Quantitative Structure–Property Relationship Predictions of Critical Properties and Acentric Factors for Pure Compounds

Wendy Hawley Carande,*‡ †Andrei Kazakov,‡ Chris Muzny,† and Michael Frenkel†

†National Institute of Standards and Technology, 325 Broadway, MS 647, Boulder, Colorado 80305-3337, United States
‡University of Colorado Boulder, Boulder, Colorado 80309, United States

ABSTRACT: Knowledge of critical constants and phase boundary pressure properties is essential to understanding thermodynamic behavior of substances and is often required in practical process design applications. Where critically evaluated data are unavailable, a quantitative structure–property relationship (QSPR) regression method can be used to relate molecular properties (descriptors) to properties of interest. The relationship is trained and tested using existing critically evaluated data and is dynamic; as new data become available, the relationship can be updated to reflect changes. In this work, we use support vector regression (SVR) to develop estimation methods for critical properties and acentric factors based on critically evaluated data for over 900 pure compounds. From three-dimensional geometry and connectivity information, we calculate over 500 descriptors for each compound. A matrix of descriptor values defines the input vectors for SVR, whereas critically evaluated data for critical temperature, the ratio of critical temperature to critical pressure, and saturation reduced pressure form the targets. We determine optimal SVR parameters by minimizing the sum of absolute deviations between the SVR outputs and the target values. We use a genetic algorithm to find the Pareto front points that optimize the output fit while reducing the number of input vectors (descriptors). We use a single Pareto front point to make a final evaluation in SVR. To define uncertainties of predicted values, we use uncertainty propagation calculations based on a Monte Carlo method that employs Latin hypercube sampling.

INTRODUCTION

As data technology evolves, our ability to access and analyze historical data has improved rapidly, and dynamic modern databases allow rapid access to up-to-date thermodynamic values and their uncertainties.1 Critical properties (such as critical temperature, Tc, and critical pressure, pc) of substances are essential to understanding their thermodynamic behavior, in particular when that behavior is predicted based on an equation of state. We use critically evaluated property values generated through assessment of experimental data.2 Hereafter, we will refer to critically evaluated values as “experimental” values for brevity. Where experimental critical property data are unavailable, various estimation techniques are utilized to obtain critical values.3 A quantitative structure–property relationship (QSPR) regression method can be used to relate molecular properties (descriptors) to critical values.4–13 The relationship is trained and tested on existing experimental data and is dynamic; as new data become available, the relationship can be updated to reflect the changes. QSPR has long been used to predict chemical properties,14 and a number of advanced methods have been developed to implement predictive QSPR algorithms.4–12 The fundamental assumption of QSPR is that there is a relationship between quantities that describe the structure of a molecule (descriptors) and the thermophysical properties that are measured experimentally (such as critical temperature). There are two types of descriptors: experimental and theoretical. Experimental descriptors are measured directly, whereas theoretical descriptors are derived from a symbolic representation of a molecule. Although many proprietary programs offer thousands of descriptors, several hundreds of descriptors can be generated using open source tools (e.g., RDKit, CDK, and Open Babel) alone.

Support vector regression (SVR) offers many advantages for computing the relationship between descriptors and thermophysical properties.15 SVR is effective in high dimensional spaces, memory-efficient, and versatile. However, when the number of features (descriptors) approaches the number of samples (experimental data), the relationship easily becomes overfit to the set of data on which it is trained. This is why a large data set is desirable, and measures are taken to trim the descriptor set to a reasonable size. Additionally, probability estimates for robustness and uncertainty are not directly provided by SVR and require additional cross-validation and uncertainty modeling steps that can be computationally costly.

The goal of this work is to combine a large set of critically evaluated data with modern computational techniques to predict thermodynamic properties. In our previous work,12 we presented the basic ideas for developing such a system and demonstrated its performance for predicting critical properties.

Received: December 1, 2014
Accepted: March 31, 2015
Published: April 14, 2015
of pure substances. Here, we develop a substantially improved version of our earlier methodology, use an updated data set for model training, and extend model development to one more property.

As previously, we target critical temperature ($T_c$) and the ratio of critical temperature to critical pressure ($T_c/p_c$). The choice of $T_c/p_c$ over just $p_c$ was explained in detail previously, but simple theoretical considerations relate this ratio to the molecular volume. Although accurate correlations do require more than one parameter to model $T_c/p_c$, many fewer parameters (11 as opposed to over 30 according to the tests performed in this work) are needed overall as compared to describing $p_c$ itself. We also consider an additional property, the acentric factor ($\omega$), as defined by eq 1. The combination of $T_c$, $p_c$, and $\omega$ represents a minimal set of properties needed for approximate equation of state, and the resulting correlations can be used for compound screening based on their performance in thermodynamic cycles.

Initial tests indicated that, for the property related to $\omega$, the saturation reduced pressure at $T_s = T/T_c = 0.7$, $p_{sat}$, better correlations can be produced than for $\omega$ itself. Therefore, $p_{sat}$ was used for an actual target, and value $\omega$ can be obtained from it following its formal definition (eq 1)

$$\omega = -\log_{10} p_{sat} |_{T_s=0.7} - 1$$

(1)

In the first section of this paper, we discuss our research methods, including how we preprocessed the data and our use of support vector machines and genetic algorithms to make predictive relationships, as well as our uncertainty propagation and cross-validation techniques. In the second section, we relay the results of our predictive models. Finally, we include concluding remarks and suggestions for further work that can be done in this field.

## METHODS

### Overview

In this section, we will discuss the methods used to make predictions of critical properties and acentric factors for pure compounds. First, three-dimensional (3D) molecular models are created for each compound to be studied. A set of descriptor values is then calculated for each compound using the 3D model. SVR is used to relate the descriptor values to the experimental data. The SVR parameters are optimized using a genetic algorithm, which includes the reduction of overall number of descriptors used in the model. An overall workflow for our process is depicted in Figure 1. The flow of the prediction algorithm is shown in Figure 2.

<table>
<thead>
<tr>
<th>Source</th>
<th>Pre-processing</th>
<th>Inputs</th>
<th>Processes</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIST/TRC SOURCE</td>
<td>calculate structural information</td>
<td>mol files</td>
<td>calculate descriptors</td>
<td>predicted values: $T_c$, $p_c$, $p_{sat}$ and uncertainties</td>
</tr>
<tr>
<td>update/evaluate experimental data</td>
<td></td>
<td></td>
<td>run prediction algorithm</td>
<td></td>
</tr>
<tr>
<td>and uncertainties</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1.** Overall workflow for generation of predictive methods.

**Creation of 3D MOL files.** 3D molecular structure representations needed for calculation of molecular descriptors were generated from two-dimensional (2D) structure definitions in MOL format stored in the NIST/TRC SOURCE data archival system. The generation procedure was developed to provide diverse sampling of low-energy conformations consistent with the original 2D representation. The initial pool of 3D conformers is generated using several available utilities. All produced structures are optimized with TINKER utility using its implementation of the MMFF94 force field. Optimized structures are further analyzed. Conformers that are either inconsistent with the original 2D representation (as established by analysis of the resulting InChI strings) or redundant are removed from the pool. In some cases, when permitted by the original 2D representations, different stereoisomers are present in the resulting pool. If present, distinct stereoisomers are separated into groups also using InChI representation. For each distinct stereoisomer, a more detailed conformational analysis is performed. Depending on the number of rotatable bonds and the number of flexible ring atoms, either systematic or stochastic searches of conformational space are performed for each stereoisomer. Systematic search is carried out with the CONFAB utility. When systematic search becomes combinatorially prohibitive, 200 simulated annealing runs with TINKER are performed instead. The number of annealing runs was chosen in our previous work and represents a compromise between the computational cost and search effectiveness for the molecules within the scope of our data set. As previously, all newly generated structures were optimized with the MMFF94, and inconsistent or redundant structures are removed. At this stage, we retain up to 200 conformations per stereoisomer; the larger structure collections are truncated based on the MMFF94 energy and diversity measured as root-mean-squared distance (RMSD) from the lowest-energy conformer. The entire procedure as described above is implemented programmatically and executed without manual intervention. A single representative 3D conformer to be used in QSPR was generated by reoptimization of all structures in final collection with the semiempirical PM6 method and choosing the structure with the lowest free energy at standard conditions, as was suggested by us previously. PM6 represents substantial improvement over previous semiempirical methods both in terms of its accuracy and the scope. PM6 use in QSPR modeling was reported to produce models of quality comparable with those generated with much more computationally expensive density
functional theory (DFT) methods. All PM6 calculations in this work were performed with Gaussian 09 package.

**Preprocessing of Data.** Critical temperatures, critical pressures, acentric factors, and the corresponding uncertainties for pure compounds were produced via dynamic evaluations of up-to-date collection of experimental data using NIST/TRC ThermoData Engine. Only original experimental data (or, for critical pressures and acentric factors, robust interpolations/extrapolations of original experimental data) were used. The final compilation contained 942 compounds for which at least one of the three properties was available experimentally. The data archive is one of the most complete (and continuously updated) collections of original experimental data, and the generated data set covers the majority of available experimental information. Larger collections of critical properties were reportedly used for correlation development. However, as we pointed out previously, these larger data sets are based on compilations that include significant fractions of nonexperimental values typically estimated with other empirical methods.

**Descriptor Calculation.** Three open source tools were used to create the descriptor set: RDKit, CDK, and Open Babel. In cases of identical descriptors generated by different tools, the RDKit values were used. A number of additional descriptors were generated in-house from the results of PM6 calculations, as described in the Supporting Information. The
combined list included 533 descriptors. Descriptors were discarded before running SVR if they met one of the following criteria: they had empty values, they had zero range (every molecule had the same value), they were highly discrete (e.g., number of aromatic nitrogens), or they were strongly correlated with other descriptors. A full listing of the descriptors used is given in the Supporting Information.

Removal of Outliers. Some molecules always exhibited behavior inconsistent with the rest of the set. Water, nitrogen, hydrogen peroxide, and carbon monoxide are notoriously difficult to model using our approach and, thus, were discarded from the data set. These outliers are well-studied molecules with associated specific models.

\( \nu \)-Support Vector Regression (\( \nu \)-SVR). In the \( \nu \)-SVR formulation,\(^{37} \) a parameter \( \nu \) controls the number of support vectors by providing a lower bound for the fraction of support vectors. Scikit-learn,\(^{38} \) a machine learning algorithm for Python, was used to perform SVR calculations. Each experimental value used has an associated combined expanded uncertainty,\(^{39} \) the reciprocal of which is used as a weighting term in SVR. We set maximum weights of \( \nu = 0.10 \) and 250 for \( T_o \), \( T_e / P_o \), and \( p_{sat} \), respectively, in order to avoid overweighting of any particular point with comparatively low uncertainty. The weights were normalized before use in the SVR calculation. The built-in normalization routine in scikit-learn was also used on the descriptor values. All SVR modeling was performed with the third-order polynomial kernel function; the advantages of polynomial kernel were discussed in our previous work.\(^{42} \)

Once the SVR parameters were selected, the SVR routine was run with all the descriptor values constituting the input vectors. SVR is done in three steps:

1. The data are split into three groups: training, validation, and testing. This selection is random and is done with a 2:1:1 ratio of training:validation:testing. In our previous work,\(^{12} \) we used rational selection for division of the data set. For the present workflow, random selection is more appropriate (due to the added part for reduction of the number of descriptors, see below). On the other hand, recent analysis\(^{40} \) suggests that rational selection presents no significant benefit over random selection in terms of predictive power of the resulting models.

2. Run SVR using the training set to create the correlations and the validation set to tune the SVR parameters. SVR is highly sensitive to three parameters: the penalty parameter (C), an upper bound on the fraction of training weights and a lower bound of the fraction of support vectors (\( \nu \)), and the independent term in the kernel function (\( r \)). A three-dimensional grid search is used to maximize the coefficient of determination (\( R^2 \)) between the predicted values and the targets.

3. Combine the training and validation sets and create the SVR correlations using the combined set, then test the correlations on the testing set.

After following these SVR steps, the next task is to reduce the number of descriptors used while maintaining a good fit to the target data. This was done by using a genetic algorithm in DEAP\(^{31} \) to find the Pareto front values that represent a trade-off between quality of the fit (the sum of absolute deviations, hereafter SAD) and the number of descriptors. Each individual in the population to be evolved was a set of descriptors to run as the input vectors in SVR, ranging from single descriptor values to the entire descriptor set. After 100 generations, a single Pareto front point was chosen and the descriptors from that set were used in a final SVR run. The Pareto front point with the lowest sum of absolute deviations is used as a baseline, and the point with the lowest number of descriptors within a range of 3–4% of the baseline is chosen for the final SVR model. This strategy for descriptor subset selection is consistent with recommendations for empirical model building based on large data sets.\(^{42} \)

Uncertainties. Uncertainties were estimated by a Monte Carlo method that employs Latin hypercube sampling, following the approach used previously.\(^{12} \) Similar methodologies were reported recently by Hajipour et al.\(^{43,44} \) Python mcerp package\(^ {45} \) was adopted in the present work. SVR itself does not provide a path to explore uncertainty propagation, which must be performed separately. This method allows the creation of uncertainty metrics based on a distribution of likely values based on the input experimental uncertainty. A total of 20,000 sample points were used for sampling using a normal distribution (with a \( \sigma \) given by half the experimental uncertainty). This means SVR was run 4000 times before a median value was taken for the predicted quantities. The uncertainties are calculated based on the spread of the predicted quantity distribution.

## RESULTS

For each target property, a description of these results, as well as accompanying figures, are provided below.

### Results for \( T_c \)

Figure 3 shows a comparison of SVR predictions with experimental values for \( T_c \) for the training (gray points) and testing and validation combined (red points) sets. The left plot is the comparison of predicted versus experimental \( T_c \) for all the descriptors available (top) and

![Figure 3](image-url)
Pareto front-selected descriptors (bottom); the right plot shows the Pareto front. The y uncertainty bars on the SVR plots indicate the uncertainty calculated from the Latin hypercube sampling method. The experimental (x axis) uncertainty bars are also provided. 106 descriptors are used in the first SVR run after initial filtering described in the previous section. After the Pareto front analysis, 33 descriptors listed in Table 1 are chosen for the final SVR model. The SAD for the final calculation is $3.69 \times 10^3$ K, above the initial run SAD of $3.31 \times 10^3$ K.

**Table 1. Descriptors Used in the Final Model for $T_c$ Ordered by the Importance Index**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gvdWV</td>
</tr>
<tr>
<td>2</td>
<td>Weighted Path 1</td>
</tr>
<tr>
<td>3</td>
<td>BCUT 5</td>
</tr>
<tr>
<td>4</td>
<td>gCOSMO-PCD</td>
</tr>
<tr>
<td>5</td>
<td>gCOSMO-NCD</td>
</tr>
<tr>
<td>6</td>
<td>Topological polar surface area</td>
</tr>
<tr>
<td>7</td>
<td>Carbon Types 3</td>
</tr>
<tr>
<td>8</td>
<td>Molecular distance edge 00</td>
</tr>
<tr>
<td>9</td>
<td>Molecular distance edge 10</td>
</tr>
<tr>
<td>10</td>
<td>Spectrophore 23</td>
</tr>
<tr>
<td>11</td>
<td>Spectrophore 47</td>
</tr>
<tr>
<td>12</td>
<td>gRAFreq</td>
</tr>
<tr>
<td>13</td>
<td>gLUMO</td>
</tr>
<tr>
<td>14</td>
<td>BCUT 4</td>
</tr>
<tr>
<td>15</td>
<td>Spectrophore 48</td>
</tr>
<tr>
<td>16</td>
<td>BCUT 1</td>
</tr>
<tr>
<td>17</td>
<td>Weighted Path 3</td>
</tr>
<tr>
<td>18</td>
<td>Balaban J value</td>
</tr>
<tr>
<td>19</td>
<td>Molecular distance edge 07</td>
</tr>
<tr>
<td>20</td>
<td>gNCD</td>
</tr>
<tr>
<td>21</td>
<td>Atom additive LogP 2</td>
</tr>
<tr>
<td>22</td>
<td>gHLgap</td>
</tr>
<tr>
<td>23</td>
<td>gCOSMO-VAR</td>
</tr>
<tr>
<td>24</td>
<td>MOE MR VSA Descriptor 06 ($2.75 \leq x &lt; 3.05$)</td>
</tr>
<tr>
<td>25</td>
<td>BCUT 0</td>
</tr>
<tr>
<td>26</td>
<td>MOE Charge VSA Descriptor 08 (0.00 $\leq x &lt; 0.05$)</td>
</tr>
<tr>
<td>27</td>
<td>Molecular distance edge 02</td>
</tr>
<tr>
<td>28</td>
<td>Carbon Types 4</td>
</tr>
<tr>
<td>29</td>
<td>Molecular distance edge 08</td>
</tr>
<tr>
<td>30</td>
<td>Molecular distance edge 12</td>
</tr>
<tr>
<td>31</td>
<td>Carbon Types 5</td>
</tr>
<tr>
<td>32</td>
<td>BCUT 3</td>
</tr>
<tr>
<td>33</td>
<td>Carbon Types 0</td>
</tr>
</tbody>
</table>

A k-fold cross-validation method is used to test the robustness of the final model. In this method, the data are randomly divided into k subsets, then the model is run with one of the folds as the testing set and the combined other folds as the training set. In our case, we used $k = 5$ and evaluate the mean absolute deviation, or MAD for each fold. For the five folds, the mean MAD is 13.4 K with a standard deviation of 1.6 K. The low standard deviation between folds indicates that the model is robust.

**Results for $T_c/p_c$.** SVR predictions of $T_c/p_c$ compared to experimental values are shown in Figure 4. The plots use the same setup and colors as Figure 3. For $T_c/p_c$, 100 descriptors are used in the first run. After the Pareto front calculation, 11 descriptors (see Table 2) remain for the final SVR calculations. The SAD for the testing set for the initial and final runs, respectively, are 2.47 and 2.65 K/kPa, indicating an increase in the SAD for the final run. However, most of that increase in SAD is due to the sparsity of values for high $T_c/p_c$; the left side of the plot remains virtually unchanged. For the k-fold cross-validation over 5 folds, the mean MAD is $1.31 \times 10^{-2}$ K/kPa and the standard deviation is $1.56 \times 10^{-3}$ K/kPa.

**Results for $p_{c}^{sat}$.** For the final target property, $p_{c}^{sat}$, Figure 5 shows predicted versus experimental values. The plots use the same setup and colors as Figures 3 and 4. For $p_{c}^{sat}$, 99 descriptors pass in to the initial calculation. The SAD for the testing set in the initial SVR run is 0.412. Although this target property displays higher scatter than the previous two, the predicted values generally fall within the limits of the experimental uncertainties. After the Pareto front calculations, 30 descriptors remain for the final SVR model accompanied by an increase in the SAD, which has a final run value of 0.479. This increase in the SAD is a reasonable trade-off for reducing the number of descriptors used in the calculation. Full listing of descriptors used in the final model for $p_{c}^{sat}$ is given in Table 3.

**Table 2. Descriptors Used in the Final Model for $T_c/p_c$ Ordered by the Importance Index**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gvdWV</td>
</tr>
<tr>
<td>2</td>
<td>gPCD</td>
</tr>
<tr>
<td>3</td>
<td>Kier and Hall Chi cluster index 5</td>
</tr>
<tr>
<td>4</td>
<td>Molecular distance edge 04</td>
</tr>
<tr>
<td>5</td>
<td>Carbon Types 5</td>
</tr>
<tr>
<td>6</td>
<td>BCUT 1</td>
</tr>
<tr>
<td>7</td>
<td>MOE logP VSA Descriptor 01 ($-\infty &lt; x &lt; -0.40$)</td>
</tr>
<tr>
<td>8</td>
<td>gCOSMO-SKW</td>
</tr>
<tr>
<td>9</td>
<td>gSumIR</td>
</tr>
<tr>
<td>10</td>
<td>Carbon Types 2</td>
</tr>
<tr>
<td>11</td>
<td>MOE MR VSA Descriptor 06 ($2.75 \leq x &lt; 3.05$)</td>
</tr>
</tbody>
</table>

Figure 4. Left-top panel: comparison of SVR predictions using all descriptors with the experimental values for the critical temperature to critical pressure ratio. The gray points: combined training and validation sets, the red points: testing set. Right panel: the Pareto front of the SAD versus the number of descriptors for the critical temperature to critical pressure ratio. The Pareto front point chosen for the final model is marked with a red X. Left-bottom panel: comparison of final SVR model predictions with the experimental values for the critical temperature to critical pressure ratio.
### DISCUSSION

**Important Descriptors.** The number of descriptors (and which ones are chosen by the genetic algorithm) differ for $T_c$, $T_c/p_c$, and $p_{crt}$. In order to access the relative importance of chosen variables (descriptors) for a given property, a first-order sensitivity analysis of the final SVR models is performed. We define sensitivity coefficients with respect to each feature (scaled descriptor) as

$$s_i = \frac{\partial y}{\partial p_i}$$

where $y_i$ is the property for the $i$th compound and $p_i$ the $j$th scaled descriptor. Then, we define the importance index for the $j$th descriptor as

$$I_j = A \sum_i |b_{ij}|$$

where the summation is performed over all compounds in the data set, and the normalization constant $A$ is chosen to set the maximum $I_j$ to 1. In actual calculations, partial derivatives in eq 2 were evaluated numerically with the central differencing scheme. The importance index introduced in such a manner provides a metric for ranking of individual descriptors included in the final model.

The results of this analysis are shown in Figure 6. Tables 1, 2, and 3 list the descriptor names linked to the $x$-axis labels in Figure 6.

Not only is each predicted quantity modeled with a different number of features but the shape of the importance index curve varies as well. $T_c$ has the greatest number of features, and $T_c/p_c$ has the least, meaning that the latter requires fewer features to model than the former. Additionally, the relative importance of features drops off rapidly for $T_c/p_c$, but shows only a slow decline for $T_c$. $T_c/p_c$ can be modeled with fewer features because the relative importance of features decreases heavily after the first few.

A full list of descriptors for $T_c$ sorted by their importance is given in Table 1. The five most important descriptors for $T_c$ are gvdWV, Weighted Path 1, BCUT 5, gCOSMO-PCD, and gCOSMO-NCD. gvdWV is the van der Waals volume, gCOSMO-PCD and gCOSMO-NCD are, respectively, the positive and negative charge densities on COSMO cavity surface. The Weighted Path descriptors are formulated by CDK and characterize molecular branching.

The BCUT descriptors are derived from a weighted version of the Burden matrix (B) calculated in CDK. B is a symmetric matrix with the dimension equal to the number of non-hydrogen atoms in the molecule. Off-diagonal elements of $B$ characterize connectivity within the molecule. Diagonal elements contain atomic properties relevant to intermolecular interactions. CDK implements three types of Burden matrices with diagonal elements defined as atomic weights ($B_1$), empirically derived atomic partial charges ($B_2$), and atomic polarizabilities ($B_3$). The BCUT descriptors are the highest and lowest eigenvalues for each of the three matrices: BCUT 0/BCUT 1, BCUT 2/BCUT 3, and BCUT 4/BCUT 5 are the highest/lowest eigenvalues of $B_1$, $B_2$, and $B_3$, respectively. Specifically, BCUT 5 is the lowest eigenvalue of the matrix weighted by polarizability. In spite of being somewhat

---

#### Table 3. Descriptors Used in the Final Model for $p_{crt}$ Ordered by the Importance Index

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gCOSMO-POS</td>
</tr>
<tr>
<td>2</td>
<td>Auto Correlation Charge 1</td>
</tr>
<tr>
<td>3</td>
<td>gvdWV</td>
</tr>
<tr>
<td>4</td>
<td>gGAFreq</td>
</tr>
<tr>
<td>5</td>
<td>gPNSA1</td>
</tr>
<tr>
<td>6</td>
<td>gNCD</td>
</tr>
<tr>
<td>7</td>
<td>MOE Charge VSA Descriptor 01 ($-\infty &lt; x &lt; -0.30$)</td>
</tr>
<tr>
<td>8</td>
<td>Molecular distance edge 04</td>
</tr>
<tr>
<td>9</td>
<td>gCOSMO-NCD</td>
</tr>
<tr>
<td>10</td>
<td>EState VSA Descriptor 09 ($4.69 \leq x &lt; 9.17$)</td>
</tr>
<tr>
<td>11</td>
<td>Weighted Path 1</td>
</tr>
<tr>
<td>12</td>
<td>Hall Kier Alpha</td>
</tr>
<tr>
<td>13</td>
<td>gCOSMO-PCD</td>
</tr>
<tr>
<td>14</td>
<td>gSumR</td>
</tr>
<tr>
<td>15</td>
<td>MOE Charge VSA Descriptor 03 ($-0.25 \leq x &lt; -0.20$)</td>
</tr>
<tr>
<td>16</td>
<td>gCOSMO-VAR</td>
</tr>
<tr>
<td>17</td>
<td>MOE MR VSA Descriptor 04 ($2.24 \leq x &lt; 2.45$)</td>
</tr>
<tr>
<td>18</td>
<td>MOE MR VSA Descriptor 07</td>
</tr>
<tr>
<td>19</td>
<td>Balaban J value</td>
</tr>
<tr>
<td>20</td>
<td>MOE Charge VSA Descriptor 09 ($0.05 \leq x &lt; 0.10$)</td>
</tr>
<tr>
<td>21</td>
<td>gDipole</td>
</tr>
<tr>
<td>22</td>
<td>MOE logP VSA Descriptor 04 ($0.00 \leq x &lt; 0.10$)</td>
</tr>
<tr>
<td>23</td>
<td>MOE logP VSA Descriptor 04 ($0.00 \leq x &lt; 0.10$)</td>
</tr>
<tr>
<td>24</td>
<td>MOE logP VSA Descriptor 01 ($-\infty &lt; x &lt; -0.40$)</td>
</tr>
<tr>
<td>25</td>
<td>Kier and Hall Chi path cluster index 5</td>
</tr>
<tr>
<td>26</td>
<td>gPCD</td>
</tr>
<tr>
<td>27</td>
<td>MOE MR VSA Descriptor 09 ($3.80 \leq x &lt; 4.00$)</td>
</tr>
<tr>
<td>28</td>
<td>Kappa3</td>
</tr>
<tr>
<td>29</td>
<td>MOE logP VSA Descriptor 08 ($0.25 \leq x &lt; 0.30$)</td>
</tr>
<tr>
<td>30</td>
<td>Molecular distance edge 14</td>
</tr>
</tbody>
</table>

For the $k$-fold cross-validation over 5 folds, the mean MAD is $3.42 \times 10^{-3}$ and the standard deviation is $4.00 \times 10^{-4}$.

![Figure 5.](image-url)
nonintuitive, BCUT descriptors effectively combine molecular information relevant to modeling of a range of different properties, including normal boiling points.\(^{49}\)

For \(T_c/p_c\), gvdWV, gPCD, Kier and Hall Chi cluster index \(S\), Molecular distance edge \(S\), and Carbon Types \(S\) are the five most important descriptors (as shown in Table 2). As expected,\(^{12}\) the van der Waals volume, gvdWV, is by far the most dominant factor in the \(T_c/p_c\) correlation. It is followed by gPCD, positive-charge surface density based on the PM6Mul- liken partial charges projected on the van der Waals surface. Molecular distance edges\(^{50}\) are calculated by CDK, which also provides the Kier and Hall descriptor that evaluates the Chi cluster index.

gCOSMO-POS, Auto Correlation Charge 1, gvdWV, gGAFreq, and gPNSA1 constitute the most important descriptors for \(p_{s}^{\text{sat}}\) (see Table 3). Auto Correlation Charge 1 is a Moreau-Broto autocorrelation descriptor\(^{51}\) that uses empirical partial charges. gCOSMO-POS is the partial positive charge on the COSMO cavity surface. gPNSA1 is a partial negatively charged surface area, and gGAFreq is a geometric-mean vibrational frequency.

Van der Waals volume is the most prominent descriptor across the board, appearing in the final cuts for all three predicted properties, ranking first for \(T_c\), \(T_c/p_c\), and third for \(p_{s}^{\text{sat}}\). As it represents the measure of molecular size (and, to some extent, atomic composition), its importance is consistent with well-established empirical observations. Also of note is strong presence of COSMO-based descriptors introduced in this work. As explained in Supporting Information, these descriptors were derived from computed charge distributions on COSMO cavity surfaces that strongly affect intermolecular interactions.

**Comparison with the Previous Work.** Our previous models\(^{12}\) have shown superior performance in predicting \(T_c\) and \(T_c/p_c\) as compared to several commonly used group contribution-based methods. Here, we compare the performance of our earlier models with those developed in this work. The comparison is done for the new, updated data set used here. The results, in a form of statistical distributions of deviations, are presented in Figure 7 for \(T_c\) and in Figure 8 for

![Figure 7. Comparison of deviation distribution between the current study and\(^{12}\) for critical temperature. \(\delta_{\text{exp}}\) is the experimental uncertainty.](image)
As seen, the new results exhibit a noticeable improvement over the previous work. The percentage of compounds for which $T_c$ was predicted within 6 K or experimental uncertainty increased from 78% to 91%. The percentage of compounds for which $T_c/p_c$ was predicted within 3% or experimental uncertainty slightly increased, from 82% to 83%. The decrease in the deviation distribution functions with an increase of the deviation for the experiment also appears faster for the present models. This is primarily due to improved methodology used here (better strategies and theoretical models used to generate 3D geometries, extended list of descriptors, improved approach to the model generation and feature selection). Another source of improvement is the updates to the data set due to addition of new data, correction of identified erroneous entries, and improvements in data-evaluation algorithms. The methodology based on dynamic model generation as both the quantity and the quality of data increase clearly represents a better strategy for development of estimation methods.

The demonstration of the method performance in the form of discrete deviation distributions when the first “bin” also includes the cases predicted within the experimental uncertainty was used by us before and represents a compromise between the rigor and the relevance in a context of practical use. On the other hand, a dominant practice is to report a single-valued metric, such as correlation coefficient, average absolute deviation or root-mean-squared (RMS) deviation. These values are typically used to compare and rank the performance of different models. However, such comparisons are only meaningful in a context of error-free data (or, in practical terms, when the deviations between the model and the data significantly exceed the data uncertainty). As the accuracy of the model increases and approaches the level of the input data, the global metrics can no longer be used. One can fit the data achieving “accuracy” better than the input data uncertainty which has no practical meaning and constitutes overfitting. Considering the dominant use of single-valued metrics and recent reports of high-accuracy correlations, it is of interest to perform a systematic comparison of the values reported in the literature for the studied properties with the actual levels of experimental uncertainties. The relevant data are compiled in Table 4.

For the experimental uncertainties, the data sets from the present work and two versions of the DIPPR database were analyzed. DIPPR represents a popular source of evaluated data frequently used in correlation development. The version from 2006 was cited by refs 35 and 53. The version from 2013 is the
most recent at the time of this writing. To maintain its standard of completeness, in addition to experimental data, the DIPPR database includes estimated values for compounds with no experimental data available. When analyzing the data, the subset based only on experimental data was used; analysis of all (combined experimental and predicted) data is also shown for illustration. DIPPR data for the acentric factor are not included due to the difficulty of separating the experiment-derived data from the estimates. The reported uncertainties are presented in the form of averages and RMS (absolute and relative). The present data sets for all three properties are larger than those based on experimental data from DIPPR due to somewhat different venues of the two projects: NIST/TRC collects any experimental data within its scope, and DIPPR focuses on the most complete characterization of the selected set of compounds while gradually expanding this selection. The averages of experimental uncertainties from the present data set and DIPPR are generally consistent within a factor of 2, and the differences between the two versions of DIPPR are very small. Combined, this information gives the scale of typical experimental uncertainty levels for large data sets. Smaller uncertainty scales can obviously be achieved within smaller subsets obtained by selective sampling based on uncertainty values. Various error metrics reported for a range of QSPR and group contribution-based estimation methods are also listed in Table 4 for all three properties. As seen, the majority of reported models does not exhibit obvious conflicts with established levels of experimental uncertainties; somewhat low error metrics observed in a few cases based on small data sets can be attributed to overall higher accuracy of these samples. However, three reports7,11,35 based on large data sets claiming the highest accuracy among other methods exhibit clear signs of overfitting across all considered properties. The problem is particularly pronounced for the critical pressure: the predictions of ref 7 show average absolute deviation that is a factor of 6 to 7 below the average experimental uncertainty,11 report average absolute deviation 2.5 to 3 times smaller than average experimental uncertainty, and35 result for the average relative deviation is a factor of 6 to 12 lower than the corresponding measure of experimental uncertainty. The data used in the study of7 were compiled from the multiple printed sources and their analysis is beyond the scope of this work. The data used by11 and35 were taken from the DIPPR database, versions 1999 and 2006, respectively. Data point statistics shown in Table 4 points to the source of the problem: both studies did not filter out nonequilibrium data, and the data sets used contained about 60% to 70% of predicted data. Formally, this makes the problem even worse since, as seen in Table 4, addition of estimated data points greatly increases the overall data uncertainties, thus resulting in even greater degree of overfitting. It should be noted that overfitting is normally expected to be detected and prevented with the statistical procedures applied in a course of machine learning (use of testing set, cross-validation, etc.). However, these procedures fail when the data set is dominated by the predicted data exhibiting errors that are mainly systematic in nature. As an additional verification of the above findings, we performed a comparison of the predictions of the present model and the model of35 using the experimental data from this study. Compounds common for both studies were selected for each property; the resulting data sets included 609, 507, and 484 compounds for $T_c$, $p_c$, and $\omega$, respectively. Detailed results are given in Supporting Information, and the main conclusions are as follows: the present model performs better against the considered set of compounds; the performance of the model of ref 35 is good but not to the extent of the accuracy that was claimed.

5 CONCLUDING REMARKS

A system for generation of empirical property-estimation methods based on large experimental data sets is presented. The system includes software-based critical evaluation of available experimental data stored in a large-scale data archival system, systematic generation of three-dimensional molecular models from stored two-dimensional structure definitions, calculation of molecular descriptors, and development of correlations using machine learning methods. Correlations for three properties, the critical temperature, the ratio of critical temperature and critical pressure, and the reduced saturation pressure (related to the acentric factor), were produced. Feature (variable) selection for the final models was performed using multiobjective genetic algorithm. Generated models exhibited excellent performance against the available experimental data as well as other estimation methods available. The predictive accuracy of the models was also confirmed with the k-fold cross-validation.

During feature selection, it was shown that, for large experimental data sets used here, one needs to use relatively large number of features to effectively utilize the available information and produce more accurate models. Traditionally, QSPR models were limited to just a few variables; here, the resulting correlations, after feature selection, included between 11 and 33 variables.

Several descriptors derived directly from quantum-chemical calculations and typically not used in QSPR were found to be useful for modeling of critical properties and acentric factors. Descriptors based on charge distributions on COSMO cavity surfaces were, expectedly, among the most important variables. Some added descriptors related to vibrational and spectroscopic analysis were also found to be important (geometric mean vibrational frequency, average vibrational frequency weighted by IR intensity, total IR intensity).

Also of note are identified cases of overly optimistic accuracy claims made in the recent literature. The problem originates from the misuse of the reference data sources that resulted in overfitting. It appears that modern machine learning methods reached the point when empirical models for property predictions can be generated very efficiently. However, the quality of the produced models is still controlled by the quantity and the quality of the input data.

5 ASSOCIATED CONTENT

Supporting Information
Table of in-house descriptors, table of all calculated descriptors, and figures comparing deviation distributions with other studies. This material is available free of charge via the Internet at http://pubs.acs.org.

Author Information

Corresponding Author
*E-mail: wendy.carande@nist.gov. Phone:(303)497-4648. Fax: (303)497-6682.

Funding
This research was funded in part by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy,
under Grant No. DE-EE0002057, and by the Professional Research Experience Program (PREP) Postdoctoral Fellowship, a collaboration between the University of Colorado at Boulder and the National Institute of Standards and Technology.

**Notes**

The authors declare no competing financial interest. 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.

**REFERENCES**


(52) Evaluated Standard Thermophysical Property Values. DIPPR (Design Institute for Physical Properties) 801. 2013


