor is well within this range. This value is identical with the value that was recently reported by Giles et al.²⁴ ($T_C^\infty = 1050$ K) and it is very close to the value recommended, for example, by Marano and Holder⁵ ($T_C^\infty = 1020.71$ K). Thus, the $IVDE$ descriptor satisfies all the requirements for a reliable representation of T_C for high n_C n -alkanes (starting at $n_C = 5$).

Table 4. Expressions for Calculating the Descriptor $IVDE$ for Various Homologous Series

Homologous Series	Formula for Calculation of $IVDE$	$\lim_{n_C \rightarrow \infty} IVDE$
1-Alkenes	$IVDE = - \left[\frac{2}{n_C} \log_2 \frac{2}{n_C} + \frac{n_C - 2}{n_C} \log_2 \left(\frac{n_C - 2}{n_C} \right) \right]$	0
1-Alcohols	$IVDE = - \left[\frac{2}{n_C + 1} \log_2 \frac{2}{n_C + 1} + \frac{n_C - 1}{n_C + 1} \log_2 \left(\frac{n_C - 1}{n_C + 1} \right) \right]$	0
<i>n</i> -Aliphatic acids	$IVDE = - \left[\frac{3}{n_C + 2} \log_2 \frac{3}{n_C + 2} + \frac{1}{n_C + 2} \log_2 \frac{1}{n_C + 2} + \frac{n_C - 2}{n_C + 2} \log_2 \left(\frac{n_C - 2}{n_C + 2} \right) \right]$	0
Aldehydes	$IVDE = - \left[\frac{2}{n_C + 1} \log_2 \frac{2}{n_C + 1} + \frac{n_C - 1}{n_C + 1} \log_2 \left(\frac{n_C - 1}{n_C + 1} \right) \right]$	0
<i>n</i> -Alkyl benzenes	$IVDE = - \left[2 \cdot \frac{1}{n_C} \log_2 \frac{1}{n_C} + \frac{n_C - 2}{n_C} \log_2 \left(\frac{n_C - 2}{n_C} \right) \right]$	0

Based on the correlation coefficient, the variance and the parameter confidence intervals of the QSPRs obtained with the other descriptors presented in Table 2, they all can be considered as potential descriptors for developing QSPRs for the representation of T_C of n -alkanes at high n_C . However, to verify their appropriateness for long-range extrapolation, their limiting value at $n_C \rightarrow \infty$ should be analyzed. Detailed description of these descriptors are available in Todeschini et al.¹⁶ and Todeschini and Consonni.²⁵

Use of the IVDE Descriptor for Prediction of T_C for Several Homologous Series

In Table 3, the experimental T_C data for five homologous series are shown. The number and the range of the available data are the following: 1-alkenes, 15 data points, up to $n_C = 20$; 1-alcohols, 13 data points, up to $n_C = 18$; n -aliphatic acids, 14 data points, up to $n_C = 22$; aldehydes, six data points, up to $n_C = 10$; and alkyl-benzenes, four data points, up to $n_C = 10$.

Expressions similar to Eq. 5 can be derived for calculation of $IVDE$ for the five series. These expressions can be used for establishing the value of $IVDE$ at $n_C \rightarrow \infty$. The expressions for calculating $IVDE$ and the limiting $IVDE$ values are shown in Table 4. Observe that the expression for $IVDE$ of 1-alkenes is identical to Eq. 5, and the expressions for 1-alcohols and aldehydes are also identical. The limiting value of $IVDE$ (at $n_C \rightarrow \infty$) is 0 for all the series.

The experimental data for the six homologous series (including n -alkanes) are plotted vs. n_C in Figure 4. Excluding the range of low n_C values, the trend of the change of T_C is similar for all the series. The abrupt change in trend at low n_C is mostly discernible for the 1-alcohol and n -aliphatic acid data.

A plot of $(-1) \times IVDE$ vs. n_C for the six homologous series is shown in Figure 5 (the values overlap for the n -alkane, 1-alkene and for the 1-alcohol, aldehyde series). Observe that the trends of the curves are similar to the trends of the T_C data except for the $n_C < 6$ range (in particular for the n -alkane and n -aliphatic acid series).

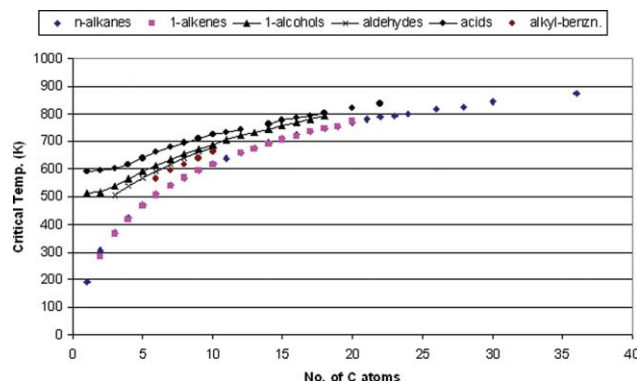


Figure 4. Plot of the critical temperatures data of six homologous series vs. the number of carbon atoms.

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

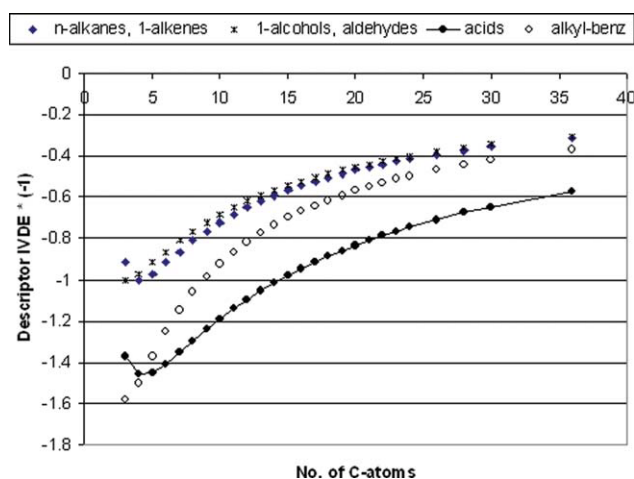


Figure 5. Plot of the $IVDE$ descriptor $\times (-1)$ of six homologous series vs. the number of carbon atoms.

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

Figure 6 shows T_C values plotted vs. $IVDE$ for the n -aliphatic acid series. Removing the first three data points (corresponding to $n_C = 3, 4$, and 5), the T_C values align along a straight line with $\beta_0 = 1043.915$ and $\beta_1 = -272.033$. The value of β_0 (thus T_C^∞) is essentially the same as for the n -alkane series, confirming that T_C^∞ must be the same for all homologous series based on the $-\text{CH}_2-$ chain. The T_C data in Figure 6 exhibits some curvature which cannot be explained by a linear equation based on one descriptor. Indeed Brauner et al.¹⁰ have shown that often more than one descriptor (typically two for T_C) are needed to represent the behavior of a property with high precision and random residual, for the range of n_C values used here. We assume, however, that the additional curvature is caused by the influence of the specific functional group ($-\text{COOH}$ in this particular case) and its effect diminishes for long-chain molecules.

QSPRs of the form of Eq. 3 were fitted to the six homologous series using the available experimental data. The T_C values of the first few members of the series were removed

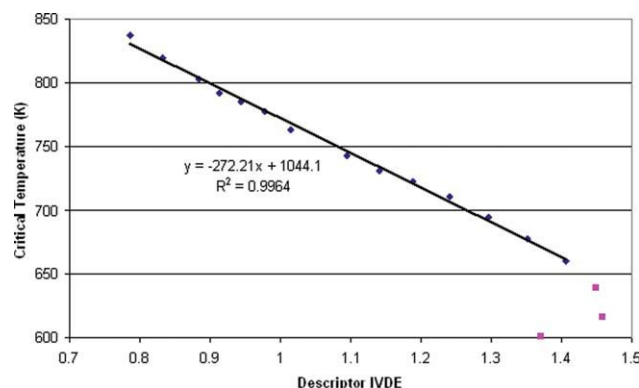


Figure 6. Plot of the critical temperatures of n -aliphatic acids vs. the molecular descriptor $IVDE$.

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

Table 5. Linear (2-Parameter) QSPR Parameter Values and Statistics for Modeling of T_C with the $IVDE$ Descriptor for Six Homologous Series

Homologous Series	Data Points Used	R^2	Variance (K)	β_0	β_1
<i>n</i> -Alkanes	$n_C = 5 - 36$ (24 points)	0.999	12.638	1045.911 ± 5.003	-591.805 ± 8.271
1-Alkenes	$n_C = 5 - 20$ (15 points)	0.999	11.487	1041.295 ± 8.137	-588.173 ± 11.871
1-Alcohols	$n_C = 6 - 18$ (13 points)	0.997	10.483	1015.897 ± 11.032	-475.353 ± 16.944
<i>n</i> -Aliphatic acids	$n_C = 6 - 22$ (14 points)	0.996	11.740	1043.915 ± 11.363	-272.033 ± 10.393
Aldehydes	$n_C = 5 - 10$ (6 points)	0.999	1.395	995.247 ± 13.313	-467.384 ± 16.689
<i>n</i> -Alkyl benzenes	$n_C = 7 - 10$ (4 points)	0.998	2.478	935.835 ± 41.291	-300.046 ± 39.972

Table 6. Linear (1-Parameter, $\beta_0 = 1046$) QSPR Parameter Values and Statistics for Five Homologous Series

Homologous Series	Data Points Used	Points	R^2	Variance (K)	β_1
1-Alkenes	$n_C = 5 - 20$, $n_C = 11$ is missing	15	0.999	11.946	-594.850 ± 2.793
1-Alcohols	$n_C = 6 - 18$	13	0.987	42.042	-520.827 ± 6.019
<i>n</i> -Aliphatic acids	$n_C = 6 - 22$	14	0.996	10.844	-273.915 ± 1.739
Aldehydes	$n_C = 5 - 10$	6	0.981	32.894	-530.582 ± 7.545
<i>n</i> -Alkyl benzenes	$n_C = 7 - 10$	4	0.873	109.562	-406.291 ± 16.120

from the regression, wherever possible, to minimize the influence of the unique functional groups. Summary of the results for the six series is presented in Table 5. The correlation coefficients in all cases are in the range of $0.996 \leq R^2 \leq 0.999$ and the confidence limits on the parameter values β_0 and β_1 are narrow (except for the *n*-alkyl benzene series), thus the $IVDE$ descriptor represents the data well. The value of β_0 (T_C^∞) is about the same (considering the confidence intervals) for the *n*-alkane, 1-alkene, and *n*-aliphatic acid series. There is however discrepancy in the β_0 values of the 1-alcohol, aldehyde and alkyl benzene series. We assume that these discrepancies are caused by larger influence of the unique functional groups, mainly due to too few $-\text{CH}_2-$ groups in the compounds included in the training set (in particular in the case of aldehydes and *n*-alkylbenzenes).

Assuming $T_C^\infty = 1046$ K in all the tested series, the experimental data was regressed by setting $\beta_0 = 1046$ (a fixed value) and determining only β_1 . The results of this study are shown in Table 6. Observe that for the 1-alkene and *n*-aliphatic acid series the statistics (variance, R^2 , confidence intervals) are even better than in the case of the two adjustable parameter models. As for the rest of the series, the quality of the fit somewhat deteriorates for the reasons mentioned above.

Developing a QSPR for the Prediction of P_C for *n*-alkanes at High n_C Values

Table 7 shows recommended P_C data (from the DIPPR database¹⁸) for the members of the *n*-alkane homologous series. The values up to $n_C = 17$ are experimental, while the values between $18 \leq n_C \leq 36$ are predicted. The uncertainties of the T_C values (as determined by the DIPPR staff) vary widely from $<0.2\%$ for the low n_C compounds up to $<25\%$ for all the compounds with $n_C \geq 14$ (Table 7).

To identify descriptors collinear with the P_C values, a training set of *n*-alkanes that includes 10 compounds of the highest n_C values for which experimental data are available is used, thus $8 \leq n_C \leq 17$. The results regarding the descriptors that are of the highest correlation with P_C for the training set, are shown in Table 8. The correlation coefficients in all cases are in the range of $0.985 \leq R^2 \leq 0.998$. These val-

ues are lower than the values for T_C (Table 2). This discrepancy can, however, be attributed to the much higher uncertainty level of the P_C data.

The descriptor with the highest correlation coefficient is the $HNar$ descriptor. This descriptor belongs to the

Table 7. Data for Modeling Critical Pressure of *n*-Alkanes with the $HNar$ Descriptor

n_C	Critical Pressure (MPa)		Descriptor	Predicted P_C	
	Value*	Uncertainty [†]		Value (MPa)	Uncertainty (%)
1	4.599	$<0.2\%$			
2	4.872	$<0.2\%$			
3	4.248	$<0.2\%$	1.00	6.04	-23.92
4	3.796	$<0.2\%$	1.2000	4.85	-14.18
5	3.37	$<1\%$	1.3333	4.06	-6.93
6	3.025	$<1\%$	1.4286	3.49	-3.67
7	2.74	$<3\%$	1.5000	3.07	-1.48
8	2.49	$<3\%$	1.5556	2.74	0.01
9	2.29	$<3\%$	1.6000	2.48	0.55
10	2.11	$<3\%$	1.6364	2.26	1.30
11	1.95	$<3\%$	1.6667	2.08	1.41
12	1.82	$<5\%$	1.6923	1.93	1.11
13	1.68	$<10\%$	1.7143	1.80	1.22
14	1.57	$<10\%$	1.7333	1.68	-0.30
15	1.48	$<25\%$	1.7500	1.59	-1.01
16	1.4	$<25\%$	1.7647	1.50	-1.26
17	1.34	$<25\%$	1.7778	1.42	-1.49
18	1.27	$<25\%$	1.7895	1.35	-0.85
19	1.21	$<25\%$	1.7895	1.29	-1.50
20	1.16	$<25\%$	1.8000	1.23	-1.88
21	1.11	$<25\%$	1.8095	1.18	-1.82
22	1.06	$<25\%$	1.8182	1.13	-2.18
23	1.02	$<25\%$	1.8261	1.09	-2.97
24	0.98	$<25\%$	1.8333	1.05	-3.11
25	0.95	$<25\%$	1.8400	1.01	-3.56
26	0.91	$<25\%$	1.8462	0.98	-3.27
27	0.883	$<25\%$	1.8519	0.95	-4.42
28	0.85	$<25\%$	1.8571	0.92	-4.25
29	0.826	$<25\%$	1.8621	0.89	-5.08
30	0.8	$<25\%$	1.8667	0.87	-5.05
32	0.75	$<25\%$	1.8710	0.84	-5.49
36	0.68	$<25\%$	1.8750	0.80	-6.67
			1.8824	0.73	-6.91
			1.8947		

*Experimental values are shown in bold, source: DIPPR database recommended value.¹⁸

[†]Source: DIPPR database.¹⁸

Table 8. Descriptors Collinear with P_C for the n -Alkane Group in the Range of $8 \leq n_C \leq 17$

Descriptor*	R^2	Variance (MPa)	β_0	β_1
<i>HNar</i>	0.998	0.0004	12.535 ± 0.403	-6.261 ± 0.235
<i>BELm7</i>	0.993	0.001	3.386 ± 0.112	-1.629 ± 0.113
<i>BELv7</i>	0.993	0.001	3.049 ± 0.088	-1.498 ± 0.102
<i>BELp7</i>	0.993	0.001	2.935 ± 0.079	-1.450 ± 0.097
<i>GNar</i>	0.996	0.001	14.912 ± 0.679	-7.362 ± 0.381
<i>VRD2, VRZ2, VRe2, VRm2, VRp2, VRv2</i>	0.996	0.001	13.995 ± 0.619	-13.813 ± 0.702
<i>AAC, IC0</i>	0.985	0.003	75.477 ± 7.296	-81.912 ± 8.113
<i>VRA2</i>	0.996	0.001	13.230 ± 0.575	-13.005 ± 0.655
<i>J, JhetZ, Jhete, Jhetm, Jhetp, Jhetv</i>	0.995	0.001	12.050 ± 0.576	-3.761 ± 0.211
<i>RBF</i>	0.995	0.001	5.970 ± 0.232	-17.179 ± 0.955
<i>AMW</i>	0.991	0.002	37.687 ± 2.846	-8.006 ± 0.635

*Definitions of the descriptors are available in the references: Todeschini et al.¹⁶ and Todeschini and Consonni.²⁵

“topological index” category (Narumi’s²⁶ harmonic topological index) and defined as

$$HNar = \frac{A}{\sum_{i=1}^A \frac{1}{\delta_i}} \quad (6)$$

where A is the number of atoms in the molecule and δ is the vertex degree in H-depleted molecule (number of adjacent atoms). For the n -alkane homologous series, the $HNar$ descriptor can be expressed as function of n_C

$$HNar = \frac{n_C}{2 + \frac{n_C - 2}{2}} \quad (7)$$

The values of the $HNar$ descriptor were calculated using Eq. 7 and entered into Table 7.

In Figure 7, the P_C values are plotted vs. the $HNar$ descriptor. Observe that for the training set the straight line $P_C = 12.535 - 6.261 \times HNar$ represents the data well. The predicted P_C data for $18 \leq n_C \leq 36$ seem to align along the same straight line. The data for low n_C values cannot be represented by a straight line, as expected.

Using Eq. 7, $\lim_{n_C \rightarrow \infty} [HNar] \rightarrow 2$. Introducing the limiting value of $HNar$ into the linear equation, and taking into account the parameter confidence intervals shown in Table 8, yields $P_C^\infty = 0.03 \approx 0$ MPa. Most of the researchers (e.g., Nikitin et al.,⁶ Kontogeorgis and Tassios¹¹) argued that $P_C^\infty = 0$ MPa. Thus, the critical pressure of n -alkanes can be satisfactorily represented by the QSPR: $P_C = 6.2 \times (2 - HNar)$ for $n_C > 8$.

Similar study was carried out for the 1-alkene, 1-alcohol, aldehyde, and the n -aliphatic acid series (the n -alkyl benzene were not included because of the lack of experimental data). In all cases, the correlation coefficient between the $HNar$ descriptor and the experimental data was in the range of $0.993 \leq R^2 \leq 0.9997$. In all these cases, the confidence intervals on the P_C^∞ are wider than the values themselves implying that $P_C^\infty = 0$ MPa.

Developing a QSPR for the Prediction of Normal Melting Point for n -alkanes at High n_C Values

Measured normal melting point (T_m) data for n -alkanes are available for up to $n_C = 100$ (Table 9). The precision of the data is usually high (uncertainty in the range of 0.2–3%), in comparison to the data for other properties of the same compounds. At lower carbon numbers n -alkanes melt from different crystalline phases: triclinic, hexagonal, and ortho-

rhombic, as shown in Table 9. Consequently, the change of T_m with n_C is nonmonotonic up to about $n_C = 20$, as shown in Figure 8. Starting at $n_C = 20$ and up to $n_C = 100$ the change appears as monotonic (see Figures 8 and 9). Therefore, the 10 compounds in the range: $21 \leq n_C \leq 30$ were used as training set in order to identify the descriptors collinear with T_m . The descriptor with highest correlation with T_m is the $IVDE$ descriptor, the same descriptor that was used for modeling T_C . The coefficients of the linear relationship (Eq. 3) and the associated statistics are shown in Table 10. Observe that the correlation coefficient is very close to one ($R^2 = 0.9997$), the confidence limits on the parameter values β_0 and β_1 are narrow and the residual plot (not shown) exhibits random distribution. Thus, the $IVDE$ descriptor represents the data of the training set well, with a linear QSPR: $T_m = 428.196 - 252.652 \cdot IVDE$. As $\lim_{n_C \rightarrow \infty} [IVDE] \rightarrow 0$ (see Eq. 5), this linear QSPR yields $T_m^\infty = 428.2$ K. Marano and Holder⁵ cite several references that put T_m^∞ within the range of 413–418 K. Those estimates are based mainly on data for the melting point of linear polyethylenes. Thus, the T_m^∞ predicted by the QSPR may be too high. Indeed, using different training sets for developing the QSPR yield lower T_m^∞ , as shown in Table 10. For the training set $31 \leq n_C \leq 52$ $T_m^\infty = 422.4$ K, for the training set $50 \leq n_C \leq 100$ $T_m^\infty = 421.6$ K and for a larger training set of 30 points in the range of set $31 \leq n_C \leq 52$ $T_m^\infty = 423.25$ K.

It is not within the scope of the present article to try to determine a compelling value for T_m^∞ . However, to adjust the QSPR to a generally accepted y^∞ value, Eq. 3 can be modified by including an additional, empirical correction term

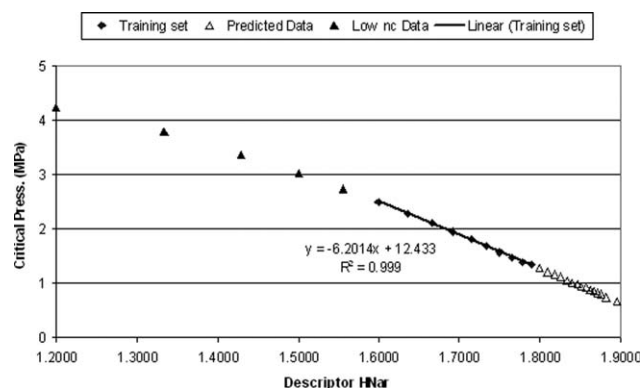


Figure 7. Plot of the critical pressures of n -alkanes vs. the molecular descriptor $HNar$.

Table 9. Normal Melting Point Data for the *n*-Alkane Homologous Series

n_C	Normal Melting Point			Crystalline
	Value (K)	Source*	Uncertainty	Phase [†]
1	90.694	A	<0.2%	T
2	90.352	A	<0.2%	T
3	85.47	A	<0.2%	T
4	134.86	A	<0.2%	T
5	143.42	A	<0.2%	T
6	177.83	A	<0.2%	T
7	182.57	A	<0.2%	T
8	216.38	A	<0.2%	T
9	219.66	A	<1%	T
10	243.51	A	<1%	T
11	247.571	A	<1%	H
12	263.568	A	<1%	T
13	267.76	A	<0.2%	H
14	278.7	B	±0.9 K	T
15	283.072	A	<0.2%	H
16	291.308	A	<0.2%	T
17	295.134	A	<0.2%	H
18	301.31	A	<0.2%	T
19	305.04	A	<0.2%	H
20	309.58	A	<0.2%	T
21	313.35	A	<1%	H
22	317.15	A	<1%	H
23	320.65	A	<1%	H
24	323.75	A	<3%	H
25	326.65	A	<1%	H
26	329.25	A	<1%	H
27	332.15	A	<1%	H
28	334.35	A	<1%	H
29	336.85	A	<1%	H
30	338.65	A	<1%	H
31	341.5	B	±0.9 K	H
32	342.35	A	<1%	H
33	345	B	±3 K	H
34	346	B	±0.7 K	H
35	348	B	±2 K	H
36	349.05	A	<1%	H
44	359.6	C		O
46	361.2	C		O
50	365.3	C		O
52	367.2	C		O
54	368.2	C		O
60	372.4	C		O
62	373.7	C		O
64	375.3	C		O
66	376.8	C		O
67	377.3	C		O
70	378.5	C		O
82	383.5	C		O
94	387	C		O
100	388.4	C		O

*Sources: A. DIPPR database recommended value¹⁵; B. NIST database¹⁶; C: Broadhurst.³³

[†]Crystalline phase: T—triclinic, H—hexagonal, and O—orthorhombic.

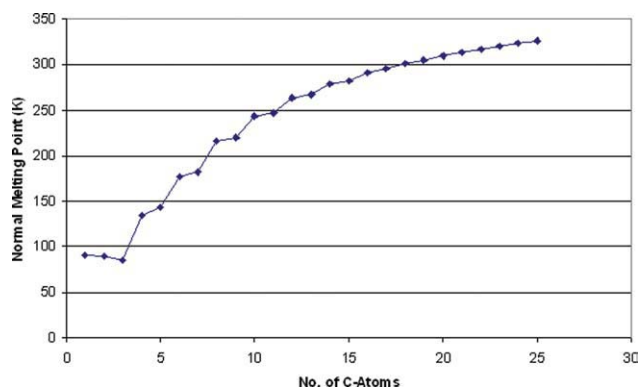


Figure 8. Plot of normal melting point of *n*-alkanes vs. the number of carbon atoms, up to $n_C = 25$.

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

$$y = \beta_0 + \beta_1 \zeta - (\beta_0 + \beta_1 \zeta - \beta^\infty)[1 - \exp(-\beta_2 n_C)] \quad (8)$$

where y is the predicted property, β^∞ is an accepted y^∞ value and β_2 is a regression parameter obtained using preferably as many experimental data points as possible (with high enough n_C values).

In the case of the prediction of T_m , we used a training set of 30 compounds in the range of $21 \leq n_C \leq 100$, the β^∞ value was set at 415 and nonlinear regression was carried out to determine the optimal values of β_0 , β_1 , and β_2 . The results obtained are $\beta_0 = 420.9248$, $\beta_1 = -239.258$, and $\beta_2 = 1/1137.644$. Using these parameter values in Eq. 8 yields the desired T_m^∞ value of 415 K and keeps the variance value at about the same as that obtained by the linear model for the same training set (see Table 10).

Conclusions

QSPRs based on the available experimental data and limiting behavior of the selected molecular descriptors were fitted to T_C and P_C data of several homologous series and T_m data of the *n*-alkane series, to enable prediction of these properties for long-chain substances.

Linear QSPRs containing the *IVDE* descriptor and two adjustable parameters found to be sufficient to represent the T_C of high n_C members of the six homologous series considered. In cases the prediction requires long-range extrapolation, the reliabilities of the QSPRs are limited due to the uncertainty in the theoretical value of T_C^∞ . The reliability of the QSPRs can be increased, by addition of a nonlinear term (if necessary) when T_C^∞ values of higher certainty will become available.

Table 10. Linear (2-Parameter) QSPR Parameter Values and Statistics for the Modeling of T_m for *n*-Alkanes with the *IVDE* Descriptor for Various Training Sets

Training Set	R^2	Variance (K)	β_0	β_1
$21 \leq n_C \leq 30$ (10 points)	0.9997	0.0278	428.196 ± 1.526	-252.652 ± 3.808
$31 \leq n_C \leq 52$ (10 points)	0.9988	0.134	422.4247 ± 2.031	-235.9398 ± 6.795
$50 \leq n_C \leq 100$ (10 points)	0.9962	0.179	421.586 ± 2.22	-231.5338 ± 11.708
$21 \leq n_C \leq 100$ (30 points)	0.9996	0.235	423.355 ± 0.605	-240.133 ± 1.96

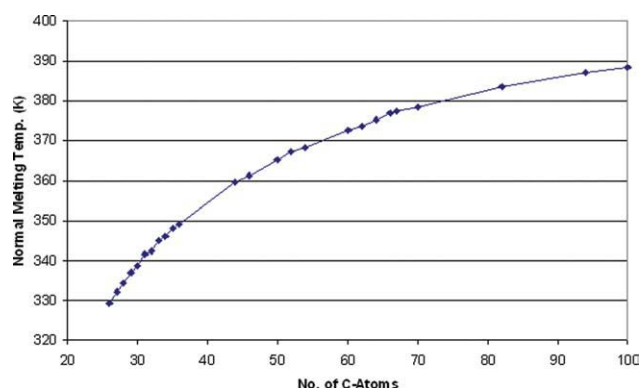


Figure 9. Plot of normal melting point of n -alkanes vs. the number of carbon atoms in the range: $26 \leq n_C \leq 100$.

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

Linear QSPRs containing the $HNar$ descriptor and two adjustable parameters found to be sufficient to represent the P_C of high n_C members of the five homologous series considered. While there is a general agreement regarding the value of P_C^∞ , the reliabilities of the QSPRs are limited in this case by the small amount and high level of uncertainty in the available experimental data. The reliability of the QSPRs can be increased when more P_C data of higher precision become available.


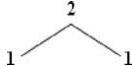
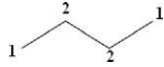
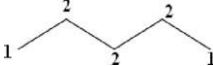
For T_m of n -alkanes large amount and high precision data are available. The currently available estimates on T_m^∞ values are based on melting points of linear polyethylenes and lie in a narrow range. To comply with those estimates, a QSPR containing the $IVDE$ descriptor in a linear term and n_C in a nonlinear term with four adjustable parameters can be developed. The higher confidence level in the data used to develop the QSPR provides more reliable prediction for T_C in case of long range extrapolation.

Thus, the adjustable QSPR that we have developed is able to match the confidence level of the predictions to the amount, precision, and reliability of the available data.

Literature Cited

- Nikitin ED, Pavlov PA, Popov AP. Critical temperatures and pressures of some alkanolic acids (C2 to C22) using the pulse-heating method. *Fluid Phase Equilibria*. 2001;189:151–161.
- Marrero J, Gani R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilibria*. 2001;183–184:183–208.
- Gasem KAM, Ross CH, Robinson RL. Prediction of ethane and CO₂ solubilities in heavy normal paraffins using generalized-parameter Soave and Peng-Robinson equation of state. *Can J Chem Eng*. 1993;71:805–816.
- Marano JJ, Holder GD. General equations for correlating the thermophysical properties of n -paraffins, n -olefins and other homologous series. 1. Formalism for developing asymptotic behavior correlations. *Ind Eng Chem Res*. 1997(a);36:1895–1907.
- Marano JJ, Holder GD. General equation for correlating the thermophysical properties of n -paraffins, n -olefins and other homologous series. 2. Asymptotic behavior correlations for PVT properties. *Ind Eng Chem Res*. 1997(b);36:1887–1894.
- Nikitin ED, Popov AP, Bogatishcheva NS. Critical properties of long-chain substances from the hypothesis of functional self-similarity. *Fluid Phase Equilibria*. 2005;235:1–6.
- Shacham M, Brauner N, Cholakov GS, Stateva RP. Property prediction by correlations based on similarity of molecular structures. *AIChE J*. 2004;50:2481–2492.
- Shacham M, Kahrs O, Cholakov GS, Stateva RP, Marquardt W, Brauner N. The role of the dominant descriptor in targeted quantitative structure property relationships. *Chem Eng Sci*. 2007;62:6222–6233.
- Cholakov GS, Stateva RP, Shacham M, Brauner N. Identifying equations that represent properties in homologous series using structure–structure relations. *AIChE J*. 2007;53:150–159.
- Brauner N, Cholakov GS, Kahrs O, Stateva RP, Shacham M. Linear QSPRs for predicting pure compound properties in homologous series. *AIChE J*. 2008;54:978–990.
- Kontogeorgis GM, Tassios DP. Critical constants and acentric factors for long-chain alkanes suitable for corresponding states applications. A critical review. *Chem Eng J*. 1997;66:35–49.
- Kurata M, Isida S. Theory of normal paraffin liquids. *J Chem Phys*. 1955;23:1126–1131.
- Sanchez IC, Lacombe RH. Statistical thermodynamics of polymer solutions. *Macromolecules*. 1978;11:1145–1156.
- Flory PJ, Orwoll RA, Vrij A. Statistical thermodynamics of chain molecule liquids. I. An equation of state for normal paraffin hydrocarbons. *J Am Chem Soc*. 1964;86:3507–3514.
- Cholakov GS, Stateva RP, Shacham M, Brauner N. Estimation of properties of homologous series with targeted quantitative structure–property relationships (TQSPRs). *J Chem Eng Data*. 2008;53:2510–2520.
- Todeschini R, Consonni V, Mauri A, Pavan M. *DRAGON User Manual*. Milano, Italy: Talete srl, 2006.
- Paster I, Shacham M, Brauner N. Investigation of the relationships between molecular structure, molecular descriptors and physical properties. *Ind Eng Chem Res*. 2009;48:9723–9734.
- Rowley RL, Wilding WV, Oscarson JL, Yang Y, Zundel NA. *DIPPR Data Compilation of Pure Chemical Properties Design Institute for Physical Properties*. Provo, UT: Brigham Young University, 2006. <http://dippr.byu.edu>.
- National Institute of Standards and Technology (NIST). In: Linstrom, PJ, Mallard WG, editors. *ChemistryWebBook, NIST Standard Reference Database Number 69*. Gaithersburg, MD, 2005. <http://webbook.nist.gov> (accessed January 2010).
- Ambrose D, Tsionopoulos C. Vapor–liquid critical properties of elements and compounds. 2. Normal alkanes. *J Chem Eng Data*. 1995;40:531–546.
- Nikitin ED, Pavlov PA, Popov AP. Vapor–liquid critical temperatures and pressures of normal alkanes with from 19 to 36 carbon atoms, naphthalene and m -terphenyl determined by the pulse-heating technique. *Fluid Phase Equilibria*. 1997;141:155–164.
- Shannon C, Weaver W. *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press, 1949.
- Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC. Discrimination of isomeric structures using information theoretic topological indexes. *J Comput Chem*. 1984;5:581–588.
- Giles NF, Taylor SW, Carn BR, Congote A, Rowley RL, Wilding WV. Property Estimation from Family Plots: A Case Study of the n -Alkane Family. 17th International Symposium on Thermophysical Properties, Boulder, CO, June 22–29, 2009.
- Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH, 2000.
- Narumi H. New topological indices for finite and infinite systems. *MATCH*. 1987;22:195–207.
- Tsionopoulos C, Ambrose D. Vapor–liquid critical properties of elements and compounds. 6. Unsaturated aliphatic hydrocarbons. *J Chem Eng Data*. 1996;41:645–656.
- Nikitin ED, Popov AP. Critical temperatures and pressures of linear alk-1-enes with 13 to 20 carbon atoms using the pulse-heating technique. *Fluid Phase Equilibria*. 1999;166:237–243.
- Ambrose D, Walton J. Vapour pressures up to their critical temperatures of normal alkanes and 1-alkanols. *Pure Appl Chem*. 1989;61:1395–1403.
- Gude M, Teja AS. Vapor–liquid critical properties of elements and compounds. 4. Aliphatic alkanols. *J Chem Eng Data*. 1995;40:1025–1036.
- Nikitin ED, Pavlov PA, Popov AP. Critical temperatures and pressures of 1-alkanols with 13 to 22 carbon atoms. *Fluid Phase Equilibria*. 1998;149:223–232.
- Tsionopoulos C, Kudchadker AP, Ambrose D. Vapor–liquid critical properties of elements and compounds. 7. Oxygen compounds other than alkanols and cycloalkanols. *J Chem Eng Data*. 2001;46:457–479.
- Broadhurst MG. Extrapolation of the orthorhombic n -paraffin melting properties to very long chain lengths. *J Chem Phys*. 1962;36:2578–2582.

Table A1. Expressions for Calculating the Descriptor *IVDM* for *n*-Alkanes: Ethane through *n*-Pentane

Compound	Vertex Degree	<i>IVDM</i>
Ethane		$-\left[\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right] = 1$
Propane		$-\left[\frac{1}{4}\log_2\frac{1}{4} + \frac{2}{4}\log_2\frac{2}{4} + \frac{1}{4}\log_2\frac{1}{4}\right] = 1.5$
<i>n</i> -Butane		$-\left[\frac{1}{6}\log_2\frac{1}{6} + \frac{2}{6}\log_2\frac{2}{6} + \frac{2}{6}\log_2\frac{2}{6} + \frac{1}{6}\log_2\frac{1}{6}\right] = 1.918$
<i>n</i> -Pentane		$-\left[\frac{1}{8}\log_2\frac{1}{8} + \frac{2}{8}\log_2\frac{2}{8} + \frac{2}{8}\log_2\frac{2}{8} + \frac{2}{8}\log_2\frac{2}{8} + \frac{1}{8}\log_2\frac{1}{8}\right] = 2.25$

Appendix A

Derivation of an expression based on n_C for calculating the *IVDM* descriptor values for *n*-alkanes and determining its limiting value at $n_C \rightarrow \infty$.

The descriptor *IVDM* is calculated using Eq. 1. For *n*-alkanes δ_a (the vertex degree) can obtain the values of one (carbon atoms at the ends of the chain) or two (all other carbon atoms). A_V (the total adjacency index) is the sum of all the vertex degrees. Calculation of *IVDM* for the first members of the *n*-alkane series, except methane (for which *IVDM* is undefined) are shown in detail in Table A1.

Introducing $A_V = 2n_C - 2$ into Eq. 1 yields

$$\begin{aligned}
 IVDM &= -\left[\frac{1}{A_V}\log_2\frac{1}{A_V} + (n_C - 2) \cdot \frac{2}{A_V}\log_2\frac{2}{A_V} + \frac{1}{A_V}\log_2\frac{1}{A_V}\right] \\
 &= -\left[\log_2\frac{1}{A_V} \cdot \left(\frac{2}{A_V} + (n_C - 2) \cdot \frac{2}{A_V}\right) + (n_C - 2) \cdot \frac{2}{A_V}\right] \\
 &= -\left[\log_2\frac{1}{2n_C - 2} + (n_C - 2) \cdot \frac{2}{2n_C - 2}\right] \\
 &= \frac{\ln(2n_C - 2)}{\ln 2} - \frac{n_C - 2}{n_C - 1} = 1 + \frac{\ln(n_C - 1)}{\ln 2} - \frac{n_C}{n_C - 1} + \frac{2}{n_C - 1}
 \end{aligned}$$

Using this expression yields $\lim_{n_C \rightarrow \infty} [IVDM] \rightarrow \infty$.

Manuscript received Jan. 30, 2010, and revision received Mar. 13, 2010.