Prediction of Normal Melting Point of Pure Substances by a Reference Series Method

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

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

DOI 10.1002/aic.14128

The "Reference Series" method of Shacham et al. is modified to enable prediction of normal melting point temperatures (T_m) of pure substances within experimental error level. A homologous series for which large amount and high precision T_m data are available is used as a "reference" series. To predict T_m for a "target" series quantitative property-property relationships (QPPRs) are derived to represent the predicted T_m values of the "target" series in terms of the T_m values of the reference series. Two QPPRs are necessary in order to match the odd and even carbon number (n_C) oscillations of T_m in the low n_C region. In the high n_C region, the QPPR is adjusted to represent correctly the asymptotic behavior of T_m. It is shown that the method is very useful for consistency analysis of T_m data and enables a reliable prediction of T_m in both the low n_C and the high n_C regions. © 2013 American Institute of Chemical Engineers *AIChE J*, 59: 3730–3740, 2013

Keywords: melting point prediction, homologous series, quantitative structure property relationship, reference series, quantitative property-property relationships

Introduction

Normal melting temperature (T_m) is an important property for assessing the environmental impact of compounds as it indicates the physical state of the chemical at ambient temperatures, thus dictates how the chemical is handled and treated. Furthermore, it is widely used in quantitative structure-activity relationships for predicting toxicity and aqueous solubility.

Methods for the prediction of physical properties of pure compounds based on their molecular structure are challenged by the prediction of solid properties, T_m in particular. This is due to the numerous factors that affect the solid state properties, but have much less (or no) effect on the liquid or gas phase properties. These factors include ionic, polar and hydrogen bonding forces, crystal packing, and positional, expansional, rotational, and conformational entropy effects.[1] Consequently, property prediction techniques are significantly less reliable when applied to solid properties compared to their reliability in predicting liquid and gas phase properties.[1–6]

The most widely used methods for predicting T_m are the "group contribution (GC)" methods.[7] Some of the GC methods have been already introduced into commercial software packages for predicting T_m on a routine basis (e.g., the Dorthmund Data Bank, DDBST, 2011 release, http://www.ddbst.de, and CRANIUM, Molecular Knowledge Systems, http://www.molecularknowledge.com/). In the DDBSP, for example, the GC methods of Joback and Reid[8] (JR), Constantinou and Gani[9] (CG), and Wen and Qiang[10] can be used for the prediction of T_m. Poling et al.[7] reported some prediction error resulting when applying two of the above GC methods. Using the JR[8] method for 307 compounds yielded an average (absolute) prediction error of 14.4%, which exceeds 10% for more than half of the sample (154 compounds). Applying the GC[9] method on a sample of 273 compounds, resulted in an average (absolute) error of 13.2%, and it exceeded 10% for 116 compounds. These error statistics are much higher than the ones obtained for predicting normal boiling temperature (T_b), for example. The theoretical lower limit on the prediction error is the uncertainty level of the available data for similar compounds. For many groups of compounds, the uncertainty associated with T_m data is <1% (or even <0.2%). Thus, there is much room for improvement of the T_m prediction techniques.

In the recent years, there has been considerable effort to improve the accuracy of T_m estimations by introducing more complex methods. For example, Marrero and Gani[11] introduced a three level GC method, Godavarthy et al.[2] developed a quantitative structure-property relationship (QSPR), which uses 16 molecular descriptors in a nonlinear model whose parameters were determined using a neural network, and Lazzus[5] suggested the use of a neural network and a particle swarm algorithm to better represent the nonlinear relationship between the contribution of the various groups to T_m. These methods reduced somewhat the average prediction errors. Yet, the errors are still considerably higher than the theoretical limit.

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

Brauner et al.[12] suggested the use of linear QSPR models for predicting T_m of members of homologous series. This method was demonstrated by developing a linear QSPR containing four molecular descriptors (five parameters) for the 1-alkanol homologous series. This QSPR can predict T_m values for the 1-alkanol series within experimental error level for cases of interpolation and short range extrapolation. The derivation of such accurate QSPR, however, requires the availability of considerable amount of high precision T_m data for the series considered. The need for large amount of data for deriving accurate prediction techniques can be explained by the irregular variation of T_m in most series for compounds with n_C (number of carbon atoms) $<\sim 25$. In most homologous series, the first few members exhibit anomalous variation of T_m, with a decreasing trend with n_C, rather than an increase, as the rest of the compounds in the series. For the remaining compounds (up to $n_C \sim 25$), oscillatory variation of T_m for odd and even n_C compounds is often encountered.

Another challenge of the property prediction techniques is the prediction of T_m for long-chain (high n_C value) substances. It is well known (see, e.g., Marano and Holder[4]) that for $n_C \to \infty$ T_m converges to a constant $\left(T_m^\infty\right)$ value. Marano and Holder[4] recommend a T_m^∞ value within the range of 413–418 K. Using the method of CG[9] for predicting the T_m of n-alkane with $n_C = 100$ yields $T_m = 462.63$ K which is considerably higher than the experimental value reported by Broadhurst[13] ($T_m = 388.4$) and the generally accepted T_m^∞ value. Paster et al.[14] developed a QSPR for predicting T_m for long-chain members of the n-alkane series. The derivation of such QSPR involves identification of molecular descriptors collinear with T_m based on available experimental data. From among these, the descriptors whose asymptotic behavior is similar to the property behavior are eventually used for prediction.

The behavior of the T_m vs. n_C curves is, however, similar for many of the homologous series. This similarity can be used for improving the prediction for series for which an insufficient amount of and/or low-precision data are available. For such situations, relying on the suggestions of Peterson[15,16], Shacham et al.[17] developed the "Reference Series" method. Using this method, experimental (if available) or predicted data of a "reference" homologous series for which the largest amount and highest precision experimental data are available, are used as the basis for prediction of properties for other (target) series. The target series are related to the reference series by means of a quantitative property-property relationship (QPPR). This method was recently used by Shacham et al.[17] to successfully predict the enthalpy, entropy, and Gibbs energy of formation of ideal gases.

In this article, we develop a modified version of the "Reference Series" method for accurate prediction of T_m. The QPPR developed for T_m prediction of various target series is combined with the QSPR proposed by Paster et al.[14] for T_m in order to predict T_m values of long-chain members of these series.

Methodology

To carry out the studies reported here mostly physical property data from the DIPPR database[18] are used. The DIPPR database contains property data for over 1800 compounds. For a particular compound-property combination, a large number of property values from various sources are often available is the DIPPR database. In such cases, the T_m values are categorized according to their source as "experimental," "predicted," "smoothed," and "unknown". From among the several available values, the DIPPR staff selects one "accepted" value, whereas other values are categorized as either "acceptable" or "rejected." In the present work, the "accepted" values were usually used and only "experimental" values were included in the training set for derivation of the QPPR models. The DIPPR staff also assigns uncertainty (U_i) to the various data points (e.g., <0.2%, <1%, etc.). These values can be considered as upper limits on experimental error in the data. The prediction accuracy cannot exceed the accuracy of the experimental data used to derive the model, thus the lower limit on the prediction error is the data uncertainty level.

Some of the T_m data were obtained from the NIST database.[19] The implementation of the CG[9] method in the DDBSP program is used for comparing the results of the proposed method with those obtained with the state-of-the-art methods. The software packages Polymath (Polymath is a product of Polymath Software, http://www.polymath-software.com) and MATLAB (MATLAB is a trademark of The Math Works, http://www.mathworks.com) were used for carrying out the linear and nonlinear regressions required for determining the parameter values of the QPPR models (Eqs. 1 or 4 below). These programs provide confidence intervals on the parameter values. The ratio of the confidence interval to the parameter value can help to determine the maximal number of parameters that can be included in the QPPR with the available T_m data for the target series.

Basic principles and an introductory example

Shacham et al.[17] have recently shown that properties of compounds in two homologous series can be represented (at least locally) by a linear QPPR

$$y_t = \beta_0 + \beta_1 y_r \qquad n_C \geq n_{C,\min} \tag{1}$$

where y_r is the property value of a compound in the *reference* series, y_t is the property of a compound (related to the reference compounds in terms of the number of carbon atoms n_C) in the *target* series, and β_0 and β_1 are parameters.

The parameter values of the QPPR (also referred to as "regression model" or "correlation") are obtained by regression of the available target series data with the corresponding reference series data. Upon determination of the QPPR's parameter values, it can be used for prediction of unknown target series property values using the corresponding known reference series data.

In case of T_m prediction, we use $y_t = (T_m)_{nC}^t$, where n_C is the number of carbon atoms in the respective member of the target series, and $y_r = (T_m)_i^r$, where i is the number of carbon atoms in the matching member of the reference series, it can obtain the values $i = n_C$, $i = n_C + 1$, or $i = n_C - 1$. Equation 1 is often not valid for the first members of the homologous series, where the strong influence of the particular functional group affects a different trend of the property change with n_C.[20] Consequently, the validity of Eq. 1 is limited to n_C greater than a certain $n_{C\min}$.

The melting point temperatures (T_m) of the n-alkane and alkanoic acid homologous series in the range of $1 \leq n_C \leq 32$ are listed in Table 1. The DIPPR database[18] recommended values are the sources of all the n-alkane and part of the alkanoic acid (up to $n_C = 20$) data. The rest of the alkanoic

Table 1. Normal Melting Point Data and Prediction Results for the *n*-Alkane and Alkanoic Acid Homologous Series

	Melting Point of *n*-Alkanes[*]			Melting Point of Alkanoic Acids[†]			QPPR Prediction	
n_C	Value (K)	Uncertainty (%)	δT_m[‡]	Value (K)	Uncertainty (%)	δT_m[‡]	Value (K)	Error (%)
1	90.69	<0.2	−0.342	281.55	<1	8.26		
2	90.35	<0.2	−4.882	289.81	<1	−37.36		
3	85.47	<0.2	49.39	252.45	<1	15.5		
4	134.86	<0.2	8.56	267.95	<1	−28.8		
5	143.42	<0.2	**34.41**	239.15	<1	**30.1**	239.41	−0.11
6	177.83	<0.2	4.74	269.25	<1	−3.42	266.59	0.99
7	182.57	<0.2	**33.81**	265.83	<1	**23.82**	263.64	0.82
8	216.38	<0.2	3.28	289.65	<1	−4.1	289.84	−0.07
9	219.66	<1	**23.85**	285.55	<1	**19.2**	286.60	−0.37
10	243.51	<1	4.061	304.75	<1	−3.12	306.21	−0.48
11	247.57	<1	**15.997**	301.63	<1	**15.35**	303.87	−0.74
12	263.57	<1	4.192	316.98	<1	−1.97	318.30	−0.42
13	267.76	<0.2	**11.25**	315.01	<1	**12.36**	316.37	−0.43
14	279.01	<0.2	4.062	327.37	<1	−1.69	327.62	−0.08
15	283.07	<0.2	**8.236**	325.68	<1	**9.98**	325.84	−0.05
16	291.31	<0.2	3.826	335.66	<1	−1.41	335.04	0.19
17	295.13	<0.2	**6.176**	334.25	<1	**8.5**	333.31	0.28
18	301.31	<0.2	3.73	342.75	<1	−1.52	341.07	0.49
19	305.04	<0.2	**4.54**	341.23	<1	**7**	339.44	0.52
20	309.58	<0.2	3.77	348.23	<1	−0.58	346.06	0.62
21	313.35	<1	**3.8**	347.65		**5.6**	344.58	0.88
22	317.15	<1	3.5	353.25		−0.5	350.62	0.74
23	320.65	<1	3.1	352.75		**5.7**	349.10	1.03
24	323.75	<3		358.45			354.61	1.07
26	329.25	<1		359.45			354.42	1.40
29	336.85	<1		363.4			362.51	0.25
32	342.35	<1		361.2			362.53	−0.37

[*] DIPPR database.[18]
[†] Up to $n_C = 20$ DIPPR database,[19] for $n_C > 20$ NIST database.
[‡] Members of the higher δT_m series are shown in bold.

acid data is from the NIST database.[19] The DIPPR data includes also uncertainties of the T_m values. The uncertainties for the *n*-alkanes are either $U_i < 0.2\%$ or $U_i < 1\%$, only in one case (*n*-tetracosane) $U_i < 3\%$. All the available uncertainties for the alkanoic acids are <1%. No uncertainty values are assigned to the NIST data.

The T_m data of Table 1 is plotted vs. n_C in Figure 1. Two distinct regions can be identified in the T_m curves of the *n*-alkanes and the alkanoic acids. In the "low n_C" region, there is a decreasing trend of the T_m values with increasing n_C. This region includes the first three members of the 1-alkane series and the first four members of the alkanoic acid series. At higher n_C, a general trend of increasing T_m values with n_C is observed in both curves. However, there are "local" oscillations in the T_m values between consecutive members with odd and even n_C values. The oscillations are the highest for low n_C compounds and diminish for $n_C > 20$ (*n*-alkanes) or $n_C > 25$ (alkanoic acids). These oscillations in the T_m values are attributed to the melting from different crystalline phases.[4,13]

The forward difference of two consecutive members of a homologous series is defined

$$(\delta T_m)_i = (T_m)_{i+1} - (T_m)_i \tag{2}$$

where $i = n_C$, can be used for numerical characterization of the oscillations. For example, for *n*-alkanes $(\delta T_m)_5 = 34.41$ K while $(\delta T_m)_6 = 4.74$ and for alkanoic acids $(\delta T_m)_5 = 30.1$ K and $(\delta T_m)_6 = -3.42$ (Table 1). Thus, the oscillations in the two homologous series are in phase: $(\delta T_m)_i$ has large positive value for odd n_C compounds and small (and negative for alkanoic acids) value for even n_C compounds. The consistency in the oscillatory behavior of the two series is demonstrated in Figure 2, showing the δT_m values of the two series vs. n_C, up to $n_C = 20$. Due to the phase-matching of the oscillations of the two series, two separate QPPRs (of the form of Eq. 1) can be derived for predicting the melting point of the alkanoic acid (target series), with $y_t = (T_m)^t_{nC}$, $y_r = (T_m)^r_{nC}$, and $n_{C\min} = 5$.

The methodology is further elaborated with reference to Figure 3, where the T_m values of the first 20 members of the *n*-alkanoic acid series are plotted vs. the T_m values of the first 20 members of the *n*-alkane series. For the first four points (the low n_C range), no particular relationship is distinguishable between the reference and target T_m values.

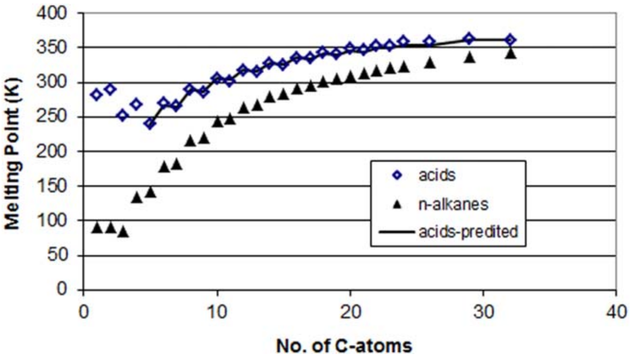

Figure 1. Plot of normal melting point data of *n*-alkanes and alkanoic acids (data of predicted values are also shown) vs. the number of carbon atoms up to $n_C = 32$.

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

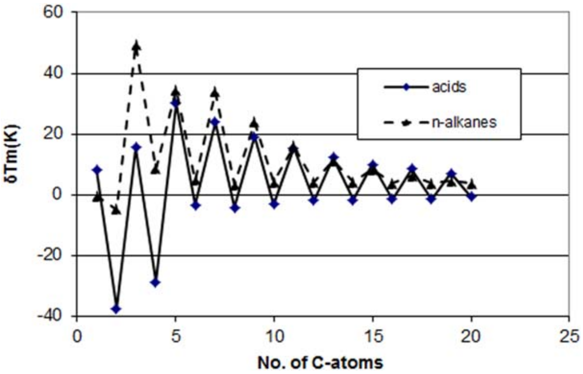

Figure 2. Plot of $\delta T_m = T_{m,nC+1} - T_{m,nC}$ of *n*-alkanes and *n*-alkanoic acids vs. the number of carbon atoms up to $n_C = 20$.

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

However, for $n_C \geq 5$ ($=n_{C\min}$) the T_m values of the odd and even n_C compounds are aligned along two separate straight lines. The parameter values and the error statistics for the respective QPPRs are shown as Case 1 in Table 2. For the odd n_C compounds $\beta_0 = 150.65$ and $\beta_1 = 0.619$ with the correlation coefficient of $R^2 = 0.998$. For the even n_C compounds, $\beta_0 = 159.32$, $\beta_1 = 0.6032$, and $R^2 = 0.9968$.

For reporting the prediction error, the average (absolute value) of the relative error (ε_{avg}) is used

$$\varepsilon_{avg} = \frac{100}{n} \sum_{j=1}^{n} \left| \frac{(T_m)_j^{exp} - (T_m)_j^{pred}}{(T_m)_j^{exp}} \right| \tag{3}$$

where $(T_m)_j^{exp}$ is the experimental and $(T_m)_j^{pred}$ is the predicted T_m value of the jth compound and n is the total number of compounds involved. The average (absolute) relative error is used in most of the publications to report and compare prediction errors. The maximal relative error, ε_{max} (the largest term included in the summation in Eq. 3) is also reported here in order to show the prediction error that can be expected in the worst case.

Using these QPPRs for predicting the *n*-alkanoic acid T_m in the $5 \leq n_C \leq 32$ range results in a maximal prediction error (ε_{max}) of 1.4% (Table 1) and average error (ε_{avg}) of 0.54%. These errors are considerably lower than what is obtained by other prediction techniques. For the same range of compounds, the CG[9] technique yields T_m values with $\varepsilon_{max} = 13.8\%$ (for pentanoic acid) and $\varepsilon_{avg} = 2.34\%$. The T_m values predicted by the proposed QPPRs are also shown in Figure 1. Observe that the predicted values match very well the experimental ones for the range where experimental data are available.

To study the influence of the number of compounds included in the training set and their location within the homologous series, the T_m predictions for *n*-alkanoic acids in the range of $5 \leq n_C \leq 32$ (applicable region) were carried out. Some of the results of this study are shown as Cases 2–5 in Table 2. There are two cases where the minimum number of training set compounds (two of odd and two of even n_C compounds) are used. In Case 2, the training set members are in the beginning of the applicable region ($5 \leq n_C \leq 8$) and in Case 3 the training set members are in the middle of the applicable region ($10 \leq n_C \leq 13$). In Cases 4 and 5, there are three compounds in the odd n_C and three compounds in the even n_C training sets, with the same locations as in Cases 1 and 2. The prediction errors are the highest for Case 2, where the longest interpolation to dotriacontanoic acid (n_C =32) is involved ($\varepsilon_{max} = 3.7\%$, $\varepsilon_{avg} = 1.86\%$). The prediction errors are the smallest for Case 4 ($\varepsilon_{max} = 2.38\%$, $\varepsilon_{avg} = 0.69\%$), where the training set is larger and the extrapolation distance is shorter.

Predicting T_m for long-chain substances

The linear QPPR of Eq. 1 may not converge to a desirable T_m^∞ value. For long-range extrapolation and prediction of T_m for long-chain substances, we adopt the modification of the linear Eq. 1, as recommended by Paster et al.,[14] by including an empirical correction term

$$y_t = \beta_0 + \beta_1 y_r - (\beta_0 + \beta_1 y_r - y^\infty)[1 - \exp(-\beta_2 n_C)] \quad n_C \geq n_{C,\min} \tag{4}$$

where y^∞ is an accepted T_m^∞ value and β_2 is an additional regression parameter.

The *n*-alkane series is most often used as the reference series due to the availability of high precision T_m data for a large number of compounds. For this series, T_m data are available up to $n_C = 100$.[13] Missing T_m values for $n_C > 20$

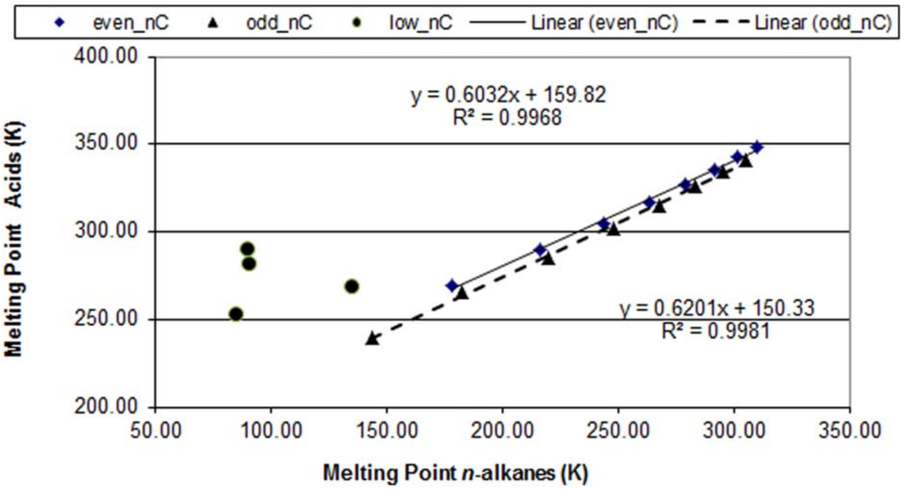

Figure 3. Plot of $(T_m)^t$ vs. $(T_m)^r$ for the *n*-alkane and alkanoic acid series.

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

Table 2. Training Sets and Prediction Errors for QPPR Modeling of T_m for the Alkanoic Acid Series

	Odd n_C Compounds			Even n_C Compounds				
Case no.	Training set n_C	β_0	β_1	Training set n_C	β_0	β_1	Maximal Error (%)*	Average Error (%)
1	5, 7 … 19	150.65	0.619	6, 8 … 20	159.32	0.6032	1.4 (26)	0.54
2	5, 7	141.41	0.681	6, 8	175.14	0.529	3.7 (32)	1.86
3	11, 13	137.56	0.66	10, 12	139.62	0.673	3.7 (6)	0.86
4	5, 7, 9	152.69	0.609	6, 8, 10	173.15	0.54	2.9 (24)	1.02
5	9, 11, 13	151.22	0.61	10, 12, 14	149.76	0.636	2.38 (6)	0.69

*n_C of compound with the maximal error is shown in parentheses.

can be predicted for the *n*-alkane series by the following QSPR proposed by Paster et al.[14]

$$y_r=\beta_0+\beta_1\text{IVDE}-(\beta_0+\beta_1\text{IVDE}-y^{\infty})[1-\exp(-\beta_2 n_C)] \quad (5)$$

where y_r is the predicted T_m value for the member of the reference (*n*-alkane) series, y^{∞} is an accepted T_m value at the limit $n_C \to \infty$, and β_0, β_1, and β_2 are the parameter values determined by nonlinear regression. The IVDE is a descriptor which was found most suitable for long-range representation of the T_m values of the *n*-alkane series. This descriptor corresponds to the mean information content of the vertex degree equality. It 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* atoms.[21] For the *n*-alkane homologous series, the IVDE descriptor can be expressed as function of n_C

$$\text{IVDE}=-\left[\frac{2}{n_C}\log_2\frac{2}{n_C}+\frac{n_C-2}{n_C}\log_2\left(\frac{n_C-2}{n_C}\right)\right] \quad (6)$$

As for the value of $y^{\infty}=T_m^{\infty}$, Paster et al.[14] indicate that there are several references that put this value within the range of 413–418 K and recommend the use of $T_m^{\infty}=415$ K, which is used throughout this work. With this value of T_m^{∞}, the parameter values obtained are $\beta_0=420.9248$, $\beta_1=-239.258$, and $\beta_2=1/1137.644$.

Equations 5 and 6 yield T_m predictions of very high accuracy for long-chain members of the *n*-alkane series. For example, for $n_C=70$, the reported experimental value[13] is $T_m=378.5$ K and the predicted value is $T_m=378.46$ K. For $n_C=100$, the reported value is $T_m=388.4$ K and the predicted value is $T_m=389.43$ K.

The use of Eqs. 4–6 for prediction of T_m of long-chain substances is demonstrated by predicting T_m of *n*-alkanoic acids in the n_C range of $33 \le n_C \le 100$. In this region, no experimental T_m data are available for these series.

Using $y^{\infty}=415$ K and the training set shown as Case 1 in Table 2 and applying nonlinear regression, yields the following optimal parameter values: for the odd n_C compounds $\beta_0=147.54$, $\beta_1=0.423$, $\beta_2=1/29.99$ with $R^2=0.9996$, whereas for even n_C compounds $\beta_0=169.76$, $\beta_1=0.383$, $\beta_2=1/30.91$ with $R^2=0.9999$. Predicting the alkanoic acid T_m in the $5 \le n_C \le 32$ using these two QPPRs yields predictions with maximal error ($\varepsilon_{\max}$) of 3.68% (for dotriacontanoic acid) and average error (ε_{avg}) of 0.38%.

Equations 4–6 with the parameters provided here can be used for predicting T_m value for a member of the alkanoic acid series for any $n_C \ge 24$ (for $n_C < 24$ the use of experimental data for the *n*-alkane series is recommended, see Table 1). The n_C value of the reference compound can be introduced into Eq. 6 to calculate the value of the IVDE descriptor. Introducing the descriptor value into Eq. 5 yields the y_r estimate, which in turn is used in Eq. 4 with the relevant set of parameter values (according to odd or even n_C) to provide the T_m prediction for the respective member of the *n*-alkanoic acid (target) series. Using this technique, the T_m values were predicted in the range of $33 \le n_C \le 100$. The results of this prediction are shown in Figure 4 together with the available T_m data[13] for the *n*-alkane series. Observe that the difference between the melting point temperatures of the two series reduces from 29.7 K at $n_C=33$ to 21.8 K at $n_C=100$. Thus, the two series converge to the same T_m^{∞} value. Also, while two different sets of parameter values are used to predict the T_m of the odd and even n_C acids, the oscillations between the two sets of T_m values become indistinguishable starting at $n_C \sim 50$.

For $n_C=100$, the nonlinear QPPR (Eq. 4) yields $T_m=411.2$ K for the member of the alkanoic acid series, compared to $T_m=388.46$ K for the corresponding member in the alkane series. The predictions obtained using the CG[9] method for $n_C=100$ are $T_m=462.63$ K and $T_m=473.8$ K for *n*-alkane and for alkanoic acid, respectively. Both values appear to be considerably higher than the accepted of T_m^{∞} values (within the range of 413–418 K).

QPPRs that require a shift of n_C between the reference and target series

Table 3 presents the T_m data available in the DIPPR database for the *n*-alkyl mercaptan (target) series (for $1 \le n_C \le 12$). The uncertainties of all the data points (not shown in the table) are $U_i < 1\%$. The reference series

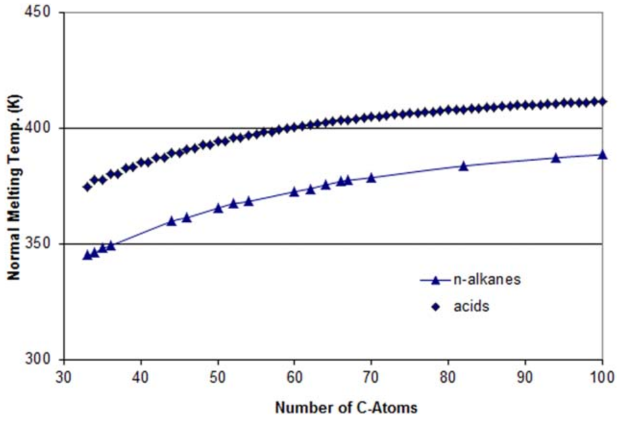

Figure 4. Plot of normal alkane T_m data and *n*-alkanoic acid predicted T_m values vs. n_C in the range $33 \le n_C \le 100$.

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

Table 3. Normal Melting Point Data and Prediction Results for the *n*-Alkane and *n*-Mercaptan Homologous Series

Melting Point of *n*-Alkanes			Melting Point of *n*-Mercaptans			Linear QPPR Prediction		Nonlinear QPPR Prediction	
n_C	Value (K)	δT_m[†]	n_C	Value[*] (K)	δT_m[†]	Value (K)	Error (%)	Value (K)	Error (%)
1	90.69	−0.342							
2	90.35	−4.882	1	150.18	−24.92				
3	85.47	**49.39**	2	125.26	**34.69**				
4	134.86	8.56	3	159.95	−2.49	160.25	0.19	160.10	0.09
5	143.42	**34.41**	4	157.46	**39.99**	158.04	0.37	157.58	0.07
6	177.83	4.74	5	197.45	−4.83	196.97	0.24	197.29	0.08
7	182.57	**33.81**	6	192.62	**37.30**	191.81	0.42	192.33	0.15
8	216.38	3.28	7	229.92	−5.97	229.92	0.00	229.83	0.04
9	219.66	**23.85**	8	223.95	**29.10**	223.79	0.07	223.65	0.13
10	243.51	4.061	9	253.05	−5.49	253.10	0.02	252.98	0.03
11	247.57	**15.997**	10	247.56	**22.59**	247.86	0.12	247.43	0.05
12	263.57	4.192	11	270.15	−5.00	270.24	0.03	270.33	0.07
13	267.76		12	265.15		265.28	0.05	265.23	0.03

[*]DIPPR database,[18] all data uncertainties <1%.
[†]Members of the higher δT_m series are shown in bold.

(*n*-alkane) data are also included in Table 3, in the range of $1 \leq n_C \leq 12$. Figure 5 shows the plot of the δT_m values of the two series vs. n_C, up to $n_C = 11$. The oscillations between pairs of T_m values in the target series are clearly noticeable; however, they are out of phase with the oscillations in the reference series. This discrepancy can be removed by "shifting" the reference series so that T_m of the n_Cth compound in the target series will be in line with the T_m of the ($n_C + 1$)th compound in the reference series. The two series are already aligned according to this principle in Table 3, and this indeed leads to synchronization between the large and small $(\delta T_m)_i$ values in the two series. Thus, for this case using $y_t = (T_m)^t_{nC}$ the appropriate definition for y_r is $y_r = (T_m)^r_{nC+1}$.

The corresponding values of y_t and y_r are depicted in Figure 6 for the first 12 members of the *n*-alkyl mercaptan series. For $n_C \geq n_{Cmin} = 3$, the y values of the odd and even n_C (target) compounds are aligned along two separate straight lines. The parameter values and the error statistics for the respective QPPRs are shown in Figure 6. For the odd n_C compounds $\beta_0 = 44.996$ and $\beta_1 = 0.8546$ with a correlation coefficient of $R^2 = 1$. For the even n_C compounds $\beta_0 = 34.359$ and $\beta_1 = 0.8624$ with $R^2 = 0.9998$. Using these QPPRs for predicting the T_m values of *n*-alkyl mercaptan in the $3 \leq n_C \leq 12$ range yields predictions with $\varepsilon_{max} = 0.42\%$ (Table 3) and $\varepsilon_{avg} = 0.25\%$. Predicting the T_m values in the same range with the CG[9] method results in $\varepsilon_{max} = 8.79\%$ and $\varepsilon_{avg} = 4.78$ %. The coefficients and the R^2 values for the nonlinear QPPRs (Eq.4) are shown in Table 6. Using the nonlinear QPPRs reduces the maximal error to $\varepsilon_{max} = 0.15\%$ and the average error to $\varepsilon_{avg} = 0.07\%$.

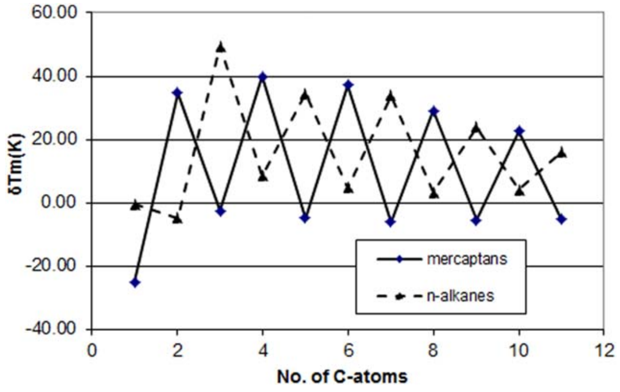

Figure 5. Plot of δT_m of *n*-alkanes and *n*-mercaptans vs. the number of carbon atoms up to $n_C = 11$.

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

Consistency analysis of the T_m data

Often several values can be found in the literature for a property of a particular compound, and sometimes there are considerable differences between those values. The common practice used to select the "correct" property value (from the several available ones) for a member in a homologous series is to prepare a "family plot" of the known property values of the members of the series vs. n_C. The values selected as "accepted" are those that exhibit a smooth variation of the family plot. This practice can be misleading when applied to T_m data, since the variation between the odd and even n_C compounds usually do not result in smooth family curves.

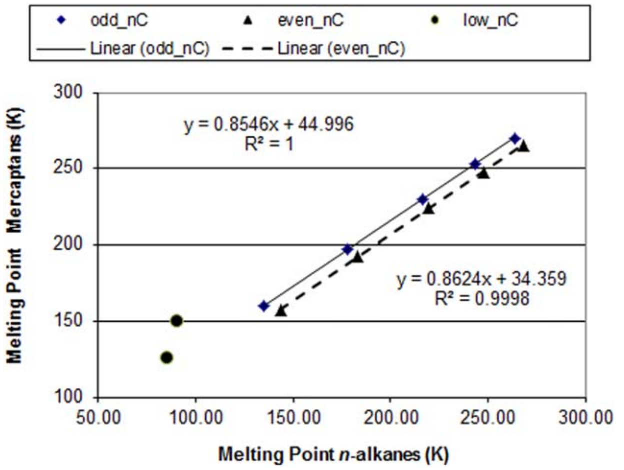

Figure 6. Plot of $(T_m)^t$ vs. $(T_m)^r$ for the *n*-alkane and mercaptan series.

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

Table 4. Range of the Available T_{m} Data of 1-Alkyl Aldehydes with Current (2013) and Earlier (2007) DIPPR "Accepted" Values

	Range of Melting Point Data[*]			DIPPR "Accepted" Values (K)[*]			
n_{C}	Min. Value (K)	Max. Value (K)	Difference %	Current[‡] (2013)	δT_{m}[+]	Earlier (2007)	δT_{m}[+]
2	148.15	152.55	2.97	149.78	15.22	149.78	20.22
3	169.85	171.55	1.00	165.00	11.80	170.00	6.75
4	173.15	177.15	2.31	176.80	14.79	176.75	14.84
5	181.15	191.59	5.76	191.59	**23.34**	191.59[†]	**25.56**
6	214.93	217.00	0.96	214.93	14.87	217.15	12.65
7	229.80	231.15	0.59	229.80	**21.85**	229.80	**16.20**
8	246.00	251.65	2.30	251.65	15.65	246.00	9.15
9	255.00	267.30	4.82	267.30	**17.70**	255.15	**12.00**
10	267.00	291.15	9.04	285.00	17.36	267.15	6.00
11	269.15	*273.15*	1.49	*302.36*	12.79	273.15	**12.00**
12	284.15	317.65	11.79	315.15	8.63	285.15	3.00
13	287.00	*288.15*	0.40	*323.78*		288.15	

[*] DIPPR database.[18]

[†] Original value of 182 K was replaced after analysis.

[‡] Predicted values greater than maximal experimental values are shown in Italic.

[+] Members of the higher δT_{m} series are shown in bold.

The problems associated with consistency analysis of T_{m} data, and the use of the reference series technique for this purpose, are demonstrated using the 1-alkyl aldehyde series.

Some of the T_m related data of the aldehyde series available in the DIPPR database are shown in Table 4. The range of the published T_{m} values for a particular compound is characterized by the minimal and maximal reported values and the percentage difference between them. For example, for butanal ($n_{\mathrm{C}} = 4$), the minimal reported value is $T_{\mathrm{m}} = 173.15$ K, while the maximal value is $T_{\mathrm{m}} = 177.15$ K, whereby the difference between the two values is 2.3%. The complete list for this compound in the DIPPR database contains 26 T_{m} values, and most of them are closer to the upper limit. One of the values in the list is completely out of range ($T_{\mathrm{m}} = 433.65$ K), which is categorized in DIPPR as "rejected." Such data are usually not included in Table 4. In the current (2013) version of DIPPR, the "accepted" value for butanal is $T_{\mathrm{m}} = 176.8$ K (Table 4), and the accepted value in an earlier release (2007) of the database was essentially the same (176.75).

The T_{m} data of dodecanal ($n_{\mathrm{C}} = 12$) is of special interest. For this compound 27 T_{m} values are available, half of them are in the 317.15–317.65 K interval. There are five values of 284–285 K and two of 374.15 K that were judged as "rejected" in the current release of DIPPR. We, however, do consider 284.15 K as the minimal value, because a close value was considered as the "accepted" value in the earlier release of the database (Table 4). Observe that for this compound, the difference between the minimal and maximal values is almost 12%.

Figure 7 shows a plot of the two sets of (current and earlier) "accepted" T_{m} values of aldehydes and the T_{m} values of n-alkanes vs. n_{C}. As the theory dictates that the T_{m} values of the two series converge to the same T_{m}^{∞}, the distance between the two curves should decrease with increasing n_{C}. This is indeed the trend for the earlier set of the recommended values, but for the current set, the distances start increasing from $n_{\mathrm{C}} = 9$. Additional indications of the inconsistency of the current set of the recommended values can be found in the values for undecanal ($n_{\mathrm{C}} = 11$) and tridecanal ($n_{\mathrm{C}} = 13$), which are considerably higher than the maximal reported experimental values (Table 4). Observe, also, that the periodic variation of the δT_{m} stops at $n_{\mathrm{C}} = 9$ for the current set of accepted values.

Because of the apparent inconsistency of the current set of accepted values, the earlier set (shown in Table 4) is used for developing the QPPR model for the 1-alkyl aldehydes (except that the original 182 K value for pentanal was replaced by 191.59 K to achieve higher accuracy). The QPPR model parameters and R^2 statistics are shown in Figure 8.

QPPR models for prediction of T_m for 9 target series

QPPR models have been developed for 13 target series. For four series lack of sufficient amount of data prevented the derivation of separate even and odd QPPRs, and these

Table 5. Summary of Linear QPPR Derivation for T_{m} of 9 Series

		Available data		Training set			Linear Model, Even			Linear Model, Odd			Linear Model	
No.	Series	n_{Cfirst}	n_{Clast}	n_{Cmin}	n_{Cmax}	Ref. Comp. Number, i	β_0	β_1	R^2	β_0	β_1	R^2	$\varepsilon_{\max}$ (%)	$\varepsilon_{\mathrm{avg}}$ (%)
1	*n*-alkanoic acid	1	32	5	20	n_{C}	159.82	0.6032	0.9968	150.33	0.6201	0.9981	0.86	0.20
2	*n*-alkyl mercaptan	1	12	3	12	$n_{\mathrm{C}} + 1$	34.36	0.8624	0.9999	45.00	0.8546	1.0000	0.43	0.05
3	aldehyde	2	13	4	13	n_{C}	66.99	0.8274	0.9981	86.49	0.7591	0.9924	2.05	0.57
4	1-alkene	2	40	7	20	$n_{\mathrm{C}} + 1$	−139.77	1.4107	0.9992	−183.52	1.5515	0.9992	1.25	0.29
5	1-alkyne	3	10	5	10	$n_{\mathrm{C}} - 1$	86.49	0.7591	0.9924	66.99	0.8274	0.9981	2.05	0.28
6	1-alcohol	1	20	4	20	n_{C}	70.29	0.8642	0.9982	81.51	0.8342	0.9953	2.86	0.47
7	*n*-alkyl amine	1	14	3	14	n_{C}	144.71	0.5940	0.9991	134.73	0.6262	0.9887	2.93	0.68
8	n-alkyl benzene	7	24	10	24	n_{C}	−171.07	1.4863	0.9993	−206.12	1.5963	0.9980	1.46	0.32
9	2-methylalkane	4	20	5	20	$n_{\mathrm{C}} - 1$	−28.29	1.0359	0.9984	−31.26	1.0528	0.9986	2.23	0.49

Table 6. Summary of Nonlinear QPPR Derivation for T_m of 9 Series

		Nonlinear Model, Even			Nonlinear Model, Odd			Nonlinear Model	
No.	Series	β_0	β_1	β_2	β_0	β_1	β_2	ε_{max} (%)	ε_{avg} (%)
1	*n*-alkanoic acid	169.76	0.3831	1/30.9083	147.54	0.4230	1/29.9862	0.55	0.09
2	*n*-alkyl mercaptan	33.67	0.7309	1/55.6926	46.45	0.7824	1/95.4536	0.10	0.03
3	aldehyde	62.71	0.6459	1/35.2489	76.76	0.5985	1/36.7039	1.75	0.51
4	1-alkene	−151.94	1.2975	1/56.2702	−174.94	1.2210	1/32.1107	1.04	0.25
5	1-alkyne	67.42	0.6150	1/36.7042	52.58	0.6645	1/35.2475	1.75	0.24
6	1-alcohol	64.14	0.7376	1/41.748	68.47	0.7455	1/45.4938	2.44	0.42
7	*n*-alkyl amine	144.04	0.4807	1/49.758	129.25	0.1660	1/16.2052	2.04	0.58
8	*n*-alkyl benzene	−186.24	1.2818	1/41.3203	−161.08	0.9475	1/24.7851	0.90	0.23
9	2-methylalkane	−47.90	0.7849	1/34.1732	−40.51	0.9369	1/62.7135	1.93	0.45

cases will be discussed in the next section. The results of the remaining nine series are summarized in Tables 5 and 6. For all the series n_{Cfirst} and n_{Clast} of the available data and n_{Cmin} and n_{Cmax} of the training set are listed (Table 5). In the "available data" category n_{Cfirst} is the n_C of the first compound that belongs to a particular series (e.g., methylacetylene who is the first member of the 1-alkyne series, has $n_C = n_{Cfirst} = 3$). In the same category, n_{Clast} is the highest n_C in the series for which data are available. In the "training set" category, n_{Cmin} is the minimal n_C in the target series beyond which a general trend of increasing T_m with n_C is observed both in the reference and in the target series. The value of n_{Cmax} for the training set is limited by the availability of experimental (rather than predicted) data for all the members of the target series between n_{Cmin} and n_{Cmax}. The largest number of compounds are included in the training set of the 1-alcohol series, where $n_{Cmin} = 4$ and $n_{Cmax} = 20$, thus total of 17 compounds in the training set. For the 1-alkyne series, there are only six compounds in the training set ($n_{Cmin} = 5$ and $n_{Cmax} = 10$). This is actually the smallest number that enables derivation of both linear and nonlinear QPPRs separately for odd and even n_C compounds. The lag or lead between the members of the reference and target series is indicated in the "Ref. Comp. number" column. Observe that for five series $y_r = (T_m)^r_{nC}$ is used in Eqs. 1 and 4, for the mercaptan and for the 1-alkene series $y_r = (T_m)^r_{nC+1}$ is to be used, while for the 1-alkyne and 2-methylalkane series the use of $y_r = (T_m)^r_{nC-1}$ is preferred.

The linear QPPR model's parameters, R^2, ε_{max}, and ε_{avg} values are also shown in Table 5. In cases where $y_t > y_r$ at $n_C = n_{Cmin}$, β_0 is positive and β_1 is in the range of $0.55 < \beta_1 < 0.9$. In cases where $y_t < y_r$ at $n_C = n_{Cmin}$, the value of β_0 is negative and $\beta_1 > 1.0$. Thus, in both cases, there is an $n_C > n_{Cmin}$ value for which $y_t = y_r$ (satisfying the requirement that the T_m values converge to a constant value for high n_C). The correlation coefficient values for all the QPPRs are $R^2 > 0.99$, indicating a good fit between the data and the predicted values. The prediction errors are the smallest for the *n*-alkyl mercaptan series ($\varepsilon_{max} = 0.43\%$, $\varepsilon_{avg} = 0.05\%$) and largest for the *n*-alkyl amine series ($\varepsilon_{max} = 2.93\%$, $\varepsilon_{avg} = 0.68\%$). The largest prediction error is for *n*-pentylamine, for which DIPPR assigns 3% uncertainty to the T_m data (as it is not clear whether the data is experimental or predicted). Thus, the high prediction error can be justified by the data uncertainty. Using the method of the CG[9] for predicting T_m for all the members of the *n*-alkyl amine series included in the training set yields the prediction errors: $\varepsilon_{max} = 13.06\%$ and $\varepsilon_{avg} = 4.34\%$, which are considerably higher than the ones obtained by the proposed method.

The nonlinear parameters shown in Table 6 can be introduced into Eq. 2 for predicting T_m of the various series at high carbon numbers. Introducing $n_C = 300$ into the nonlinear QPPR yields T_m values for the various series in the range of 413–415 K, as expected. The nonlinear QPPRs yield also lower ε_{max} and ε_{avg} values than the linear QPPRs (see Tables 5 and 6).

QPPR models for series where odd and even n_C compounds are not separated

Table 7 presents the T_m data available in the DIPPR database for the *n*-alkyl acetate (target) series (for $1 \le n_C \le 12$). The uncertainties of the data points are $U_i < 1\%$ for most of the compounds and $U_i < 3\%$ for *n*-butyl and *n*-hexyl acetate. For the *n*-hexyl acetate, we replaced the DIPPR "accepted" T_m value with a "rejected" value, which was proven to be more consistent with the rest of the training set values. The values of δT_m for the acetate series are shown in Table 7 and are plotted vs. n_C in Figure 9. Observe that in the low n_C region ($3 \le n_C \le 6$), there are oscillations between the T_m values of the odd and even n_C compounds, however starting at $n_C = 7$ no such oscillations exist. In such cases, there is no need to fit two separate QPPRs (to the odd and even n_C compounds) and one QPPR for the whole training set is sufficient. However, in this case the *n*-alkane series is not suitable to serve as reference series because of the high level of the oscillations of the T_m values. A reference series with low level of oscillations (or no oscillations) need to be used. An appropriate reference series in such cases is the 1-alkene series, for which high precision T_m data are available (Table 7)

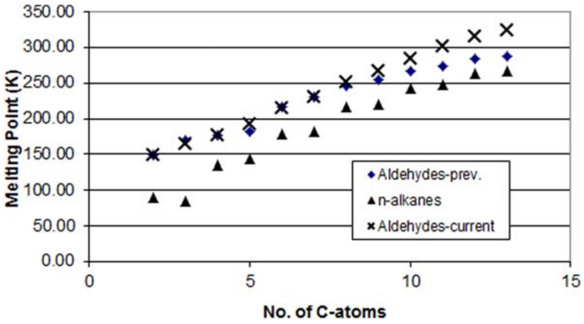

Figure 7. Plot of the T_m data of *n*-alkanes and 1-alkyl aldehydes (current: 2013 and earlier: 2007 accepted values) vs. the number of carbon atoms up to $n_C = 13$.

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

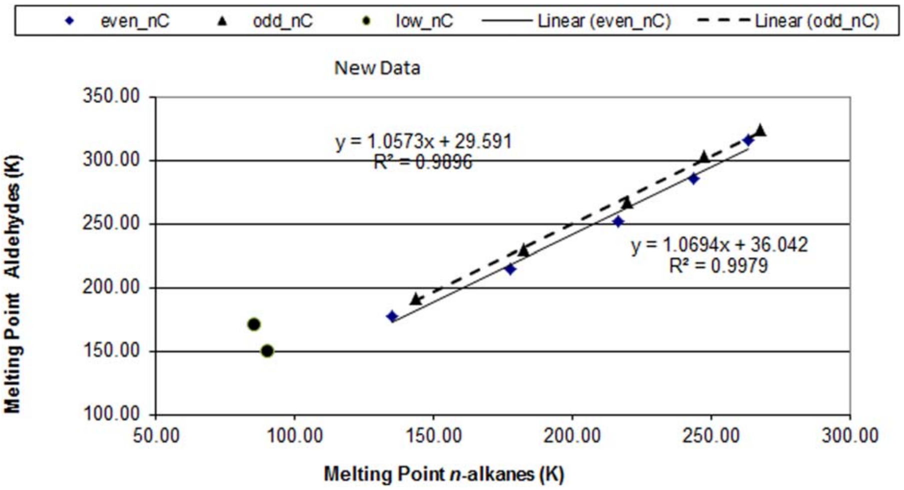

Figure 8. Plot of $(T_m)^t$ vs. $(T_m)^r$ for the *n*-alkane and 1-alkyl aldehyde (earlier, 2007 "accepted" values) series.

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

Table 7. Normal Melting Point Data and Prediction Results for the 1-Alkene and *n*-Alkyl Acetate Homologous Series

	Melting Point of 1-Alkenes*			Melting Point of Acetates*			Linear QPPR Prediction	
n_C	Value (K)	Uncertainty (%)	δT_m‡	Value (K)	Uncertainty	δT_m‡	Value (K)	Error (%)
3	87.90	<1	−0.10	175.15	<1	14.45		
4	87.80	<1	20.22	189.60	<1	−11.45		
5	108.02	<1	25.37	178.15	<1	21.50		
6	133.39	<0.2	**20.73**	199.65	<3	2.70		
7	154.12	<0.2	17.33	202.35	<1	9.90	200.58	0.88
8	171.45	<1	**20.46**	212.25†	<3	10.70	212.10	0.07
9	191.91	<0.2	14.99	222.95	<1	11.70	225.71	1.24
10	206.90	<0.2	**17.09**	234.65	<1	12.50	235.68	0.44
11	223.99	<0.2	13.96	247.15	<1	10.97	247.05	0.04
12	237.95	<1		258.12	<1		256.33	0.69

*DIPPR database.[18]

†"Accepted" value of 192.25 K was replaced after analysis.

‡Members of the higher δT_m series are shown in bold.

and the δT_m oscillations are fairly small (starting at $n_C = 6$, see Figure 9). Therefore, the 1-alkene series is used as reference series in cases where the variation between the odd and even n_C compounds in the target series is not distinguishable. Equations 5 and 6 can be used to predict T_m of the members of the 1-akene series (for $n_C > 10$). For such predictions, the parameter values $\beta_0 = 430.092$, $\beta_1 = -340.903$, and $\beta_2 = 1/77.463$ should be used in Eq. 5.

The coefficients of the linear and nonlinear QPPR models for predicting T_m of *n*-alkyl acetates with the 1-alkenes as reference series are shown in Table 8. The correlation coefficient of the linear QPPR is $R^2 = 0.9933$ and the prediction errors are $\varepsilon_{max} = 1.24\%$ and $\varepsilon_{avg} = 0.56\%$, significantly lower than the respective data uncertainties (Table 7). The nonlinear QPPR yields slightly lower prediction errors. Using the CG[9] method for predicting T_m for the members of the *n*-alkyl acetate series included in the training set yields the prediction errors: $\varepsilon_{max} = 6.93\%$ and $\varepsilon_{avg} = 4.8\%$, which are considerably higher than the ones obtained by the proposed method.

For the *n*-alkyl formate series, only five measured data points could be used in the training set. These data points enabled derivation of both the linear and the nonlinear QPPR's (Case 2 in Table 8); however, the R^2 in this case is lower and the prediction errors are slightly higher than for the acetate series.

For the *cis*- and *trans*−2-alkene series, only three data points (for each) were available for inclusion in the training set. The three data points enabled derivation of only the linear QPPRs with $\beta_0 = 0$ (the confidence interval on the calculated β_0 was larger in absolute value than the parameter

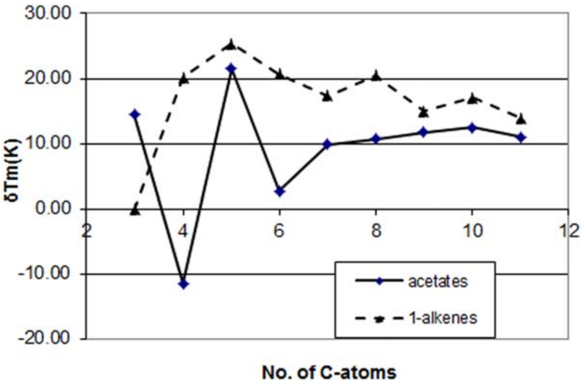

Figure 9. Plot of δT_m of 1-alkenes and 1-alkyl-acetates vs. the number of carbon atoms up to $n_C = 11$.

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

Table 8. Summary of Linear and Nonlinear QPPR Derivation for T_m of 4 Series (1-Alkenes the Reference Series)

		Available Data		Training Set			Linear Model			Linear Model		Nonlinear Model			Nonlinear Model	
No.	Series	n_{Cfirst}	n_{Clast}	n_{Cmin}	n_{Cmax}	Reference Number, *i*	β_0	β_1	R^2	ε_{max} (%)	ε_{avg} (%)	β_0	β_1	β_2	ε_{max} (%)	ε_{avg} (%)
1	*n*-alkyl acetates	3	12	7	12	n_C	98.07	0.6651	0.9933	1.24	0.56	112.16	−0.2804	1/13.809	0.95	0.39
2	*n*-alkyl formate*	2	11	5	9	n_C	120.72	0.5860	0.9865	2.45	0.94	90.66	0.3383	1/28.964	1.53	0.82
3	*cis*−2-alkenes†	4	12	5	6,8	n_C		1.0264	0.8967	8.93	4.80					
4	*trans*−2-alkenes	4	20	6	8	n_C		1.0673	0.9887	1.57	1.13					

*T_m data for $n_C = 10$ and 11 are predicted.
†T_m data for $n_C = 7$, 9–12 are predicted.

value). For the *trans*−2-alkenes, the respective QPPR yields predictive errors only slightly above the data uncertainty ($U_i = 1\%$). However, for the *cis*−2-alkenes, the three data points in the training set are not consecutive ($n_C = 5$, 6, and 8) and possibly because of that the prediction errors are considerably higher.

Conclusions

A new "reference series" method is proposed for analyzing T_m data and prediction of T_m values for members of homologous series. The method is based on modeling separately the relationship between the T_m values of the odd and even n_C compounds of the target and reference series, by two distinct QPPRs. Both converge to an accepted T_m^∞ value for long-chain substances. It was shown that using two separate models yield considerably more accurate results than state-of–the-art GC methods, where the same model is used for odd and even n_C compounds. In contrast to the QSPR methods, which require considerable amount and high precision T_m data for the members of the target series in order to identify the appropriate descriptors and the QSPR parameter values, it was shown that the same quality of predictions can be obtained using the "reference series" method, while relying on much smaller amount of data for the target series.

Matching the phases of T_m oscillations in the reference and target series often require introducing a lag or a lead in the n_C of the members of the reference series. It has been shown that the forward difference between two consecutive members of the reference series can be used for identifying the need for adjustment of the n_C values of the reference and target series.

The combination of the QSPR proposed by Paster et al.[14] for predicting T_m for long-chain members of the *n*-alkane series with the proposed QPPRs has proven to represent correctly the available data for the target series, and to provide long-range extrapolations consistent with the accepted T_m^∞ value of ~415 K. The predictions obtained by GC methods exceed this value for even moderately high n_C value (such as $n_C = 100$).

It is well known that experimental data are associated with some noise (e.g., measurement errors). In particular, the melting point is known to be very sensitive to the presence of small amounts of impurities, which may have a significant effect on the experimental T_m values. Therefore, it is important to establish reliable prediction methods for T_m even if experimental data are available as a means for consistency analysis of the data. The selection of an "accepted" T_m value of the highest probability from a large set of reported values can be a difficult task, as "family plots" may not yield smooth curves, as in the case of similar properties (e.g., T_C and T_b). Using two sets of "accepted" T_m values for the 1-alkyl aldehyde series, it was demonstrated that the proposed method is very appropriate for selecting the most consistent set of values.

The proposed method was applied to nine series, where the training set data could be separated to odd and even n_C compounds. For all of these series, ε_{avg} was less than 1% and ε_{max} was less than 3%. These values are considerably smaller than the prediction errors obtained with the GC methods. In the absence of noticeable T_m oscillations in the data of the target series, or insufficient data (with $n_C > n_{Cmin}$), separation to odd and even n_C compounds cannot be justified. Four such series where identified. Only in one case (where three nonconsecutive data points were used for the QPPR derivation), excessive prediction errors of $\varepsilon_{avg} = 4.8\%$ and $\varepsilon_{max} = 8.93\%$ were obtained.

The proposed method can be easily applied for developing QPPRs for additional target series for which some T_m data are available. The application of the method involves the use of the data and equations provided in the article for the reference series and application of a linear (and nonlinear) regression program for determining the QPPRs coefficients.

Literature Cited

1. Dearden JC. Quantitative structure-property relationships for prediction of boiling point, vapor pressure, and melting point. *Environ Toxicol Chem.* 2003;22:1696–1709.
2. Godavarthy SS, Robinson RL, Gasem KAM. An improved structure-property model for predicting melting point temperatures, *Ind Eng Chem Res.* 2006;45:5117–5126.
3. Hughes LD, Palmer DS, Nigsch F, Mitchell JBO. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log P. *J Chem Inf Model.* 2008;48:220–232.
4. 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;36:1887–1894.
5. Lazzus JA. Hybrid method to predict melting points of organic compounds using group contribution plus neural network plus particle swarm algorithm. *Ind Eng Chem Res.* 2009;48;8760–8766.
6. Preiss UP, Beichel W, Erle AMT, Paulechka YU, Krossing I. Is universal, simple melting point prediction possible? *Chem Phys Chem.* 2011;12:2959–2972.
7. Poling BE, Prausnitz JM, O'Connell JP. Properties of Gases and Liquids, 5th ed. New York: McGraw-Hill, 2001.
8. Joback KG, Reid RC. Estimation of pure-component properties from group-contributions. *Chem Eng Commun.* 1987;57:233–243.
9. Constantinou L, Gani R. New group-contribution method for estimating properties of pure compounds. *AIChE J.* 1994;40:1697–1710.
10. Wen X, Qiang Y. Group vector space (GVS) method for estimating melting point of hydrocarbons. *J Chem Eng Data.* 2002;47:286–288.
11. Marrero J, Gani R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilibria.* 2001;183–184:183–208.

12. Brauner N, Cholakov GSt, Kahrs O, Stateva RP, Shacham M. Linear QSPRs for predicting pure compound properties in homologous series. *AIChE J*. 2008;54:978–990.
13. Broadhurst MG. Extrapolation of the orthorhombic n-paraffin melting properties to very long chain lengths. *J Chem Phys*. 1962;36:2578–2582.
14. Paster I, Shacham M, Brauner N. Adjustable QSPRs for prediction of properties of long-chain substances. *AIChE J*. 2011;57:423–433.
15. Peterson BK. Relationships between the properties of families of materials. *Ind Eng Chem Res*. 2010;49:3492–3495.
16. Peterson BK. Property estimation using analogous series. *Ind Eng Chem Res*. 2011;50(12):7696–7704.
17. Shacham M, Paster I, Brauner N. Property prediction and consistency analysis by a reference series method. *AIChE J*. 2013;59:420–428.
18. 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, 2010, Available at: http//www.aiche.org/dippr, accessed May 2013.
19. National Institute of Standards and Technology (NIST). In: Linstrom, PJ, Mallard WG, editors. ChemistryWebBook, NIST Standard Reference Database Number 69. Gaithersburg, MD, 2005, Available at: http://webbook.nist.gov, accessed January 2013.
20. Sellers P, Stridth G, Sunner S. An examination of the constancy of the CH_2 increment in homologous series. *J Chem Eng Data*. 1978;23:250–258.
21. Todeschini R, Consonni V. Handbook of Molecular Descriptors. Weinheim: Wiley-VCH, 2000.

Manuscript received Feb. 6, 2013, and revision received April 13, 2013.