COMBINING CONSTANT PROPERTY PREDICTION TECHNIQUES FOR WIDER APPLICABILITY AND IMPROVED ACCURACY

Neima Brauner*, Mordechai Shacham°

* School of Engineering, Tel-Aviv University, Tel-Aviv, Israel
° Chem. Eng. Dept., Ben-Gurion University, Beer-Sheva, Israel

ABSTRACT

The property prediction field is a continuously evolving field where the objective is to extend the capabilities to additional groups of compounds and to additional properties, and to reach the ultimate goal of prediction accuracy within the uncertainty level of the experimental data. We have shown that these goals cannot be reached by using one type of technique (like group contribution, QSPR or QPPR). A broader algorithm should be developed which can fit the most adequate prediction technique/s to a wide range of compound-property combinations.

A general procedure for development a constant property prediction system is proposed. The procedure includes continuous update and maintenance of the physical property database, mapping the range of applicability of the descriptors in the descriptor database and checking the self-consistency of both the descriptor and property values. Included are also general guidelines and recommendations for the selection of the prediction method, the most appropriate training set and the regression technique to be used for a particular target compound-property combination.

INTRODUCTION

Current state of the art property prediction techniques rely on the use of databases which contain molecular structure related information and the corresponding physical property data for deriving property prediction models. These models are consequently used for predicting properties of "target" compounds for which only molecular structure related information is available. For example, the QSPR (Quantitative Structure Property Relationship) methods use molecular descriptors for representation of the molecular structure. Prediction models, in terms of molecular descriptors, are derived using stepwise linear or nonlinear regressions, artificial neural networks, particle swarm algorithms etc. The accuracy of the QSPR predictions is often limited because the "training set" used for deriving the QSPR may not adequately represent the structure and properties of some of the target compounds.

In contrast to the QSPRs, the "Group Contribution"(GC) methods use training sets "similar" to the target compound in order to derive models which represent the property variation in terms of the functional groups of the target compound. The compounds included in the training set (i.e., "predictive compounds") are supposed to represent the functional groups of the target compound. State of the art GC methods also discriminate between properties that change linearly with the molecular mass (M_W), such as liquid molar volume, enthalpy and entropy of formation, and properties that change nonlinearly with M_W (e.g., critical temperature and pressure and normal boiling point). Nonlinear change of some properties requires derivation of nonlinear prediction models, which are less reliable in extrapolation and involve the use of less rigorous nonlinear regression techniques. Independent studies (Poling et. al., [1]) have shown that the presently used GC methods yield predictions of satisfactory accuracy for several properties and various groups of compounds. However, there are groups of compounds (e.g. low M_W or very high M_W) and certain properties (such as solid properties) for which the prediction accuracies are unsatisfactory. By satisfactory accuracy we mean that the prediction error is within the experimental uncertainty level of the pertaining property data of the training set members.

Brauner et. al.,[2] have developed the so called Targeted QSPR (TQSPR) method. Using this method a unique linear QSPR is derived for a particular property of a selected target compound. The training set that is used for the derivation of the TQSPR is automatically selected by the TQSPR algorithm, and essentially includes compounds that contain the functional groups present in the target compound. The selection of descriptors to the QSPR model is performed by applying a step wise regression algorithm, which considers also the noise in the data. Since the functional groups are included in the descriptor database, the TQSPR method actually represents a generalization of the GC method. The availability of additional large number of descriptors that are related to the chemical structures of the compounds included in the training set enables the representation of the property variation within the selected training set by a linear model (rather than a nonlinear model which is often required in the GC method). The larger the training set size is, larger number of descriptors required to adequately represent the property variation. The number of descriptors that can be selected to the TQSPR model is limited only by the noise level in the data due to experiemtnal errors in the property values of the training set members [2]. However, larger training sets and more descriptors in the TQSPR model cause diffculties in estimating the prediction errors for the target compound and for the other (structurally similar) compounds which are not members the training set.

To reach the goal of minimizing the error associated with the predicted property value of the target compound (to the data uncertainly level) Shacham et al.,[3,4] suggested to tailor a tight training set to the target compound, which includes a limited number of compounds with high structural similarity to the target compound. Such a tight training set enables derivation of a linear TQSPR model in terms of a single descriptor (TQSPR1 model). The use of a single descriptor TQSPR1 model enabled development of several statistical indicators associated with the information pertaining only to the training set that enable reliable estimation of the prediction error for the desired property value of the target compound.

We have employed the TQSPR1 method (Shacham et al., [3]) for predicting 15 constant properties for 80 groups of compounds in order to discriminate between the property-compound combinations for which predictions of satisfactory accuracy are obtained and those associated with excessive prediction errors. Causes of excessive prediction errors that were identified, which are common to all QSPR methods, will be discussed below. These types of errors can be easily prevented by a careful selection of the training-set compounds and/or the descriptor for the TQSPR1 model.

There are however certain compound/property combinations for which neither the TQSPR/TQSPR1 methods (nor the GC methods) can provide predictions of satisfactory accuracy. One such combination involves the prediction of properties that change nonlinearly with M_W for long chain substances by extrapolation. Another combination involves prediction of solid properties (normal melting point and heat of fusion) in the region where differentiation between odd and even n_C compounds is required. In the following the three methods that we have developed: TQSPR1 (Shacham and Brauner[4]), long chain extrapolation (Paster et al.[5]) and reference series (Shacham et al.[6], Brauner and Shacham[7]) will be briefly reviewed and their combination for extending the range of a satisfactory prediction accuracy will be described.

RECENTLY DEVEOPED PROPERTY PREDICTION TECHNIQUES

The Dominant Descriptor Targeted QSPR Method (TQSPR1)

The Dominant Descriptor Targeted QSPR (TQSPR1) method was introduced by Shacham and Brauner [4]. The first stage of the method involves identification of the similarity group and a training set that is structurally related to the target compound. The similarity between the target compound (the compound for which the property needs to be predicted) and a potential predictive compound is measured by the partial correlation coefficient between their molecular-descriptor vectors. The training set is established by selecting from the similarity group the *p* predictive compounds (usually *p* = 10) with the highest correlation with the target compound for which experimental property values y_i are available.

The selected training set is used for the development of a TQSPR1 model for a particular property of the target compound. A linear structure-property relation is assumed of the form:

$$\mathbf{y} = \beta_0 + \beta_1 \boldsymbol{\zeta}_D + \boldsymbol{\varepsilon} \qquad (1)$$

In this equation **y** is a *p*-dimensional vector of the respective property values, $\boldsymbol{\zeta}_D$ is a *p*-dimensional vector of the (dominant) molecular descriptor (to be selected via a stepwise regression algorithm), β_0 and β_1 are the corresponding model parameters to be estimated, and $\boldsymbol{\varepsilon}$ is a *p*-dimensional vector of random errors.

To identify the dominant descriptor (DD), we examine the partial correlation coefficient between the vector of the particular property values of the compounds included in the training set (i.e., **y**) and the vector of descriptor values for these compounds, for all descriptors available in the database. These correlation coefficients will be referred to as the *descriptor-property* (*D-P*) correlation for the particular property. The DD, $\boldsymbol{\xi}_D$, is the descriptor that is associated with the highest value of the D-P correlation.

The so-obtained TQSPR1 can be subsequently employed for estimating the property value for the target compound, $\hat{y}_t$ in terms of the (known) dominant-descriptor value , ζ_{Dt}:

$$\tilde{y}_t = \beta_0 + \beta_1 \zeta_{Dt} \qquad (2)$$

Equation (2) is applicable, also, to additional compounds in the similarity group for which no property data are available.

There may be circumstances where there is a need to replace one or more members of the training set in order to improve the accuracy of the prediction. In such cases the next predictive compound to enter the training set is the one with the highest correlation with the target compound that also fulfills some additional requirements (like being in the same phase condition as the target compound at standard state). Similarly, the DD need to be occasionally replaced. In such cases the new DD is the one with the highest D-P correlation coefficient which also fulfills some additional requirements (e.g., preferring a non- 3D descriptor).

Considering the property value at $n_C \rightarrow \infty$ in derivation of the TQSPR1 model

Experimental property data are available, most often, for low n_C compounds as many of the properties of high n_C compounds cannot be measured due to thermal instability. Thus, property prediction of high n_C compounds often involves extrapolation (to higher n_C). Since the selection of the DD is carried out using only the available experimental data, extrapolation to higher n_C may yield inaccurate predictions. Recently, Paster et al. [5] presented a technique where the property value at $n_C \rightarrow \infty$ (y^∞) is also considered when selecting the DD, whereby the asymptotic behaviour of the DD should match the asymptotic behaviour of the property considered. For example, for properties that converge to a constant y^∞ value (such as normal boiling and melting temperatures), a DD is sought which also converges to a constant value for $n_C \rightarrow \infty$ ($\boldsymbol{\zeta}_D^\infty$). If the use of this descriptor in Eq.(1) enables accurate representation of the training set property data (**y**), as well as convergence to a generally accepted y^∞ value, the linear structure-property relationship is used. Otherwise, if Eq.(1) converges to a different y^∞ value, Eq. (1) is modified by including an additional non-linear correction term (with an additional regression parameter, β_2):

$$\mathbf{y} = \beta_0 + \beta_1 \boldsymbol{\zeta}_D - \left(\beta_0 + \beta_1 \boldsymbol{\zeta}_D - y^\infty\right)[1 - \exp(-\beta_2 n_C)] + \boldsymbol{\varepsilon}$$

and

$$y_t = \beta_0 + \beta_1 \zeta_{Dt} - \left(\beta_0 + \beta_1 \zeta_{Dt} - y^\infty\right)[1 - \exp(-\beta_2 n_C)] \qquad (3)$$

For properties which are additive in nature (like molar volume, enthalpy and entropy), in the limit of large n_C , each additional carbon unit contributes a fixed increment to the property value. For representation of such properties it is preferable to use n_C as a DD for extrapolation to high n_C compounds.

The reference series method

Shacham et al. [6] and Brauner and Shacham [7] introduced the "reference series" method for improving the prediction accuracy in homologous series of properties for which insufficient data are available. A two-stage procedure is used, whereby a linear (or nonlinear) Quantitative Structure-Property Relationship (QSPR) is fitted to a "reference" series, for which an adequate amount of precise data is available. This QSPR should represent correctly both the available data and the asymptotic behavior of the property. In the second stage a Quantitative Property-Property Relationship (QPPR) is derived to represent the predicted property values of a "target" series in terms of the property values of the reference series.

It has been shown that properties of compounds in two homologous series can be represented (at least locally) by a linear QPPR:

$$y_t = B_0 + B_1 y_r \qquad n_C \geq n_{C,min} \qquad (4)$$

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 n_C) in the target series, and B_0 and B_1 are parameters obtained by regression of the experimental data. The *n*-alkane series, for which the largest amount and highest precision property data are available, is most often used as the reference series. This method does not require computation of molecular descriptors, which can be considered as an advantage if no molecular structure files and/or descriptor computational programs are available. For predicting properties for long chain substances, a nonlinear version of Eq. (4) (similar to Eq. 3) can be used.

METHODOLOGY

The TQSPR1 method was evaluated by predicting 15 constant properties for a set of 471 compounds (Shacham et al.[3]). The database used in that study contains physical property data for 1798 compounds. Included in this data base are numerical values and data uncertainties (U_i) for 34 properties (e.g., critical properties, normal melting and boiling temperatures, heat of formation, flammability limits etc.). All the property data is from the DIPPR database (Rowley et al., [8]). The DIPPR database often contains a large number of experimental, predicted or smoothened values for a particular compound-property combination, while one particular value is designated as the "recommended" value. For the evaluation of the TQSPR1 method usually the recommended values were used.

Our database contains also 3224 molecular descriptors generated by the Dragon, version 5.5 software (DRAGON is copyrighted by TALETE srl, http://www.talete.mi.it) from minimized 3D molecular structures that were obtained from Rowley[9].

A Visual Basic program that uses the TQSPR1 method was developed. The program can operate in batch mode and attempts to predict all the properties for all the compounds in the data base. Most of the computations reported were carried out using this program. For the examples presented here, some additional details were computed with a special version of the SROV (MATLAB) program of Shacham and Brauner [10], which was revised in order to fit the needs of the TQSPR1 algorithm.

The identification of the similarity group and the training set is performed by using a subset of 294 selected descriptors, mostly 1-D descriptors from the "atom centred fragment" and "functional group count" categories. This subset was selected based on its ability to identify compounds belonging to the target compound's homologous series (if available), and to discriminate between compounds according to the number and location of branches and double bonds (Paster [11]). Some of the results were compared with predictions obtained by various GC methods. To this aim, the implementation of the GC methods in the Dorthmund Data Bank (DDBST, 2011 release, http://www.ddbst.de) was used for the predictions.

The selected 15 constant properties included in the study from the DIPPR database are listed in Table 1. Included in the Table are the symbols of the properties (as defined by DIPPR) and short descriptions of the properties. Out of the 34 properties (included in DIPPR), 19 were excluded for several reasons. Some were excluded because they are categorized as "defined", namely, they are calculated from other properties and/or from the molecular structure (i.e., the molecular weight, or critical compressibility factor that is calculated based on the critical properties). Excluded are also properties for which most of the values in the database are predicted and/or are associated with very high uncertainties. The 471 compounds included in the study could be associated with 84 groups (mostly homologous series).

Table 1. Constant properties included in the study from the DIPPR database

No.	Symbol	Property description
1	ENT	Absolute Entropy of Ideal Gas at 298.15 K and 100000 Pa
2	FP	Flash Point
3	HCOM	Net Enthalpy of Combustion Standard State (298.15 K)
4	HFOR	Enthalpy of Formation of Ideal gas at 298.15 K and 100000 Pa
5	HFUS	Enthalpy of Fusion at Melting Point
6	HSTD	Enthalpy of Formation in Standard State at 298.15 K and 100000 Pa
7	HSUB	Heat of Sublimation at the triple point
8	LVOL	Liquid Molar Volume at 298.15 K
9	MP	Melting Point (1 atm)
10	NBP	Normal Boiling Point (1 atm)
11	PC	Critical Pressure
12	RI	Refractive Index at 298.15 K
13	SSTD	Absolute Entropy in Standard State at 298.15 K and 100000 Pa
14	TC	Critical Temperature
15	VC	Critical Volume

SOURCES OF EXCESSIVE PREDICTION ERRORS AND RECOMMENDATIONS FOR IMPROVEMENTS

The results of the evaluation of the TQSPR1 method are reported in detail by Shacham et al., [3]. For the great majority of the compound/property combinations, predictions of acceptable accuracy (meaning that the prediction error is lower than the highest data uncertainty) were obtained. For many of the compound/property combinations the TQSPR1 predictions were compared with predictions obtained with state of the art GC methods (Constantinou and Gani, [12], Wen and Qiang , [13]) and the TQSPR1 predictions were of comparable or higher accuracy than the GC methods. There were, however, cases where the prediction error exceeded the data uncertainty. In the following the causes of excessive prediction errors are discussed. These are common to the QSPR methods that we are familiar with. Our recommendations for improving the prediction accuracy are also outlined

"Recommended" property value of a compound is inconsistent with values for similar compounds.

For many compound-property combinations several experimental values are reported in the literature, often with a considerable difference between them. It is not an easy task to categorize some of the values as "unacceptable" and to recommend the value which is most probably the correct one. For example, Shacham et al. [6] discuss a case where excessive prediction error the heat of formation of 1-octene is caused by use of the recommended value in the DIPPR database, which is however found to be inconsistent with the recommended values of the rest of the members of the 1-alkene series. Shacham et al. [6] show that selection of a different value as "recommended" from the available experimental data reduces the error below the data uncertainty.

To prevent errors that may be caused by miss-selection of the "recommended" property value, an iterative prediction-replacement process can be carried out in the database. For a group of similar compounds (like members of a homologous series), the prediction method is applied by targeting each of the members in turn, while using the others as a training set. If the prediction error exceeds the data uncertainty, another "recommended" value closer to the predicted value for the target compound is selected. This prediction-replacement process is carried out for all the members of the group until no more replacements need to (or can) be made.

Use of 2D descriptors whose range of definition does not cover the entire compound space

There are many 2D descriptors whose range of definition is limited in terms of the number of non-Hydrogen molecule atoms. Shacham et al. [3] mentioned for example the descriptor *EEig13d,* which is defined only for molecules that contain more the 12 non-Hydrogen atoms. Usually a (pseudo) zero value is assigned by Dragon to the descriptors when their calculation is attempted outside their range of definition. Obviously if the descriptor used in the QSPR model is associated with pseudo zero for the target compound and/or for some members of the training set, excessive prediction errors can be expected.

To prevent prediction errors of this source, the descriptors with limited range of definition should be clearly marked in the database (in order not to confuse real zeros with pseudo zeros). Compounds associated with a pseudo-zero value of the DD, should be removed from the training set (and replaced by other compounds from the similarity group, if necessary). When the DD value for the target compound is a pseudo zero, another DD must be selected according to the principles outlined in the description of the TQSPR1 method.

Inconsistency in the 3D molecular structure files and or 3D descriptors

We have studied extensively the potential sources of excessive prediction errors when 3D descriptors are used in QSPRs or TQSPRs ([3], [14], [15]). Paster et al. [14] demonstrated that the 3D descriptors are very sensitive to the method, parameters and the initial configuration used to generate the 3D optimized molecular structure. 3D structure (MOL) files available from different sources may be non-optimized, or partially optimized configurations, with no documentation on how these were obtained. Many computational algorithms may converge to a local minimum and finding a global optimum structure will depend upon the starting configuration. For flexible molecules (e.g., long chain hydrocarbons) the 2D to 3D conversion by different software can render different conformers, resulting in variation of the calculated value for the same 3D descriptor. Thus, if QSPR or TQSPR that contains 3D descriptors are used for prediction, it is very important to get the 3D structures of all the compounds involved (predictive and target) from the same reliable source.

Shacham et al., [3] presented an example where the critical volume of 1-decanol is predicted using a 3D (*Ds*) descriptor as DD. The prediction error in that case was very high. Examination of the 3D molecular structure (MOL) files of 1-decanol and its two immediate neighbours in the homologous series (1-nonanol and 1-undecanol) revealed inconsistency of the 1-decanol's MOL file, and consequently, also in the value of the *Ds* descriptor. The inconsistency in that case was rotation of the -CH_2OH group in the 1-decanol structure file compared to its position in the structure files of the other members of the 1-alkanol series. Such a difference is hardly noticeable, but it can affect a considerable variation in the value of the DD of the target compound, causing excessive prediction errors.

Methods for debugging of the descriptor data base have been developed [14, 15]. These should reveal inconsistent representation of some of the molecular structures and the associated 3D descriptors which are included in the database, and identify noisy 3D descriptors that exhibit extremely high sensitivity to insignificant variations in the 3D representation of the molecular structure. The use of the latter in QSPR models should be avoided. These methods should be applied also whenever new compounds are added to the database.

Phase change at standard state within the training set

There are several properties listed in Table 1 for which the reported values are for the "standard state", which is defined as the stable phase at 298.15 K and 1 bar. These properties include standard state enthalpy of formation (*HSTD*), entropy of formation (*SSTD*) and enthalpy (heat) of combustion (*HCOM*). Refractive indexes (*RI*) are reported usually for the liquid phase, which may not be the valid phase at the "standard state" (i.e., the compound is in gas phase at the standard state). Liquid molar volume (*LVOL*) is usually reported at standard state temperature, unless $T_C < 298.15$, in which case *LVOL* is reported at the normal boiling point (T_b), or at the triple point temperature if T_{tp} >298.15. Obviously, phase and/or temperature variations maybe reflected also in the property variation within the group of similar compounds (i.e., homologous series).

Shacham et al.[3] showed an example of prediction of *SSTD* of *n*-tridecane using *n*-alkanes in the $12 \leq n_C \leq 22$ region as a training set. The target compound and the members of the training set are in liquid phase at standard state, while for $n_C \geq 18$ the compounds are in solid phase. The TQSPR1 model obtained when using the full training set yields prediction error of 15 %. Removing the five compounds with $n_C \geq 18$ from the training set yields prediction with negligible error of 0.022 %.

Thus, for prediction of properties whose value is determined at the standard state, the phase condition of the target compound at standard state must be first determined. Only compounds which are in the same phase condition as the target (at the standard state) should be included in the training set.

Change of property values by orders of magnitude within the training set

Shacham et al., [3] demonstrated excessive prediction errors caused by order of magnitude change in the property values within the training set. For example, predicting the ideal gas enthalpy of formation (*HFOR*) for the first 11 members of the 1-alkene series: the DIPPR recommended values of *HFOR* are 2.023e7 for propylene, -5.0e5 for 1-butene and -2.162e7 for 1-pentene. Thus, there are at least two orders of magnitude difference between the *HFOR* value of 1-butene and the rest of the members of the 1-alkene series. Consequently the TQSPR1 prediction of the *HFOR* of 1-butene is very poor (86 % prediction error). However, if the coefficients of the TQSPR1 model are determined by minimization of the relative error (rather than the absolute error least squares) the prediction errors are reduced to acceptable levels for all the compounds.

Thus, in cases where there are order of magnitudes differences between the property values of the training set members (or a change of sign in the property value), it is recommended to use relative error least squares for determining the QSPR/TQSPR1 model parameters.

Long range extrapolation to higher n_C compounds

The DD descriptor selected based on property values of relatively low n_C compounds does not necessarily comply with the asymptotic behaviour of the property at large n_C. In such cases, applying the TQSPR for prediction the property value of a high n_C compound may result in excessive prediction errors. For example, the highest n_C compound for which property data are available for the 1-alkene series in the DIPPR database is 1-triacontene (with n_C= 30). Shacham et al., [16] reported *LVOL* prediction results for 1-triacontene using 10 1-alkenes in the range of $11 \leq n_C \leq 20$ as the training set. Relying on the training set *LVOL* data for identification of the DD, yields *LVOL* prediction which differ by 21% from the value reported by DIPPR. Selecting the DD while matching its asymptotic behaviour at $n_C \rightarrow \infty$ with the asymptotic behaviour of *LVOL* (as explained in the " Considering the property value at $n_C \rightarrow \infty$ …" section) leads to a TQSPR1 model which gives LVOL value with 1.7 % difference from the value reported by DIPPR. This difference is considerably lower than the DIPPR data uncertainty (10%).

Thus, in cases of extrapolation to higher n_C compounds it is always advisable to match the asymptotic behaviour of the DD with the asymptotic behaviour of the target property.

Irregularities in the solid properties related to the crystalline structure

The irregularities with regard to solid properties (e.g., normal melting temperature, T_m and heat of fusion, *HFUS*) are demonstrated in reference to the T_m of members of the *n*-alkane and the *n*-alkanoic acid homologous series (Fig. 1). Three distinct regions can be identified in the T_m curves. 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 3 members of the *n*-alkane series and the first 4 members of the alkanoic acid series. At medium range 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 lower 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 (Marano and Holder, [17]). In the high n_C region, there is a smooth increase of T_m with a diminishing slope converging to the asymptotic T_m = 415 K value for $n_C \rightarrow \infty$ (Paster et al., [5]).

Brauner and Shacham, [7] reported very good results for predicting T_m in the medium and high n_C regions using the "reference series" method. In the medium n_C region two versions of Eq. 4 are used: one for odd n_C target compounds and one for even n_C target compounds. In this case we use in Eq. 4 the definition of $y_t = (T_m)_{t,nC}$, where n_C is the number of carbon atoms in the respective member of the target series, and $y_r = (T_m)_{r,i}$ where *i* is the number of carbon atoms in the matching member of the reference series, it can obtain the values $i = n_C$ or $i = n_C +1$ or $i = n_C -1$ (see details in Brauner and Shacham, [7]).

For predicting T_m in the high n_C region, Brauner and Shacham, [7] provided a QSPR (of the form of Eq. 3) for predicting T_m of the members of the n-alkane (reference) series, and recommended the use of a nonlinear version of the QPPR (Eq. 4) for members of other target series.

Further development of the "reference series" method is underway to enable reliable prediction of other solid properties of compounds belonging to of homogenous series, (such as *HFUS*) and to extend its applicability for property predction of other groups of compounds.

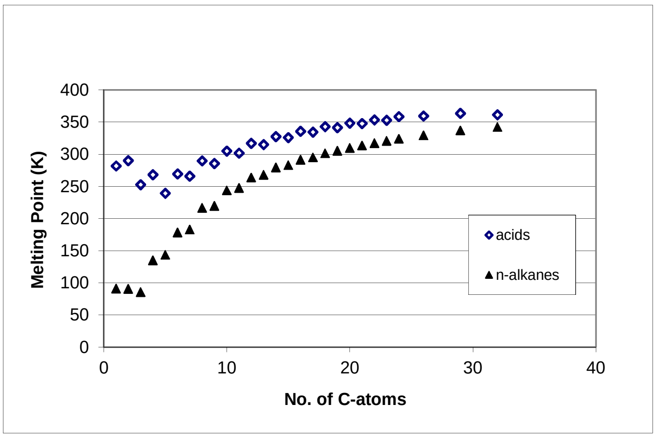

Fig. 1. Plot of normal melting point data of *n*-alkanes and *n*-alkanoic acids vs. n_C up to n_C = 32.

Prediction errors for low carbon number compounds and first members of homologous series

Shacham et al. [3] list extrapolation to a low n_C target compound as the most common cause for excessive prediction error. In Fig. 1, for example, the T_m values of *n*-alkanes with $n_C > 3$ have very little relevance when predicting T_m for methane, as the trend of change of T_m in the low n_C region differs from the trend in the medium n_C region. This change in trend is caused by the influence of the functional groups that is attached to the methyl group (the dominant group in higher n_C compounds) on the property value. For low n_C compounds the effect of the functional group is dominant, but it diminishes with the addition of methyl groups. Because of that experimental information on the property values of low n_C compounds (or first members of homologous series) is essential for correctly assessing the functional groups contribution to the property values.

Phase change at standard state for the members of the training set can be another reason for excessive prediction errors for low n_C compounds. Shacham et al. [3] used the first 11 members of the 1-alkene series to demonstrate this difficulty. For this series the first three members: ethylene, propylene and 1-butene, are in the gaseous phase at standard state, while the rest of the compounds are in the liquid phase. The liquid phase compounds cannot provide reliable information regarding the standard state properties of the gas phase compounds.

The absolute value of some properties of low n_C compounds is often much smaller than the absolute values of the same properties of the high n_C compounds. This may inflate the relative errors in prediction of the property for a low n_C compound. Shacham et al. [3] provided an example where similar absolute prediction error for T_b of ethylene (T_b= 169.41 K) yields much higher relative error than for T_b of 1-dodecene (T_b = 486.15). As discussed above, minimization of the sum of squares of the relative error should be considered in such cases.

CONCLUSIONS

The property prediction field is a constantly evolving field where the objective is to extend its applicability to additional groups of compounds and additional properties, and to reach the ultimate goal of prediction accuracy within the experimental data uncertainty level. We have shown that these goals cannot be reached by using one type of technique (like group contribution, QSPR or QPPR), but a broader algorithm should be developed, which can fit the most adequate prediction technique/s to a wide range of compound-property combinations.

Based on our experience in developing several new prediction techniques and applying them to a wide variety of compound-property combinations, we propose the following procedure and principles in developing a general property prediction system.

1. Continuous update of the property data base with new experimental property data that becomes available. Continuous maintenance of this database by verifying the consistency of the "recommended" property values with such values of similar compounds.
2. Continuous maintenance of the descriptor database by mapping the range of applicability of certain 2D descriptors and checking the consistency of 3D structure (MOL) files and 3D descriptors with MOL files and descriptors of similar compounds.
3. When predicting properties for a new target compound, the prediction technique needs to be adjusted to the specific compound/property combination and the available training set. For fluid properties the use of the TQSPR1 method is recommended provided that enough similar predictive compounds are available. The selection of the DD should consider both the training set property data and the asymptotic property behaviour (in case of extrapolation to larger n_C). For properties defined at standard state, first the phase condition of the target at standard state should be predicted. Only predictive compounds of the target's phase condition can be included in the training set for predicting the desired property. If the property values change by orders of magnitude within the training set, the minimization of the relative (instead of absolute) errors is recommended. For solid properties (and fluid properties with insufficient amount of predictive compounds) the use of the reference series method should be preferred.

Prediction of properties by extrapolation to low n_C compounds (or first members of homologous series) so as to keep the prediction error below experimental uncertainty level requires further research.

The extension of TQSPR1 method to predict temperature – dependent properties (e.g., vapor pressure) and other phase equilibrium related properties (e.g., interaction parameters for phase equilibrium calculations, when applying EoS) was demonstrated in Shacham et al., [18] and Paster et al., [19]). Preliminary results show that our methods can be used also for predicting other important parameters (e. g., activity coefficients for non-ideal binary systems).

REFERENCES

[1] B.E. Poling, J.M. Prausnitz and J.P. O'Connel, Properties of Gases and Liquids, fifth ed. McGraw-Hill, New York, 2001.

[2] N. Brauner, R. P. Stateva, G. St. Cholakov and M. Shacham, A structurally ‘‘targeted’’ QSPR method for property prediction,. Ind. Eng. Chem. Res., vol. 45, pp. 8430–8437, 2006.

[3] M. Shacham, M. Elly, I. Paster and N. Brauner, Evaluation and refinement of the property prediction stage of the targeted QSPR method for 15 constant properties and 80 groups of compounds, *Chem.Eng.Sci* , In Press, Accepted Manuscript, Available online 18 April 2013.

[4] M. Shacham and N. Brauner, Analysis and Refinement of the Training set in Predicting a Variety of Constant Pure Compound Properties by the Targeted QSPR Method , *Chem. Eng. Sci.*, vol. 66, pp. 2606–2615, 2011

[5] I. Paster, M. Shacham and N. Brauner, Adjustable QSPRs for prediction of properties of long-chain substances, *AIChE J.*, vol 57, pp. 423–433, 2011.

[6] M. Shacham, I. Paster and N. Brauner, Property Prediction and Consistency Analysis by a Reference Series Method,

AIChE J., vol. 59, pp. 420–428, 2013.

[7] N. Brauner and M. Shacham, Prediction of Normal Melting Point of Pure Substances by a Reference Series Method, *AIChE J.*, DOI 10.1002/aic.14128, Accepted manuscript, online: 28 APR 2013.

[8] R. L. Rowley, W. V. Wilding, J. L. Oscarson, Y. Yang and, N. A. Zundel, DIPPR Data Compilation of Pure Chemical Properties Design Institute for Physical Properties. Brigham Young University Provo Utah, 2010. ‹http//dippr.byu.edu›.

[9] R. L. Rowley, Personal Communication, 2010.

[10] M. Shacham, and N. Brauner, The SROV Program for Data Analysis and Regression Model Identification, *Comp. Chem. Eng.*, vol. 27, pp. 701-714, 2003.

[11] I. Paster, Prediction of Properties Using Molecular Descriptors, PhD Thesis, Ben-Gurion University of the Negev, Beer-Sheva, 2013.

[12] L. Constantinou, and R. Gani, New Group-Contribution Method for Estimating Properties of Pure Compounds. *AIChE J.*, vol. 10, pp. 1697-1710, 1994.

[13] X. Wen and Y. Qiang, Group vector space (GVS) method for estimating boiling and melting points of hydrocarbons. *Chem. Eng. Data*, vol. 47, pp. 286-288, 2002.

[14] I. Paster, M. Shacham and N. Brauner, Investigation of the Relationships between Molecular Structure, Molecular Descriptors and Physical Properties, *Ind. Eng. Chem. Res.*, vol. 48, pp. 9723–9734, 2009.

[15] I. Paster, G. Tovarovski, M. Shacham and N. Brauner, Combining Statistical and Physical Considerations in Deriving Targeted QSPRs Using Very Large Molecular Descriptor Databases, pp. 61 - 66 in S. Pierucci and G. Buzzi Ferraris (Eds), 20th European Symposium on Computer Aided Process Engineering – ESCAPE 20, June 6- 9, 2010, Ischia, Naples, Italy.

[16] M. Shacham, M. Elly, I. Paster, and N. Brauner, Self-Consistency Analysis of Physical Property and Molecular Descriptor Databases Using a Variety of Prediction Techniques, paper 181b, Presented at the 2012 AIChE Annual Meeting, Pittsburgh, PA, Oct. 28 – Nov. 2, 2012.

[17] J. J. Marano and G. D. Holder, General equations 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., vol. 36, pp. 1895-1907, 1997.

[18] M. Shacham, G. St. Cholakov, R. P. Stateva and N. Brauner, Quantitative structure-property relationships for prediction of phase equilibrium related properties. Ind. Eng. Chem. Res., vol. 49, pp. 900–912, 2010.

[19] I. Paster, N. Brauner and M. Shacham, Analysis and refinment of the TRC-QSPR method for vapor pressure prediction. The open Thermodynamic Journal, vol. 5, pp. 29-39, 2011