Fluid Phase Equilibria 298 (2010) 131–142

Contents lists available at ScienceDirect

Fluid Phase Equilibria

journal homepage: www.elsevier.com/locate/fluid

Predictive correlations based on large experimental datasets: Critical constants for pure compounds☆

Andrei Kazakov*, Chris D. Muzny, Vladimir Diky, Robert D. Chirico, Michael Frenkel

Thermophysical Properties Division, National Institute of Standards and Technology, 325 Broadway, Boulder, CO 80305-3328, USA

ARTICLE INFO

Article history:
Received 10 November 2009
Received in revised form 2 July 2010
Accepted 14 July 2010
Available online 27 July 2010

Keywords:
Correlation
Critical properties
Empirical modeling
Property estimation
Quantitative Structure–Property Relationships
Support Vector Machines

ABSTRACT

A framework for development of estimation methods is demonstrated using prediction of critical constants for pure compounds as an example. The dataset of critical temperature T_c and critical pressure p_c for over 850 compounds used in the present work was extracted from the TRC SOURCE data archival system and is based exclusively on experimental values taken from the literature. Experimental T_c and p_c values were critically evaluated using the methods of robust regression and their uncertainties were assigned in a rigorous manner. The correlations for critical constants were developed based on Quantitative Structure–Property Relationships (QSPR) methodology combined with the Support Vector Machines (SVM) regression. The propagation of the experimental uncertainties into the predictions produced by the correlations was also assessed using a procedure based on stochastic sampling. The new method is shown to perform significantly better than a number of commonly used estimation methods.

Published by Elsevier B.V.

1. Introduction

In the modern information-driven world, large comprehensive experimental data collections have not only become more readily available, but are also more accessible to systematic studies due to rapid advances in database technologies. These electronic data collections are also becoming increasingly dynamic; i.e., they respond much more promptly to changes as new data become available or some old data are corrected. This is especially true for well established, high-demand fields such as thermophysical property data for which computer-based databases were adopted early on. Due to their practical importance, thermophysical data have been collected intensively throughout the history of thermodynamics, leading to massive amounts of available information, and this trend becomes even more apparent at present [1]. Unfortunately, even such a high rate of experimental data accumulation is not likely to ever match the growing demand in various engineering applications. Therefore, in most practical situations, it is necessary to complement existing experimental data with property estimation methods [2]. While methods based on theoretical considerations are generally preferable, in many cases, an underlying theoretical foundation is too complex to be used in practical estimation methods. Consequently, empirical correlations derived from existing experimental data have always played an important role in thermophysical property estimation. These empirical correlations were traditionally developed on the basis of some reliable but often limited data compilations. At the present time, under conditions of a rapidly changing "data landscape", the next logical step is to allow *empirical correlations to evolve with the data that they are based on*. Furthermore, the availability of large electronic data collections opens possibilities for the development of new, more general empirical correlations that are based on much larger datasets than those used historically and that take advantage of modern data mining technologies [3]. The principles of the methodology for correlation development formulated here are as follows:

1. Input property values used for correlation development must be evaluated based exclusively on original experimental data;
2. Input property values must include estimated uncertainties that should be accounted for in the process of the correlation development;
3. Provisions for interpolation and/or limited extrapolation of the experimental data must be implemented to obtain input property value estimates that are either not available or are poorly defined from direct measurements;

☆ Contribution of the U.S. National Institute of Standards and Technology and not subject to copyright in the United States. Trade names are provided only to specify procedures adequately and do not imply endorsement by the National Institute of Standards and Technology. Similar products by other manufacturers may be found to work as well or better.

* Corresponding author.
E-mail address: andrei.kazakov@nist.gov (A. Kazakov).

0378-3812/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.fluid.2010.07.014

4. There should be a mechanism for estimating uncertainties of predictions generated by the correlation [4]; these uncertainties can also serve as a criterion for the applicability of a correlation to a system under consideration;
5. The overall procedure should be extendable to different properties with no or minor modifications;
6. The overall procedure should have robust algorithmic implementation, thus allowing rapid modification of a correlation following changes in stored data.

Only partial implementation of the above rules can be found in the existing approaches and some of the items are commonly ignored. The present work describes a systematic practical implementation of the above principles using correlation of critical constants for pure compounds as an example. First, we discuss the key points that necessitate the formulated principles and present the practical implementation that meets them. New correlations for critical constants are produced as a result of this effort. A detailed analysis of the produced correlations (including comparison of their performance with that of a number of existing estimation methods) is also presented.

Individual procedures and numerical algorithms used in this work were chosen based on a careful screening of the literature for the most suitable approach as well as extensive testing. An implementation framework based on Quantitative Structure–Property Relationships (QSPR) was selected. QSPR (also referred to as Quantitative Structure-Activity Relationships, QSAR, when applied to biological activities) is well established, and is one of the most widely used methodologies for data mining and correlation development in pharmaceutical and agricultural applications [5]; it is also gaining use in empirical modeling of a wide variety of physical properties [6–8]. The basic idea of QSPR is to relate a property of interest to molecular numerical features derived theoretically from chemical structures. These numerical features are referred to as *descriptors*. The relationship is established through regression analysis using large (i.e., statistically significant) collections of data, which makes this approach particularly attractive for data mining.

As noted earlier, the present work is focused on prediction of critical constants for pure compounds: the critical temperature and the critical pressure. Critical constants represent one of the most practically important thermodynamic properties, and development of methods for their estimation has a long history [2]. Most presently accepted approaches are based on Group-Contribution (GC) methodologies (e.g., [9–13]); however, QSPR-based methods are gaining recognition from the early work of Grigoras [14] to extensive recent developments [15–22]. The latest edition of the popular reference book, "The Properties of Gases and Liquids" [2], includes coverage of QSPR-based approaches.

The following sections will describe in detail the individual steps involved in the proposed implementation: processing of original experimental information, generation of molecular structures and descriptors, and regression analysis.

2. Experimental data

A principal goal of this work was to develop correlations based *exclusively on experimental data*. In most literature studies, "experimental" values are normally taken from some reputable reference compilation without detailed analysis of their origins. In many cases, however, the compiled values are not experimental but rather obtained either directly from, or with the aid of, other correlations. For example, in a recent study [22], the authors used, seemingly, the largest dataset (over 1230 compounds) ever considered in QSPR modeling of critical constants. All property values were taken from the 1999 version of the DIPPR compilation. However, at the time of this writing, even the most recent version of the DIPPR database [23] contains only 575 compounds with critical temperatures labeled as "experimental", and even fewer (430) compounds with critical pressures. As a result, the majority of property values in the dataset used to develop correlations were in fact estimated from other correlations. Furthermore, the problem goes even deeper: compilations may erroneously list some values as "experimental" due to misinterpretation of the original sources. These errors tend to go unnoticed and are commonly propagated to other data collections.

While it is certainly tempting to use the largest dataset possible by including "good" recommended values regardless of their origin, this action has deleterious effects on any empirical correlation produced *via* statistical treatment of the data. Inclusion of estimated values introduces artificial bias, which, in extreme cases, may even dominate the result, yielding what appears to be a good, low-noise, accurate correlation, but factually is a meaningless reproduction of one empirical relationship with another [24].

The direct use of experimental data from the original sources, as advocated here, bypasses contamination of datasets with estimated values. However, it does pose challenges of its own. Specifically, it requires a robust and rigorous protocol for pre-processing of the primary data, as described next.

All original experimental data used in the present study were taken from the NIST/TRC SOURCE data archival system [25–27]. SOURCE is one of the largest collections of experimental thermophysical property data in the world and, at the time of this writing, contains nearly 4 million data points. SOURCE is also actively maintained to be concurrent with newly published information; the number of data points is currently increasing at the rate of about 0.5 million per year. The data collection procedures developed over many years facilitate enforcement of data quality. Every experimental value is stored along with its estimated combined uncertainty [4] obtained based on the information from the original source as well as using custom in-house software [28] and data expert analysis. Among other properties, SOURCE contains one of the world's most comprehensive collections of critical constants.

The present study is focused on two critical constants: the critical temperature T_c and the critical pressure p_c. In cases when only one datum was available for a given compound, the value and its uncertainty were taken directly from SOURCE. When multiple data points were present, they were evaluated to produce a single recommended value and its uncertainty, as described next. One of the major problems of any large data collection is the presence of outliers caused primarily by either erroneous or erroneously entered data. Treatment of data contaminated with outliers has been a subject of extensive studies, and numerous approaches have been proposed. Here, the methods of *robust estimation* [29] were chosen. Evaluation of the critical temperature was performed with the maximum likelihood M-estimate based on the Cauchy distribution [29]. In cases when only multiple p_c points were available without or with only limited other vapor pressure data, the critical pressure was evaluated with the same M-estimate approach as was used for T_c evaluation. If sufficient vapor pressure data are available, a more general procedure, similar to the one used by NIST/TRC ThermoData Engine (TDE) [30], was adopted. All relevant vapor pressure measurements (inclusive of triple and boiling points) for a given compound were retrieved from the database and fitted with the Wagner equation [31,32]. The critical pressure p_c is obtained as one of the fitted parameters. Use of the entire vapor pressure curve not only improves the accuracy of the evaluated p_c, but also allows estimation of critical pressures *via* limited extrapolation when experimental p_c data are not available [33]. Fitting of data was carried out with MM-estimate methodology [34] based on the FAST-LTS formulation [35] of Least-Trimmed-Squares (LTS) method and M-estimate step using the Tukey's bisquare objective

Table 1
Statistical distribution of compounds with available experimental data.

Composition	Count	Percentage
H/C/O	337	40.0
H/C	193	22.5
H/C/halogen	90	10.4
H/C/N	59	6.8
H/C/O/halogen	45	5.2
C/halogen	40	4.6
H/C/O/Si	23	2.7
H/C/S	15	1.7
H/C/O/N	10	1.2
C/O/halogen	9	1.0
Si/halogen	8	0.9
H/C/Si/halogen	5	0.6
N/halogen	4	0.5
Contains P	3	0.4
H/C/O/S	2	0.2
H/C/Si	2	0.2
Other[a]	20	2.3

[a] Mostly small inorganic compounds such as SF_6, H_2S, etc.

function [29]. To improve stability of the fitting procedure, several numerical constraints derived from the Waring criterion [32,33,36] were applied.

Upon completion of the evaluation procedures, the final number of compounds with at least one evaluated critical constant was 920. A subset of compounds was removed from further consideration based on the following criteria:

- complexes (salts, organometallic complexes, etc.);
- compounds containing elements not supported by the quantum-chemical methods used;
- compounds containing isotopes that are different from the naturally occurring compositions;
- open-shell compounds;
- compounds associated with references containing known erroneous data.

Most of the elimination criteria were dictated by the basic consideration of maintaining molecular structural integrity and similarity in the gas and liquid phases. The final compound count was 865 (i.e., cases for which at least T_c was available); among them, 677 compounds had both T_c and p_c evaluated. To the best of our knowledge, this dataset represents the largest collection of critical constants for pure compounds derived exclusively from experimental data of traceable origins. Evaluated critical temperatures for this set range from 126.19 K (molecular nitrogen) to 913 K (*p*-terphenyl); critical pressures vary from 472 kPa (*n*-hexatriacontane) to 22064 kPa (water). A statistical summary for the resulting set with respect to atomic composition of included compounds is given in Table 1. While this collection appears reasonably diverse, it is also strongly biased toward hydrocarbons and oxygenated hydrocarbons that represent the majority (62.5%) of the population. This is expected and is unavoidable from a practical point of view, as hydrocarbons and their oxygen derivatives are generally the most well studied classes of compounds. Compound distributions according to the chemical families are provided in Supplementary material.

3. Generation of molecular structures and descriptors

As a starting point, a database of two-dimensional (2D) structure representations (inclusive of stereo assignments where applicable) was compiled for all compounds from the final set described in the previous section. Subsequent generation of three-dimensional (3D) representations followed a two-step procedure. First, initial 3D structures were produced from 2D representations using a range of software tools [37–41]. These tools produce 3D molecular geometries by use of stored template information and/or distance geometry treatment [42] and further refine it with molecular mechanics-based optimization. Some of the tools used (i.e., BALLOON [39] and Open Babel [40]) also provide low-energy conformer search capabilities. Final 3D geometries were generated by optimization of the initial 3D representations at the semiempirical PM3 level [43] using MOPAC [44]. Semiempirical methods offer a compromise between computational speed and chemical accuracy because the sizes of the molecules of interest can be quite large making the usage of higher-level quantum-chemical methods too computationally expensive.

Special attention was paid to large flexible molecules (defined here as compounds with 15 or more rotatable bonds). It is generally recommended to consistently use the lowest energy conformers in QSPR studies [45]; the search for low-energy conformations was conducted with simulated annealing using the TINKER package [46] and its implementation of the MM3 forcefield. In a limited number of cases, when molecules contained atoms poorly supported by molecular mechanics, annealing cycles were performed with the much more computationally expensive PM3 potential using an in-house computer code based on the DYNAMO library [47].

Molecular descriptors were computed with CODESSA [48], a package designed to process MOPAC outputs produced after 3D structure generation directly. CODESSA generates over 450 constitutional, topological, geometrical, electrostatic, quantum-chemical, and thermodynamic descriptors. A number of descriptors were manually removed from further consideration. Rejection of certain descriptors was driven by the intention to produce a more general model that depends on fundamental molecular properties and, as such, might have a wider applicability. It is a common practice to reject descriptors based solely on statistical considerations; however, when dealing with statistically-unbalanced (biased) data, such as the dataset used here, manual removal guided by physical considerations was chosen. The rejected descriptors included those

- associated with specific atoms (e.g., “relative number of Br atoms”);
- that involve ambiguous definitions and/or are strongly discrete (e.g., “number of rings”);
- with duplicated physical meanings.

The latter primarily concerns the descriptors that represent similar or identical properties computed using either empirical partial charges [49–51] or the partial charges derived directly from quantum-chemical (PM3) analysis. From the empirical modeling point of view, mixing empirical and quantum-chemical charges may be considered beneficial and is performed routinely (e.g., [22]). The rationale behind using these duplicated descriptors is that it diversifies the descriptor pool. Additionally, one may hope for possible “cancellation of errors” as different approaches may have different deficiencies. The disadvantage of this approach is that it may obfuscate the analysis of the final results, i.e., problems related to these descriptors are harder to trace to their origin (the method). Here, by rejecting the descriptors with duplicated physical meanings, descriptor consistency was chosen over a possible gain in their diversity. The final set of descriptors used in this study included 175 entries; a complete list is given in Supplementary material.

4. Regression analysis

To obtain a correlation that relates computed descriptors to the property values, one needs to analyze the data using regression analysis. This step is the most critical part of QSPR, and various approaches have been tested and used in recent years [52]. The

choices normally include multiple linear regression (MLR), partial least squares (PLS), various implementations of neural networks (NN), k nearest neighbors (kNN), and support vector machines (SVM). SVM methodology [53] is a relatively recent addition to the nonlinear regression methods used in QSPR, however, it has been reported [20,54,55] to yield similar, and often superior, performance as compared with other approaches. In addition to the empirical evidence of its good performance for QSPR (and many other real-world problems), SVM also exhibits several numerical advantages. Numerical solution of SVM regression is reduced to solving a constrained quadratic programming problem that has a unique global minimum, and a number of efficient standard solution algorithms are available. Finally, SVM is extremely robust in dealing with high-dimensional data, a feature of particular practical importance for QSPR problems when compounds are associated with many descriptors. The SVM approach was, therefore, chosen for regression analysis in the present study.

The problem of nonlinear SVM regression (ε-SVR formulation) is stated as follows [56]: for a given set of N data points, $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $\mathbf{x}_i$ is a vector of variables and y_i is the value of the observable for the ith point, minimize

$$\frac{1}{2}\|\mathbf{w}\| + C\sum_{i=1}^{N}(\xi_i + \xi_i^*), \tag{1}$$

subject to

$$y_i - \mathbf{w}\cdot\phi(\mathbf{x}_i) - d \le \varepsilon + \xi_i,$$
$$\mathbf{w}\cdot\phi(\mathbf{x}_i) + d - y_i \le \varepsilon + \xi_i^*,$$
$$\xi_i, \xi_i^* \ge 0, i \in [1, N].$$

Minimization is performed over the vector of coefficients $\mathbf{w}$, vectors of slack variables $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$, and d. The basic idea of SVM regression is to define a "tube" of radius ε that best fits the data $\boldsymbol{y}$ and has the centerline defined as a function of $\mathbf{x}$ that is as flat as possible. A preset parameter ε defines the nominal accuracy of the SVM approximation, i.e., the deviations below ε are considered unimportant and do not penalize the objective function (1). The first term in (1) enforces the "flatness" of the tube's centerline, and the second term controls the deviations from the data; preset parameter C defines the tradeoff between the flatness of the function and the accuracy of the approximation. The function $\phi(\mathbf{x})$ provides a nonlinear mapping of the original variable vector $\mathbf{x}$, thus making an overall SVM approximation (i.e., the tube's centerline),

$$f(\mathbf{x}) = \mathbf{w}\cdot\phi(\mathbf{x}) + d,$$

a more general, nonlinear function of $\mathbf{x}$. Transformation of the minimization problem (1) into a dual formulation by introducing the Lagrangian multipliers shows that explicit knowledge of the nonlinear mapping function $\phi(\mathbf{x})$ is not required for the actual solution; instead, the *kernel*,

$$K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j), \tag{2}$$

needs to be defined. The nonlinear choices of kernel function normally include polynomial,

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i\cdot\mathbf{x}_j + 1)^{\omega}, \tag{3}$$

and Gaussian, more commonly referred to as radial basis function (RBF),

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2), \tag{4}$$

kernels, where ω and γ are the kernel parameters. The RBF kernel (4) appears to be the most popular choice for the vast majority of practical SVM regression applications, and for QSPR problems in particular [54,55]. As was pointed out earlier, the actual numerical solution of the SVM regression problem is obtained in terms of the kernel:

$$f(\mathbf{x}) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + d, \tag{5}$$

where α_i and α_i^* are the Lagrangian multipliers produced from the solution of the optimization problem. The SVM expansion (5) is known to be sparse, i.e., the coefficients $(\alpha_i - \alpha_i^*)$ take nonzero values only for a subset of $i \in [1, N]$. The data points associated with nonzero terms in Eq. (5) are called the *support vectors*.

For the purposes of the present work, the optimization problem (1) was formulated in a slightly modified form [57]:

$$\frac{1}{2}\|\mathbf{w}\| + C\sum_{i=1}^{N}c_i(\xi_i + \xi_i^*), \tag{6}$$

i.e., the extra weights c_i were introduced for each point to account for individual experimental uncertainties.

In applications of SVM regression to QSPR problems, all descriptors and the property values are scaled to [0,1] intervals, and the values computed with Eq. (5) are subsequently mapped back to the physical space yielding the predicted property value. To complete the definition of the problem, one needs to set the SVM parameter C and, if applicable, the kernel parameters. The optimal settings for these parameters are not known *a priori* and depend on a particular dataset. Therefore, the practical implementation of the SVM regression is carried out in several steps [57]. First, the dataset is split into three parts: the *training* set, the *validation* set, and the *testing* set. The SVM parameters are then optimized to achieve the best accuracy in predicting the data from the validation set with the model generated (or *trained*) using only the training set. Once the optimal SVM parameters are established, training and validation sets are combined, and the final model is produced with the optimized SVM parameters using this combined set. Finally, the fidelity of the model is confirmed by comparing its predictions with the data from the remaining (unused) testing set.

Special consideration was given to selection of the testing and validation sets. In most applications, this selection is performed randomly. However, considering the bias toward certain classes of compounds that is present in the dataset under consideration (and this situation is likely to be the case for any dataset of experimental data from the literature), random selection is not the best strategy. Golbraikh et al. [58] recognized this problem and suggested several stratified selection strategies based on sphere exclusion algorithms shown to be superior to the approaches based on random selection or property value rankings. Leonard and Roy [59] used a conceptually similar approach based on *k-means clustering* and reached similar conclusions. Here, the method referred to as "3M" by Golbraikh et al. [58] was adopted.

All SVM regression analysis in the present work was performed with a customized version of the LIBSVM package [57].

5. Uncertainty analysis

For practical use of any model, one must recognize its accuracy and limitations, and there should be a mechanism for quantitative estimation of uncertainties along with the predicted values. Uncertainty analysis of QSPR models is rarely conducted beyond computing a global error metric, such as the mean absolute deviation or root-mean-squared error. Use of the global error measure is certainly helpful, as it does provide a sense of the model performance. However, the fact that the experimental values have uncertainties of their own [4] is generally overlooked (i.e., the experimental measurements are considered error-free) in this global analysis. Furthermore, one should expect that the model

accuracy would not be uniform across the dataset, and there is a clear benefit of more fine-grained uncertainty estimation depending on a specific case (compound). To perform a more detailed uncertainty analysis, sources of potential model errors need to be identified first, as listed below:

1. uncertainties of the experimental data used to produce the model;
2. extrapolation errors, i.e., the case when the compound of interest is outside the model training domain;
3. nominal accuracy of the model ε (i.e., the "width" of the SVM regression "tube");
4. missing physics: descriptor(s) that should account for the relevant physical behavior is(are) not included;
5. change in the physical mechanism that controls the property of interest as compared to the cases used to generate the model (e.g., compound association);
6. deficiency of the theoretical molecular structure: "wrong" conformer, errors of the quantum-chemical method used, etc.

Items 4, 5, and 6 are, generally, very difficult to address *a priori*. They are normally encountered and identified during the practical model use. However, these errors are mostly *correctable*, i.e., the model can be adjusted by including missing descriptors, using more appropriate theory to produce molecular structures, etc. Items 1 and 2, on the other hand, will always be present in any empirical model, and item 3 is inherent to the SVM regression approach. Fortunately, the uncertainties associated with items 1–3 can also, to some extent, be accessed numerically as part of the model development process.

The nominal accuracy of the SVM regression ε (item 3) is available by definition. The methodology for evaluation of uncertainties due to experimental errors is developed and implemented in the present study. From the mathematical point of view, the model predictions depend on the experimental measurements *via* coefficients $(\alpha_i - \alpha_i^*)$ in Eq. (5). In general, this relationship cannot be expressed analytically. Furthermore, as was mentioned previously, the coefficients $(\alpha_i - \alpha_i^*)$ take nonzero values only for a subset of experimental points (support vectors); however, the makeup of this subset will also vary with the experimental values $\mathbf{y}$. This makes the conventional use of numerical differentiation as a means to access the accuracy *via* differential analysis highly impractical. Therefore, the situation calls for the use of Monte Carlo sampling procedures, an approach commonly adopted to evaluate the propagation of uncertainty through systems of such complexity [60]. Implementation of this technique as applied to the SVM model (5) put forward in this study is as follows. Given the vector of the experimental values $\mathbf{y}$, a sample of size M, $(\mathbf{y}^1, \ldots, \mathbf{y}^M)$, is generated. Each ith component y_i^j of the vector $\mathbf{y}^j$, where $j \in [1, M]$, represents a random variable distributed according to the probability density function (assumed normal) defined by the experimental uncertainty of the original experimental value y_i. Each vector $\mathbf{y}^j$ is then used to generate a separate SVM model, f^j. As a result, a collection of SVM models, $(f^1, \ldots, f^M)$, is produced and stored for further use. From this point, the generation of the final model prediction becomes more involved. Instead of producing a single value from a single model for a given vector of scaled descriptors $\mathbf{x}$, one generates a *sample of predictions* of size M, one for each model f^j. From this sample, both the final predicted value (i.e., the median of the sample) and its confidence interval are derived. Monte Carlo sampling was performed using the latin hypercube strategy [61], which allows achievement of high efficiency for a limited sample size. Sample size $M = 3000$ was found to be more than sufficient for convergence of both the median and the confidence interval.

Finally, the issue of extrapolation (item 2) should be mentioned. While the methodologies for defining the *applicability domain* for QSPR/QSAR models are emerging [62], they still lack generalization. However, the extrapolation errors, to some extent, are also accounted for by the propagation of uncertainty procedure described above. Experimental uncertainties, when propagated to unconstrained regions outside of the training domain, are expected to cause numerical instabilities, resulting in increased uncertainties of the model predictions.

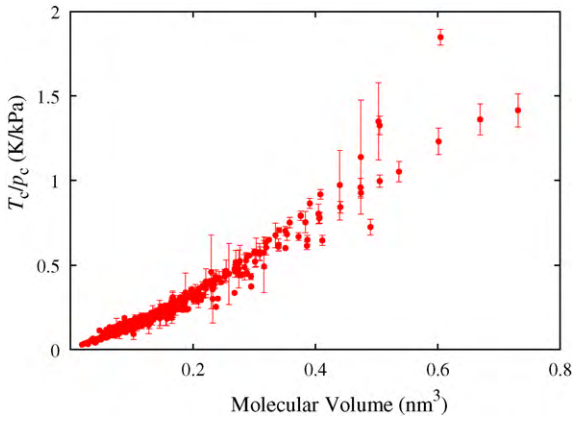

Fig. 1. Experimental ratio of critical temperature and critical pressure as a function of computed molecular volume for 677 compounds.

6. Development of QSPR-SVM correlations

Prior to the development of QSPR correlations using the evaluated experimental data and the SVM apparatus described in the previous sections, one needs to consider the specific properties to be correlated. While the end goal is to obtain estimates for T_c and p_c, all previous efforts [15–22] indicate that both of these properties are complex and nontrivial functions of molecular parameters. Therefore, property transformations that can potentially simplify the resulting correlations were considered first. The property transformations are generally deemed unnecessary when nonlinear regression is used. However, in practical situations, when one deals with unbalanced data, it is beneficial to supplement experimental data with an additional information coming from the physical insights to the problem, such as the property transformation suggested below.

One of the obvious choices is to use the ratios T_c/p_c and T_c^2/p_c. This proposal is motivated by the fact that, from the van der Waals equation of state, $T_c/p_c \propto b$ and $T_c^2/p_c \propto a$, where a and b are the van der Waals constants. Constants a and b are closely related to molecular parameters, i.e., a is the measure of intermolecular attractions, and b is related to molecular volume. The fact that the ratio T_c/p_c can be correlated better (i.e., with fewer and easily computed molecular parameters) than either property individually was reported previously [63–66]. In particular, Kontogeorgis and co-workers [64–66] advocated a simple expression that related T_c/p_c to a single parameter, the compound's van der Waals surface area computed from tabulated group contributions. To further illustrate the advantage of using T_c/p_c, this ratio is plotted against the van der Waals volume (consistent with its expected relationship to b) in Fig. 1 for all 677 compounds with evaluated T_c and p_c. The molecular van der Waals volume is one of the descriptors computed by CODESSA as a volume of overlapping spheres [67] defined by the atomic van der Waals radii and the optimized molecular geometry. As can be seen, the data indeed exhibit a strong correlation with this single independently-computed parameter in confirmation of these very simple theoretical considerations. While the correlation shown in Fig. 1 is not quantitative (i.e., it cannot be represented with a single curve of the desired accuracy), introduction of additional variables

Table 2
SVM regression data.

	Training	Validation	Testing
Set statistics			
T_c	574	191	100
T_c/p_c	450	150	77

		Kernel		
		Polynomial	RBF	
	ε	C	C	γ
SVM parameters				
T_c	4×10^{-3}	0.40	7088	4.16×10^{-3}
T_c/p_c	1×10^{-3}	3.21	17.86	2.28×10^{-1}

(descriptors) is expected to improve the quantitative agreement. Therefore, T_c/p_c was chosen as one of the properties for the generation of correlations. Relative experimental uncertainties for p_c are typically an order of magnitude greater than those for T_c. Consequently, the uncertainty of the ratio T_c/p_c (or T_c^2/p_c) is controlled by uncertainties in p_c. If one is to derive critical temperature estimates from correlations for T_c/p_c and T_c^2/p_c, it will be contaminated with much greater uncertainties than those of the original experiments leading to an unacceptable loss of accuracy. Therefore, in spite of its potential complexity, it is more beneficial to correlate T_c directly (and p_c indirectly through calculation of T_c/p_c).

Details of the numerical implementation as applied to the properties T_c and T_c/p_c are given below. The nominal accuracy parameter ε (i.e., the "width" of the SVM "tube") was set to 4×10^{-3} and 1×10^{-3} for T_c and T_c/p_c correlations, respectively. In the case of the critical temperature, the chosen value corresponds to approximately 3 K in physical space, i.e., before scaling. For T_c/p_c, the preset value of ε translates to a deviation of about 5% for the lowest value of T_c/p_c in the dataset. The weighting factors c_i in Eq. (6) were defined as

$$c_i = \frac{\delta_{\min}}{\delta_i}, \tag{7}$$

where δ_i is the experimental uncertainty for the ith point, and $\delta_{\min}$ the smallest uncertainty observed in the dataset. Linear weight scaling (7) is consistent with the loss function used in the SVM formulation (1).

It is often recommended to reduce the number of descriptors considered, leaving only the ones that have the strongest effect on the property of interest. The criteria for descriptor elimination are commonly based on statistical analysis of the dataset; however, their application to unbalanced (i.e., biased) data is problematic. Therefore, it was decided not to reduce the descriptor set beyond that discussed in Section 3, and the full set of 175 descriptors was used in all correlation development work presented. Detailed discussion of this issue is presented in Section 8.1.

The SVM regression analysis in this study was primarily tested using second-order polynomial (Eq. (3) with $\omega = 2$) and RBF (4) kernels. Limited initial tests were also conducted with linear and the recently suggested Pearson VII function [68] kernels; the former exhibited worse performance compared to all nonlinear kernels tested, while the latter produced results nearly identical to those of the RBF kernel. For both T_c and T_c/p_c correlations, about 25% and 12% of the total compound count were used for the validation and testing sets, respectively. The exact numbers are given in Table 2.

SVM parameters, as explained in Section 4, were optimized by minimizing the objective function

$$F = \sum_i \left| \frac{y_i - y_i^{\mathrm{p}}}{\delta_i} \right|, \tag{8}$$

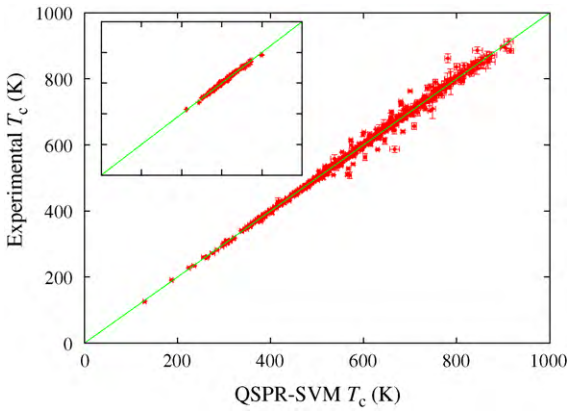

Fig. 2. Comparison of experimental and predicted critical temperatures. Model predictions shown were obtained using the second-order polynomial kernel and SVM parameters listed in Table 2. The main figure shows data from the combined training and validation set. The insert displays independent testing set data with the same axis scales as those of the main figure.

where the summation is performed over all property values from the validation set, y_i^{p} is the value predicted with the model obtained using the training dataset and a trial set of the SVM parameters, y_i the experimental value, and δ_i the experimental uncertainty. For the second-order polynomial kernel, a single parameter C was optimized using the grid search. For the RBF kernel (4), two parameters C and γ were optimized simultaneously using the differential evolution algorithm [69]. The resulting values of parameters for all cases are also given in Table 2.

7. Results

The comparison of experimental and predicted critical temperatures is shown in Fig. 2 for the polynomial kernel-based SVM model (the results obtained with the RBF kernel are visually similar and not shown for brevity).

The error bars represent the evaluated uncertainty for the experimental data; the error bars for predictions combine the SVM's ε and uncertainties estimated *via* Monte Carlo sampling as described in Section 5. As can be seen, the experimental data are described very well, and the majority of the points cluster around the centerline that denotes "the perfect fit". Furthermore, the testing set (shown in insert), an independent measure of the model's predictive capability, is also well described. Shown in Fig. 3 are the distributions of absolute deviations between the model and the experimental val-

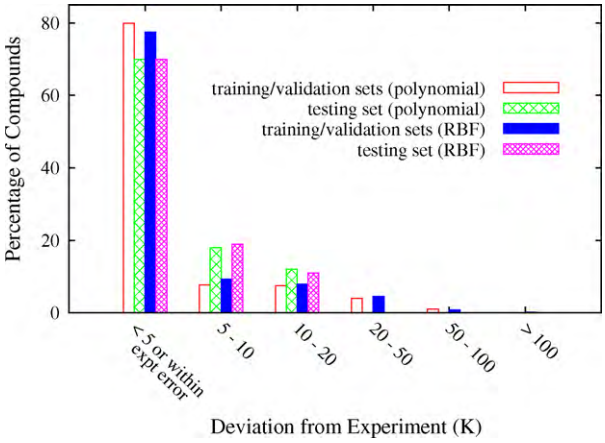

Fig. 3. Statistical distribution of absolute deviations between SVM models based on different kernels and the experimental data for critical temperature.

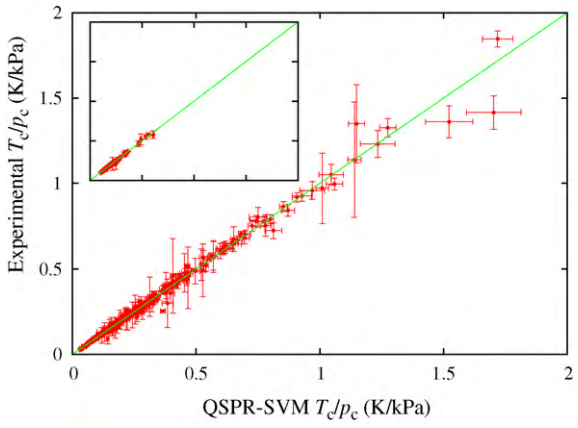

Fig. 4. Comparison of experimental and predicted ratios of critical temperatures and critical pressures. Model predictions shown were obtained using the second-order polynomial kernel and SVM parameters listed in Table 2. The main figure shows data from the combined training and validation set. The insert displays independent testing set data with the same axis scales as those of the main figure.

ues computed independently for combined training/validation and testing sets. As an acceptable level of performance, the first "bin" of this discretized distribution is defined as deviations below 5 K or *within the evaluated experimental uncertainty*. About 80% of values from the combined training/validation sets belong to this category for either functional form of the SVM kernel; about 70% of points for the independent testing set occupy this first "bin". For all cases shown, the deviations between the model and the experimental data are less than 10 K or within the experimental uncertainty for about 90% of the T_c values. The error distributions for the combined training/validation and testing sets are very similar, giving further support to the model's fidelity. The results also show that the two forms of the SVM kernel function perform approximately the same.

A similar analysis was performed for the second property, T_c/p_c, as shown in Figs. 4 and 5. The experimental uncertainties are much higher for this property, due to higher propagated uncertainties of p_c. This is especially pronounced for the large-sized molecules (high values of T_c/p_c). Taking into account these larger uncertainty levels, the model again displays good performance, and, as previously, the testing set is described well. The error analysis is done here in terms of the relative deviation (Fig. 5), a more appropriate error metric for this property. An acceptable accuracy is defined as "less than 3% or within the evaluated experimental uncertainty". This value is

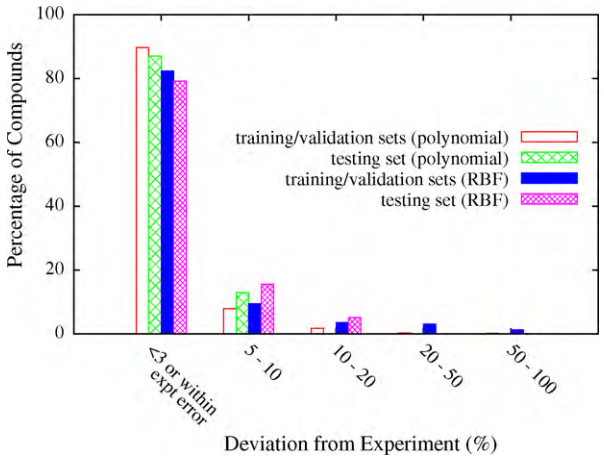

Fig. 5. Statistical distribution of relative deviations between SVM models based on different kernels and the experimental data for the ratio of critical temperature and critical pressure.

consistent with a typical experimental uncertainty for p_c measurements that normally controls T_c/p_c uncertainty. Approximately 90% of compounds belong to this category for the model based on the polynomial kernel. The RBF-based model performs slightly worse with more than 80% of cases displaying acceptable accuracy. For all cases presented, between 92% and 100% of the points are predicted to better than 10% or within the experimental uncertainty.

The RBF kernel is generally a preferred choice in most practical SVM regression applications. Based on the results presented so far, a conclusive recommendation regarding the preferred functional form of kernel cannot be made: the polynomial form seems to yield slightly better results, but the differences are not sufficient for generalization. However, the polynomial kernel has another advantage over the RBF function. Specifically, RBF, being a Gaussian-based function, is *local* in nature. As such, RBF generally works very well for data interpolation, but its ability to extrapolate deteriorates rapidly once the point of interest moves away from its center. Furthermore, the numerical instabilities will also decay with the function itself, which may lead to erroneous conclusions during uncertainty analysis. The polynomial kernel, on the other hand, typically works better in situations when limited extrapolation is needed and was therefore adopted here.

As seen in Figs. 2 and 4, a small number of points appear as outliers, i.e., the deviations between the experimental data and the model predictions significantly exceed the uncertainties. Comprehensive analysis of these cases was also performed and, due to the lack of space, is given in Supplementary material. The results revealed anomalous cases, typically where significant compound association is expected to take place. Multiple experimental errors were also detected.

8. Discussion

8.1. Number of descriptors in QSPR-SVM correlations

As mentioned earlier, the reduction of the number of variables in QSPR-SVM models was not considered beyond that discussed in Section 3, and the full set of 175 descriptors was used in correlation development. This deliberate decision, generally, goes against traditional QSPR methodology and requires some additional discussion.

In empirical modeling, one can distinguish two extreme cases. The first case can be described as smoothing interpolation, i.e., an interpolating surface is drawn through a number of nodes in multidimensional space, usually keeping all variables. This approach can provide high accuracy in the close vicinity of the nodes, but interpolation between the nodes or extrapolation outside of the domain may be of poor quality and may exhibit numerical instabilities depending on the positioning of the nodes and the functional form of the interpolating surface. Traditional GC-based estimation methods are close to this extreme situation in that they may involve hundreds of variables (groups), and group contribution values are often derived on the basis of a single point (node).

The second case can be represented by traditional QSAR/QSPR modeling. Here, only few variables are selected from a large pool, and a linear or nonlinear approximation of the data is developed on their basis. Variable selection is governed by statistical considerations, and it is expected to keep the number of variables as low as possible. The accuracy of these models varies, but is generally not very high as seen in practical use (e.g., see the 3-variable model for T_c from Ref. [17]). In addition, these models are expected to describe trends (interpolation or extrapolation) without the problems typical of the smoothing interpolation approach. It must be noted that QSPR descriptors do not represent true state variables, and one cannot claim that a finite set would contain all infor-

mation required for accurate description of the property under consideration. In fact, it is entirely possible that this information would be distributed over a very large set of descriptors, and consequently, keeping the number of variables low would not result in an acceptably accurate model. The traditional tendency to limit the number of variables to just a few probably originates from the fact that the training datasets used historically in QSAR/QSPR modeling were also very limited. Provided that a sufficiently large dataset is available, a larger number of variables can be selected following the same statistical considerations. In doing so, however, one faces another challenge: available experimental datasets are almost always extremely unbalanced, and, as discussed previously, the dataset for critical constants is no exception. Experimental measurements are not performed in a remotely systematic manner with regard to compound family or molecular size; and in fact, this is not physically possible. When dealing with unbalanced (biased) datasets, the statistical approaches used in QSPR tend to lose valuable information from poorly represented data in the process of descriptor selection. This situation may be remedied in the future with the emerging data balancing technologies [70]. At present, however, there is no established solution.

The SVM regression approach adopted in this work is very robust and efficient in dealing with high-dimensional data and variable redundancy. Although the resulting models are expected to have some interpolative features and may have problems outlined at the beginning of this section, the empirical evidence suggests no noticeable numerical instabilities, i.e., the testing sets for both properties are well-predicted. Further analysis was performed by generating predictions and their estimated uncertainties for about 8000 additional compounds from SOURCE with no experimental critical constants available (these estimates are now accessible in TDE [71]). If the predictions are affected by numerical instabilities, one should expect that the corresponding uncertainties computed *via* the stochastic sampling procedure would have wide distributions. This was not observed. For 95% of the compounds, the uncertainties for T_c and T_c/p_c were less than 15 K and 16%, respectively.

8.2. Extrapolative ability of QSPR-SVM correlations

QSPR-based property prediction methods are often criticized for poor established extrapolative abilities as compared with GC-based approaches (e.g., [13]). Generally, one cannot expect good extrapolative accuracy from an empirical correlation unless the property in question exhibits a well-defined asymptotic behavior that is enforced by the mathematical form of the correlation. None of the existing empirical methods for prediction of critical constants possess this feature. Some GC-based methods that correlate T_c/T_b may be considered to have reasonable asymptotic behavior, but only if a reliable T_b value is available; otherwise, no extrapolative ability can be claimed with certainty. Furthermore, even the concept of extrapolation cannot be defined in conventional terms, as the problem is intrinsically multidimensional regardless of the types of variables used (i.e., descriptors in QSPR or group counts in GC). The common practice to demonstrate the method's extrapolative abilities is to use the longest homologous series with available experimental data (i.e., straight-chain alkanes) as an example; however, it must be realized that this is a special case amongst numerous possibilities.

What is arguably more important than the reasonable functional behavior outside of the domain with available data is *the ability to assess the model's applicability domain*. In other words, one should be able to recognize when the model predictions become unreliable due to extrapolation error rather than to have the comfort of a seemingly "reasonable", yet baseless, estimate.

As discussed previously, the SVM regression with the polynomial kernel used in the present QSPR-SVM approach provides some degree of extrapolation by design. Furthermore, combined with the stochastic uncertainty analysis, this approach leads to increased estimated uncertainties outside of the domain constrained by the training data. This feature allows a more informed judgement regarding the reliability of the estimates produced by the model. To illustrate this, the conventional example of the normal alkane homologous series is considered (Fig. 6). As can be seen, the critical temperatures estimated with QSPR-SVM agree well, as expected, with the experimental data that are available up to C_{36}, and the estimated uncertainty generally follows the uncertainty of the experimental data. As predictions are extrapolated beyond the available data, the estimated uncertainty gradually increases, reaching about 85 K at 55 carbon atoms. One can immediately conclude that the estimates produced for larger alkanes are of low fidelity. Also shown in Fig. 6 are the results of Monte Carlo simulations from Nath et al. [72]. Their last value was computed for *n*-octatetracontane ($C_{48}H_{98}$), and agrees with the QSPR-SVM estimate.

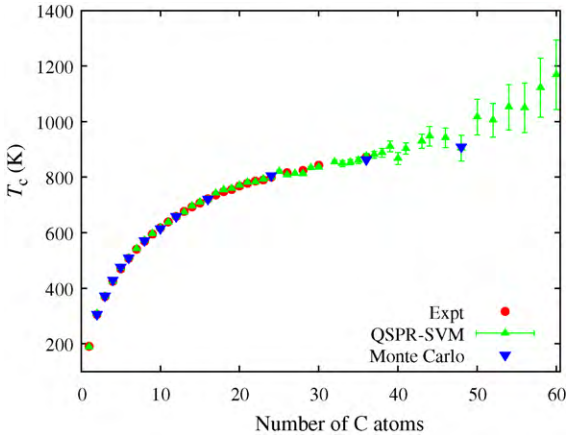

Fig. 6. Critical temperatures of normal alkanes. The points labeled "Monte Carlo" are obtained from Monte Carlo simulations [72].

8.3. Interpretation of QSPR-SVM correlations

The complexity of the SVM expansion, Eq. (5), makes its direct interpretation difficult; however, it is possible to gain a qualitative understanding of the dominant factors controlling the predicted critical constants. As noted earlier, prior to their use in the SVM regression, all descriptors are scaled to [0,1] intervals. Therefore, the linear terms in the expansion of Eq. (5) with respect to the scaled descriptor values x_k,

$$f(\mathbf{x}) = d + \sum_k a_k x_k + \cdots, \tag{9}$$

are expected to have the most influence on the predicted value. Ranking of the descriptors by absolute values of coefficients a_k in Eq. (9) can be used as a qualitative measure of their importance. Table 3 lists the descriptors with highest rankings for both correlations. Only descriptors with absolute values of a_k that are greater than half of the highest absolute value observed for the entire set are shown. For T_c, there are 12 descriptors that have $|a_k|$ above the chosen cutoff. As expected, most of them characterize polar interactions and size of compound. This list is generally consistent with descriptor sets identified as important for modeling T_c in previous QSPR studies [15–17,21,22]. Particular overlap is observed with the set reported by Katritzky et al. [16] who also used CODESSA for generation of descriptors. Inclusion of descriptors associated with the number of occupied electronic levels is a new feature that was not reported previously. It appears to reflect and quantify the pres-

Table 3
Descriptors associated with dominant linear terms in QSPR-SVM expansions.

$\lvert a_k\rvert/\lvert a_k\rvert_{max}$	Descriptor[a]
	T_c
1.00	α polarizability
0.96	Number of occupied electronic levels/number of atoms
0.84	Relative molecular weight
0.82	Kier and Hall index (order 3)
0.81	Total molecular electrostatic interaction
0.81	Total molecular one-center electron-electron repulsion/number of atoms
0.79	*XY* shadow
0.77	Molecular weight
0.67	Average information content (order 0)
0.60	Number of occupied electronic levels
0.56	Gravitation index (all bonds)
0.56	Zero point vibrational energy/number of atoms
0.54	Average complementary information content (order 1)
	T_c/p_c
1.00	Wiener index
0.57	Molecular volume
0.50	Total charge-weighted partial positive surface area
0.50	Difference between total charge-weighted partial positive and negative surface areas

[a] Rigorous definitions of all descriptors are given in Ref. [48].

ence of heteroatoms that are known to have a significant effect on T_c. It is likely that, in previous studies, compounds with heavy heteroatoms were either underrepresented in the training sets or their influence was accounted for by the use of less general, atom-specific descriptors that were excluded in the present work.

The list for T_c/p_c contains only 4 descriptors that satisfy the cutoff criterion. Based on the discussion in Section 6, molecular volume, expectedly, has a very high ranking (number two in the list). The highest ranking descriptor for this property is the Wiener index. The Wiener index [73] is one the oldest topological descriptors and was successfully used in numerous correlations. It has been shown to exhibit a strong nonlinear correlation with the van der Waals surface area [74] for a compound, which is a likely reason of its high ranking for T_c/p_c correlation.

8.4. Comparison of QSPR-SVM with other estimation methods

To demonstrate the performance of the QSPR-SVM approach, a systematic comparison with other methods commonly used for critical constant estimation was performed. The methods of Joback [9], Constantinou–Gani (CG) [10], Wilson–Jasperson (WJ) [11], and Marrero–Pardillo (MP) [12], all of which are based on Group-Contribution (GC) methodology, were considered. In all calculations, the implementation of these methods in the TDE software [30] was used. A number of issues were addressed to conduct the comparisons in a consistent manner. First, during the development of QSPR-SVM correlations, the experimental data were split into two sets: a large training/validation set and a smaller, testing set. Because the QSPR-SVM correlations were developed using the information from the training/validation set, they have an advantage over other methods, when compounds from this set are considered. Therefore, when comparisons are made, the information for the training/validation and testing sets was processed and presented separately. Second, GC-based methods have a narrower coverage than QSPR-SVM, as they depend on data availability for each functional group. The scope of the considered methods, as compared to the present approach, is illustrated in Fig. 7. As can be seen, Joback, CG, and WJ methods cover 83–92% of the compounds from the training/validation set and nearly all compounds from the testing set. The MP method has a substantially narrower scope, covering only 60% for critical temperature and 48% for both

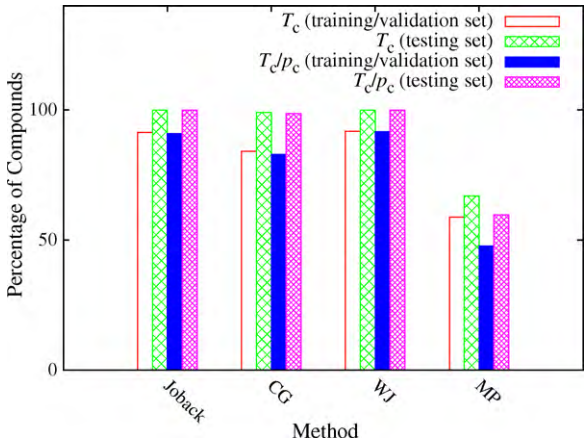

Fig. 7. Scope of Joback [9], Constantinou–Gani (CG) [10], Wilson–Jasperson (WJ) [11], and Marrero–Pardillo (MP) [12] estimation methods. The bars indicate the percentage of compounds from QSPR-SVM sets supported by the respective method.

critical temperature and critical pressure for the training/validation set, and 67% and 60% for T_c and for both T_c and p_c, respectively, for the testing set. It follows that, when performing the comparisons, only compounds supported by a specific GC method were considered. Third, with the exception of the CG method, the GC methods require knowledge of normal boiling points T_b. The common practice that leads to more accurate predictions is to use experimentally measured T_b; for example, Nannoolal et al. [13] report a two-fold increase in average absolute error in T_c predictions when experimentally measured T_b values were replaced with predicted ones. Here, however, the objective was to compare *a priori* estimates without any addition of experimental information, so the needed normal boiling points for the GC methods were also estimated. The TDE capabilities were used for these estimates, as well. In doing so, whenever possible, the method from the same family was used to estimate T_b (i.e, the Joback or MP method was used to estimate both T_b and T_c). If the method for normal boiling point was unavailable, as in the case of WJ, or was recognized by TDE as inaccurate for a particular compound, the best alternative estimation method implemented in TDE was used. Finally, for consistency with the present work, the ratio T_c/p_c, rather than p_c itself, was used for comparisons. In addition to the evaluation of QSPR-SVM performance as compared to the other approaches, the following analysis also provides a new validation of the GC-based methods against the large dataset of experimental values developed in this work. The statistical distributions of deviations from the experimental data obtained with the different estimation methods are shown in Figs. 8 and 9 for T_c and T_c/p_c, respectively. As seen in Fig. 8, the QSPR-SVM approach performs substantially better in prediction of critical temperature than any of the GC-based methods (within their respective scopes) not only for the compounds from the training/validation set, as one may have expected, but, more importantly, for the compounds from the testing set as well. Among the GC-based methods, the method of Joback exhibits the worst performance; the distribution of absolute deviations from the experimental data appears rather flat and peaks between 10 and 50 K (Fig. 8). The CG and WJ methods show somewhat better performance, and MP displays the best accuracy among the GC methods, with results similar to those of QSPR-SVM for the testing set; however, as noted previously, the scope of MP is substantially narrower.

The advantage of QSPR-SVM is yet more apparent for prediction of T_c/p_c (Fig. 9), where about twice as many compounds are predicted within an accuracy of 3% (or experimental uncertainty) as for any of the other methods. As for T_c, MP shows the best performance (within its limited coverage) among the GC-based methods, followed by the WJ, CG, and Joback methods.

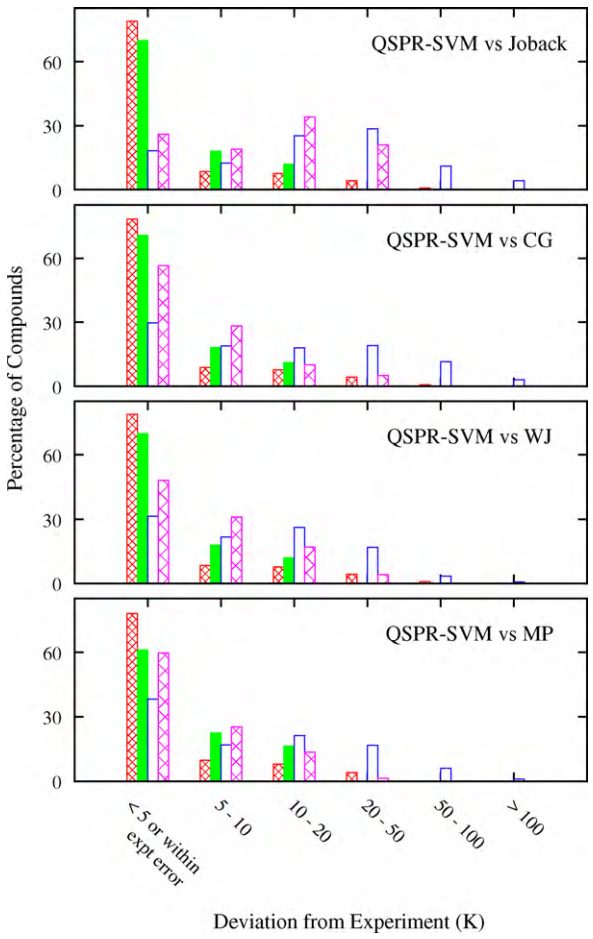

Fig. 8. Comparison of the present QSPR-SVM and GC-based methods for prediction of critical temperature, T_c. For each group of bars defined by the error range, the order of bars is as follows: QSPR-SVM (training/validation set), QSPR-SVM (testing set), GC method (training/validation set), and GC method (testing set). The GC method considered is shown in each figure.

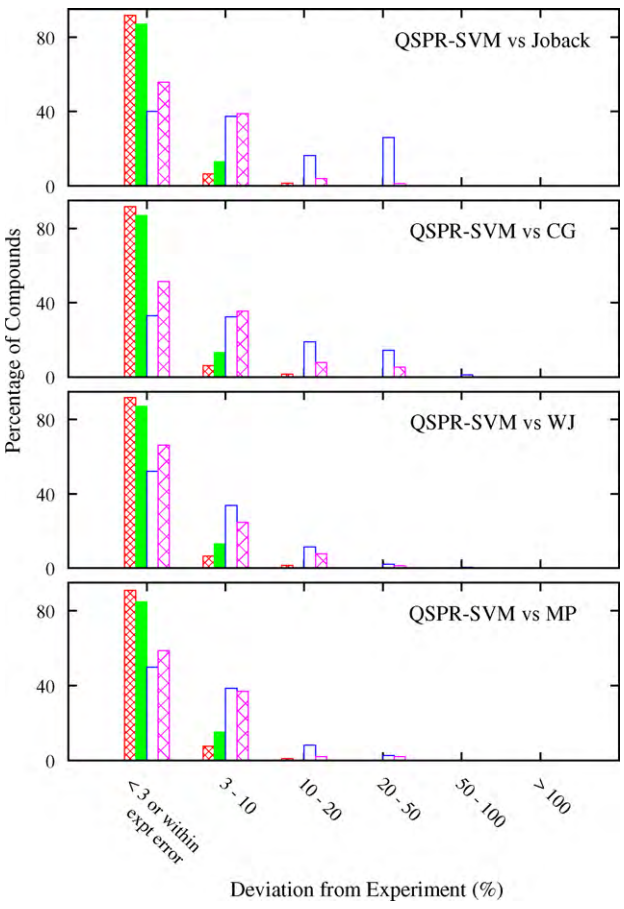

Fig. 9. Comparison of the QSPR-SVM and GC-based methods for prediction of the ratio of critical temperature to critical pressure T_c/p_c. The bar markings are the same as in Fig. 8.

It can be concluded from the above analysis that the present QSPR-SVM approach exhibits substantially better performance in predicting critical constants for pure compounds as compared to several commonly used GC-based methods applied in an *a priori* manner (i.e., without the use of the experimentally measured normal boiling temperatures).

9. Summary

A framework for the development of predictive correlations for thermophysical properties was formulated and successfully demonstrated using critical constants for pure compounds as an example. The procedures implemented in this work include rigorous evaluation of the original experimental data from the literature and development of empirical models based on the QSPR methodology combined with the SVM regression analysis. Evaluation of experimentally measured critical constants was performed with the methods of robust regression, and generated a dataset that included 865 compounds with evaluated critical temperatures; among them, 677 compounds had both critical temperature and critical pressure. Experimental uncertainties were also evaluated and explicitly taken into account during correlation development. A procedure based on stochastic sampling that allows uncertainty assessment for predicted values was also presented. The resulting correlations exhibited good performance, as evidenced by the comparison of predicted and experimental values for subsets of compounds that were not used in the correlation development. The findings of this work also indicate that the QSPR-SVM approach generally performs better both in terms of accuracy and scope than several popular Group-Contribution-based estimation methods when applied in an *a priori* manner (without experimental T_b information). Additional predictions of critical constants were generated for about 8000 compounds with no experimental data and were made available in NIST/TRC ThermoData Engine software.

Acknowledgments

The authors thank Mr. Jake Bouricius for his contributions during the initial stages of this work. Helpful suggestions of Prof. John P. O'Connell of the University of Virginia are also greatly appreciated. Some calculations in this study utilized the high-performance computational capabilities of the Helix Systems at the National Institutes of Health, Bethesda, MD (http://helix.nih.gov).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fluid.2010.07.014.

References

[1] M. Frenkel, Global information systems in science: application to the field of thermodynamics, J. Chem. Eng. Data 54 (9) (2009) 2411–2428.
[2] B.E. Poling, J.M. Prausnitz, J.P. O'Connell, The Properties of Gases and Liquids, 5th edition, McGraw-Hill, New York, 2000.

[3] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition, Morgan Kaufmann, San Francisco, 2005.
[4] R.D. Chirico, M. Frenkel, V.V. Diky, K.N. Marsh, R.C. Wilhoit, ThermoML-an XML-based approach for storage and exchange of experimental and critically evaluated thermophysical and thermochemical property data. 2. Uncertainties, J. Chem. Eng. Data 48 (5) (2003) 1344–1359.
[5] C. Hansch, A. Leo, Exploring QSAR. Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, DC, 1995.
[6] P. Jurs, Quantitative structure–property relationships, in: J. Gasteiger (Ed.), Handbook of Chemoinformatics, vol. 3, Wiley-VCH, Weinheim, 2007, pp. 1314–1335.
[7] A.R. Katritzky, U. Maran, V.S. Lobanov, M. Karelson, Structurally diverse quantitative structure–property relationship correlations of technologically relevant physical properties, J. Chem. Inf. Comput. Sci. 40 (1) (2000) 1–18.
[8] A.R. Katritzky, D.A. Dobchev, M. Karelson, Physical, chemical, and technological property correlation with chemical structure: the potential of QSPR, Z. Naturforsch. B: Chem. Sci. 61 (4) (2006) 373–384.
[9] K.G. Joback, R.C. Reid, Estimation of pure-component properties from group-contributions, Chem. Eng. Commun. 57 (1–6) (1987) 233–243.
[10] L. Constantinou, R. Gani, New group-contribution method for estimating properties of pure compounds, AIChE J. 40 (10) (1994) 1697–1710.
[11] G.M. Wilson, L.V. Jasperson, Critical constants T_c, p_c, estimation based on zero, first, and second order methods, in: AIChE Spring Meeting, New Orleans, LA, 1996.
[12] J. Marrero-Morejon, E. Pardillo-Fontdevila, Estimation of pure compound properties using group-interaction contributions, AIChE J. 45 (3) (1999) 615–621.
[13] Y. Nannoolal, J. Rarey, D. Ramjugernath, Estimation of pure component properties: Part 2. Estimation of critical property data by group contribution, Fluid Phase Equilibr. 252 (1–2) (2007) 1–27.
[14] S. Grigoras, A structural approach to calculate physical-properties of pure organic-substances – the critical-temperature, critical volume and related properties, J. Comput. Chem. 11 (4) (1990) 493–510.
[15] L.M. Egolf, M.D. Wessel, P.C. Jurs, Prediction of boiling points and critical temperatures of industrially important organic compounds from molecular structure, J. Chem. Inf. Comput. Sci. 34 (4) (1994) 947–956.
[16] A.R. Katritzky, L. Mu, M. Karelson, Relationships of critical temperatures to calculated molecular properties, J. Chem. Inf. Comput. Sci. 38 (2) (1998) 293–299.
[17] B.E. Turner, C.L. Costello, P.C. Jurs, Prediction of critical temperatures and pressures of industrially important organic compounds from molecular structure, J. Chem. Inf. Comput. Sci. 38 (4) (1998) 639–645.
[18] G. Espinosa, D. Yaffe, A. Arenas, Y. Cohen, F. Giralt, A fuzzy ARTMAP-based quantitative structure–property relationship (QSPR) for predicting physical properties of organic compounds, Ind. Eng. Chem. Res. 40 (12) (2001) 2757–2766.
[19] X.J. Yao, Y.W. Wang, X.Y. Zhang, R.S. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, Radial basis function neural network-based QSPR for the prediction of critical temperature, Chemom. Intell. Lab. Syst. 62 (2) (2002) 217–225.
[20] S.S. Yang, W.C. Lu, N.Y. Chen, Q.N. Hu, Support vector regression based QSPR for the prediction of some physicochemical properties of alkyl benzenes, Theochem. J. Mol. Struct. 719 (1–3) (2005) 119–127.
[21] D. Sola, A. Ferri, M. Banchero, L. Manna, S. Sicardi, QSPR prediction of N-boiling point and critical properties of organic compounds and comparison with a group-contribution method, Fluid Phase Equilibr. 263 (1) (2008) 33–42.
[22] S.S. Godavarthy, R.L.J. Robinson, K.A.M. Gasem, Improved structure–property relationship models for prediction of critical properties, Fluid Phase Equilibr. 264 (2008) 122–136.
[23] Evaluated standard thermophysical property values, DIPPR (Design Institute for Physical Properties) 801 (September) (2008).
[24] A. Kerber, R. Laue, M. Meringer, C. Rücker, MOLGEN-QSPR, a software package for the study of quantitative structure property relationships, Match Comm. Math. Comput. Chem. 51 (2004) 187–204.
[25] M. Frenkel, Q. Dong, R.C. Wilhoit, K.R. Hall, TRC SOURCE database: a unique tool for automatic production of data compilations, Int. J. Thermophys. 22 (1) (2001) 215–226.
[26] Q. Dong, X.J. Yan, R.C. Wilhoit, X.G. Hong, R.D. Chirico, V.V. Diky, M. Frenkel, Data quality assurance for thermophysical property databases – applications to the TRC SOURCE data system, J. Chem. Inf. Comput. Sci. 42 (3) (2002) 473–480.
[27] Q. Dong, R.D. Chirico, X.J. Yan, X.R. Hong, M. Frenkel, Uncertainty reporting for experimental thermodynamic properties, J. Chem. Eng. Data 50 (2) (2005) 546–550.
[28] V.V. Diky, R.D. Chirico, R.C. Wilhoit, Q. Dong, M. Frenkel, Windows-based guided data capture software for mass-scale thermophysical and thermochemical property data collection, J. Chem. Inf. Comput. Sci. 43 (1) (2003) 15–24.
[29] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes: The Art of Scientific Computing, 3rd edition, Cambridge University Press, 2007.
[30] M. Frenkel, R.D. Chirico, V. Diky, X.J. Yan, Q. Dong, C. Muzny, ThermoData Engine (TDE): software implementation of the dynamic data evaluation concept, J. Chem. Inf. Model. 45 (4) (2005) 816–838.
[31] W. Wagner, New vapor-pressure measurements for argon and nitrogen and a new method for establishing rational vapor-pressure equations, Cryogenics 13 (8) (1973) 470–482.
[32] D. Ambrose, J. Walton, Vapour pressures up to their critical temperatures of normal alkanes and 1-alkanols, Pure Appl. Chem. 61 (8) (1989) 1395–1403.
[33] W.V. Steele, 50 years of thermodynamics research at Bartlesville – the Hugh M. Huffman legacy, J. Chem. Thermodyn. 27 (2) (1995) 135–162.
[34] V.J. Yohai, High breakdown-point and high-efficiency robust estimates for regression, Ann. Stat. 15 (2) (1987) 642–656.
[35] P.J. Rousseeuw, K. van Driessen, Computing LTS regression for large data sets, Data Min. Knowl. Discov. 12 (1) (2006) 29–45.
[36] W. Waring, Form of a wide-range vapor pressure equation, Ind. Eng. Chem. 46 (4) (1954) 762–763.
[37] CambridgeSoft, Chem3D Pro. Version 11.0, 2007.
[38] Advanced Chemistry Development, Inc., ACD/ChemSketch. Version 11.0 for Microsoft Windows. Reference Manual, 2007.
[39] M.J. Vainio, M.S. Johnson, Generating conformer ensembles using a multiobjective genetic algorithm, J. Chem. Inf. Model. 47 (6) (2007) 2462–2474.
[40] The Open Babel Package, Version 2.2.0, Software Available at http://openbabel.sourceforge.net (July 2008).
[41] K. Gilbert, R. Guha, SMI23D – 3D Coordinate Generation, Software Available at http://www.chembiogrid.org/cheminfo/smi23d (June 2008).
[42] G.M. Crippen, T.F. Havel, Distance Geometry and Molecular Conformation, Research Studies Press, Taunton, 1988.
[43] J.J.P. Stewart, Optimization of parameters for semiempirical methods. 1. Method, J. Comput. Chem. 10 (2) (1989) 209–220.
[44] J.J.P. Stewart, MOPAC. Manual, J. Frank, Seiler Research Laboratory, 6th edition, United States Air Force Academy, Colorado Springs, CO, 1990, October.
[45] M. Karelson, Molecular Descriptors in QSAR/QSPR, Wiley-Interscience, New York, 2000.
[46] J.W. Ponder, TINKER: Software Tools for Molecular Design. Version 4.2, 2004, June.
[47] M.J. Field, M. Albe, C. Bret, F. Proust-De Martin, A. Thomas, The DYNAMO library for molecular simulations using hybrid quantum mechanical and molecular mechanical potentials, J. Comput. Chem. 21 (12) (2000) 1088–1100.
[48] SEMICHEM and University of Florida, CODESSA: Comprehensive Descriptors for Structural and Statistical Analysis. User Manual, 1995–1997.
[49] N.S. Zefirov, M.A. Kirpichenok, F.F. Ismailov, M.I. Trofimov, Calculation schemes for atomic electronegativities in molecular graphs within the framework of Sanderson principle, Dokl. Akad. Nauk. SSSR 296 (4) (1987) 883–887.
[50] M.A. Kirpichenok, N.S. Zefirov, Electronegativity and geometry of molecules. 1. Principles of developed approach and analysis of the effect of nearest electrostatic interactions on the bond length in organic-molecules, Zh. Org. Khim. 23 (4) (1987) 673–691.
[51] M.A. Kirpichenok, N.S. Zefirov, Electronegativity and geometry of molecules. 2. Concept of a freely relaxed molecule and analysis of its geometry based on the electrostatic approach, Zh. Org. Khim. 23 (4) (1987) 691–703.
[52] J. Gasteiger (Ed.), Handbook of Chemoinformatics, vol. 3, Wiley-VCH, Weinheim, 2007.
[53] O. Ivanciuc, Applications of support vector machines in chemistry, in: K.B. Lipkowitz, T.R. Cundari (Eds.), Reviews in Computational Chemistry, vol. 23, Wiley-VCH, Weinheim, 2007, pp. 291–400.
[54] X. Yao, A. Panaye, J. Doucet, R. Zhang, H. Chen, M. Liu, Z. Hu, B. Fan, Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression, J Chem. Inf. Comput. Sci. 44 (4) (2004) 1257–1266.
[55] A. Varnek, N. Kireeva, I.V. Tetko, I.I. Baskin, V.P. Solov'ev, Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points? J. Chem. Inf. Model. 47 (3) (2007) 1111–1122.
[56] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[57] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, software available at http://www.csie.ntu.edu.tw/cjlin/libsvm (2001).
[58] A. Golbraikh, M. Shen, Z.Y. Xiao, Y.D. Xiao, K.H. Lee, A. Tropsha, Rational selection of training and test sets for the development of validated QSAR models, J. Comput. Aided Mol. Des. 17 (2) (2003) 241–253.
[59] J.T. Leonard, K. Roy, On selection of training and test sets for the development of predictive QSAR models, QSAR Comb. Sci. 25 (3) (2006) 235–251.
[60] M.E. Reed, W.B. Whiting, Sensitivity and uncertainty of process designs to thermodynamic model parameters – a Monte Carlo approach, Chem. Eng. Commun. 124 (1993) 39–48.
[61] J. Helton, F. Davis, Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, Reliab. Eng. Syst. Saf. 81 (1) (2003) 23–69.
[62] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, QSAR applicability domain estimation by projection of the training set in descriptor space: a review, ATLA 33 (5) (2005) 445–459.
[63] A.E. Elhassan, M.A. Barrufet, P.T. Eubank, Correlation of the critical properties of normal alkanes and alkanols, Fluid Phase Equilibr. 78 (1992) 139–155.
[64] G.M. Kontogeorgis, I.V. Yakoumis, P. Coutsikos, D.P. Tassios, A generalized expression for the ratio of the critical temperature to the critical pressure with the van der Waals surface area, Fluid Phase Equilibr. 140 (1–2) (1997) 145–156.
[65] I.V. Yakoumis, E. Nikitin, G.M. Kontogeorgis, Validation of a recent generalized expression of T_c/p_c vs. the van der Waals surface area according to recent measurements, Fluid Phase Equilibr. 153 (1) (1998) 23–27.
[66] A. Zbogar, F.V.D. Lopes, G.M. Kontogeorgis, Approach suitable for screening estimation methods for critical properties of heavy compounds, Ind. Eng. Chem. Res. 45 (1) (2006) 476–480.
[67] F.M. Richards, Areas, volumes, packing, and protein-structure, Ann. Rev. Biophys. Bioeng. 6 (1977) 151–176.
[68] B. Üstün, W.J. Melssen, L.M.C. Buydens, Facilitating the application of support vector regression by using a universal Pearson VII function based kernel, Chemom. Intell. Lab. Syst. 81 (1) (2006) 29–40.
[69] K.V. Price, R.M. Storn, J.A. Lampinen, Differential Evolution: A Practical Approach to Global Optimization, Springer-Verlag, Heidelberg, 2005.

[70] E. Vladislavleva, G. Smits, D. den Hertog, On the importance of data balancing for symbolic regression, IEEE Trans. Evol. Comput. 14 (2) (2010) 252–277.
[71] V. Diky, R.D. Chirico, A. Kazakov, C.D. Muzny, M. Frenkel, ThermoData Engine (TDE): software implementation of the dynamic data evaluation concept. 4. Chemical reactions, J. Chem. Inf. Model. 49 (12) (2009) 2883–2896.
[72] S.K. Nath, F.A. Escobedo, J.J. de Pablo, On the simulation of vapor-liquid equilibria for alkanes, J. Chem. Phys. 108 (23) (1998) 9905–9911.
[73] H. Wiener, Structural determination of paraffin boiling points, J. Am. Chem. Soc. 69 (1) (1947) 17–20.
[74] I. Gutman, T. Körtvèlyesi, Wiener indexes and molecular surfaces, Z. Naturforsch. A: Phys. Sci. 50 (7) (1995) 669–671.