Journal of
Chemical Information
and Modeling

Article

ThermoData Engine (TDE): Software Implementation of the Dynamic Data Evaluation Concept

Michael Frenkel, Robert D. Chirico, Vladimir Diky, Xinjian Yan, Qian Dong, and Chris Muzny

J. Chem. Inf. Model., **2005**, 45 (4), 816-838• DOI: 10.1021/ci050067b • Publication Date (Web): 18 May 2005

Downloaded from http://pubs.acs.org on February 26, 2009

ThermoData Engine (TDE): Software Implementation of the Dynamic Data Evaluation Concept

Michael Frenkel,* Robert D. Chirico, Vladimir Diky, Xinjian Yan,† Qian Dong, and Chris Muzny

Physical and Chemical Properties Division, National Institute of Standards and Technology,
Boulder, Colorado 80305-3328

Received February 23, 2005

The first full-scale software implementation of the dynamic data evaluation concept {ThermoData Engine (TDE)} is described for thermophysical property data. This concept requires the development of large electronic databases capable of storing essentially all experimental data known to date with detailed descriptions of relevant metadata and uncertainties. The combination of these electronic databases with expert-system software, designed to automatically generate recommended data based on available experimental data, leads to the ability to produce critically evaluated data dynamically or 'to order'. Six major design tasks are described with emphasis on the software architecture for automated critical evaluation including dynamic selection and application of prediction methods and enforcement of thermodynamic consistency. The direction of future enhancements is discussed.

1. INTRODUCTION

The NIST ThermoData Engine[1] (TDE) represents the first full-scale software implementation of the dynamic data evaluation concept for thermophysical property data. Below we shall discuss briefly the principal differences between static and dynamic data evaluation concepts.[2,3]

Traditionally, critical data evaluation is an extremely time- and resource-consuming process, which includes extensive use of manpower in data collection, data mining, analysis, fitting, etc. Because of this, it must be performed far in advance of a need within an industrial or scientific application. As a result, despite the enormous cost associated with the critical data-evaluation process, a very significant part of the existing recommended data has never been used in any meaningful application. This is because data requirements often shift between the initiation and completion of an evaluation project. In addition, it is quite common that by the time the critical data-evaluation process for a particular chemical system or property group is complete (sometimes after years of data evaluation involving highly skilled data experts), it must be reinitiated because significant new data have become available. This type of slow and inflexible critical data evaluation is defined here as 'static'. Essentially, all existing data evaluation projects fall into this category. Moreover, the static data evaluation process for thermodynamic data has been unable to provide adequate conceptual solutions for chemical process design in rapidly developing fields such as biotechnology, where there is a demand for simulation of hundreds of new technologies every year.

The new concept of dynamic data evaluation was developed at NIST/Thermodynamics Research Center (NIST/TRC).[2,3] This concept requires the development of large electronic databases capable of storing essentially all relevant experimental data known to date with detailed descriptions of relevant metadata and uncertainties. The combination of these electronic databases with artificial intellectual (expert-system) software, designed to automatically generate recommended data based on available experimental data, leads to the ability to produce critically evaluated data dynamically or 'to order'. This concept contrasts sharply with static critical data evaluation, which must be initiated far in advance of need. The dynamic data evaluation process dramatically reduces the effort and costs associated with anticipating future needs and keeping static evaluations current.

Implementation of the dynamic data evaluation concept consists of a number of major tasks:[3] (1) design and development of a comprehensive database system structured on the principles of physical chemistry and capable of supporting a large-scale data entry operation for the complete set of thermophysical (including transport) and thermochemical properties for chemical systems, including pure compounds, mixtures, and chemical reactions; (2) development of software tools for automation of the data-entry process with robust and internally consistent mechanisms for automatic assessments of data uncertainty; (3) design and development of algorithms and software tools to ensure quality control at all stages of data entry and analysis; (4) development of algorithms and computer codes to implement the stages of the dynamic data-evaluation concept; (5) development of algorithms to implement, target, and apply prediction methods depending on the nature of the chemical system and property, including automatic chemical structure recognition mechanisms; and (6) development of procedures allowing generation of output in a format interoperable with major engineering applications, including commercial simulation engines for chemical-process design.

2. SOURCE ARCHIVAL SYSTEM

Types of "Data". The term data is very general and is commonly applied to a wide variety of information. Within

* Corresponding author e-mail: frenkel@boulder.nist.gov.
† Current address: Scientific Information Center, Thornton, CO 80241.

10.1021/ci050067b
Published on Web 05/18/2005

the context of TDE, it is necessary to define several distinct types of data, as each plays a specific role within the overall structure. In a recent article[4] we provided working definitions for the basic types of thermodynamic data; true data (hypothetical); experimental data; predicted data; and critically evaluated data. Abbreviated definitions are provided here, but the reader is referred the original descriptions for more complete discussion.

True Data (Hypothetical). True data are exact property values for a system of defined chemical composition in a specified state. These data are (1) unique and permanent, (2) independent of any experiment or sample, and (3) hypothetical concepts with no known values. Because they are hypothetical, true data values are not represented within TDE; however, the property values generated by TDE (critically evaluated data) are approximations to the true values.

Experimental Data. Experimental data are defined as those obtained as the result of a particular experiment on a particular sample by a particular investigator. The feature that distinguishes experimental data from predicted and critically evaluated data is the use of a chemical sample including characterization of its origin and purity. Experimental data only are stored in the TDE-SOURCE archival system and serve as one of three data sources for processing by TDE; the others types are predicted data and user-supplied data.

Predicted Data. Predicted data are defined here as those obtained through application of a predictive model or method, such as a particular molecular dynamics, corresponding states, group contribution method, etc.

Critically Evaluated Data. Like predicted data, there is no particular sample involved with critically evaluated data. The feature that distinguishes critically evaluated data from predicted data is the involvement of the judgment of a data evaluator or evaluation system, such as TDE. Critically evaluated data are recommended property values generated through assessment of available experimental and predicted data.

TDE-SOURCE Archive of Experimental Data. The TDE-SOURCE archive of experimental data is a subset of the TRC SOURCE archive. The TRC SOURCE[5,6] was designed and built as an extensive relational data archival system for experimental thermophysical and thermochemical properties reported in the world's scientific literature. The SOURCE archive now includes over 1 600 000 numerical property values and their uncertainties on more than 17 200 pure compounds, 17 000 binary and ternary mixtures, and 4000 reaction systems. Its structure is based principally on the Gibbs phase rule and complies with all the requirements necessary to serve as a comprehensive data storage facility of experimental property data for implementation of the dynamic data evaluation concept. At present, the rate of collection of numerical property values is near 300 000 per year. It is estimated that TRC SOURCE will contain 80% of the available experimental thermodynamic data for organic materials by the end of 2006. Version 1.0 of TDE includes the TDE-SOURCE archive, which contains all experimental data for pure compounds contained in TRC SOURCE at the end of 2004.

3. PROCESSING OF EXPERIMENTAL DATA AND UNCERTAINTY ASSIGNMENTS

The purpose of the TDE software can be summarized as follows: use experimental data (contained in TDE-SOURCE), predicted data (generated with algorithms in TDE), plus any user-supplied data, as input to an expert system to generate critically evaluated data that are approximations to true data. The difference between experimental, predicted, or critically evaluated values and true values can be defined as an error. The error is never known; however, its mathematical expectation is never zero. A measure of the quality or confidence in an experimental, predicted, or critically evaluated value is expressed in terms of the uncertainty, which is a range of values believed to include the true value with a certain probability. A distinguishing feature of the TDE software is that all data types associated with TDE include estimates of uncertainties. Uncertainties for the experimental and predicted values form the basis of uncertainties for the critically evaluated values. It is important to emphasize that only comprehensive formulations of uncertainties (combined uncertainties that include uncertainty estimates for all error sources) provide the full measure of data quality. Only combined uncertainties with a level of confidence of approximately 95% are included in TDE. If these are propagated into uncertainties for properties related to industrial streams,[7] this can lead to enormous economic benefits in the implementation of results of chemical process simulations, particularly for optimal equipment selection. Implementation of this possibility can change fundamentally the nature of future chemical process modeling and design.

To serve as a basis for implementation of the dynamic data evaluation concept, an archive of experimental property data must meet several criteria:

- Full traceability from numerical values to bibliographic sources
- Unambiguous data definitions
- Minimal data transcription errors
- Consistent and reliable assignment of uncertainties

All experimental data stored in the TDE-SOURCE archive originate from the traditional archival thermodynamic literature (journal articles, reports, and theses). Data for the TDE-SOURCE archive are compiled using Guided Data Capture (GDC) software that was described previously in this journal.[8] Property values are captured with a strictly hierarchical system based upon rigorous application of the thermodynamic constraints of the Gibbs phase rule with full traceability to source documents. Use of the GDC software ensures that captured data meet the above criteria.

All data selection, capture, and archiving in the TRC SOURCE data system are completed within the TRC Data Entry Facility at NIST. Personnel of the NIST/TRC Data Entry Facility are responsible for managing all contributions to TRC SOURCE including those from in-house compilers and from NIST/TRC collaborators worldwide. NIST/TRC operates a large in-house data-capture effort staffed chiefly by undergraduate students of chemistry and chemical engineering. Collaborators from outside NIST/TRC are involved with focused data-capture projects such as those related to specific compound types, properties, lingual sources, or contributions to the TRC Tables project.[9] In 2003, these collaborations were expanded to include authors of articles published in major peer-reviewed journals, as indicated in recent announcements in the *Journal of Chemical and Engineering Data*,[10] *The Journal of Chemical Thermodynamics*,[11] *Fluid Phase Equilibria*,[12] and *Thermochimica Acta*.[13] All experimental data contributed by authors are available free of charge from the Web.[14]

All experimental data in TRC SOURCE, whether captured in-house or through outside collaborations, are processed and validated through the same procedures. Target data sources (articles, reports, etc.) are selected by expert thermodynamicists at NIST/TRC. All data and relevant metadata are captured with the GDC software. Senior NIST/TRC personnel review the collected information for completeness and general validity. The information is then archived in the TRC SOURCE data system. Once archived, subjecting each compound for which new data were added to the critical evaluation process of TDE provides an additional strict validity check. Large deviations between experimental and recommended values generated by TDE are carefully reviewed for typographical, compound identification, and other types of common errors. These procedures constitute the NIST/TRC data quality assurance program described previously.[15]

A key application of the information gathered with GDC is generation of an estimated combined standard uncertainty for each numerical property value. The expression of uncertainty requires clear definition of a variety of quantities and terms. Definitions and descriptions of all quantities related to the expression of uncertainty in this paper conform to the *Guide to the Expression of Uncertainty in Measurement*, ISO (International Organization for Standardization), October, 1993.[16] Reference 16 is commonly referred to by its abbreviation; the GUM. Additional information and related references can be found in ref 17. The recommendations have been summarized in Guidelines for the Evaluation and Expression of Uncertainty in NIST Measurement Results,[18] which is available via free download from the Internet (http://physics.nist.gov/cuu/).

Recently, we summarized the recommendations of the GUM with particular application to the reporting of experimental thermodynamic property data.[17] The most comprehensive expressions of uncertainty are the Combined Standard Uncertainty u_x and the Combined Expanded Uncertainty U_x. The Combined Standard Uncertainty u_x can be represented as a mathematical expression

$$u_x = \mathrm{f}(x_1, x_2, x_3, ...) \tag{1}$$

where the symbols x_i represent various contributions to the uncertainty that are propagated to estimate u_x. For example, the estimated uncertainty for a temperature value might be a function of the method and traceability of the sensor calibration, the instrument used to read its response, estimated gradients in the apparatus, effects of thermal inertia, and so forth. A well-designed experiment (i.e., one that includes the identification and control of the largest contributions x_i in eq 1 through determination of values of $\partial u_x/\partial x_i$) will improve the quality of the uncertainty estimates, but some scientific judgment is always involved in estimating u_x.

The combined standard uncertainty u_x represents one standard deviation and is related to the combined expanded uncertainty U_x through the expression

$$U_x = u_x \cdot k_x \tag{2}$$

where k_x is the coverage factor. The coverage factor is a numerical multiplier used to expand the combined standard uncertainty u_x with a specified level of confidence (usually 95%), which is an estimate of the probability that the measurand is within a specified range. The measurand is sometimes referred to as the 'true value', the exact value of which is unknowable, as noted earlier.

Recently, we reviewed practices in the expression of uncertainty in the experimental literature for thermodynamic property measurements with determinations of the critical temperature T_c for pure compounds used as a case study.[19] In that article it was shown that although gradual and continuous progress has been made in the reporting of uncertainty information, comprehensive uncertainty analyses remain rare, particularly with regard to consideration of contributions arising from sample impurities. Examples were provided of dramatic underreporting of uncertainty magnitudes due to failure to consider this important component. In the time period since 1990, approximately 42% of the articles reporting experimental T_c values listed only some type of precision information rather than a comprehensive combined uncertainty. Information on precision provides only a lower bound for the combined uncertainty and is of limited value to data evaluators and application engineers. Because reported uncertainties are so often poorly defined, a method for generation of independent estimates for combined uncertainties was developed at NIST/TRC.

The scheme developed by NIST/TRC for estimation of the combined standard uncertainty u_{comb} for a given property p as a function of constraints c and variables v is based upon a summation of terms:

$$u_{\mathrm{comb}}{}^2 = u_p{}^2 + \Sigma\{u_c(\partial p/\partial c)_v\}^2 + \Sigma\{u_v(\partial p/\partial v)_c\}^2 \tag{3}$$

The partial derivatives $(\partial p/\partial c)_v$ and $(\partial p/\partial v)_c$ are calculated approximately based upon the reported property values, if possible, or are estimated based upon approximate models for the property. The summations are over all constraints and variables. The standard uncertainty for the property u_p is rarely provided in a document and is estimated at NIST/TRC based upon the following general relationship:

$$u_p{}^2 = \{u_{\mathrm{method}}{}^2 + \Sigma(f_{\mathrm{m}}\cdot u_{\mathrm{method-details}}{}^2)\} + \{u_{\mathrm{sample}}{}^2 + \Sigma(f_s\cdot u_{\mathrm{sample-details}}{}^2)\} \tag{4}$$

This relationship involves two major contributions to u_p: uncertainties associated with the experimental method and those associated with the sample.

The term u_{method} is a default contribution to u_p and is based on the particular experimental method only. For example, a heat capacity $C_{\mathrm{sat,m}}$ determined with high-precision adiabatic calorimetry might have a default value for u_{method} of $0.002\cdot C_{\mathrm{sat,m}}$, while the same property determined with a differential-scanning calorimeter might have a default value 10 times larger. Some details related to particular methods are also considered, such as the method of calibration for a vibratingtube densimeter. These adjustments are indicated as $u_{\mathrm{method-details}}$ in eq 4 and can increase or decrease u_p based on the value of f_{m}, which is 1 or -1.

The term u_{sample} in eq 4 represents a default contribution to u_p related directly to the purity of the sample. Additional contributions to u_p related to the sample are indicated as $u_{\mathrm{sample-details}}$ in eq 4. The magnitude of $u_{\mathrm{sample-details}}$ is a function of several items, including the property, special characteristics of the material (e.g., thermal stability or hygroscopicity), and the experimental conditions (e.g., pressure or temperature range). This formulation is required to take into account the fact that impurities do not affect all properties or experimental methods to the same extent.

Values for the standard uncertainties u_c and u_v (and u_p, if appropriate) are taken from the original document, if provided and supported in the text. Default values are substituted for those not provided. Default values are based upon the general method used and are larger than those reported typically in the literature for the method. Incomplete reporting or the absence of this information in a document is considered indicative of the general quality of the work. Consequently, results reported with incomplete uncertainty descriptions are assigned uncertainties at NIST/TRC, which are commonly larger than those with well-supported estimates.

If estimates of u_{comb} are provided in the document, these are checked against the estimates calculated with eq 4. Large discrepancies are reviewed carefully and can form the basis for modification of default values. Because various indicators of precision (repeatabilities, deviations from fitted curves, etc.) provide only a lower limit for any uncertainty estimate, these are considered only if they are larger than the default uncertainties for the particular variable, constraint, or property.

The approach described here for the estimation of combined standard uncertainties provides the basis for consistent evaluations of the numerous data types encountered. This brief overview demonstrates that even approximate estimates of u_{comb} require careful consideration of a wide variety of contributions to the uncertainty. All numerical property values used in the TDE evaluation process are accompanied by uncertainties expressed as expanded combined uncertainties (level of confidence $\approx$ 95%), where independent variables and constraints (if any) are considered exact, and all uncertainty contributions are propagated to the property. Uncertainties are used for relative weighting of data points in regression procedures and for propagation to uncertainties for the program output: critically evaluated property data. In the absence of uncertainty information (e.g., if a user supplies data without uncertainties), the program assigns conservative default uncertainties based on the identity of the property.

4. SOFTWARE ARCHITECTURE

Property Groups or 'Blocks'. The properties that are evaluated dynamically within this first version of the TDE software are thermophysical properties of pure compounds. The focus is primarily on organic compounds containing the elements C, H, N, O, S, F, Cl, Br, I and to a lesser degree Si. Future developments will include expansions to include properties of reactions and mixtures, including phase equilibria.

Table 1. Blocks of Related Properties Used as Source Data for Critical Evaluation by the TDE Software

block name	properties
phase diagram block	triple point temperature
	critical temperature
	normal melting temperature
	boiling temperature
	normal boiling temperature
	phase boundary pressure[a]
	critical pressure
volumetric block	density
	molar density
	specific volume
	molar volume
	compressibility factor
	second virial coefficient
	third virial coefficient
	critical density
	critical volume
	critical compressibility
energy block	enthalpy of phase transition
	cryoscopic constant
	enthalpy of vaporization or sublimation
	heat capacity at constant pressure
	heat capacity at saturation pressure
	heat capacity at constant volume
	speed of sound
other property block	refractive index
	NaD-refractive index
	viscosity
	kinematic viscosity
	fluidity
	surface tension
	thermal conductivity

[a] The property phase boundary pressure includes pressures associated with all phase boundaries including vapor pressures, sublimation pressures, crystal-liquid boundary pressures, and crystal-crystal boundary pressures.

Table 2. List of All Properties that Are Critically Evaluated by TDE

block name	properties
phase diagram block	triple point temperature
	critical temperature
	phase boundary pressure (all phase boundaries)
volumetric block	single-phase density
	saturated density
	second virial coefficient
	critical density
energy block	enthalpy of phase transition
	enthalpy of vaporization or sublimation
	heat capacity at constant pressure (ideal gas)
	heat capacity at saturation pressure
	speed of sound
other property block	refractive index
	viscosity
	surface tension
	thermal conductivity

The complete list of properties considered within TDE is given in Table 1. The properties are classified into four property groups or blocks: phase diagram properties, volumetric properties, energy properties, and other properties. Table 2 lists all properties that are critically evaluated by TDE. The list in Table 2 is shorter because some properties listed in Table 1 are closely related through simple algebraic calculations (such as molar volume and specific density) or through reversal of properties and variables (such as boiling temperatures and vapor pressures). The combining of closely

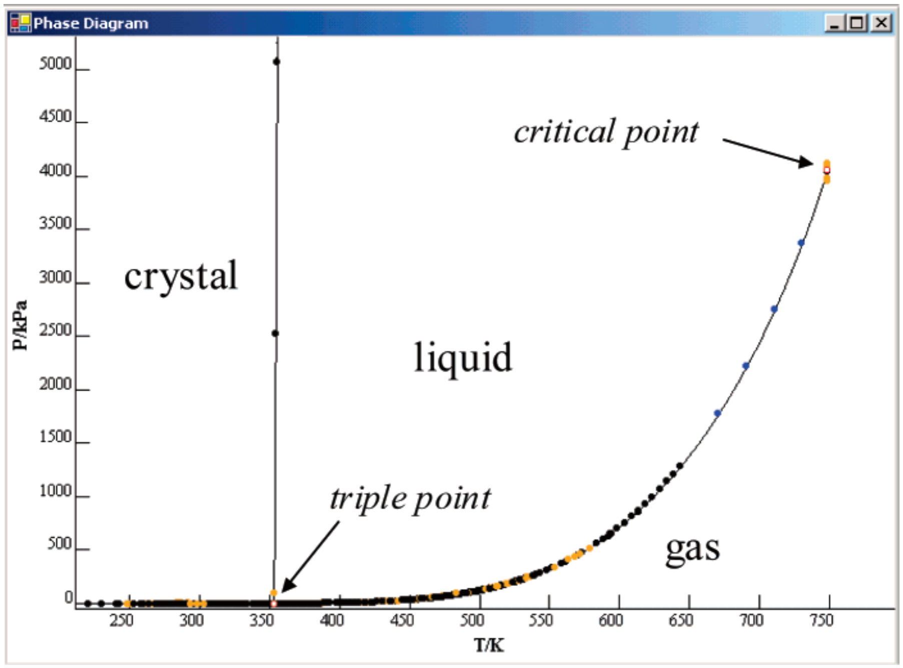


Figure 1. Phase diagram for a typical compound. The specific example is the phase diagram for naphthalene. The critical point, a triple point, and three subcritical phases (crystal, liquid, and gas) are indicated.

related properties into single properties is termed normalization and is described in the next section. After normalization, the properties within a block are evaluated together with subsequent enforcement of interblock property consistency, as described later.

Relationships between the property blocks are complex, but several generalizations can be made. The phase-diagram block is used to delineate the phase regions (crystal forms, liquid, and gas) and their boundaries (Figure 1). Properties in the other three blocks are associated with single phases, phase boundaries, or special points (triple or critical) defined by the phase diagram. The phase-diagram, volumetric, and energy blocks are tied by thermodynamic consistency conditions. That is, properties within the various blocks are related through mathematical thermodynamic identities. After initial evaluation within a block, enforcement of thermodynamic consistency conditions is one of the most important features of TDE and is described later in this paper. The fourth property block (other) has no influence on the first three blocks and is evaluated last because properties evaluated in the first three blocks are used for processing the properties in this block.

Representation of Properties in TDE. The number of variables F associated with a thermodynamic property is defined by the well-known Gibbs phase rule. For pure compounds, this rule reduces to the simple form $F = 3 - nPhase$, where *nPhase* is the number of phases present. Consequently, properties of pure compounds are (1) single valued, if they are associated with triple points; (2) functions of one variable, if they are associated with phase boundaries (or are limiting values such as virial coefficients), or (3) functions of two variables, if they are single-phase properties. Certain properties, which are not thermodynamic, may have more independent variables, such as refractive index, which is also a function of wavelength. All properties (other than those associated with triple points; i.e., three-phase equilibrium) are represented by defined equations with specified ranges for the independent variables.

There are three general methods for representation of evaluated properties that are based on (1) an equation of state (EOS), (2) separate equations for particular subsets of related properties, and (3) separate equations for all properties. The second approach is a hybrid of the first and third and is impractical for application in TDE because of the wide variety of properties and data scenarios addressed. The advantage of an EOS approach is that representation and intrinsic consistency of all thermodynamic properties with a single equation is inherent in the method.

There are myriads of alternative EOS formulations in the thermodynamic literature. These are mostly empirical in nature and are often applicable to specific families of compounds with, at best, a tenuous connection of parameters to physical quantities. A recent advanced implementation of the EOS approach is that based on the Helmholtz energy.[20] Widespread application of this advanced EOS is severely limited by the lack of required extensive high-quality experimental data for properties spanning the gas−liquid saturation lines and single-phase (gas, liquid, and supercritical fluid) regions. In practice, advanced EOS representations are constructed for data that were evaluated previously. Consequently, the advanced EOS approach is suitable for a limited number (≈100) of extensively studied compounds. Less-exacting ‘cubic’ EOS formulations have been developed for approximate property representations; however, these are incapable of representation of properties close to the limits of experimental uncertainty, making them generally unsuitable for TDE.

A key goal in the development of TDE was good representation of property values for a wide variety of data scenarios from extensive data for all phases and boundaries to the common case of limited or low-quality experimental data. Application of the second general method for property

Table 3. Normalization Procedures Used in TDE

property	normalized property	information required for normalization	comment
normal melting temperature	triple point temperature T_{tp}	none	increased uncertainty
enthalpy of fusion in air	enthalpy of fusion at T_{tp}	none	increased uncertainty
cryoscopic constant	enthalpy of fusion at T_{tp}	T_{tp}	
normal boiling temperature and boiling temperature at pressure p	vapor pressure, p	approximate dp/dT for uncertainty transformation	
triple point pressure	$p(T_{tp})$	T_{tp}	
critical pressure p_c	$p(T_c)$	critical temperature T_c	
critical compressibility	critical density	T_c, p_c, molar mass	
critical density ρ_c	$\rho(T_c)$	T_c	
specific volume	density		
molar volume, molar density, compressibility factor	density	molar mass	
virial coefficients	density	none	
fluidity	dynamic viscosity	none	
kinematic viscosity	dynamic viscosity	density	

representation (separate equations relating property subsets) was not chosen because such equations are effective for only very specific data scenarios. To achieve the goal of applicability to the widest array of data scenarios, representation of all properties by separate equations was chosen in development of TDE. This approach requires explicit application of thermodynamic consistency conditions during the evaluation process and is therefore very challenging mathematically. However, this method of representation has the important advantage that additional equations can be used to constrain derived property curves to only 'valid' shapes, thus increasing the quality of the evaluation in the event of an inadvertently bad data source.

Data Sources. Data used in the TDE evaluation process are experimental data stored in the TDE-SOURCE database, predicted data for filling gaps in experimental data, and property values entered by the user. This last data type allows inclusion of proprietary or other data not available with the TDE-SOURCE program database. User data are processed in the same way as that of the database. Data predictions in the first version of TDE (Version 1.0) are based on group-contribution and corresponding-states methods. Methods and algorithms for data prediction are discussed later in section 6.

Data used for evaluation (source data) within TDE are organized in a hierarchy for a particular compound. This organization allows easy display of traceability to the prediction method or bibliographic source for any numerical value. For a given compound, data are organized as data sets within properties and data points within data sets. Data sets join data from one literature source, for one sample, and for each experimental method. Data generated with a particular prediction method are also combined into a data set. Data points are property values with accompanying information: values of independent variables, a numerical property value, and an estimated combined uncertainty. Properties are distinguished by name, independent variables, and the identities of all phases present. For data involving two phases, the phase associated with the property (the primary phase) is identified. For example, densities on the saturation line are identified as those of the gas (in equilibrium with the liquid) or those of the liquid (in equilibrium with the gas). The provenance of every data point used in TDE can be traced to its source, whether it is a bibliographic citation, prediction method, or user-supplied data.

Data Normalization. Certain properties are commonly expressed in different, but closely related, formulations. For example, density can be expressed as specific density (mass/volume), molar density (moles/volume), specific volume (volume/mass), molar volume (volume/mole), or compressibility factor. In these cases, it is impractical to apply separate equations for each property with common parameters. Instead, the property data are normalized, i.e., reduced to a single property selected for output representation. Normalization is trivial, if it does not require data beyond the molecular mass and fundamental physical constants, such as those for density listed above. Nontrivial normalization requires other evaluated data. For example, conversion of kinematic viscosity into dynamic viscosity requires an evaluated density at the same conditions of temperature and pressure. Table 3 lists all of the normalizations carried out in TDE together with the properties required in the case of nontrivial normalizations. The initial and normalized properties are listed. Critical pressures and critical densities involve a special type of normalization in which the properties are converted to saturation properties at the critical temperature. This provides consistency between the saturation line and evaluated critical properties. Single-phase and saturated phase properties are described by separate equations in TDE; however, both kinds are used near phase boundaries as source data for the equations.

Evaluation Sequence (Overall). The sequence of the overall evaluation process is shown in Figure 2. Once the user selects a compound (step 1), there are three major steps in the process. The first major step has three substeps and involves user participation. The substeps are (1) compound selection, (2) data gathering from the TDE-SOURCE database (and from the user), and (3) optional data review by the user, as shown in Figure 2. The second major step has four substeps and does not involve user participation. The four substeps (as numbered in Figure 2) are (4) trivial normalization, (5) completion of the initial critical evaluation process within the first three blocks, (6) enforcement of interblock thermodynamic consistency, and (7) completion of the critical evaluation process for the final block properties not involved in interblock consistency enforcement. The third major step again involves optional user involvement and has three substeps. The three substeps (as numbered in Figure 2) are (8) review of results including various deviation plots for all source data, (9) selection of alternative fitting

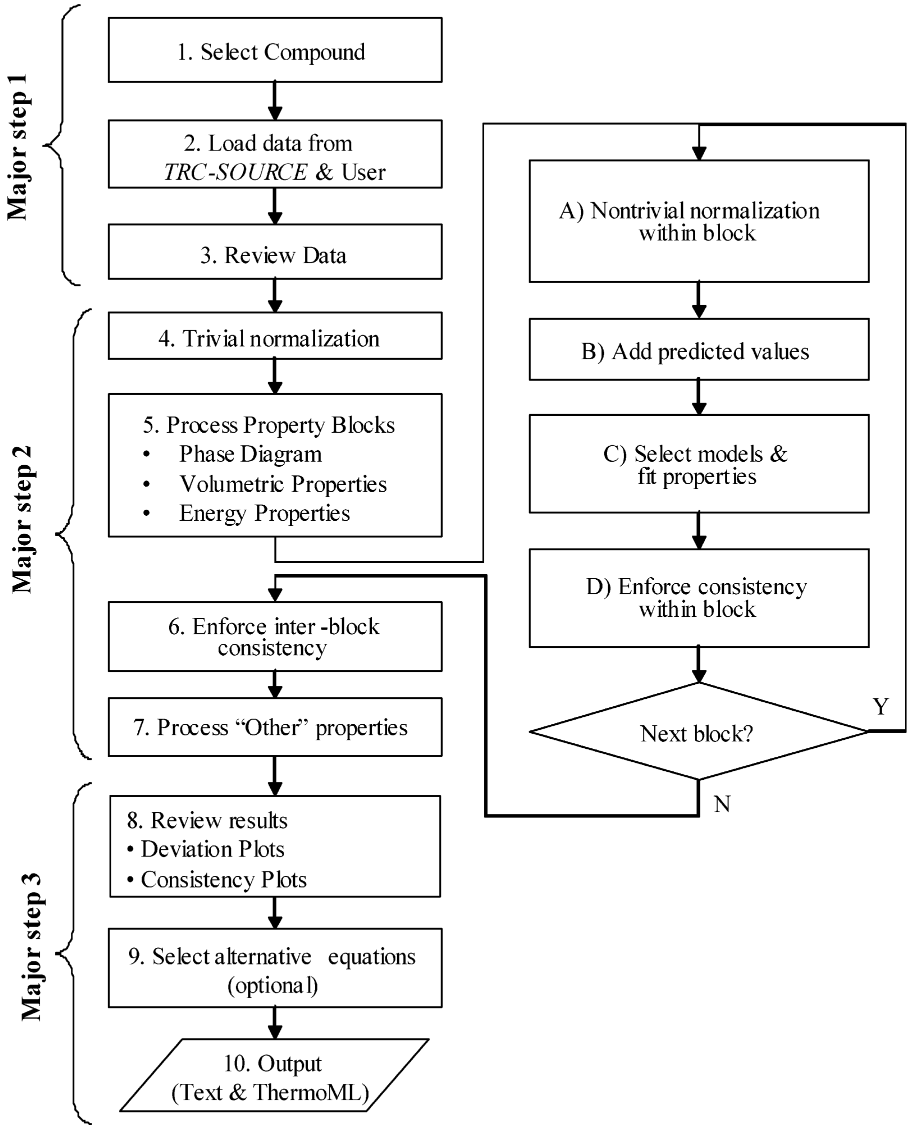

Major step 1
1. Select Compound
2. Load data from TRC-SOURCE & User
3. Review Data
Major step 2
4. Trivial normalization
5. Process Property Blocks
• Phase Diagram
• Volumetric Properties
• Energy Properties
6. Enforce inter-block consistency
7. Process "Other" properties
Major step 3
8. Review results
• Deviation Plots
• Consistency Plots
9. Select alternative equations (optional)
10. Output (Text & ThermoML)
A) Nontrivial normalization within block
B) Add predicted values
C) Select models & fit properties
D) Enforce consistency within block
Next block?
Y
N

Figure 2. Sequence of the overall evaluation process.

equations, and (10) output in text and ThermoML format. Because the foundation of TDE is enforcement of thermodynamic consistency, all properties are always evaluated to the extent possible. This is why there is no step involving property selection by the user before evaluation.

The sequence in which the four property blocks are evaluated (as well as the sequence of property evaluations within each block) is critical. Each block requires data evaluated in the previous block for evaluation of the properties it contains. Interblock consistency is enforced through iterative processes described later. The fourth property block (other) is dependent on the first three but is not involved in consistency enforcement, which is why it is last in the sequence.

Although the user is not directly involved in the automated critical evaluation process, there are three methods by which the user can affect the results obtained: (1) addition of data, (2) forced rejection of data, and (3) modification of estimated experimental uncertainties. These are completed ahead of the automated evaluation in the first major step at the top of Figure 2. The user has extensive control over the format of the program output in the third major step (bottom of Figure 2) including selection of alternative equations for data representation (i.e., equations other than those selected automatically by NIST ThermoData Engine) and specification of ranges for independent variables.

Evaluation Sequence (Single Property). Although the evaluation process for each property often includes some unique aspects, a general sequence can be described that demonstrates some key functions of the program. The following paragraphs describe each step in this general sequence, which is shown in Figure 3.

The first step, normalization, was described earlier. Once initial values for properties in the phase-diagram block are evaluated, phase adjustment for subsequent single-phase properties is possible (step 2 of Figure 3). Phase adjustment involves redistribution of data between phases and rejection of data with invalid phase specifications. Phases supported by TDE are gas (which includes the supercritical region), liquid, and various condensed phases, such as crystals of different types, glasses, and liquid crystals. Within the TDE-SOURCE data, the phase specification "fluid" is often applied to the gas, single-phase liquid, and supercritical regions. This is necessary because the phase regions are not

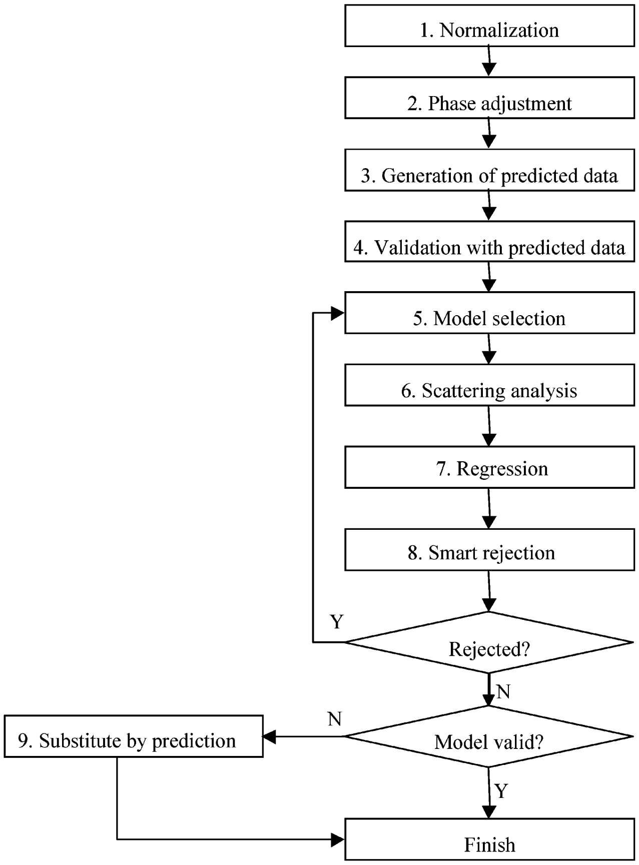


Figure 3. General evaluation sequence for a property.

defined until the phase boundaries are determined. During phase adjustment, all single-phase data are distributed between the gas and liquid phases. All values with $T > T_c$ or $p < p_{sat}$ are associated with gas. All other fluid-phase data are considered liquid. For saturated properties, data outside the defined extent of the phase boundaries are rejected and not used in evaluation. For example, all vapor pressure values with T greater than the evaluated T_c are rejected.

In step 3 of Figure 3, predicted values and their estimated uncertainties are generated where possible. Prediction methods and their selection are described later in section 6. The predicted values are then used in step 4 (Figure 3) for rough validation of the experimental data. Values deviating from the predictions by more than triple the conservatively estimated uncertainties for the predictions are rejected. Uncertainties for the predictions are relatively large, so this validation step is primarily to catch large errors that are typically typographical either from the original source documents or generated during data processing.

Flexible and automated model selection (steps 5–8 of Figure 3) is a key feature of the TDE software and is based on the extent and quality of the experimental data available. For example, the 5-parameter Wagner equation[21] is selected for vapor pressure representation, if the critical temperature is available; otherwise, an expansion of ln(p) vs T is selected. Scattering analysis (step 6 of Figure 3) is applied to each data set (typically data from a particular bibliographic source). This analysis checks for large deviations within a particular data set and checks the validity of the estimated experimental uncertainties by fitting the selected equation with an appropriate number of terms to all of the available data and determining both the deviations for individual data points and the overall scattering. This information is used in statistical weighting of data, as described later. In step 7 (Regression), the number of terms is selected (if supported by the mathematical form of the equation), and the equation parameters are fit to the experimental data. Linear and nonlinear (Levenberg–Marquardt, simplex, Powell) least-squares fitting methods are employed by TDE.[22] After fitting, data that show relatively large deviations are detected and rejected through the smart rejection procedure (step 8 of Figure 3). This procedure is similar to that described by Wilhoit et al.[23] The tolerance level Γ_i for rejection is based on the data quality in the neighborhood of each data point

$$\Gamma_i = f \cdot \sum_j \mathrm{w}_j \Delta_j \cdot e^{-|X_j - X_i|/k} \qquad (5)$$

where Γ_i is the tolerance for ith data point, f is the tolerance factor (usually 3), w_j is the weight of the jth data point, Δ_j is the square of the deviation of the jth data point from the equation, X is the independent variable (eq 5 is shown for the case of a property as a function of one variable), and k is the propagation distance parameter (usually 10–20 K for temperature as an independent variable). Parameters controlling rejection are f and k and the weights w. The ith data point is rejected if the square of its deviation from the equation exceeds the Γ_i criterion. Setting $f = 3$ and $k = \infty$ reduces the equation to the conventional 3σ-rejection. Use of the smart rejection procedure increases the tolerance in regions with lower data quality. Figure 4 illustrates the results of smart rejection with the experimental vapor pressures of propane as an example. Extensive, high-quality data are available for $T > 175$ K, but only data with relatively large uncertainties are available between 175 K and T_{tp} (near 85 K). Data points shown in gray are outside the tolerance limits.

After step 8 of Figure 3, if any data were rejected, steps 5–8 are repeated (with the rejected data removed). This cycle is repeated until no further data are rejected. As data are rejected, the number of parameters or even the selected equation may be changed during the cycle, depending on the property. In step 10 of Figure 3, the resulting equation is tested for validity of shape using criteria individual to each property and equation. The equation is also tested to determine whether it adequately describes the experimental data. If deviations exceed 10 times the estimated uncertainty of the experimental data points, the equation is rejected. If the equation shape is invalid or does not adequately describe the experimental data, they are substituted by predicted values (step 9 of Figure 3).

Application of Predictions. As noted earlier, predictions are used for validation of source data and for filling gaps in experimental data. The prediction methods used in the first release of TDE are group contribution (GC), corresponding-states (CS), and combined (GC-CS) methods. Details related to automated selection of prediction methods are described later in section 6. If multiple prediction methods are available for a given property, the most reliable (lowest estimated uncertainty) method is chosen. This approach allows simple addition of new or higher-level methods, such as those based on molecular dynamics or various types of ab initio calculations. Unfortunately, it is generally the case that higher-level

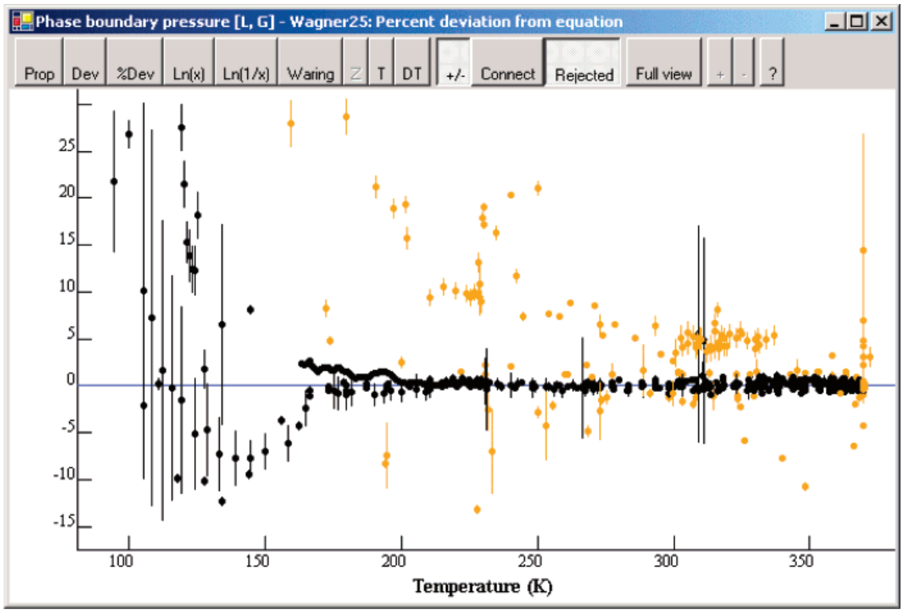


Figure 4. Percentage deviations from the fitted Wagner equation[21] for experimental vapor pressures for propane. Data points in gray were rejected with the smart rejection procedure that includes consideration of local data quality.

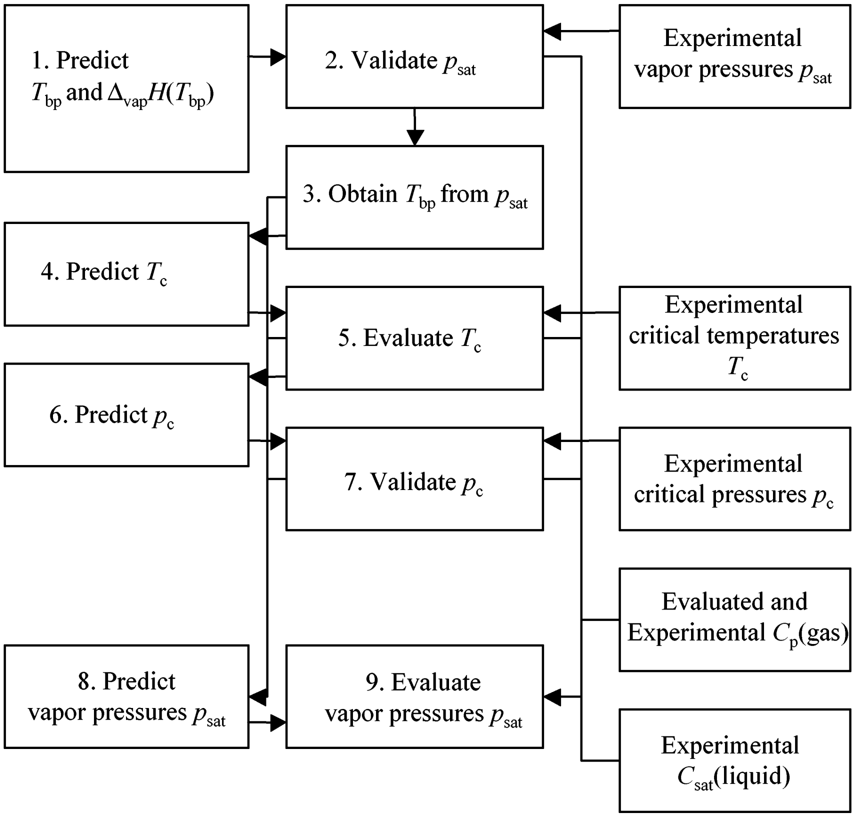

1. Predict T_{bp} and $\Delta_{vap}H(T_{bp})$
2. Validate p_{sat}
Experimental vapor pressures p_{sat}
3. Obtain T_{bp} from p_{sat}
4. Predict T_c
5. Evaluate T_c
Experimental critical temperatures T_c
6. Predict p_c
7. Validate p_c
Experimental critical pressures p_c
Evaluated and Experimental C_p(gas)
8. Predict vapor pressures p_{sat}
9. Evaluate vapor pressures p_{sat}
Experimental C_{sat}(liquid)

Figure 5. Use of predictions for the normal boiling temperature T_{bp}, enthalpy of vaporization at the normal boiling temperature $\Delta_{vap}H(T_{bp})$, the critical temperature T_c, the critical pressure p_c, and vapor pressure p_{sat} in evaluation of vapor pressure. Most of the prediction methods used are of the group-contribution type and require the molecular structure.

methods have been only poorly validated, resulting in property values with ill-determined uncertainties.

An example of the use of property prediction is shown in Figure 5. In step 1 of Figure 5, the normal boiling point T_{bp} and enthalpy of vaporization at that temperature $\Delta_{vap}H(T_{bp})$ are predicted with the molecular structure only (a group-contribution method) and are used for validation of experimental vapor pressures p_{sat} (step 2). A value of T_{bp} (now based on experimental data) is derived from the validated p_{sat} data (step 3) and is used in prediction of the critical temperature T_c (step 4). The predicted T_c is then used for rough validation of experimental T_c values and the evaluated T_c is generated (step 5). The evaluated T_c is used for critical pressure p_c prediction (step 6), and any experimental p_c data are validated (step 7). Predicted p_{sat} values are generated from T_{bp}, T_c, and p_c (step 8). Finally, p_c (converted to p_{sat} at T_c) is processed together with experimental and predicted (if needed) vapor pressure data to generate the evaluated vapor-

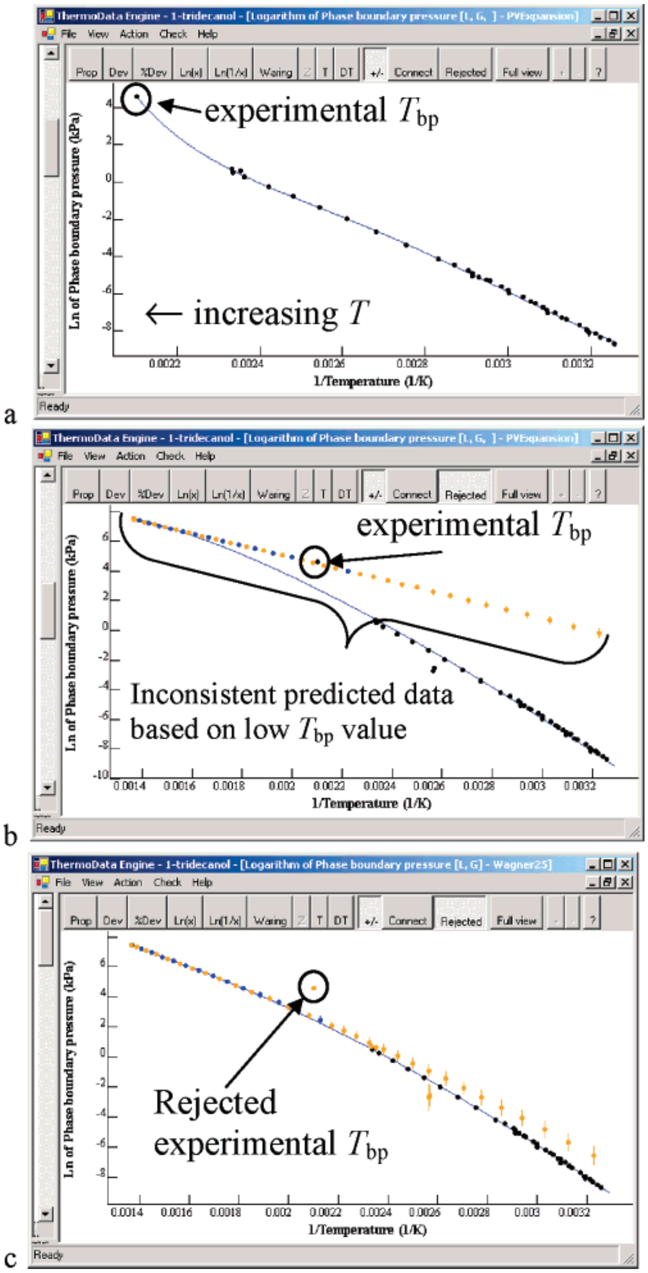

a
ThermoData Engine - 1-tridecanol - [Logarithm of Phase boundary pressure [L, G,] - PVExpansion]
File View Action Check Help
Prop Dev %Dev Ln(x) Ln(1/x) Waring Z T DT +/- Connect Rejected Full view + - ?
experimental T_{bp}
← increasing T
Ln of Phase boundary pressure (kPa)
4 2 0 -2 -4 -6 -8
0.0022 0.0024 0.0026 0.0028 0.003 0.0032
1/Temperature (1/K)
Ready
b
ThermoData Engine - 1-tridecanol - [Logarithm of Phase boundary pressure [L, G,] - PVExpansion]
File View Action Check Help
Prop Dev %Dev Ln(x) Ln(1/x) Waring Z T DT +/- Connect Rejected Full view + - ?
experimental T_{bp}
Inconsistent predicted data
based on low T_{bp} value
Ln of Phase boundary pressure (kPa)
6 4 2 0 -2 -4 -6 -8 -10
0.0014 0.0016 0.0018 0.002 0.0022 0.0024 0.0026 0.0028 0.003 0.0032
1/Temperature (1/K)
Ready
c
ThermoData Engine - 1-tridecanol - [Logarithm of Phase boundary pressure [L, G] - Wagner25]
File View Action Check Help
Prop Dev %Dev Ln(x) Ln(1/x) Waring Z T DT +/- Connect Rejected Full view + - ?
Rejected
experimental T_{bp}
Ln of Phase boundary pressure (kPa)
6 4 2 0 -2 -4 -6 -8
0.0014 0.0016 0.0018 0.002 0.0022 0.0024 0.0026 0.0028 0.003 0.0032
1/Temperature (1/K)
Ready

Figure 6. The top plot (a) shows the results after completion of step 3 of Figure 5 with an anomalously low experimental T_{bp} value included. Plot (b) shows that vapor pressures predicted (step 8 of Figure 5) with this anomalous value are inconsistent with all other values. In plot (c), the anomalous T_{bp} value was rejected and replaced with one generated from the vapor pressures at lower temperatures with much improved consistency apparent.

pressure equation (step 9). Liquid and ideal-gas heat capacities, if available, are used to constrain extrapolation of p_{sat} down to the triple point temperature T_{tp}.

The value of T_{bp} can be significantly changed during the vapor pressure evaluation, if the initially derived value (step 3 of Figure 5) included previously undetected large errors in the source data. In those cases, steps 3−9 are iterated until convergence is reached. An example of the effect of iterations is illustrated in Figure 6. The example shows experimental p_{sat} data for 1-tridecanol that includes an anomalously low value of T_{bp} (probably due to undetected sample decomposition). Figure 6a shows the results after completion of step 3 of Figure 5. The shape of the curve of ln(p_{sat}) against 1/T shown in Figure 6a is anomalous, and results from the experimental T_{bp} being too low. Use of this value in the vapor pressure prediction (step 8 of Figure 5) yields values that are inconsistent with all experimental data, as shown in Figure 6b. The experimental T_{bp} value is rejected, and a new T_{bp} is generated from the vapor pressures at lower temperature. The final vapor pressure curve is shown in Figure 6c, where the rejected T_{bp} value is apparent.

Default and Alternative Equations. The modular structure of TDE allows flexibility in equation selection for each property. For each equation a class is defined that provides required functionality to TDE for calculation of property values and their derivatives by state variables and equation parameters, determining the number of parameters, assigning statistical weights to source data, and assessing validity of parameters. Models (i.e., equations) can be added or substituted without changes to the evaluation code. The ‘default’ equations are the set of equations selected by TDE for the properties during the evaluation process. If multiple equations are available, TDE selects the most suitable or better-fitting equation. For example, if the critical temperature is available, the Wagner equation[21] is selected for vapor pressure representation; otherwise, another expansion function is used. The complete list of equations supported by TDE is provided in the Supporting Information.

Alternative equations are defined as those needed by users but not selected by TDE as default equations in the evaluation process. The user can request refitting evaluated data by any alternative equation. The alternative equations are fit to critically evaluated property values generated by TDE. The alternative equations are not used in the evaluation process. TDE supports three sets of alternative equations that are commonly used in engineering applications: Yaws,[24] DIPPR,[25] and PPDS[26] equations as well as some other common equations such as the Antoine equation for vapor pressures.

Statistical Weighting of Data. Statistical weights are used for scaling the contribution of each data point to the objective function[22] during fitting and in generation of uncertainties for evaluated values. Generally, weights are based on the reciprocal square of the uncertainties of the property values. As described earlier in section 3, uncertainties for experimental property values obtained from the archival literature are estimated at NIST/TRC based upon information provided by the article authors for metadata, which is often incomplete. The adjustments described here are intended to further refine the weighting factors in the data evaluation process.

The statistical weight w for a data point used in fitting is calculated by the equation

$$w = (U^2 + \delta^2 + s^2 + S^2)^{-1} \tag{6}$$

where U is the source data uncertainty (the expanded combined uncertainty estimated as described in section 3), δ is the deviation from an appropriate equation fitted to the data set, s is a data set quality factor (average scattering from the equation fitting the data set), and S are adjustments for data sets comprised of smoothed values or values calculated by equations. The uncertainty U is revised if the original uncertainty is missing (possibly with user data) or is unrealistically small. Reasonable defaults are used in both cases for all properties. δ and s contributions characterize data set quality and are significant when the scattering within the data set exceeds the stated uncertainties *U*. The *S* contribution decreases weights for smoothed data sets because smoothed values show little scattering by definition. This scattering does not reflect the quality of the underlying experimental data. We assume that the best reported data are available as original experimental values. For single-value properties, such as triple point temperature, *w* reduces to $1/U^2$.

In certain situations, property values are transformed before fitting for making equations linear with respect to the parameters and allowing direct computation of the parameters. For example, vapor pressure is commonly fitted with a logarithmic equation form, as is saturated density, when fitted with the Rackett equation. In such cases, all contributions to *w* are transformed in the way the uncertainty would undergo during property transformation. When the transformation is logarithmic, uncertainties and deviations contribute to *w* as relative (divided by the property value) rather than absolute values. Weights are dynamically recalculated before each fitting procedure.

Data Quality Assurance for TDE-SOURCE Data. Automated data processing may give meaningless results if erroneous data are not detected and corrected or discarded. Such data are unavoidably present in any data source to some extent. TDE-SOURCE database data entry tools (Guided Data Capture)[8,27] ensure that all information entered is correctly defined and completely specified in terms of the Gibbs phase rule (phases, independent variables). The only remaining kinds of errors are numerical. Regardless of the source of erroneous values, whether it is a poorly designed experiment, typographical errors in publications, or misidentified phases or substances, three types of errors can be recognized:[15] (1) invalid property values (e.g., negative temperatures; subcritical gas densities corresponding to $Z > 1$; single-phase liquid densities lower than the saturated density at the same temperature), (2) out-of-range variable values (e.g., saturated vapor pressure reported for $T > T_c$; single-phase liquid density at a pressure lower than the saturated vapor pressure), and (3) large deviations from the fitting equation, taking into account the local data quality using eq 5.

Invalid values are detected before fitting each property based on the information available at that time in the evaluation process. Phase diagram properties are processed first, which allows use of the range of existence for each phase to invalidate any data outside the appropriate region. Property values then undergo additional validation against predicted values using the methods listed in Table 4, plus the group-contribution method of Ruzicka and Domalski[28] for values of C_{sat}(liquid) and the Peng−Robinson equation of state[29] for gas-phase densities.

Table 4. Major Properties Currently Predicted by TDE

property	methods	type[a]
normal boiling temperature	JR,[40b] CG,[41c] MP[42d]	GC
critical temperature	JR,[b] CG,[c] MP,[d] WJ[43e]	GC
critical pressure	JR,[b] CG,[c] MP,[d] WJ[e]	GC
critical density	JR,[b] CG,[c] MP[d]	GC
saturated liquid density	Yamada and Gunn[44]	CS
ideal gas heat capacity	JR[b, f]	GC
liquid heat capacity	Bondi[45]	CS
second virial coeffs	Xiang[46]	CS
third virial coeffs	Liu and Xiang[47] Orbey and Vera[48]	CS
vapor pressure	Ambrose and Walton[49]	CS
gas viscosity	Lucas[50]	CS
liquid viscosity	Sastri and Rao[51]	GC and CS
gas thermal conductivity	Chung[52]	CS
liquid thermal conductivity	Chung[53]	CS

[a] GC = group contribution; CS = corresponding states. [b] Joback and Reid. [c] Constantinou and Gani [d] Marrero and Pardillo. [e] Wilson and Jasperson. [f] Method of Joback and Reid with group parameters reevaluated at NIST/TRC.

Any data, even high-quality data, may be harmful in the evaluation process, if the uncertainties associated with them are excessively small. The smart rejection technique described earlier is not effective in such cases. (Excessively small ‘uncertainties’ are common in the archival literature because of incomplete assessment of uncertainty or ambiguous metadata, as described earlier in section 3. This results commonly in highly inconsistent property data.) The presence of data inconsistency or inadequate uncertainties for a data set can be revealed when a fitting equation has an invalid shape (validity criteria are discussed below) or fails to fit the data within the estimated uncertainties. Sometimes, the question of which data are incorrect can be resolved during the enforcement of thermodynamic consistency between properties. Once data inconsistency is detected, the algorithm attempts to resolve the problem by revising relative weights of the data sets. The basis for locating a problem data set is deviations from a rough fitting equation that ensures a valid shape (e.g., the Rackett equation for saturated liquid density) or the fact that elimination of one data set resolves the inconsistency. If all attempts to resolve inconsistency are unsuccessful, the experimental data are substituted by predicted values.

Uncertainties. Uncertainties are calculated based on the covariance method for variable dependent properties. The covariance matrix for an equation is obtained by multiplication of the reciprocal least-squares matrix built from sums of contributing data points (rather than average values) by the sum of the squares of the adjusted uncertainties U_A of data points, described below. When the least-squares task is nonlinear, the least-squares matrix is taken from the last iteration. If the least-squares matrix combines more than one property, an appropriate fragment of the reciprocal matrix may be taken.

Adjusted uncertainties U_A are calculated through combination of the estimated TDE-SOURCE data uncertainties *U* and curve deviations δ:

$$U_A^{\ 2} = {}^1/_2 \cdot (U^2 + \delta^2) \tag{7}$$

Curve deviations reflect both data errors and model limitations to some extent. However, it would be incorrect to rely entirely on curve deviations, which can be very small for smoothed data and cannot reflect nonrandom errors. Uncertainties of evaluated property values *U* calculated with TDE equations are calculated by the conventional formula

$$U = \left(\sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij} \cdot \frac{\partial F}{\partial p_i} \cdot \frac{\partial F}{\partial p_j} \right)^{1/2} \tag{8}$$

where C_{ij} are elements of the covariance matrix, N is the number of parameters, and $\partial F/\partial p$ are the first derivatives of the property with respect to the equation parameters.

During construction of a covariance matrix, some adjustments to the formal procedure are applied. If the only source of errors is random scattering with a normal distribution, and the model is fully adequate, the covariance matrix elements are divided by the number of data points reduced by the number of parameters in order to reflect the dispersion of the mean values, and the Student factor (usually corresponding to a 95% level of confidence) is applied to the calculated uncertainties. The true distribution of errors is usually far from normal, and all empirical equations are approximate by definition. To describe uncertainties more adequately, some restrictions are made through application of an effective number of data points based on the statistical weight used in fitting, i.e., the weights of all selected data points are summed and divided by the largest weight. Then, the effective number of data points is restricted to not greater than 25, and the Student coefficient is applied. This procedure is also used for uncertainty assessment of single-valued properties, such as triple-point T_{tp} and critical temperatures T_{c}. As noted earlier, critical pressures p_{c} and critical densities ρ_{c} are treated as part of the saturation curves and not as single-valued properties. Uncertainties for these are calculated from the covariance matrices for the saturation equations.

The covariance matrix is a function of the mathematical form of an equation, source data distribution, and the general deviation of the data from the equation. If the number of parameters in an equation is small, the calculated uncertainties are nearly uniform in the range of the independent variables and tend to be underestimated. If the number of parameters is large, the calculated uncertainties are excessively small in the middle and unrealistically high at the edges. To counteract this effect, TDE covariance matrices are generally reduced to rank 3 or 4. This is done after fitting by setting some of the parameters as constants with values obtained in the unrestricted fit (and covariance terms equal to zero) and recalculating the covariance for the remaining parameters. In polynomial-like equations, parameters corresponding to terms of higher power are fixed for the covariance adjustment.

Quality, Validity, and Success of Evaluation. At the end of the evaluation process additional checks are applied to reduce the probability of seriously erroneous results caused by errors in the source data or highly unusual data scenarios.

The quality of fit Q represents how well an equation fits the source data. It is calculated with the formula

$$Q = \sum_{i=1}^{N}\left\{w_i\cdot\left(\frac{\delta_i}{U_i}\right)^2\right\}/\sum_{i=1}^{N}w_i \qquad (9)$$

where N is the number of data points, w_i are their statistical weights, U_i are estimated source data uncertainties, and δ_i are deviations from the curve. If Q is significantly greater than 1, the data are not adequately fit by the equation, and the evaluation may be of no value. Substitution by predicted values is usually done by TDE in such cases.

A second parameter that is used to check the validity of an evaluation is the relative assessed uncertainty R

$$R = U/\mathrm{V} \qquad (10)$$

where V is the evaluated property value, and U is the evaluated uncertainty. If R is too high, depending on the property, it may indicate that the evaluation is not valid. Large R may be reasonable for evaluated vapor pressures, where an order-of-magnitude value might be useful but are less acceptable for densities or temperatures of phase transitions.

A third test for each property involving independent variables involves tests for validity of the curve shape. There are individual validity criteria for essentially every property evaluated by TDE. For example, first and second derivatives (with respect to temperature) of saturated vapor pressure curves must be positive, and the second derivative of the saturated liquid density curves must be negative (also, the first derivative for essentially all compounds except water). Some of the curve-shape criteria are strict, and if not satisfied, the equation is invalidated and substituted with predicted values. Other, less strict, criteria can be expressed numerically and are used for comparative assessment of different equations or parameter sets.

Through application of these three criteria, TDE gives evaluated results consistent with available source data, predictions, and physical principles. In the worst case, when the source data are internally inconsistent or do not pass validation, the uncertainty of the evaluated values is determined by that of the prediction methods used, but the evaluation basis is always clear. A key goal in development of TDE was creation of a system for critical evaluation of property data that could act autonomously. In the absence of significant data errors, decision making is relatively straightforward and involves selection of default equations (i.e., model selection), number of parameters, application of predictions for filling data gaps, and detection of low-quality data points for rejection. To ensure that the results of an evaluation would not be highly erroneous, the program was developed to not rely on perfect source data, and practically every major operation such as evaluation of a property or consistency enforcement includes assessments of success and validity checks for derived equations. If a failure is detected, a variety of remedial actions are tried, including reduction in the number of equation parameters, change of the selected default equation, additional source data validation, rejection of inconsistent data sets, rollback of consistency enforcement, substitution of property data by prediction, or, in the worst case, complete exclusion of a property from the evaluation results.

Program Structure and Interface. The program core is written in the C++ language and, therefore, is highly portable to different platforms. It is based on the C++ class concept. The program is made up of a user interface and computational core that includes the Compound, Prediction, Property, and Model classes. The central class is the Compound class, which holds all information about a particular compound and supports operations for loading compound source data from TDE-SOURCE, performing evaluations and accessing the primary (default equations and numerical values) and secondary (alternative equations)

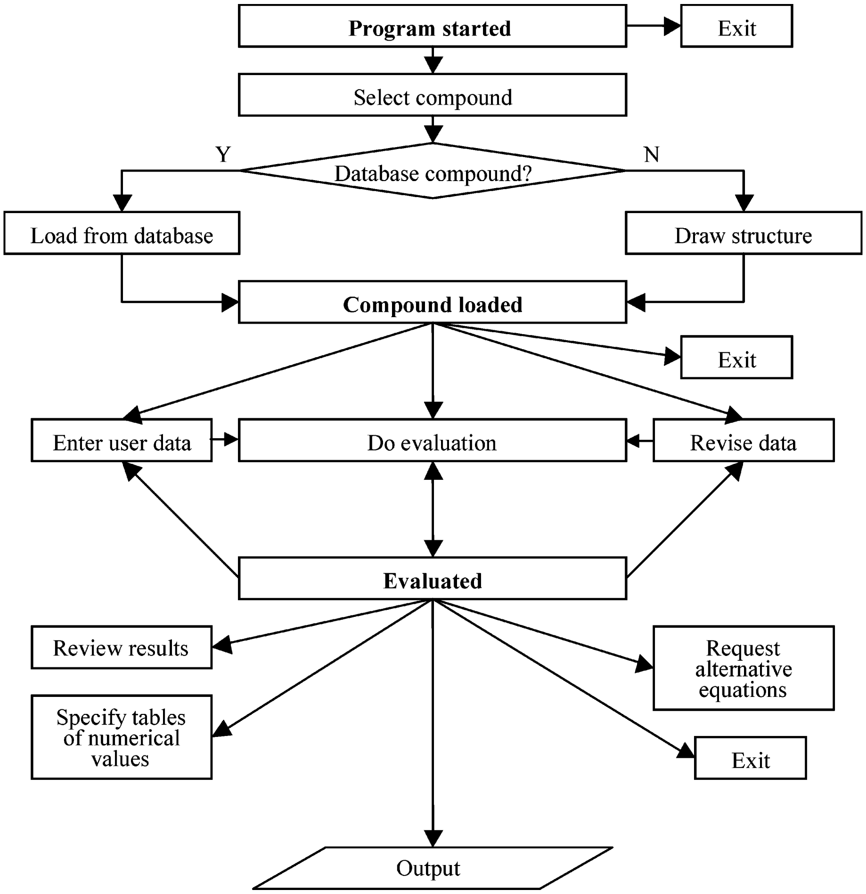


Figure 7. Operational model of the TDE program. Text in bold indicates the three steady states of the program, as described in the text.

evaluation results. The Prediction class contains all data and methods necessary for property predictions, including the molecular structure, methods, and method parameters. The Property class contains all data on a particular property and the fitting equations. The Model class provides full implementation of an equation, and each instance of the class stores parameters for a particular property. The modular structure of the program allows equations to be added and interchanged without changes to the evaluation procedures.

The interface model (Figure 7) is based on three steady states. The first appears when the program is started, and the only user option is to select a compound for evaluation. After selection, all experimental data for the selected compound are extracted from the TDE-SOURCE database or from a file saved previously. If structural information is not in TDE-SOURCE, the user is requested to draw the structure. After selection of a compound (second steady state), the user can revise source data, add proprietary (‘user’) data, and start evaluation. After evaluation, in the third state, the user can review evaluation results and request alternative equations. All the functions of the second state remain available, and the user can modify data and repeat the evaluation as desired.

Data Communication. NIST/TRC in cooperation with DIPPR (the Design Institute for Physical Properties of the American Institute of Chemical Engineers) and IUPAC (the International Union of Pure and Applied Chemistry) developed ThermoML, an XML (Extensible Markup Language)-based approach for storage and exchange of thermophysical and thermochemical property data.[4,17,30] XML technology[31] provides significant advantages for data exchange such as its ‘native’ interoperability based on ASCII code, modular structure, and transparent readability by both humans and computers. Both the software and hardware communities support this technology extensively. In 2002, IUPAC approved project 2002-055-3-024, XML-based IUPAC Standard for Experimental and Critically Evaluated Thermodynamic Property Data Storage and Capture,[32] and established a Task Group to conduct it as one of the activities of the Committee on Printed and Electronic Publications.[33] At its meeting in January 2004,[34] the Task Group accepted ThermoML as the framework of an emerging IUPAC standard and approved the establishment of the ThermoML namespace for it.[35]

ThermoML is fully implemented in TDE as the primary method of data communication. TDE accepts user data files in ThermoML format, and all evaluation results are available in the form of a ThermoML file. That file includes compound identification, numerical values of evaluated properties (with defined uncertainties), and default and alternative equations with their parameters and covariance matrices. ThermoML output can be automatically parsed and processed by applications such as process simulation engines. Required components are the ThermoML schema and definition files for equations, both of which are available at the NIST/TRC Web site (www.trc.nist.gov/ThermoML.html). Equation definition files contain the mathematical definitions for equations through importation of the MathML schema[36] and define symbolic representation of equation variables and parameters. The ThermoML schema can be used by validating parsers when reading ThermoML data files. Equation definition files provide Supporting Information about the equation that may assist development of readers for ThermoML equations. ThermoML supports all TDE equations explicitly, and the ThermoML definitions of all TDE equations are available at the NIST/TRC web site noted above.

5. ENFORCEMENT OF THERMODYNAMIC CONSISTENCY

Levels of Consistency Enforcement. Three different levels of consistency enforcement are employed within TDE: single-property enforcement (constraining a single fitting equation to be consistent with other properties), in-block enforcement (constraining several equations within one property block to be mutually consistent), and interblock enforcement involving properties from different blocks.

Single-property procedures are applied for saturated vapor pressures to ensure consistency with heat capacity differences between the liquid and gas phases at low pressures and to ensure that condensed-state phase boundary lines converge at triple points. In-block procedures are used for all vapor and sublimation pressures (phase diagram block) as well as for saturated liquid and gas densities (volumetric block) and single-phase and saturated gas densities (volumetric block). An interblock procedure refines the gas density and enthalpy of vaporization through improved consistency with vapor pressures and liquid densities. Consistency conditions are either introduced with appropriately high weights to the objective functions or are implied by common parameters. Details of each example are provided in the following paragraphs.

Vapor Pressure Constraint. Constraint of vapor-pressure extrapolations to low temperatures with liquid−gas heat capacity differences is a common technique.[37] TDE uses an analogous approach. Assuming the difference in compressibility factors Z for the gas and liquid $(Z_g - Z_l) = 1$ at low temperatures where $p_{sat} < 10$ kPa, the heat capacity difference ΔC_{sat} between the saturated liquid and gas can be derived from the temperature dependence of the vapor pressure

$$\Delta C_{sat} = C_{sat}(g) - C_{sat}(l) = \mathrm{R}\cdot(T^2\cdot d^2\ln(p_{sat})/dT^2 + 2\cdot T\cdot d\ln(p_{sat})/dT) \quad (11)$$

$C_{sat}(g)$ can be approximated by the heat capacity of the ideal gas at such low pressures. Available validated ideal-gas heat capacities[38] are stored in the TDE database. These evaluated values are used in preference to any experimental values stored in TDE-SOURCE. The heat capacity conditions represented in eq 11 with the weights based on heat capacity uncertainties are added to the objective function for vapor pressure for $p_{sat} < 10$ kPa with steps of 10 K.

Condensed-State Phase Boundary Convergence. Equations describing condensed-state phase boundaries (crystal−liquid and crystal−crystal) are constrained to converge at triple points. The polynomial form of the equations used for these boundaries allows simple fulfillment of this constraint through the representation equation

$$p - p_{tp} = \sum_{i=1}^{N} a_i \cdot (T - T_{tp})^i \quad (12)$$

where T_{tp} and p_{tp} are respectively the triple-point temperature and pressure and a_i are polynomial coefficients.

Vapor and Sublimation Pressure Convergence. When one or more sublimation curves are known, in addition to the vaporization curve, they must be consistent at the triple point(s). Two consistency conditions are enforced in TDE. First, adjacent vapor pressure and sublimation curves must converge to the same triple-point pressure p_{tp}. This is expressed by the following equation

$$p_{sat}(l) = p_{sat}(cr) \text{ at } T = T_{tp} \quad (13)$$

The second condition is that the difference between the enthalpies of sublimation and vaporization derived from the pressure equations at T_{tp} must yield the enthalpy of fusion. If $p_{tp} < 10$ kPa and ΔZ is assumed to be 1, this second condition can be expressed as

$$-\mathrm{R}\cdot T^2\cdot[d\ln\{p_{sat}(l)\}/dT - d\ln\{p_{sat}(cr)\}/dT] = \Delta_{fus}H \quad (14)$$

All of the fitted equations are accompanied by their relative weights. Those for experimental and vapor pressure data are the same as those used to fit the properties individually. The weights for the two enforcement conditions (eqs 13 and 14) are calculated with the following equation. For eq 13, the weight w_A is

$$w_A = \kappa_A\cdot(1/W_{vp} + 1/W_{sub})^{-1} = \kappa_A\cdot W \quad (15)$$

where W_{vp} is the sum of the weights for the vapor pressure data and W_{sub} is the sum of the weights for the sublimation pressure data. The formulation results in W being closely related to the lesser of W_{vp} and W_{sub}. For eq 14, the weight w_B is

$$w_B = \kappa_B\cdot(1/W_{vp} + 1/W_{sub})^{-1}/U_{\Delta fusH}^{\;2} = \kappa_B\cdot W/U_{\Delta fusH}^{\;2} \quad (16)$$

where $U_{\Delta fusH}$ is the expanded combined uncertainty for the enthalpy of fusion $\Delta_{fus}H$. This approach is needed because, generally, there are very different amounts of experimental data in each phase region. The constants κ_A and κ_B are determined empirically to optimize the fitting performance.

The results of enforcement of thermodynamic consistency at T_{tp} are illustrated in Figure 8. Applying these conditions can lead to an invalid shape of the vapor pressure and sublimation curves, if the source data are highly erroneous and inconsistent. In that case, it is necessary to decide which property is most likely erroneous and to derive the equation for it from the other properties. Generally, sublimation pressures are more prone to measurement error and, consequently, are rejected in such an evaluation.

ρ_{sat}(Liquid) and ρ_{sat}(Gas) Consistency at T_c. Consistency of saturated liquid and gas densities at the critical temperature T_c is satisfied through a common parameter in the equations (the critical density ρ_c) by solving a joint least-squares system of equations. Figure 9 shows a demonstration of this

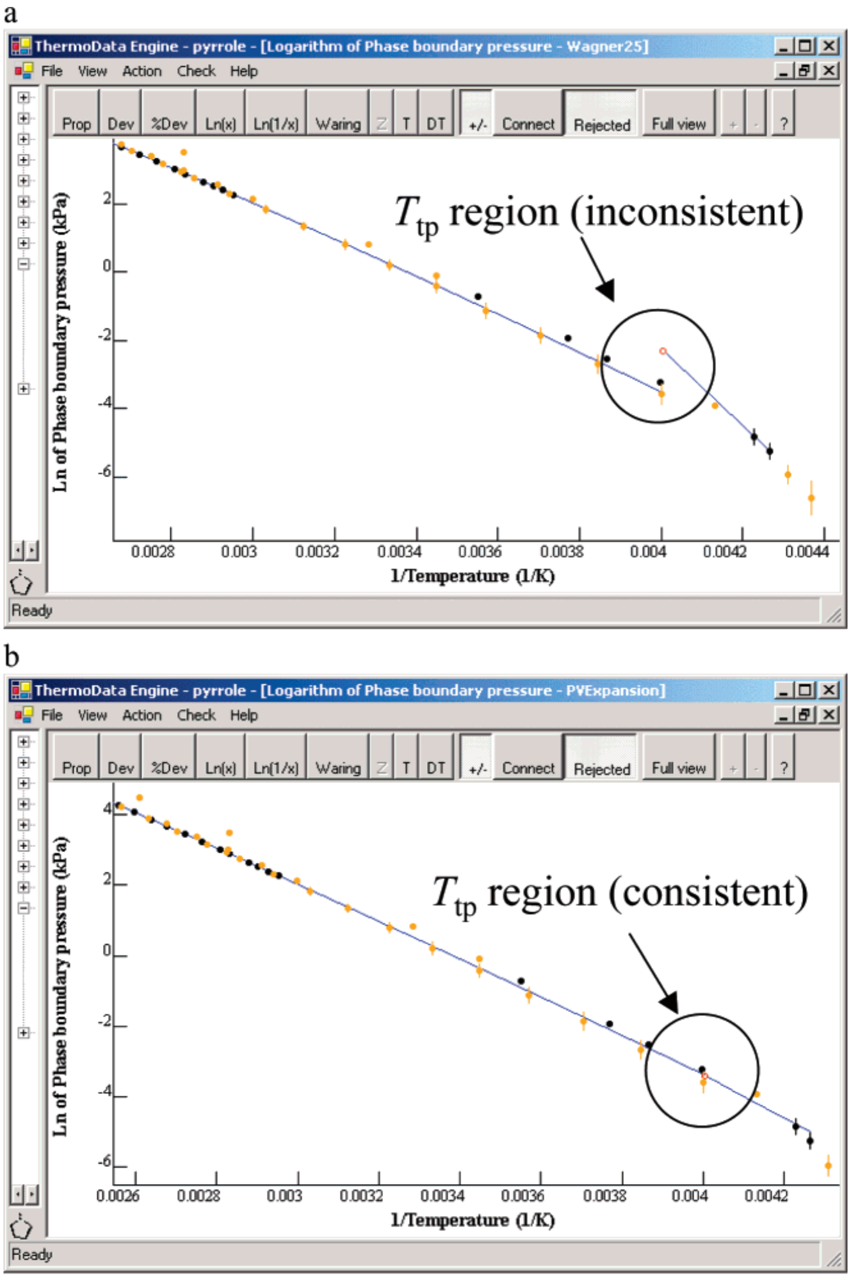

a
ThermoData Engine - pyrrole - [Logarithm of Phase boundary pressure - Wagner25]
File View Action Check Help
Prop Dev %Dev Ln(x) Ln(1/x) Waring Z T DT +/- Connect Rejected Full view + - ?
T_{tp} region (inconsistent)
Ln of Phase boundary pressure (kPa)
2
0
-2
-4
-6
0.0028
0.003
0.0032
0.0034
0.0036
0.0038
0.004
0.0042
0.0044
1/Temperature (1/K)
Ready
b
ThermoData Engine - pyrrole - [Logarithm of Phase boundary pressure - PVExpansion]
File View Action Check Help
Prop Dev %Dev Ln(x) Ln(1/x) Waring Z T DT +/- Connect Rejected Full view + - ?
T_{tp} region (consistent)
Ln of Phase boundary pressure (kPa)
4
2
0
-2
-4
-6
0.0026
0.0028
0.003
0.0032
0.0034
0.0036
0.0038
0.004
0.0042
1/Temperature (1/K)
Ready

Figure 8. (a) Vapor pressure and sublimation curve for pyrrole without enforcement of consistency at T_{tp}. (b) The same data scenario with automatic enforcement of thermodynamic consistency in terms of value and relative slope at T_{tp}. The relative slopes are directly related to the enthalpy of the phase transition ($\Delta_{fus}H$ for the melting transition).

enforcement. The mathematical forms of the equations ensure an infinite slope at T_c.

Gas-Phase Densities: ρ_{sat}(Gas) and the Virial Equation. The authors are unaware of an equation suitable for representation of low- and high-pressure gas densities with low experimental data coverage. To address this problem, a combination of the virial equation (as function of molar volume) at low pressures and the saturated-phase density equation at high pressures is used for the gas phase. The smooth connection point is below 500 kPa if the virial equation is limited to the second virial coefficient or above $0.85 \cdot T_c$ if the third virial coefficient is included. The third virial coefficient is used, if a prediction can be made or experimental data are available for ρ(gas) for $p > 500$ kPa. (*Note:* Densities for $T > 0.85 \cdot T_c$ in the single-phase gas are not represented in version 1.0 of TDE. Also, all single-phase densities for $T > T_c$ (i.e., the fluid region) are not represented. Representation of these regions is planned for a future release of the program.) The consistency conditions at the connection temperature T_Θ are as follows: (1) the saturated density and (2) its first derivative by temperature derived from the virial and vapor pressure equations must be equal to those derived from the saturated gas density equation:

$$\rho_{virial}(T_\Theta, p_{sat}(T_\Theta)) = \rho_{sat}(T_\Theta) \qquad (17)$$

$$d\rho_{virial}(T_\Theta, p_{sat}(T_\Theta))/dT = d\rho_{sat}(T_\Theta)/dT \qquad (18)$$

Weights w_C for the enforcement conditions (eqs 17 and 18) are the same and are

$$w_C = \kappa_C \cdot \{1/W_{\rho(gas)} + 1/W_{\rho sat(gas)}\}^{-1} = \kappa_C \cdot W \qquad (19)$$

where $W_{\rho(gas)}$ is the sum of the weights for the single-phase

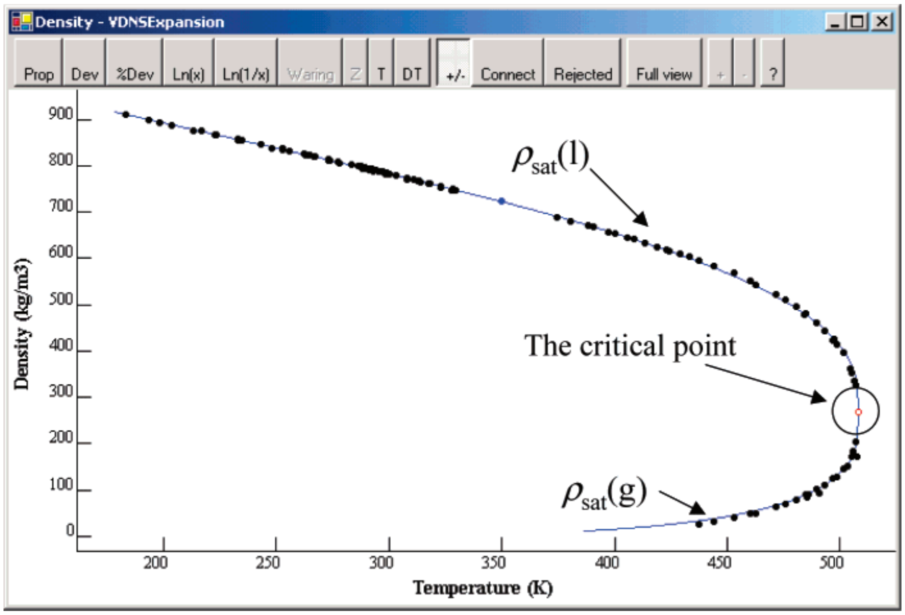


Figure 9. Demonstration of enforced consistency between the saturated density for the liquid ρ_{sat}(l) and gas ρ_{sat}(g) curves at the critical point.

gas density data and $W_{\rho sat(gas)}$ is the sum of the weights for the saturated gas density data. As for eq 15 above, the formulation results in W being closely related to the lesser of $W_{\rho(gas)}$ and $W_{\rho sat(gas)}$ with the constant κ_C determined empirically.

Consistency of ρ_{sat}(Gas) with Enthalpies of Vaporization Derived from Vapor Pressures p_{sat} and the Clapeyron Equation. A similar procedure is used for interblock consistency enforcement before the end of an evaluation. Revised equations are a virial equation for single-phase gas density at low pressures, a saturated gas density equation at high pressures, and an enthalpy of vaporization equation. Other involved properties are vapor pressure and liquid density. The objective function is the sum of squares of the deviations from the fitted equations (both the property representations and enforcement conditions) multiplied by appropriate relative weights. The objective sum is minimized by a nonlinear optimization technique. The virial equation is represented in the nonlinear form $\rho(\text{gas}) = f(T, p)$. This form is used to ensure that the equation gives a solution at the saturation pressure, even with little or no experimental data near the saturation line.

In addition to the source data and seamless connection conditions described in the previous section (eqs 17 and 18), consistency of the enthalpy of vaporization $\Delta_{vap}H$ derived from the p_{sat} curve together with the gas and liquid densities (converted to molar volumes V_m) through the Clapeyron equation

$$\Delta_{vap}H = T\cdot(\mathrm{d}p_{sat}/\mathrm{d}T)\cdot\{V_m(\mathrm{g}) - V_m(\mathrm{l})\} \qquad (20)$$

is enforced with $\Delta_{vap}H$ represented by the equation

$$\ln(\Delta_{vap}H) = a_1 + \sum_{i=2}^{nTerms} a_i\cdot T_r^{\;i-2}\cdot\ln(1\text{-}T_r) \qquad (21)$$

where a_i and $T_r = T/T_c$ are fitted parameters. Equation 21 ensures a valid shape for the $\Delta_{vap}H$ values derived with eq 20 even in the absence of experimental $\Delta_{vap}H$ values. This condition is constrained at separate temperatures with a step of 5 K. The weight w_D for the consistency equation (eq 20) is

$$w_D = \kappa_D\cdot\{1/W_{\rho(gas)} + 1/W_{\rho sat(gas)}\}^{-1}/U_{\Delta vapH}^{\;\;2} = \kappa_D\cdot W/U_{\Delta vapH}^{\;\;2} \qquad (22)$$

where $U_{\Delta vapH}$ is the uncertainty calculated with the covariance matrix for eq 21. The effect of this interblock consistency enforcement is shown in Figure 10.

6. APPLICATION OF PREDICTION METHODS

Background. The aim in development of predictions in TDE was not to invent new methods but to implement an algorithm for applying existing methods to create an approach that is broadly applicable across all classes of organic compounds. The most important aspect of this approach is that all methods be validated against critically evaluated experimental data. This provides the basis for estimations of uncertainty for all predicted values. The requirement of uncertainty estimation is essential for integrating predicted data into the TDE evaluation process. As described earlier, all data used by TDE have associated estimates of their expanded combined uncertainties with approximate levels of confidence of 95%. The initial release of the TDE software (version 1.0) includes methods based on the principles of corresponding states (CS) and group contribution (GC). Additional, more computationally intensive prediction schemes, such as molecular dynamics or ab initio quantum methods, will be considered for future releases, particularly as reliable bases for estimates of combined uncertainties for these methods are developed. Though applied specifically to CS and GC methods, the general approach described here can be applied universally.

The general approach for validation of predictive methods and estimation of their combined uncertainty is described here. For this, it is necessary to give quantitative definitions to two terms used in molecule comparisons that are qualita-

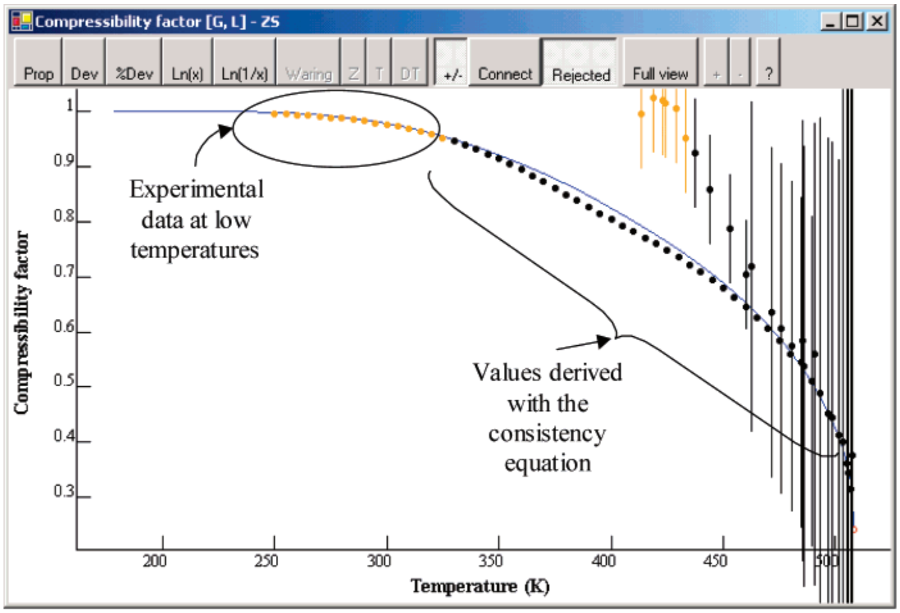


Figure 10. Action of interblock consistency enforcement on the saturated gas density ρ_{sat}(gas) in an unfavorable data situation. Data are shown in terms of the compressibility factor Z. Values with large error bars at high temperature are low-quality experimental values. The curve represents the evaluated values for ρ_{sat}(gas) expressed in terms of Z_{sat}(gas).

tive in nature: similarity and complexity. The mathematical definitions of these are not universal and can vary from property to property or method to method, depending on the sensitivity of the property or method to the presence of particular substructures or functional groups.

General Prediction Approach. The key issue in developing a property prediction system is how to recognize strengths and shortcomings for a particular method so that the best method can be selected for a given molecule and property. Generally, there are two serious problems with predictive methods. First, the predictive ability is often quite limited, due to an inadequate experimental data set against which the parameters of the method were optimized. Parameters optimized with a small data set tend to have a bias toward the structural features represented in that set. This bias can cause serious errors when the parameters are used to predict properties of compounds with features beyond those in the small set. Second, during development, there is often little performance evaluation to investigate the predictive capabilities of the method. In part, this problem can be traced to a lack of experimental data. Only predictions for compounds not included in the original development can provide a valid test for that method, which in turn requires the availability of new data on previously unmeasured compounds. Consequently, knowledge of the predictive ability of a method is often poor.

The approach used within TDE was developed to address these problems and to provide reliable property predictions by making the best use of available experimental data and correlations. The design principles for this approach follow.

- Establish a physicochemical property database that contains critically evaluated experimental data with reliable estimates of uncertainty.
- Create a predictive method performance database by evaluating methods against the critically evaluated data.
- Examine the validity (against established data correlations) of predicted property values for a large number and variety of compounds for which no experimental data are available and develop rejection criteria based on these results.
- Implement an automated, case-based method selection scheme with the ability to select the method with the best performance for compounds most similar to the query compound.
- Estimate uncertainties for the predicted property values based on (a) the performance of the method for similar compounds and (b) the complexity of the query compound. (Mathematical definitions for similarity and complexity are given later in this section.)

The overall algorithm implementing these design criteria is shown in Figure 11. The collection of critically evaluated properties used with the prediction scheme for TDE contains about 880 normal boiling temperatures T_b, 550 critical temperatures T_c, 450 critical pressures p_c, 350 critical volumes V_c, 680 liquid density values at T_b, and 270 acentric factors. This evaluated data set was created through single-property and multiproperty consistency checks and provides a collection that exceeds that used in the production of any specific prediction method, thus allowing for a performance analysis based on data outside of the original set used in parameter optimization.

TDE (version 1.0) uses two types of predictive methods: group contribution (GC),[29,39] corresponding states (CS),[29] and their combination (GC-CS). CS methods require knowledge of the critical properties, which must be estimated by a GC method, when no experimental critical properties are available, as is the case for most compounds. The only input requirement for a GC method is structural information, which can be expressed as an atom connectivity table. Based on the connectivity table, the TDE software extracts groups defined for each GC method. Currently, TDE includes atom connectivity tables and images for 14 000 organic compounds. A simple structure drawing interface is provided for input of structures not included.

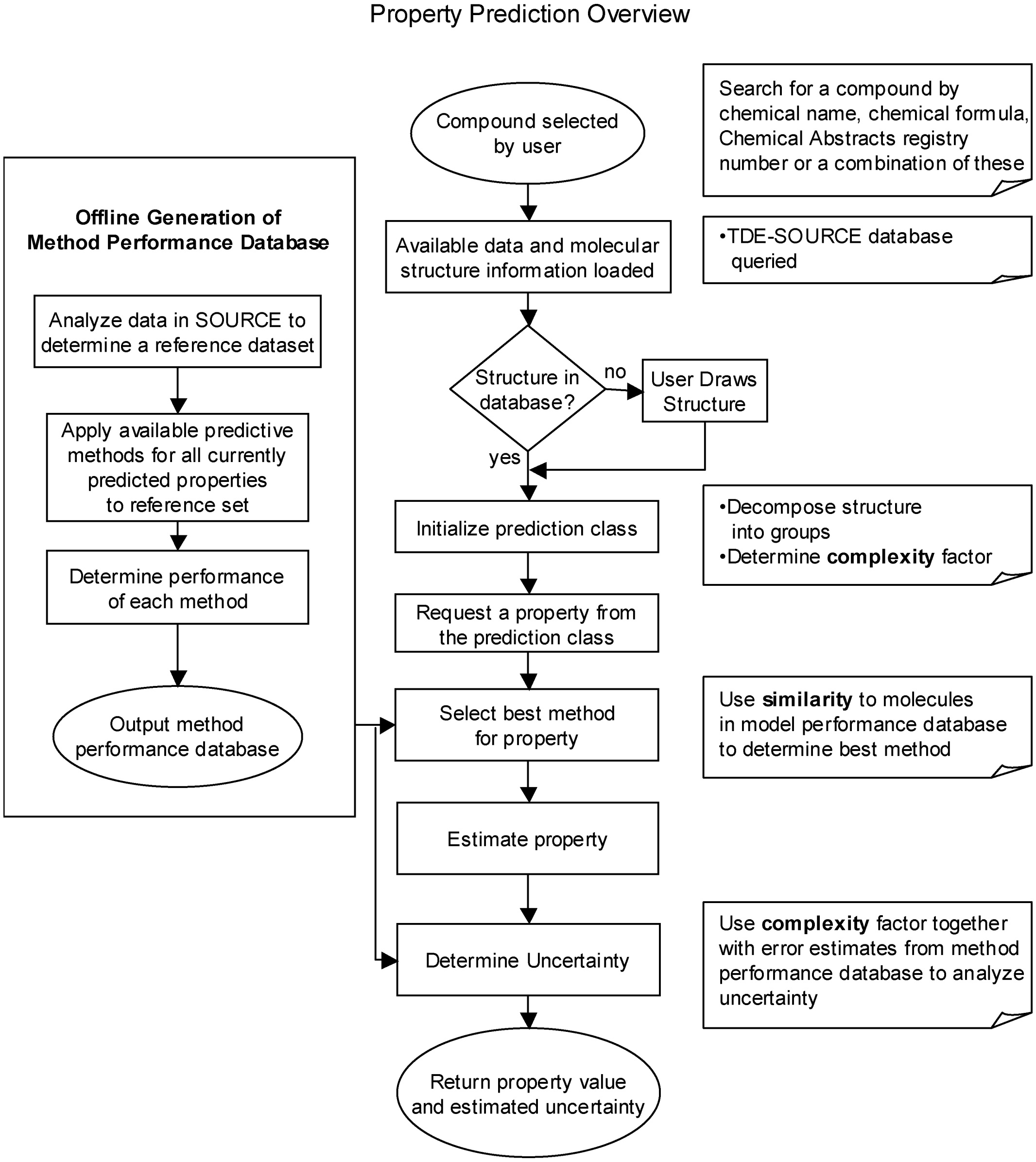

Property Prediction Overview
Offline Generation of Method Performance Database
Analyze data in SOURCE to determine a reference dataset
Apply available predictive methods for all currently predicted properties to reference set
Determine performance of each method
Output method performance database
Compound selected by user
Search for a compound by chemical name, chemical formula, Chemical Abstracts registry number or a combination of these
Available data and molecular structure information loaded
•TDE-SOURCE database queried
Structure in database?
no
User Draws Structure
yes
Initialize prediction class
•Decompose structure into groups
•Determine complexity factor
Request a property from the prediction class
Select best method for property
Use similarity to molecules in model performance database to determine best method
Estimate property
Determine Uncertainty
Use complexity factor together with error estimates from method performance database to analyze uncertainty
Return property value and estimated uncertainty

Figure 11. Outline of the property prediction approach.

Table 4 shows the complete list of properties for which predictive methods are available in TDE 1.0. Expansions to include additional properties and those of mixtures and reactions are planned. The workflow for the predictive approach in TDE can be summarized as follows. For any particular compound, TDE searches the TDE-SOURCE database for relevant experimental data, does preliminary data processing, and provides the predictive method with all pertinent experimental data values and atom-connectivity information. When the structural information for a compound is not available, the user is allowed to draw the structure, and the connectivity table is produced by TDE. Group parameters are then extracted and used to compute complexity and similarity information needed for method selection and uncertainty estimation.

Except for a few properties, such as normal boiling temperature T_b, estimation requires other property values as input. For example, most estimation methods for critical temperature require T_b, while for the method of Sastri and Rao[51] (for liquid viscosity), T_b, T_c, p_c, and the acentric factor are needed. To address this issue, the estimation of properties must be carried out in a carefully constructed sequence, which is shown in Figure 12.

Prediction Method Selection. The case-based method selection algorithm used in TDE is based on performance information compiled for all methods for compounds with well-established properties that are similar to the query compound. For properties having multiple prediction methods (T_{bp}, T_c, p_c, and V_c in TDE 1.0) automated selection is based on the best performance for similar molecules (BPSM). This approach was chosen after analysis of various others, including averaging of results from all available methods, weighted average of all methods based on performance against similar molecules, and an average of the two or three best methods. To clarify and quantify how the BPSM approach works, an example for prediction of critical temperature T_c is given here.

In this example, four group contribution methods for T_c prediction (listed in Table 4) are tested against 399 organic compounds for which critically evaluated T_c values are available with small uncertainties. Table 5 shows the average absolute deviations and standard deviations of predicted T_c

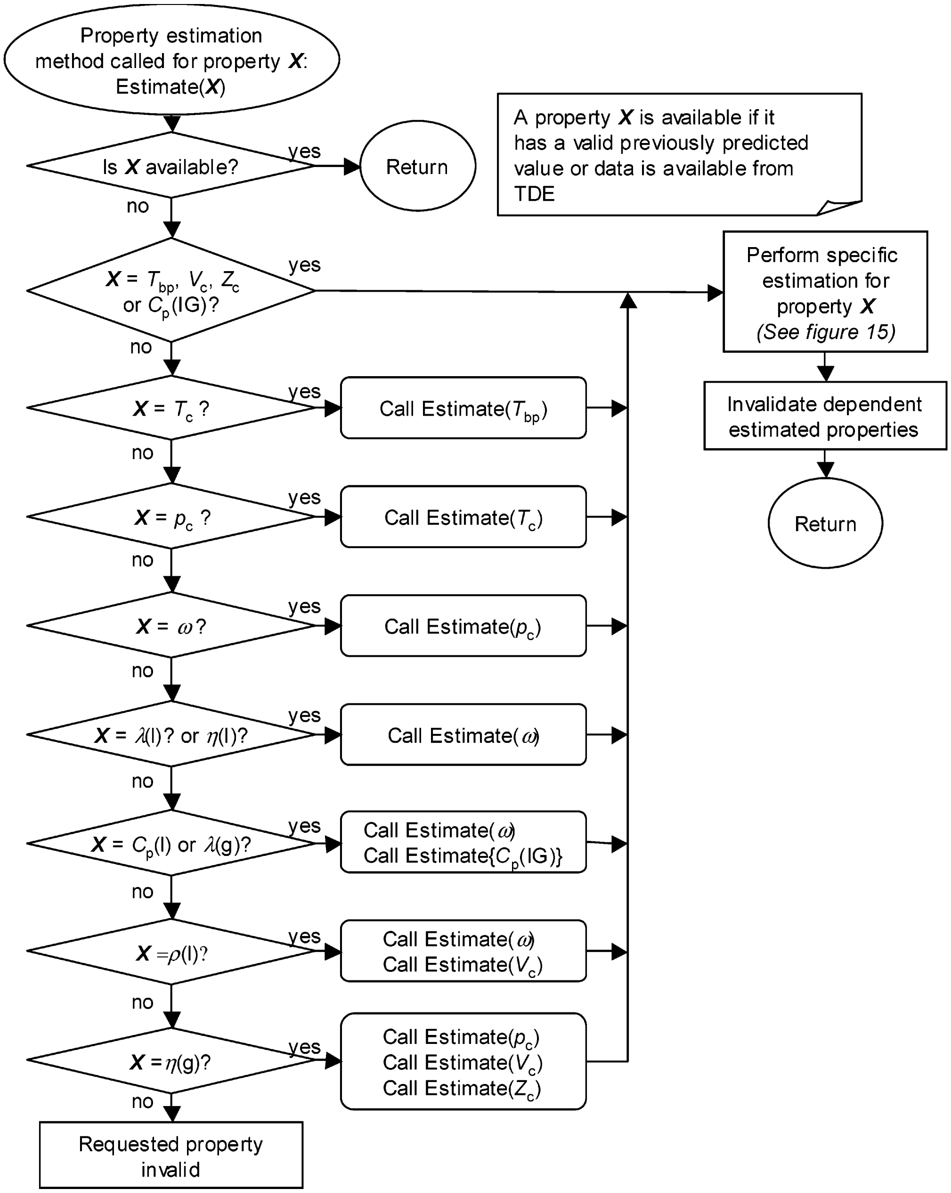


Figure 12. Workflow for property prediction. Properties predicted are the normal boiling temperature, T_{bp}; critical temperature, T_c; critical pressure, p_c; critical volume, V_c; critical compressibility, Z_c; acentric factor, ω; liquid density, ρ(l); ideal gas heat capacity, C_p(IG); liquid heat capacity, C_p(l); gas viscosity, η(g); liquid viscosity, η(l); gas thermal conductivity, λ(g); and liquid thermal conductivity, λ(l).

Table 5. Average Absolute Deviation $\bar{\Delta}$ and Standard Deviation σ of Calculated T_c from Critically Evaluated T_c for 399 Organic Compounds by Various Prediction Methods and the Best Performance on Similar Molecules (BPSM) Approach

method	$\bar{\Delta}$/K	σ/K	$N_\Delta > 10$ K[a]
Joback and Reid[40]	7.1	11.3	90
Wilson and Jasperson[43]	5.8	7.9	75
Constantinou and Gani[41]	15.4	29.2	132
Marrero and Pardillo[42]	5.2	9.1	59
BPSM	3.5	5.5	26

[a] The final column represents the number of compounds N_Δ for which the deviation of the predicted value from the critically evaluated value is greater than 10 K.

Table 6. Correctness of Model Selection Using the BPSM (Best Performance on Similar Molecules) Approach for Prediction of T_c

correctness[a]	compounds	percent
A	203	51
B	113	28
C	64	16
D	19	5
total	399	100

[a] Correctness is defined as follows: A = the best method was selected, B = the second best method was selected, C = the third best method was selected, D = the poorest method was selected.

for the methods and the BPSM technique relative to the critically evaluated T_c values for the test compounds. Table 5 also lists the number of compounds with deviations greater than 10 K between predicted and critically evaluated T_c values. The results indicate that the BPSM approach produces far fewer large deviations than any of the individual methods. Collectively, the results shown in Table 5 demonstrate that the BPSM approach is superior to any one of the four methods alone.

Another measure of the success of the BPSM approach is shown in Table 6, which lists the "correctness" of method selection for compounds considered in the test. For the 399 compounds in the test, the method providing the T_c prediction

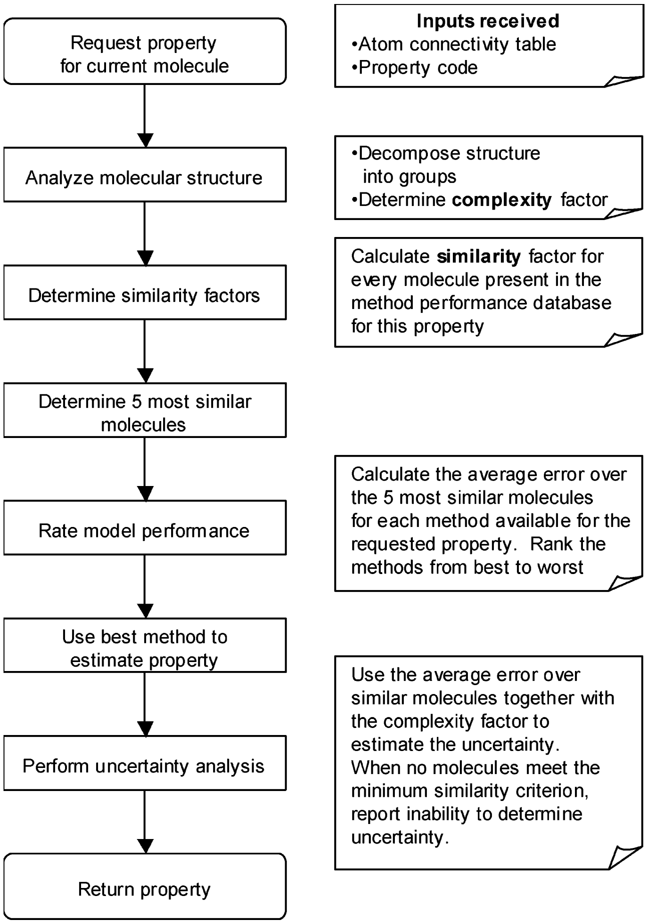

Request property for current molecule
Inputs received
•Atom connectivity table
•Property code
Analyze molecular structure
•Decompose structure into groups
•Determine **complexity** factor
Determine similarity factors
Calculate **similarity** factor for every molecule present in the method performance database for this property
Determine 5 most similar molecules
Rate model performance
Calculate the average error over the 5 most similar molecules for each method available for the requested property. Rank the methods from best to worst
Use best method to estimate property
Perform uncertainty analysis
Use the average error over similar molecules together with the complexity factor to estimate the uncertainty. When no molecules meet the minimum similarity criterion, report inability to determine uncertainty.
Return property

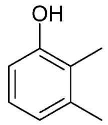
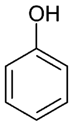
Figure 13. Workflow for prediction method selection based on performance against critically evaluated property data.

closest to the critically evaluated value was chosen for 203 compounds (51%). For 113 compounds the method that produced the second smallest deviation was selected, for 64 compounds the method producing the third smallest deviation was selected, and the worst available method was selected for 19 compounds. Of these 19 compounds, only seven have deviations from the critically evaluated values greater than 10 K. This performance is far better than any single method and is the best of any of the composite methods tested.

Figure 13 shows the general algorithm of the prediction method selection scheme. For each query compound, TDE analyzes its structural components, gathers information from the related method performance table, finds the five compounds that are most similar to the query compound (as defined in the next section), and uses the method that produces the smallest errors for the five similar compounds to predict the property for the query compound. Currently, method performance information has been established for normal boiling temperature T_b, critical temperature T_c, critical pressure p_c, and critical volume V_c.

Similarity of Organic Compounds. The concept of compound similarity plays a key role in TDE both in selection and error estimation for prediction methods. There is no unique definition for similarity because the attributes that make molecules similar for one application will not necessarily be the same for another. For example, similar compounds in Quantitative Structure–Property Relationships (QSPR) can be very different from those in Quantitative Structure–Activity Relationships (QSAR), because of differing sensitivities to stereostructural features, substructure size, and functional group types.

Table 7. Identification of Structural Groups Used for Compound Similarity Analysis[a]

		
rCH=	3	5
rC=	3	1
-CH$_3$	2	0
-OH	1	1

[a] *r* indicates that the group is part of a ring structure.

Within TDE, molecule similarity is used for selecting suitable prediction methods for each query compound based on results for similar molecules. Considering that the number and type of functional groups in a compound play a major role in determining its thermodynamic properties, the similarity definition used in TDE is based on differences in functional group numbers and types. We define similarity $\mathscr{S}$ as

$$\mathscr{S} = 1/(1 + \mathscr{D}) \tag{23}$$

where $\mathscr{D}$ represents differences between two compounds (A and B) and is expressed as

$$\mathscr{D}(\mathrm{A,B}) = \sum w_i \cdot \{N^{\mathrm{A}} - N^{\mathrm{B}}\}_i \tag{24}$$

In this equation, w_i is the weighting factor for the ith type structural group, N^A and N^B are the number of groups of type i in each compound, and the summation is over all structural groups present in the two molecules. The weighting factors are necessary because, for example, highly polar groups, such as −OH, −COOH, and −NH$_2$, can have a larger impact for certain properties (such as T_b) than nonpolar groups.

The two molecules shown in Table 7 are used here in a sample calculation of the quantitative measure of similarity. The difference between the molecules is expressed as

$$\mathscr{D}(\mathrm{A,B}) = 2w(\mathrm{rCH}{=}) + 2w(\mathrm{rC}{=}) + 2w(-\mathrm{CH}_3) \tag{25}$$

Various weighting factors w for each functional group are defined and used in TDE. The factors are assigned according to the influence of each functional group on the particular property being predicted.

Complexity. The complexity of a compound is an important factor that can have a substantial effect on the reliability of property estimations. Experience shows that predictions for complex compounds, such as molecules with multiple polar groups or fused rings, generally have larger deviations from reliable experimental values than those for simpler molecules such as short-chain alkanes or compounds with single functional groups. Therefore, it is necessary to define molecular complexity mathematically so that uncertainties of predicted properties can be estimated with confidence.

Like similarity, it is not possible to define molecular complexity in a generalized way for all compounds, because the complexity of a compound is closely related to the

Table 8. Rules for Calculating Complexity $\mathcal{C}$ [a]

condition	effect on complexity (C)
beginning complexity for all molecules	$C = 1$
for each >CH− group	+1
for each >C< group	+2
if molecule has groups from class 1	$+ 2 \cdot (N_1 - 1)$[b]
if molecule has =C= groups	$+ 2 + 4 \cdot (N_d - 1)$[c]
if molecule has groups from class 2	$+ 3 + 5 \cdot (N_2 - 1)$[b]
if molecule has groups from class 3	$+ 4 + 10 \cdot (N_3 - 1)$[b]
if molecule has exactly one carbon group from class 4	+ 5
if molecule has exactly two carbon groups from class 4	+ 3
if molecule contains a single ring with 3, 4, or 5 atoms	+ 30
if molecule contains a single ring with 6 or 7 atoms	+ 10
if molecule contains a single ring with 8 or 9 atoms	+ 20
if molecule contains a single ring with more than 9 atoms	+ 30
if molecule contains two rings	+ 30
if molecule contains more than two rings	+ 50

[a] Complexity is determined based on the following classes of groups: Class 1: carbons with double or triple bonds not connected to O or N except =C=. Class 2: F, Cl, Br, I, −N<, −CN, −SH, −S−. Class 3: −OH, −O−, >CO, −CHO, −(C=O)OH, −(O=C)O−, $-NH_2$, >NH, −N=, NH=, $-NO_2$, >SO, $>SO_2$. Class 4 − class 1 plus $-CH_3$, $>CH_2$, >CH−, >C<. [b] N_1, N_2, and N_3 are the number of groups from class 1, 2, or 3, respectively. [c] N_d is the number of =C= groups.

property being considered. Therefore, the definition of complexity is problem specific. We have defined a set of rules (Table 8) for calculating complexity for a compound. For any compound, its complexity can be obtained by combining the group complexity values in Table 8. The complexity of a compound is related to the group types and numbers of groups in a compound. This set of rules was compiled based on analysis of results for normal boiling temperatures T_b and critical properties and is expected to be suitable for most other pure compound properties. The rules connect the features of a compound as described by its structural components (i.e., groups) with the reliability of related group contribution methods in the prediction of T_b and critical properties. For example, the *n*-alkanes of 30 or 40 carbon atoms may not be generally considered as complex compounds, but within TDE, their calculated complexity is relatively high because the predictive methods cannot predict their critical properties with high reliability. The concept of compound complexity is used in TDE to help provide the most reliable uncertainties estimates possible. Our understanding of the relationships between molecular structure and a method's reliability is limited, and modifications to the definition of complexity in Table 8 may be made as analyses for additional property types and prediction methods are completed.

Uncertainty Estimation for Predicted Values. Estimates of the combined expanded uncertainty U for predicted properties in TDE are a function of the average deviations from the critically evaluated data set for similar molecules $\bar{\Delta}$ and the complexity $\mathcal{C}$ of the query compound, based on the rules listed in Table 8. Different specific formulas may be employed for different properties. For T_b and T_c, the following formula is adopted:

$$U = c_1 \cdot \bar{\Delta} + c_2 \cdot \mathcal{C}^{0.5} \qquad (26)$$

Here, c_1 and c_2 are coefficients that are obtained by minimizing the difference between estimated uncertainties and prediction deviations from critically evaluated values. For example, the coefficients in eq 26 for T_c were determined to be $c_1 = 1.2$ and $c_2 = 2.2$. The estimated uncertainties for predicted T_c by the BPSM method for the 399 organic compounds in the critically evaluated data set were calculated and were found to be greater than or equal to the observed difference in 95% of the cases with an average value of 9.8 K, which is not excessively large.

In summary, an approach for property prediction has been implemented in TDE that relies on a method performance database and quantitative measures of the similarity and complexity of compounds. With this information a robust prediction scheme is provided with realistic estimates of expanded combined uncertainties for all values.

7. CONCLUSIONS AND FUTURE DEVELOPMENT

(1) NIST ThermoData Engine (TDE)[1] complies with all the requirements necessary for implementation of the dynamic data evaluation concept. The scope of the first version of TDE is limited to pure compounds.

(2) Predictive capabilities of the first TDE version consist of 28 predictive methods based on group contribution and corresponding states principles and limited to the compounds whose molecules consist of carbon, hydrogen, oxygen, nitrogen, sulfur, fluorine, chlorine, bromine, iodine, and, in some instances, silicon.

(3) Further development will include incorporation of computational tools for generating equations of state on-demand depending on the data ‘scenario’ as well as implementation of daily updates of the TDE-SOURCE local data storage facility for TDE using a Web multitier dissemination architecture. Longer-term plans include expansion of TDE to include critical data evaluation for binary mixtures and incorporation of predictive methods beyond group contribution and corresponding states techniques.

ACKNOWLEDGMENT

The authors express their appreciation to Drs. S. Watanasiri (AspenTech, Cambridge, MA, U.S.A.), M. Satyro (Virtual Materials Group, Calgary, Canada), and A. I. Johns (National Engineering Laboratory, Glasgow, U.K.), whose advice was very valuable during development of TDE. Special thanks are given to colleagues from DIPPR: Drs. T. L. Teague (ePlantData), D. L. Embry (ConocoPhillips), A. K. Dewan (Shell), and IUPAC Task Group members Dr. J. H. Dymond (University of Glasgow, U.K.), Prof. E. Königsberger (Murdoch University, Australia), Prof. K. N. Marsh (University of Canterbury, New Zealand), Dr. S. E. Stein (NIST/Gaithersburg, MD, U.S.A.), and Prof. W. A. Wakeham (University of Southampton, U.K.), who have been working closely with us in the development of the emerging IUPAC ThermoML standard. In addition, the authors thank Drs. D. G. Friend, E. W. Lemmon, J. W. Magee, R. A. Perkins (NIST/Boulder, CO, U.S.A.), G. J. Rosasco, P. Linstrom, V. K. Shen (NIST/Gaithersburg, MD, U.S.A.), and especially M. L. Huber and A. H. Harvey (NIST/Boulder, CO, U.S.A.) for their valuable advice and practical suggestions during testing of TDE. The authors wish to acknowledge the late Dr. Randolph Wilhoit of Texas A & M University who was an inspiration for the implementation of the dynamic data evaluation concept.

Supporting Information Available: Complete list of equations supported by TDE. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Frenkel, M.; Chirico, R. D.; Diky, V. V.; Yan, X.; Dong, Q.; Muzny, C. NIST ThermoData Engine, NIST Standard Reference Database 103, National Institute of Standards and Technology, Gaithersburg, MD, 2004. http://www.nist.gov/srd/nist103.htm

(2) (a) Wilhoit, R. C.; Marsh, K. N. Future Directions for Data Compilation. *Int. J. Thermophys.* **1999**, *20*, 247−255. (b) Frenkel, M. Dynamic Compilation: A Key Concept for Future Thermophysical Data Evaluation. In *Forum 2000: Fluid Properties for New Technologies − Connecting Virtual Design with Physical Reality*; Rainwater, J. C., Friend, D. G., Hanley, H. J. M., Harvey, A. H., Holcomb, C. D., Laesecke, A., Magee, J. W., Muzny, C., Eds.; NIST Special Publication 975, Gaithersburg, MD, 2001; pp 83−84.

(3) Frenkel, M. Global Communications and Expert Systems in Thermodynamics: Connecting Property Measurement and Chemical Process Design. *Pure Appl. Chem.* **2005**, *77*, in press.

(4) Frenkel, M.; Chirico, R. D.; Diky, V. V.; Marsh, K. N.; Dymond, J. H.; Wakeham, W. A. ThermoML − An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 3. Critically Evaluated Data, Predicted Data, and Equation Representation. *J. Chem. Eng. Data* **2004**, *49*, 381−393.

(5) Frenkel, M.; Dong, Q.; Wilhoit, R. C.; Hall, K. R. TRC SOURCE Database: A Unique Tool for Automatic Production of Data Compilations. *Int. J. Thermophys*. **2001**, *22*, 215−226.

(6) Yan, X.; Dong, Q.; Frenkel, M.; Hall, K. R. Windows-Based Applications of TRC Databases: Structure and Internet Distribution. *Int. J. Thermophysics* **2001,** *22*, 227−241.

(7) (a) Whiting, W. B. Effects of Uncertainties in Thermodynamic Data and Models on Process Calculations. *J. Chem. Eng. Data* **1996**, *41*, 935−941. (b) Vasquez, V. R.; Whiting, W. B. Uncertainty and Sensitivity Analysis of Thermodynamic Models Using Equal Probability Sampling (EPS). *Comput. Chem. Eng.* **2000**, *23*(11/12), 1825−1838.

(8) Diky, V. V.; Chirico, R. D.; Wilhoit, R. C.; Dong, Q.; Frenkel, M. Windows-Based Guided Data Capture Software for Mass-Scale Thermophysical and Thermochemical Property Data Collection. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 15−24.

(9) TRC Thermodynamic Tables − Hydrocarbons, Thermodynamics Research Center: National Institute of Standards and Technology, Boulder, CO, 1942−2002. TRC−Thermodynamic Tables − Non-Hydrocarbons, Thermodynamics Research Center: National Institute of Standards and Technology, Boulder, CO, 1942−2004.

(10) Marsh, K. N. New Process for Data Submission and Dissemination. *J. Chem. Eng. Data* **2004**, *48*, 1.

(11) Electronic data submission to NIST Thermodynamics Research Center. *J. Chem. Thermodyn.* **2004**, *36(1)*, iv.

(12) Electronic data submission to NIST Thermodynamics Research Center. *Fluid Phase Equilib.* **2004**, *226*, v.

(13) Electronic data submission to NIST Thermodynamics Research Center. *Thermochim. Acta* **2004**, *421*, 241.

(14) http://www.trc.nist.gov/ThermoML.html.

(15) Dong, Q.; Yan, X.; Wilhoit, R. C.; Hong, X.; Chirico, R. D.; Diky, V. V.; Frenkel, M. Data Quality Assurance for Thermophysical Property Databases: Applications to the TRC SOURCE Data System. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 473−480.

(16) *Guide to the Expression of Uncertainty in Measurement* (International Organization for Standardization, Geneva, Switzerland, 1993). This *Guide* was prepared by ISO Technical Advisory Group 4 (TAG 4), Working Group 3 (WG 3). ISO/TAG 4 has as its sponsors the BIPM, IEC, IFCC (International Federation of Clinical Chemistry), ISO, IUPAC (International Union of Pure and Applied Chemistry), IUPAP (International Union of Pure and Applied Physics), and OIML. Although the individual members of WG 3 were nominated by the BIPM, IEC, ISO, or OIML, the *Guide* is published by ISO in the name of all seven organizations.

(17) Chirico, R. D.; Frenkel, M.; Diky, V. V.; Marsh, K. N.; Wilhoit, R. C. 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* **2003**, *48*, 1344−1359.

(18) Taylor, B. N.; Kuyatt, C. E. *Guidelines for the Evaluation and Expression of Uncertainty in NIST Measurement Results*, NIST Technical Note 1297; NIST: Gaithersburg, MD, 1994. http://physics.nist.gov/cuu/Uncertainty/bibliography.html.

(19) Dong, Q.; Chirico, R. D.; Yan, X.; Hong, X.; Frenkel, M. Uncertainty Reporting for Experimental Thermodynamic Properties. *J. Chem. Eng. Data* **2005**, in press.

(20) (a) Span, R.; Wagner, W. Equations of state for technical applications. I. Simultaneously optimized functional forms for nonpolar and polar fluids. *Int. J. Thermophys.* **2003**, *24*, 1−39. (b) Span, R.; Wagner, W. Equations of state for technical applications. II. Results for nonpolar fluids. *Int. J. Thermophys.* **2003**, *24*, 41−109. (c) Span, R.; Wagner, W. Equations of state for technical applications. III. Results for polar fluids. *Int. J. Thermophys.* **2003**, *24*, 111−162.

(21) Wagner, W. New Vapor Pressure Measurements for Argon and Nitrogen and a New Method of Establishing Rational Vapor Pressure Equations. *Cryogenics* **1973**, *13*, 470−482.

(22) Press: W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C*; Cambridge University Press: New York, 1998.

(23) Wilhoit, R. C.; Marsh, K. N.; Hong, X.; Gadalla, N.; Frenkel, M. *Thermodynamic Properties of Organic Compounds and their Mixtures. Volume 8B. Densities of Aliphatic Hydrocarbons: Alkanes*; Springer: Berlin, 1996.

(24) Yaws, C. L. *Chemical properties handbook: physical, thermodynamic, environmental, transport, safety, and health related properties for organic and inorganic chemicals;* McGraw-Hill: New York, 1999.

(25) *DIPPR 801 Policies and Procedures Manual.* BYU-DIPPR Thermophysical Properties Laboratory. Brigham Young University: Provo, 2003. http://www.ppds.co.uk/library/pdf/PPDS_EquationForms.pdf

(26) *PPDS2 Temperature-Dependent Equation Forms.* National Engineering Laboratory, Glasgow U.K., 1998. http://www.ppds.co.uk/library/pdf/PPDS_EquationForms.pdf.

(27) http://www.trc.nist.gov/GDC.html.

(28) (a) Ruzicka, V., Jr.; Domalski, E. S. Estimation of heat capacities of organic liquids as a function of temperature using group additivity. I. Hydrocarbon compounds. *J. Phys. Chem. Ref. Data* **1993**, *22*, 597−619. (b) Zabransky, M.; Ruzicka, V., Jr. Estimation of heat capacities of organic liquids as a function of temperature using group additivity. An amendment. *J. Phys. Chem. Ref. Data* **2004**, *33*, 1071−1081.

(29) Poling, B. E.; Prausnitz, J. M.; O'Connell, J. P. *The Properties of Gases and Liquids*, 5th ed.; McGraw-Hill: New York, 2001.

(30) Frenkel, M.; Chirico, R. D.; Diky, V. V.; Dong, Q.; Frenkel, S.; Franchois, P. R.; Embry, D. L.; Teague, T. L.; Marsh, K. N.; Wilhoit, R. C. ThermoML − An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 1. Experimental Data. *J. Chem. Eng. Data* **2003**, *48*, 2−13.

(31) (a) Finkelstein, C.; Aiken, P. *Building Corporate Portals with XML*; McGraw-Hill: New York, 1999. (b) http://www.w3.org/XML/.

(32) http://www.iupac.org/projects/2002/2002-055-3-024.html.

(33) http://www.iupac.org/standing/cpep.html.

(34) *Chem. Int*. *26*(4), 2004.

(35) www.iupac.org/namespaces/ThermoML/.

(36) Sandhu, P. *The MathML Handbook*; Charles River Media, Inc.: Hingham, MA, 2003. See, also: www.w3.org/Math/

(37) Rohac, V.; Ruzicka, K.; Ruzicka, V.; Zaitsau, D. H.; Kabo, G. J.; Diky, V.; Aim, K. Vapour pressure of diethyl phthalate. *J. Chem. Thermodyn.* **2004**, *36*, 929−937.

(38) Frenkel, M.; Kabo, G. J.; Marsh, K. N.; Roganov, G. N.; Wilhoit, R. C. *Thermodynamics of Organic Compounds in the Gas State, Volumes I and II*; Thermodynamics Research Center: College Station, TX, 1994.

(39) *Thermochemistry and Equilibria of Organic Compounds*; Frenkel, M., Ed.; VCH: New York, 1993.

(40) Joback, K. G.; Reid, R. C. Estimation of Pure-Component Properties from Group-Contributions. *Chem. Eng. Comm.* **1987**, *57*, 233−243.

(41) Constantinou, L.; Gani, R. New Group-Contribution Method for Estimating Properties of Pure Compounds. *AIChE J.* **1994**, *40*, 1697−1710.

(42) Marrero-Morejon, J.; Pardillo-Fontdevila, E. Estimation of Pure Compound Properties Using Group-Interaction Contributions. *AIChE J.* **1999**, *45*, 615−621.

(43) Wilson, G. M.; Jasperson, L. V. Critical Constants T_c, p_c, Estimation Based on Zero, First, Second-Order Methods. *AIChE Meeting*, New Orleans, LA, 1996.

(44) Yamada T.; Gunn R. D. Saturated Liquid Molar Volumes. The Rackett Equation. *J. Chem. Eng. Data* **1973**, *18*(2), 234−236.

(45) The method of Bondi is described on page 6.19 ref 29.

(46) Xiang H. W. The New Simple Extended Corresponding-States Principle: Vapor Pressure and Second Virial Coefficient. *Chem. Eng. Sci.* **2002**, *57*, 1439−1449.

(47) Liu, D. X.; Xiang, H. W. Corresponding-States Correlation and Prediction of Third Virial Coefficients for a Wide Range of Substances. *Int. J. Thermophys.* **2003**, *24*, 1667.

(48) Orbey, H.; Vera, J. H. Correlation for the Third Virial Coefficient using T_c, p_c, and ω as Parameters. *AIChE J.* **1983**, *29*, 107−113.

(49) Ambrose D.; Walton, J. Vapor-Pressures up to their Critical Temperatures of Normal Alkanes and Alkanols. *Pure Appl. Chem.* **1989**, *61*(8), 1395−1403.

(50) Methods are described in ref 29: page 9.9 for the low-pressure gas and page 9.35 for the Lucas method for high-pressure gas.

(51) Sastri, S. R. S.; Rao, K. K. A New Group Contribution Method for Predicting Viscosity of Organic Liquids. *Chem. Eng. J. Biochem. Eng. J.* **1992**, *50*, 9−25.

(52) Chung, T. H.; Lee, L. L.; Starling, K. E. Applications of Kinetic Gas Theories and Multiparameter Correlation for Prediction of Dilute Gas Viscosity and Thermal-Conductivity. *Ind. Eng. Chem. Fundam.* **1984**, *23*, 8−13.

(53) Chung, T. H.; Ajlan, M.; Lee, L. L.; Starling, K. E. Generalized Multiparameter Correlation for Nonpolar and Polar Fluid Transport-Properties. *Ind. Eng. Chem. Res.* **1988**, *27*, 671−679.

CI050067B