2003-11-11

Thermodynamic Property Prediction for Solid Organic Compounds Based on Molecular Structure

Benjamin T. Goodman
Brigham Young University - Provo

THERMODYNAMIC PROPERTY PREDICTION FOR SOLID ORGANIC COMPOUNDS BASED ON MOLECULAR STRUCTURE

by

Benjamin T. Goodman

A dissertation submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemical Engineering

Brigham Young University

December 2003

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a dissertation submitted by

Benjamin T. Goodman

This dissertation has been read by each member of the following graduate committee and by a majority vote has been found satisfactory.

_____	_____
Date	Richard L. Rowley, Chair
_____	_____
Date	W. Vincent Wilding
_____	_____
Date	John L. Oscarson
_____	_____
Date	Ronald E. Terry
_____	_____
Date	Kenneth A. Solen

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the dissertation of Benjamin T. Goodman in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university.

_____ _____

Date

Richard L. Rowley
Chair, Graduate Committee

Accepted for the Department

_____

W. Vincent Wilding
Department Chair

Accepted for the College

_____

Douglas M. Chabries
Dean, College of Engineering and Technology

ABSTRACT

THERMODYNAMIC PROPERTY PREDICTION FOR SOLID ORGANIC COMPOUNDS BASED ON MOLECULAR STRUCTURE

Benjamin T. Goodman

Department of Chemical Engineering

Doctor of Philosophy

A knowledge of thermophysical properties is necessary for the design of all process units. Reliable property prediction methods are essential because reliable experimental data are often not available due to concerns about measurement difficulty, cost, scarcity, safety, or environment. In particular, there is a lack of prediction methods for solid properties. Predicted property values can also be used to fill holes in property databases to understand more fully compound characteristics. This work is a comprehensive analysis of the prediction methods available for five commonly needed solid properties. Where satisfactory methods are available, recommendations are made; where methods are unsatisfactory in scope or accuracy, improvements have been made or new methods have been developed. In the latter case, the following general scheme has been used to develop correlations: extraction of a training set of experimental data of a specific accuracy from the DIPPR 801 database, selection of a class of equations to use in the correlation, refinement of the form of the equation through least squares regression, selection of the chemical groups and/or molecular descriptors to be used as independent variables, calculation of

coefficient values using the training set, addition of groups where refinement is needed, and a final testing of the resultant correlation against an independent test set of experimental data.

Two new methods for predicting crystalline heat capacity were created. The first is a simple power law method (PL) that uses first-order functional groups. The second is derived as a modification of the Einstein-Debye canonical partition function (PF) that uses the same groups as the PL method with other descriptors to account for molecule size and multiple halogens. The PL method is intended for the temperature range of 50 to 250 K; the PF method is intended for temperatures above 250 K. Both the PL and PF methods have been assigned an uncertainty of 13% in their preferred temperature ranges based on comparisons to experimental data.

A method for estimating heat of sublimation at the triple point was created using the same groups as used in the heat capacity PF method (estimated to have an error of 13%). This method can be used in conjunction with the Clausius-Clapeyron equation to predict solid vapor pressure. Errors in predicted solid vapor pressures averaged about 44.9%. As most solid vapor pressures are extremely small, on the order of one Pascal, this error is small on an absolute scale.

An improvement was developed for an existing DIPPR correlation between solid and liquid densities at the triple point. The new correlation improves the prediction of solid density at the triple point and permits calculation of solid densities over a wide range of temperatures with an uncertainty of 6.3%.

Based on the analysis of melting points performed in this study, Marrero and Gani's method is recommended as the primary method of predicting melting points for organic compounds (deviation from experimental values of 12.5%). This method can be unwieldy due to the large number of groups it employs, so the method of Yalkowsky et al. (13.9% deviation) is given a secondary recommendation due to its broad applicability with few input requirements.

ACKNOWLEDGMENTS

I would like to thank Richard Rowley for his advice and contributions as well as the DIPPR 801 project for financial and informational support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

A	compound-specific parameter used in PL method (see Eq. 2.3); also used generically in Eq. 2.1 along with B and C
AAD	average absolute deviation (see Eq. 2.18)
AALD	average absolute logarithmic deviation (see Eq. 3.13)
AAPD	average absolute percent deviation (see Eq. 2.17)
a_i, b_i	group increments used in the PL method (see Eq. 2.15)
BRANCH1	number of branch atoms (does not count H or F) (see Eq. 5.1)
BRANCH2	number of branch atoms (does not count H or F; Br, S, and I weighted more) (see Eq. 5.2)
C_p	heat capacity (constant pressure)
C_p^s	solid heat capacity (constant pressure)
C_p°	ideal gas heat capacity (constant pressure)
d_i, f_i, g_i	group increments used to predict ΔH_s (see Eq. 3.12)
DIPPR	Design Institute for Physical Properties
FLEX1	number of flexible atoms (does not count H or F) (see Eq. 5.1)
FLEX2	number of flexible atoms (does not count H or F; Br, S, and I weighted more) (see Eq. 5.2)
$g(\nu)$	distribution of vibrational frequencies in partition function

h	Planck's constant
k	Boltzmann's constant
k^s	solid thermal conductivity
m	parameter used in derivation of PL method (see Eq. 2.2); also the mass of a molecule
N	number of molecules
N_a	number of atoms in a molecule
n_i	number of times a group appears in a molecule
NG	complete range of groups for a method
n_X	number of C and Si valences occupied by a halogen or hydrogen in a molecule
P	pressure
p_i	descriptor coefficient used to predict ρ^s at $T/T_{TP} = 0.85$ (see Eq. 4.2)
PF	partition function derived equation for C_p^s (see Eq. 2.14)
PL	power law equation for C_p^s (see Eq. 2.3)
P_{TP}	triple point pressure
P_v^l	liquid vapor pressure
P_v^s	solid vapor pressure; sublimation pressure
Q	Einstein canonical partition function (see Eq. 2.4)
q_i	descriptor coefficient used to predict ρ^s at $T/T_{TP} = 1.0$ (see Eq. 4.3)
QSPR	quantitative structure-property relationship

r	parameter used in generalizing Debye vibrational distribution (see Eq. 2.12)
R	universal ideal gas constant
RIGID1	number of rigid atoms (does not count H or F) (see Eq. 5.1)
RIGID2	number of rigid atoms (does not count H or F; Br, S, and I weighted more) (see Eq. 5.2)
r_g	radius of gyration
SD	standard deviation
$S°$	ideal gas entropy
T	temperature
T_b	normal boiling point
T_f	melting point; fusion temperature
TOTAL1	number of atoms (does not count H or F) (see Eq. 5.3)
TOTAL2	number of atoms (does not count H or F; Br, S, and I weighted more) (see Eq. 5.4)
T_{TP}	triple point temperature
U_0	zero point energy of equilibrium lattice sites in partition function
$U_{0,\text{classic}}$	zero point energy of equilibrium lattice sites in partition function using strict definition (no hybridization)
x, x_E, x_D, x_G	vibration non-dimensionalised by temperature (see Eq. 2.6) for generalized, Einstein, Debye, and the presented partition function models, respectively

α_i, β_i, γ_i	group increments used in the PF method (see Eq. 2.16)
δ	Dirac delta function
ΔH_f°	ideal gas heat of fusion (melting)
ΔH_f	heat of fusion (melting)
ΔH_s	heat of sublimation
ΔH_v	heat of liquid vaporization
ΔS_f	entropy of fusion (melting)
ΔV_s	change in molar volume upon sublimation
ΔZ_s	change in compressibility upon sublimation
Θ, Θ_E, Θ_D, Θ_G	vibration as a characteristic temperature (see Eq. 2.6) for generalized, Einstein, Debye, and the presented partition function models, respectively
θ_A, θ_B, θ_C	moments of inertia for a non-linear molecule expressed as characteristic temperatures
θ_v	ideal gas vibration expressed as a characteristic temperature
μ	chemical potential
ν	vibrational frequencies in partition function
ρ^l	liquid density
ρ^s	solid density
σ	symmetry number

CHAPTER 1 - INTRODUCTION

Thermodynamic Properties

Thermodynamic properties are those that deal with interrelationships between temperature, pressure, volume, and energy. Common thermodynamic properties for solids are heat capacity, density, sublimation pressure, melting point temperature, heat of melting , and heat of sublimation. Heat capacity is a measure of how the temperature of a substance changes when heat is added to or drawn from it and is a function of temperature and, to a lesser extent, pressure. Sublimation pressure (also referred to as solid vapor pressure) is the pressure a solid exerts on its environment as it vaporizes and is solely a function of temperature. Density is the ratio of a material's mass to its volume and it is affected by temperature and pressure. Heat of melting and heat of sublimation are latent enthalpy changes required to make a phase change to the liquid and vapor phases, respectively. Thermodynamic properties have a high practical significance because they describe how physical changes in a compound's environment such as pressure and temperature affect the compound and how the compound in return changes the pressure and temperature of its surroundings. As temperature and pressure affect the kinetics and extent of chemical reactions as well as the flow characteristics of chemicals, these properties are important to the chemical industry.

Solid Thermodynamic Property Availability

Unfortunately, experimental values cannot be measured and tabulated for every property and for every compound, let alone for every temperature and pressure. As industry finds new applications for chemicals, properties are needed to evaluate the usefulness of the compound for the application. While using experimental property values is preferred, these are often not available. Using correlations to predict properties is generally more feasible with respect to both time and money than measuring the properties for the exploratory phase of development when large numbers of chemicals are evaluated.

The DIPPR 801 project maintains an evaluated database for physical properties.[1] The DIPPR 801 staff evaluates the data available for each compound and assigns an uncertainty to the data based on the reliability of the measurement technique, compound purity, source, and other factors. Where experimental data are unavailable, a prediction method is used. While accurate prediction methods are available for most of the properties included in the database, a current void in solid-property prediction methods is reflected in the number of compounds in the DIPPR database with no estimate of solid properties. The DIPPR database includes temperature-dependent correlations (at 1 atm) for solid heat capacity (C_p^s), solid vapor pressure (P_v^s), and solid thermal conductivity (k^s) whenever possible. Of the over 1800 compounds in the database, 28% have no C_p^s data and 38% have a single value listed (usually at the triple point). The situation is much worse for P_v^s and k^s; 87% of the compounds have no P_v^s values and 96% have no k^s values. The purpose of this study was to address the paucity of solid property estimation methods

by evaluating current correlations for solid properties and developing new, reliable correlations where none exist or where improvements can be made.

Quantitative Structure-Property Relationships

Quantitative structure-property relationships (QSPRs) are empirical or semi-empirical correlations between a chemical or physical property and a set of molecular structural descriptors. The types of descriptors used in QSPRs can vary greatly, and include constitutional, topological, electrostatic, geometrical, or quantum-chemical descriptors.[2] QSPR methods can be as simple as a group-contribution method (which uses basic functional groups) or as complex as using quantum-chemical descriptors such as HOMO (highest occupied molecular orbital) or LUMO (lowest unoccupied molecular orbital) energies. Properties correlated by QSPR methods range from gas octonol-water partition coefficients to boiling points.[2,3] While QSPRs are developed using regression techniques, the true power of QSPR is the extension of the regression to predict properties for compounds not included in the original regression. Cross-validation techniques can give an indication of how well the properties can be predicted for unknown compounds.

Databases in which the experimental data are evaluated for quality, such as the DIPPR database, are essential for development of data correlations for use in estimation and prediction of physical properties. An evaluated database provides a reliable training set of values that can be used to determine the independent properties and variables that strongly correlate with the property to be estimated and then to regress values for the

coefficients in the new correlation. Data from the database not used in the development of a correlation can also be used for testing new correlations or extending current correlations. This rationale has fostered the use of the DIPPR database for development of new prediction methods using QSPR correlations. Previously, others have used this methodology to develop prediction methods for the normal boiling point[4] and surface tension.[5]

There are numerous QSPR software packages available.[2] QSPR software narrows the initial set of descriptors to the set that is statistically significant. The significant descriptors are then used in a regression to make a prediction model. Cross-validation routines are also included so that the prediction power of the regression can be estimated. In addition, QSPR software typically has built-in methods for quickly calculating descriptors and tools for visualizing molecules. The QSPR software used in this research is the Oxford Molecular Group's Tsar 3.2 (upgraded in the course of the research to version 3.3).[6,7]

Group-Contributions

Group-contribution methods for correlating physical properties have existed longer than the QSPR approach has been called such, but group contributions can be considered a subset of QSPR. They are constitutional (structural) descriptors that constitute the most widely used descriptors in property correlations. Group contributions base the regression on linear combinations of chemically unique groups. Examples of chemical functional groups are methylene groups (-CH_2-) and carboxylic acid groups

(-COOH). Functional groups can be first order, such as the preceding examples (where every atom is only counted once), or second order. Second-order groups include all the bonded neighbors of the center group. Therefore, all atoms are counted at least twice, once as a center atom and at least once as a neighbor. Most group-contribution methods are first order. One of the more commonly known second-order methods is the Benson method for correlating the following ideal gas properties: ΔH_f° (298 K), S° (298 K), and C_p° (300, 400, 500, 600, 800, and 1000 K).[8] The main advantage of group contributions over other QSPR methods is simplicity. Functional groups can easily be counted without resorting to complex quantum mechanical calculations; only knowledge of the basic structural formula of the molecule is required. However, the arrangement of the atoms in the molecule may not give sufficient structural details for an accurate QSPR correlation.

There are advantages and disadvantages to both first- and second-order group methods. The main advantage of first-order groups is that there are fewer groups. This means that it is more likely that there will be enough experimental data to produce reliable values for all of the required groups in the correlation. It also means that calculations will be quicker and simpler. While second-order group methods are more complex, they also tend to be more accurate. Some second-order group methods, such as the Benson method mentioned above, have easier computer implementation because the method for counting the groups naturally excludes double counting or overlapping groups (such as counting a carboxylic acid as a carboxylic acid, an ester, and an alcohol). Second-order methods are also more likely to take into account the difference between isomers. Constantinou and Gani developed a group-contribution method for several

properties that contained both first- and second-order groups.[9] In this scheme, the first-order contributions are calculated first. If a more accurate correlation is needed or desired, second-order corrections are then added to the first-order building blocks. This scheme has the advantage of having a simple method and a more accurate method that use the same foundation. This allows the user of the correlation to decide between ease of use and accuracy. The disadvantage of this scheme is that the second-order level is harder and more time consuming to calculate than a stand-alone, second-order method, such as that of Benson, because the groups have to be calculated twice.

Specific Objectives

The general purpose of this work is to examine the property prediction methods available for solid compounds and develop new methods where needed. As this is a rather large goal, it needs to be narrowed based on need and feasibility. For example, the lack of experimental data for k^s has already been mentioned. This indicates that there are insufficient data to develop a QSPR-type prediction method. Specifically, this study will examine C_p^s, P_v^s, heat of sublimation (ΔH_s), solid density (ρ^s), and normal melting point (T_f).

Chapter 2 examines available methods for predicting C_p^s and proposes two prediction methods for organic compounds. Chapter 3 looks at the related properties P_v^s and ΔH_s. A method for predicting ΔH_s of organic compounds is presented and the efficacy of using it and the Clausius-Clapeyron equation to estimate P_v^s is examined. Chapter 4 examines ρ^s and presents a simple method for predicting ρ^s based on liquid

density. Chapter 5 evaluates the various T_f prediction methods (the most studied of any of the solid properties).

References

(1) R. L. Rowley, W. V. Wilding, J. L. Oscarson, Y. Yang, N. A. Zundel, T. E. Daubert, and R. P. Danner, *DIPPR® Data Compilation of Pure Compound Properties*, Design Institute for Physical Properties, AIChE: New York (2003).

(2) A. R. Katritzky, V. S. Lobanov, and M. Karelson, "QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure," *Chem. Soc. Rev.*, 1995, **24**, 279-287.

(3) R. Murugan, M. P. Grendze, J. E. Toomey, A. R. Katritzky, M. Karelson, V. Lobanov, and P. Rachwal, "Predicting physical properties from molecular structure," *Chemtech*, 1994, **24**, 17-23.

(4) D. Ericksen, W. V. Wilding, J. L. Oscarson, and R. L. Rowley, "Use of the DIPPR Database for the Development of QSPR Correlations: Normal Boiling Point," *J. Chem. Eng. Data* 2002, **47**, 1293-1302.

(5) T. A. Knotts, W. V. Wilding, J. L. Oscarson, and R. L. Rowley, "Use of the DIPPR Database for the Development of QSPR Correlations: Surface Tension," *J. Chem. Eng. Data* 2001, **46**, 1007-1012.

(6) *TSAR Version 3.2*, Oxford Molecular Group, Oxford Molecular Limited: Oxford, 1998.

(7) *TSAR Version 3.3*, Oxford Molecular Group, Oxford Molecular Limited: Oxford, 2000.

(8) R. C. Reid, J. M. Prausnitz, and B. E. Poling, *The Properties of Gases and Liquids*, 4th Edition, McGraw-Hill, Inc.: San Francisco, 1974.

(9) L. Constantinou and R. Gani, “New Group Contributions Method for Estimating Properties of Pure Compounds,” *AIChE J.*, 1994, **40**, 1697-1710.

CHAPTER 2 - SOLID HEAT CAPACITY

Previously Published Solid Heat Capacity Prediction Methods

All of the QSPR correlations for C_p^s published to date have used group-contribution methods. These group-contribution methods can be broken up into two categories: those that predict the solid heat capacity at room temperature (298 K) and those that include some temperature dependence, but are limited in scope in some other way.

Kopp's Rule is a simple method for estimating the heat capacity of a solid at 298 K. It is based on counting the number of atoms in a molecule and adding a contribution for each type of atom. Kopp's Rule is actually, therefore, an element-based correlation wherein the "group" effect is assumed to be additive over atoms rather than over chemically distinct bonds. Kopp's Rule has contributions for 7 elements plus a miscellaneous contribution for those elements without a specific contribution. This has been modified by Hurst and Harrison to include contributions for 32 elements plus a miscellaneous contribution.[1] The modified Kopp's Rule has an expected average absolute error of 9.6% compared to 11.8% for the original Kopp's Rule for the 721 compounds used by Hurst and Harrison.[1] The main advantage of Kopp's Rule is that it can be used to estimate the heat capacity of any solid. A severe limitation to this method is that it is only

applicable at 298 K and this is above the melting point of many solids. Additionally, its accuracy is limited by the base assumption of elemental contributions which do not account for electron distribution differences due to different bonding interactions with neighboring atoms.

Brock's method[2] is another general correlation for computing solid heat capacity at 298 K. In Brock's method, the number of linear, nonlinear, and ionic groups are counted, and the vibrational degrees of freedom for each group are used to obtain the heat capacity. Brock also accounts for the restricted degrees of vibrational freedom for rings and aromatic groups and offers examples on how different types of functional groups fit into his scheme. While on the surface this method appears simple, there is a subtle complexity involved in partitioning molecules into different groups that requires considerable knowledge of and experience with the method to render it effective. Brock did not include an estimation of the method's accuracy. While Brock's method has a breadth of applicability comparable to Kopp's Rule in that it can be applied to any compound, no distinction is made between liquid and solid heat capacities, suggesting that it does not have a great deal of accuracy for solid values and making its use for crystalline phases suspect.

Domalski and Hearing[3] used a Benson-like second-order group-contribution method to correlate, among other properties, the solid heat capacity of organic compounds at 298 K. This method is a more sophisticated alternative to the methods of Kopp or Brock, but has a limited number of groups available for solid heat capacity due to the limited data set used in group parameter regression. Second-order group methods

require significantly more experimental data than first-order methods because it is not just group values that must be regressed, but every combination of a group with possible bonded neighboring groups. Hurst and Harrison showed that the Domalski and Hearing method gives lower errors than their modified Kopp's Rule for the compounds for which all of its constituent groups are available.[1]

Kubaschewski and Ünal[4] use a group-contribution method based on a simple linear temperature dependence with a T^{-2} curvature term,

$$C_p = A + BT + CT^{-2} \quad \textbf{(2.1)}$$

to estimate solid heat capacity. In Eq. 2.1, C_p is heat capacity (at constant pressure), T is temperature, and A, B, and C are compound-specific coefficients. For this method, C is based on the number of ionic radicals (i.e., Na^+, SO_4^{2-}) in the molecule. The parameters A and B are calculated from two values of C_p: at the melting point, where C_p is estimated based on the number of ionic radicals, and at 298 K where a group-contribution method (with each ion constituting a group) is used. The authors recommend that experimental values, when available, be used to calculate A and B. This is especially true for the value at 298 K. Of course, using experimental values makes this method an extrapolation/interpolation tool, which has value, but the focus of this work is on predictive tools.

Mostafa et al.[5] similarly used the T^{-2} temperature dependence in Eq. 2.1 for ionic crystalline solids. They used a group-contribution method with individual ionic radicals, and two additional ligand groups, H_2O and CO, to obtain each of the three parameters in Eq. 2.1: A, B, and C. Mostafa et al. compared their method to Kopp's Rule and found that

the predictions at 298 K were more accurate than Kopp's rule for most compounds. They calculated a mean error of 8.17% for 649 of the 664 salts at the highest temperature for which the correlation was valid for each compound. The methodology of this correlation is "cleaner" than that of Kubaschewski and Ünal, because it uses group contributions for all three parameters rather than just one, with crude estimates based on the number of ions supplying the other two.

Kabo et al.[6] have developed two simple methods for predicting the C_p^s of alkanes, alkenes, alkanols, and alkanones at 10 K increments from 10 - 150 K. The first method is for estimating the heat capacity of solid alkanes and is based upon the number of neighboring carbon atoms attached to each individual carbon. The second correlation is applicable to alkanes, alkenes, alkanols, and alkanones and is based upon "effective bonds" where each carbon-carbon bond pair (influenced by its neighbors) is the basic additive unit. Kabo et al. stated that the accuracy of these methods is within 3-5 times the estimated experimental error. Kabo et al. also have published even more specialized group contribution methods for the estimation of the heat capacities of solid alkyl[7] and phenyl[8] ureas at increments from 5 - 320 K with a proposed accuracy of 5%.

While Mostafa et al. provided an acceptable prediction method for inorganic salts, the available methods for organic compounds all have temperature limitations (298 K for Kopp's Rule, Brock, and Domalski and Hearing, and less than 150 K for Kabo et al.). A more general prediction method that accurately represents the temperature dependence of organic compounds is needed.

Temperature Dependence of Solid Heat Capacity

To develop a more general correlation for the heat capacity of organic solids, its temperature dependence must be sought from either theory or experiment. While Eq. 2.1 correlates well with C_p^s data, it has a disadvantage from a development standpoint: the three compound-dependent parameters must be correlated. The more compound-dependent parameters a correlation has, the larger the data set needed to make the regression significant. A preliminary investigation of using Eq. 2.1 also showed that consistency among the parameters was hard to achieve. Plots of the three coefficients regressed from Eq. 2.1 for the *n*-alkane family showed that a steady trend in *A* as the number of carbons increased would be interrupted as one compound favored an increase in *B* at the expense of the trend in *A*. As this problem does not exist with one-parameter equations, two equations with one adjustable parameter each were selected as the basis for new C_p^s correlations.

A simple, empirical, power-law form for the temperature-dependence of C_p^s,

$$C_p^s = AT^m \qquad \textbf{(2.2)}$$

has been used previously for solid hydrocarbons.[9] In Eq. 2.2, *A* and *m* are empirical coefficients with *m* less than one. To develop a simple, first-order prediction method, we applied this same temperature functionality to all solid organic compounds and treated *m* as a universal constant. We then used group contributions to obtain compound-specific values for *A*. Correlation of *m* for various chemical families or for individual compounds would be a logical starting point for development of a second-order method. However,

we felt at this time that the limited amount of experimental C_p^s data was inadequate to simultaneously correlate A and m for individual compounds.

To find an optimum value of m, 455 compounds with a total of 7967 data points at various temperatures were extracted from the DIPPR database (having a DIPPR quality code of estimated accuracy better than 5%) and used to optimize m and regress individual values for A in Eq. 2.2. The optimum m in a least-squares sense was found to be 0.793. We therefore used

$$C_p^s = \frac{A}{1000} T^{0.793} , \qquad \textbf{(2.3)}$$

where C_p^s is in J/mol·K and T is in Kelvin, as the starting point for development of the first of our two predictive methods, which we designate as the power-law (PL) method. The factor of 1000 in Eq. 2.3 converts units to J/mol·K from those used in the DIPPR database and in the correlation of A, J/kmol·K.

The second method developed in this work is based on the Einstein canonical partition function, Q. In the Einstein theory, atomic motions within the solid crystal are modeled as vibrations relative to equilibrium crystalline lattice positions. Expressed in terms of normal frequencies, ν, and the zero of energy for the crystal with all atoms at their equilibrium lattice sites, U_0, the partition function is[10]

$$\ln Q = -\frac{U_0}{2kT} - \int_0^\infty \left[\ln\left(1 - e^{-h\nu/kT}\right) + \frac{h\nu}{2kT} \right] g(\nu) d\nu \qquad \textbf{(2.4)}$$

where k is Boltzmann's constant, h is Planck's constant, and $g(\nu)d\nu$ gives the number of normal frequencies between ν and $\nu + d\nu$. In this model, the partition function and

consequently the thermodynamic properties of the crystal are determined by the choice of the frequency distribution function, $g(\nu)$. For crystals of atomic species, the frequency distribution must be constrained to $3N$ total normal frequencies, where N is the number of atoms in the crystal; i.e.,

$$\int_0^\infty g(\nu)d\nu = 3N \ . \tag{2.5}$$

It is convenient to use dimensionless frequency, x, and characteristic temperature, Θ, in the C_p^s expression. These frequency variables are defined by

$$x = \frac{h\nu}{kT} = \frac{\Theta}{T} \ . \tag{2.6}$$

The molar heat capacity, obtained from Eq. 2.4 through standard thermodynamic identities, when expressed in terms of dimensionless frequency, is[10]

$$C_p^s = k\int_0^\infty \frac{x^2 e^x g(x)}{\left(e^x - 1\right)^2} dx \ , \tag{2.7}$$

where it has been assumed that constant-pressure and constant-volume heat capacities are approximately equivalent for solids.

The form of the temperature dependence for C_p^s is thus determined by the distribution function model used to represent the internal frequencies in the crystal. Einstein chose to set all $3N$ frequencies to an identical value of Θ_E, or

$$g(x) = 3N\delta(x - x_E) \ , \tag{2.8}$$

where $\delta(x)$ is the Dirac delta function and x_E is related to Θ_E by the definition shown in Eq. 2.6. By substituting Eq. 2.8 into Eq. 2.7, one obtains for the molar C_p^s

$$C_p^s = \frac{3Rx_E^2 e^{x_E}}{\left(e^{x_E} - 1\right)^2}, \tag{2.9}$$

where R is the gas constant. To improve the performance of the Einstein theory, Debye chose

$$g(x) = \begin{bmatrix} 9Nx_D^{-3}x^2 & 0 \le x \le x_D \\ 0 & x > x_D \end{bmatrix} \tag{2.10}$$

for the frequency distribution, where x_D is defined analogously to x_E but now in terms of Θ_D, the so-called Debye temperature at which the frequency distribution is truncated to conserve modes (Eq. 2.5). Using Eq. 2.10 in Eq. 2.7, one obtains the Debye equation for the molar C_p^s

$$C_p^s = 9Rx_D^{-3}\int_0^{x_D} \frac{x^4 e^x}{\left(e^x - 1\right)^2} dx . \tag{2.11}$$

Debye's choice for the vibrational frequency distribution function assumes that the quadratic temperature dependency, known to be true near absolute zero, can be used over the whole temperature domain. However, vibrational frequency distributions vary significantly at higher temperatures and are quite complex as shown in Figure 2.1 taken from McQuarrie.[10]

Here, the quadratic constraint on the frequency distribution used by Debye is relaxed but the power-law relationship shown in Eq. 2.10 is retained in the form

$$g(x) = \begin{bmatrix} 3(r+1)NN_a x_G^{-(r+1)}x^r & 0 \le x \le x_G \\ 0 & x > x_G \end{bmatrix}, \tag{2.12}$$

where r is a real number greater than -1. Again, the frequency distribution is truncated at

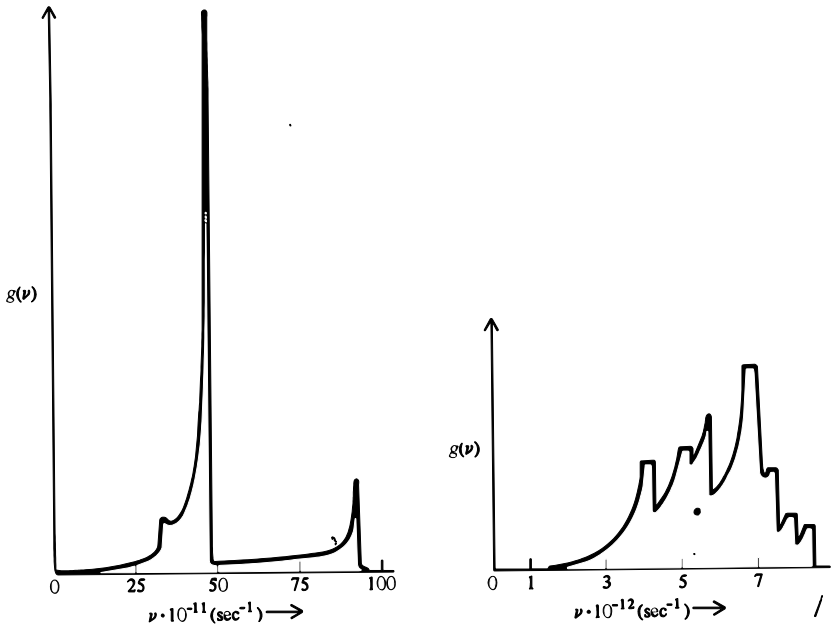

Figure 2.1. Experimental frequency spectrum of aluminum (left) and iron (right) [from McQuarrie[10]].

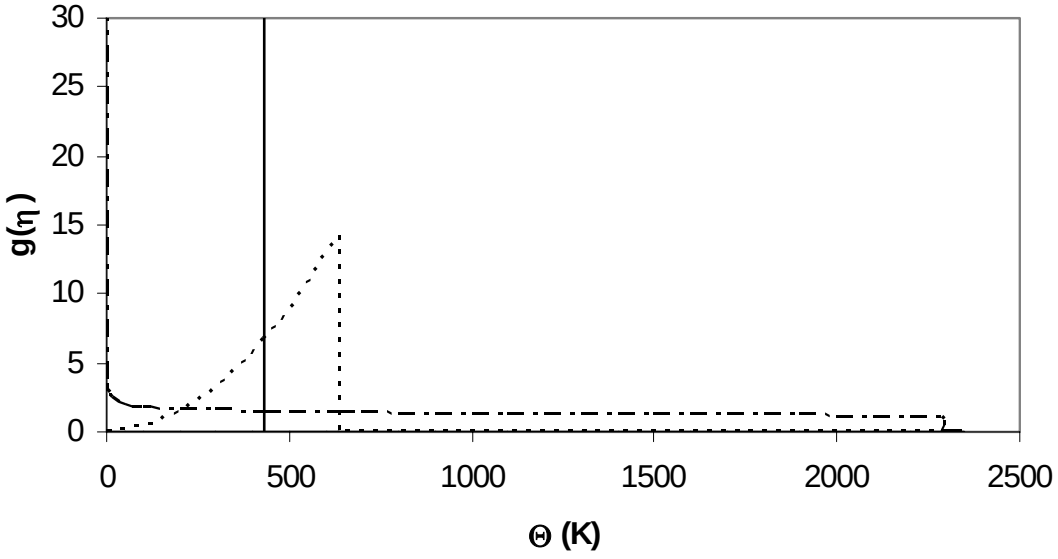

Figure 2.2. Frequency distributions for *n*-butane at 100 K for models based on Einstein (—), Debye (- - - -), and Eq. 2.12 with r = -0.15 (– - – -).

a generalized characteristic temperature, Θ_G, to conserve the total number of modes. In this application of the Einstein theory to complex organic solids, multi-atom molecules are considered such that the number of modes is now $3NN_a$ where N is the number of molecules and N_a is the number of atoms per molecule. This approach hybridizes the molecular and atomic vibrations of the crystal, essentially giving each atom in the crystal an average vibration that is based upon the vibration of each atom in the molecule. This yields for the molar C_p^s

$$C_p^s = 3N_a R(r+1)x_G^{-(r+1)} \int_0^{x_G} \frac{x^{r+2}e^x}{\left(e^x - 1\right)^2} dx \,. \qquad \textbf{(2.13)}$$

Equation 2.13 reduces to the Debye equation (Eq. 2.11) for monatomic species when $r = 2.0$; likewise, it reduces to the Einstein equation (Eq. 2.9) for monatomic species in the limit as r approaches infinity. As with the simple empirical equation, the temperature functionality is assumed to be the same for all compounds and r is treated as a universal constant. This again reduces the C_p^s equation to an equation in a single variable, Θ_G, which is correlated in terms of group contributions using the DIPPR database. The optimum value determined for r was -0.15. The form of $g(x_G)$ is compared to the Einstein and Debye models in Figure 2.2 for *n*-butane at 100 K. As the actual frequency distribution is complex (see Figure 2.1), $g(x_G)$ simply is the best empirical fit of C_p^s values for a wide range of organic compounds. The starting point for developing a second predictive equation for C_p^s, analogous to Eq. 2.3, is therefore

$$C_p^s = 2.55N_a R x_G^{-0.85} \int_0^{x_G} \frac{x^{1.85}e^x}{\left(e^x - 1\right)^2} dx \,. \qquad \textbf{(2.14)}$$

Equation 2.14 is designated as the partition-function (PF) method for calculating C_p^s. To utilize either of these two methods, PL or PF, for SCP prediction, correlations for the constants A and Θ_G that appear in Eqs. 2.3 and 2.14, respectively, have been developed.

Using Group Contributions to Select Parameters

By the choice of m in Eq. 2.3, the known T^3 low-temperature limit for C_p^s has intentionally been sacrificed. Similarly, the functional form of Eq. 2.14 with $r \neq 2$ deviates from the T^3 low-temperature limit. In essence, agreement at very low temperatures has been lost to provide a simple equation that optimizes predictive capability over the temperature range of interest in most engineering applications. The training set used to obtain group contributions for A in Eq. 2.3 and Θ_G in Eq. 2.14 is a subset of the original data set and includes 455 compounds and 7967 C_p^s data points.

For compounds that have more than one solid phase, the crystalline phase stable at the lowest temperature was chosen. This lowest-temperature phase is the most influenced by the interactions of functional groups, and it is consistent with Bondi's definition for a "standard" heat of sublimation. Bondi uses the lowest first-order phase transition as the standard because "solids above this transition often exhibit sharply reduced lattice energy and are unrepresentative of the 'typical' solid."[11] Because the correlations have been developed based only on this "standard" phase, they should not be applied to other crystalline phases.

Table 2.1. Linear group values for C_p^s prediction methods

Group	Description	Example	SMILES formula	*a* Eq. 2.15	*α* Eq. 2.16
–CH_3	Methyl	n-butane	**C**CCC	0.20184	-241.7
>CH_2	Methylene	1-heptanol	OC**C**CCCCC	0.11644	17.929
>CH–	Secondary carbon	2,3-dimethylpentane	C**C**(C)C(C)CC	0.030492	229.47
>C<	Tertiary carbon	2,2-dimethylbutane	C**C**(C)(C)CC	-0.04064	529.76
CH_2=	Terminal alkene	1-octene	**C**=CCCCCCC	0.18511	-387.3
–CH=	Alkene	1,3-butadiene	C=**C**C=C	0.11224	-118.89
>C=	Substituted alkene	Isobutene	C=**C**(C)C	0.028794	191.3
=C=	Allene	1,2-butadiene	C=**C**=CC	0.053464	-154.12
#CH	Terminal alkyne	Ethylacetylene	CCC#**C**	-0.02914	-538.85
#C–	Alkyne	Dimethylacetylene	C**C**#CC	0.13298	-225.13
Ar –CH=	Aromatic carbon	Biphenyl	c1**c**cccc1(c2ccccc2)	0.082478	-36.615
Ar >C=	Substituted aromatic C	Toluene	c1cccc**c**1(C)	0.012958	148.32
Ar –O–	Furan oxygen	Furan	C1=C**O**C=C1	0.066027	-70.693
Ar –N=	Pyridine nitrogen	Quinoline	c1(ccc**n**2)c2cccc1	0.056641	-229.57
Ar >N–	Substituted pyrrole N	N-methylpyrrole	**n**1(C)cccc1	0.008938	215.18
Ar –NH–	Pyrrole nitrogen	Pyrrole	C1=CC=C**N**1	-0.05246	178.85
Ar –S–	Thiophene sulfur	Thiophene	**S**1C=CC=C1	0.090926	-492.78
–O–	Ether	Dimethyl ether	C**O**C	0.064068	-154.96
–OH	Alcohol	1-pentanol	CCCCC**O**	0.10341	-286.75
–COH	Aldehyde	1-butanal	CCC**C=O**	0.15699	-451.8
>C=O	Ketone	3-hexanone	CC**C(=O)**CCC	0.12939	-252.22
–COO–	Ester	Methyl methacrylate	C=C(C)**C(=O)O**C	0.13686	-530.27
–COOH	Acid	*n*-butyric acid	CCC**C(=O)O**	0.21019	-498.54
–COOCO–	Anhydride	Maleic anhydride	**O1C(=O)**C=CC1**(=O)**	0.33091	-1321.5
–CO_3–	Carbonate	Ethylene carbonate	C1**OC(=O)O**C1	0.2517	-639.94
–NH_2	Primary amine	Methylamine	C**N**	0.056138	-53.298
>NH	Secondary amine	Piperidine	C1CCCC**N**1	-0.00717	363.75
>N–	Tertiary amine	Trimethylamine	C**N**(C)C	-0.01661	377.78
=NH		Dicyandiamide	N#CNC**(=N)**N	0.17689	-568.75
#N	Nitrile	Acetonitrile	CC**#N**	0.015355	-515.66
–N=N–	Diazide	*p*-aminoazobenzene	Nc1ccc(cc1)**N=N**c2ccccc2	0.3687	-761.63
–NO_2	Nitro	Nitrobenzene	c1(**N(=O)=O)**ccccc1	0.23327	-619.91
–N=C=O	Isocyanate	Phenyl isocyanate	c1**(N=C=O)**ccccc1	0.2698	-703.05
–SH	Thiol/mercaptan	*n*-hexyl mercaptan	CCCCCC**S**	0.21123	-594.12
–S–	Sulfide	Diethyl sulfide	CC**S**CC	0.14232	-391.13
–SS–	Disulfide	Di-*n*-propyl disulfide	CCC**SS**CCC	0.31457	-734.81
=S	Sulfur double bond	Thiourea	NC**(=S)**N	0.13753	-949.61
>S=O	Sulfoxide	Dimethyl sulfoxide	C**S(=O)**C	0.040002	-251.27
–F	Fluoride	Benzotrifluoride	c1(C(F)**(F)**F)ccccc1	0.15511	-320.76
–Cl	Chloride	Ethyl chloride	CC**[Cl]**	0.16995	-429.06
–Br	Bromide	Bromobenzene	c1**(Br)**ccccc1	0.19112	-70.347
–I	Iodide	Iodobenzene	c1**(I)**ccccc1	0.11318	-589
>Si<	Silane	Tetramethylsilane	C**[Si]**(C)(C)C	0.12213	140.96
>Si(O–)–	Siloxane	Hexamethyldisiloxane	C[Si](C)(C)**O[Si]**(C)(C)C	0.10125	77.804
cyc >Si(O–)–	cyclic Siloxane	Octamethylcyclotetra-siloxane	[Si]1(C)(C)**O[Si]**(C)(C)O [Si](C)(C)O[Si](C)(C)O1	0.063438	77.804
P(=O)(O-)$_3$	Phosphate	Triphenyl phosphate	c1ccccc1(**O[P](=O)(O**c2ccccc2)**O**c3ccccc3)	0.15016	-520.71
>P–	Phosphine	Triphenylphosphine	**P**(c1ccccc1)(c2ccccc2)(c3ccccc3)	0.069602	489.97
>P(=O)–	Phosphine oxide	Triphenylphosphine oxide	**P(=O)**(c1ccccc1)(c2ccccc2)(c3ccccc3)	0.21875	-242.12

The functional group definitions chosen are similar to those used in the Joback method for boiling points.[12] These group definitions are common, simple to use, and are available in many automated software prediction packages. A limited QSPR analysis indicated that group contributions adequately correlated the values of *A* in Eq. 2.3. However, the QSPR analysis indicated that the radius of gyration, r_g, was statistically significant in addition to group contributions for the correlation of Θ_G values in Eq. 2.14. In addition to the standard Joback groups, quadratic terms were found to be statistically important for the two most common groups: methylene and aromatic carbon groups. A correction for multiple halogen groups was also necessary and is included as a correction term based on the fraction of C or Si terminal valences occupied by halogen atoms. The final correlations obtained were

$$A = \exp\left(6.7796 + \sum_{i}^{NG} a_i n_i + \sum_{i}^{NG} b_i n_i^2 \right) \qquad \textbf{(2.15)}$$

$$\Theta_G = 1886.2 + 3.3626 \times 10^{12} r_g + \sum_{i}^{NG} \alpha_i n_i + \sum_{i}^{NG} \beta_i n_i^2 + \sum_{i}^{NG} \gamma_i \frac{n_i}{n_X} \qquad \textbf{(2.16)}$$

where a_i, b_i, α_i, β_i, and γ_i are values for group *i* regressed from the training set, n_i is the number of times that group *i* appears in the molecule, n_X is the total number of all halogen and hydrogen atoms attached to C and Si atoms in the molecule, *NG* is the total number of groups in the molecule, and r_g is the radius of gyration of the molecule in meters. Values of the radius of gyration are obtainable from several sources including the DIPPR 801 database. These equations should not be used for temperatures below 50 K for reasons mentioned above.

Table 2.2. Nonlinear and halogen group values for C_p^s prediction methods

Group	Description	Eq. 2.15	Eq. 2.16
A. Nonlinear terms		b	β
$-CH_2$	Methylene	-0.00188	-2.9045
Ar =CH–	Aromatic carbon	-0.00033	-2.9616
B. Halogen fraction terms			γ
–Cl	Cl fraction		-1361.4
–F	F fraction		-1231.3
–Br	Br fraction		-3864.5

The parameters for Eqs. 2.15 and 2.16 were obtained using the multiple regression package in Oxford Molecular Tsar 3.2.[13] Tables 2.1 and 2.2 contain the values of the group contributions obtained. Linear group terms are given in Table 2.1; the nonlinear terms for methylene and aromatic carbon groups and the correction terms for the halogen fractions are given in Table 2.2. Table 2.1 also illustrates group definitions. The designated group is highlighted with a bold typeface in the SMILES formula[14] for the compound. SMILES (Simplified Molecular Input Line Entry Specification) is a simple in-line chemical notation for the structure of a compound. SMILES formulas are compiled in the DIPPR database and are very convenient for use in software such as Tsarthat automates the parsing of molecular structure into groups. A simple SMILES tutorial can be found on the world wide web.[15]

It is useful to evaluate the correlation of the training set data in terms of an average absolute percent deviation (AAPD)

$$AAPD = \sum_{i=1}^{n} \left| \frac{x_{pred,i} - x_{\exp,i}}{x_{\exp,i}} \right| \frac{100\%}{n}, \qquad \textbf{(2.17)}$$

the standard deviation (SD), and the average absolute deviation (AAD)

$$AAD = \frac{\sum_{i=1}^{n} \left| x_{pred,i} - x_{\exp,i} \right|}{n}. \qquad \textbf{(2.18)}$$

As C_p^s values may range over more than an order of magnitude, these three measures provide useful statistics in different parts of the temperature range. AAPD emphasizes absolute errors in the region where the magnitude of C_p^s is small, at low temperatures, while SD weights more heavily the larger absolute errors expected at higher temperatures where C_p^s is larger.

Table 2.3 shows the AAPD, AAD, and SD results of the correlation for the training sets used to obtain the values in Tables 2.1 and 2.2. Fractional deviations of the correlated values from the training set values are shown for the PL method (Eqs. 2.3 and 2.15) in Figure 2.3. The PL method performs well at lower temperatures, but there is a negative bias to the residuals at higher temperatures. A similar plot for the PF method (Eqs. 2.14 and 2.16) is shown in Figure 2.5. There is no noticeable bias in this correlation. Most of the deviation from experimental values is due to

Table 2.3. Quality of C_p^s training set correlation

	PL method	PF method
Training Set Compounds	455	455
Training Set C_p^s Values	7967	7967
AAPD (%)	6.84	7.96
AAD (J/mol·K)	9.30	9.43
SD (J/mol·K)	19.2	16.5

inadequacies in the assumed temperature dependence of the model; i.e., the assumption of a universal constant *m* in Eq. 2.2 and *r* in Eq. 2.13. This can be observed by comparing Figures 2.4 and 2.5. Figure 2.4 utilizes the values of Θ_G regressed from the experimental data using Eq. 2.13 rather than the values correlated in terms of groups using Eq. 2.16. Little degradation of prediction values occurs in the estimation of Θ_G from the group-contribution correlation. Table 2.3 indicates that the PL equations correlate the data overall slightly better, but the SD for the PF method is lower suggesting that it may be preferred at higher temperatures (above 250 K). The high-temperature bias noted in

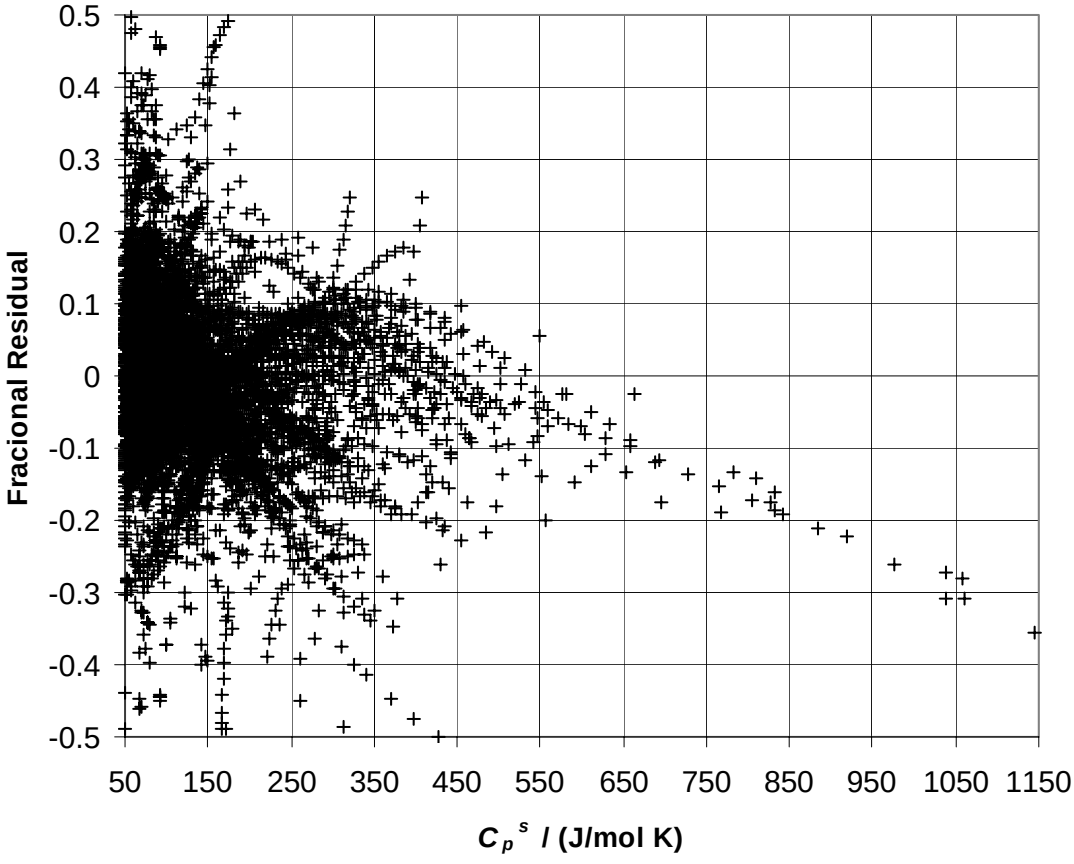

Figure 2.3. Fractional deviations of correlated C_p^s values from the training set values when correlated with the PL method.

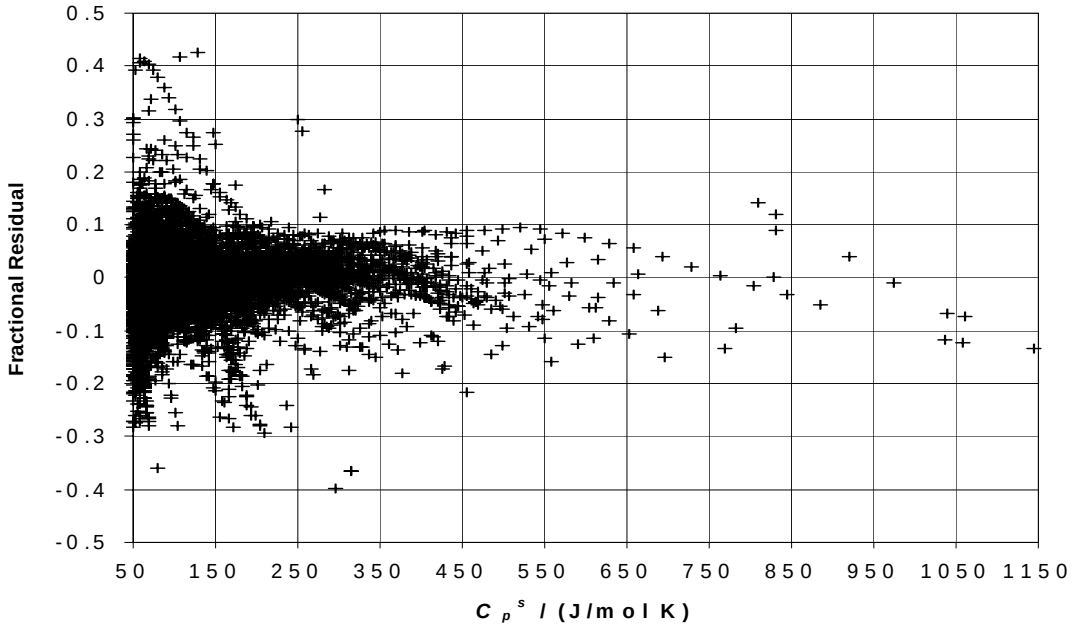

Figure 2.4. Fractional deviations of C_p^s values from the training set values when the PF method is used with regressed values of Θ_G.

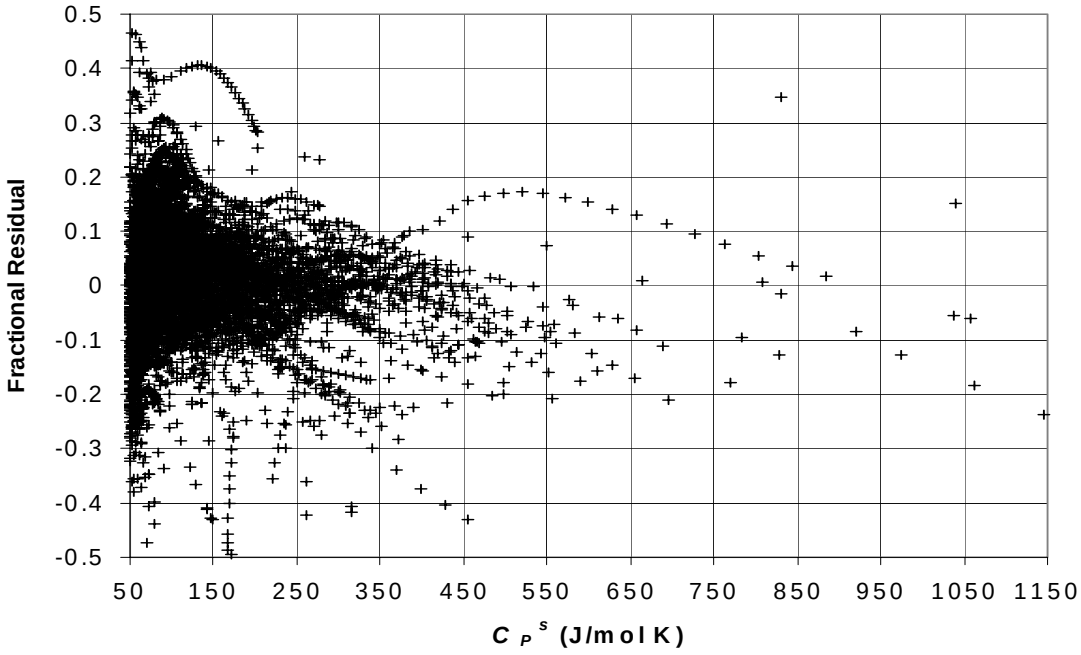

Figure 2.5. Fractional deviations of C_p^s values from the training set values when the PF method is used with values of Θ_G from group contributions.

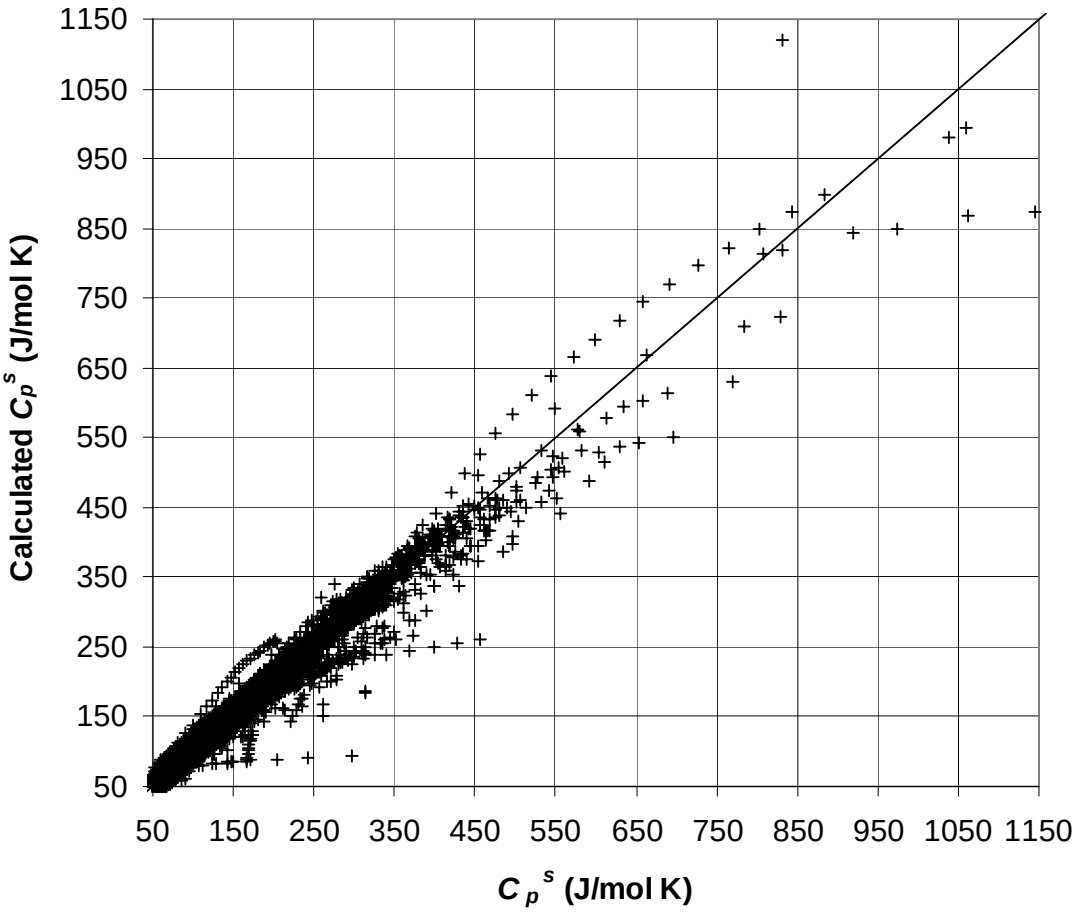

Figure 2.6. Comparison of correlated and experimental C_p^s values obtained in the regression of the PF method from the training set.

Figure 2.3 is consistent with this suggestion. The lower SD may also suggest that the PF method is less susceptible to larger errors. For example, some diester compounds produced a noticeably larger error with PL than with PF.

Discussion of the Correlations

Figure 2.6 shows the overall correlation for the PF method in terms of absolute deviation of the correlated C_p^s values from the accepted experimental data in the DIPPR

Table 2.4. Comparison of predicted C_p^s values to those from the test set

T	# compounds	# C_p^s values	Statistic	PL	PF
> 50 K	45	948	AAPD (%)	13.0	20.7
			AAD (J/mol·K)	25.2	21.9
			SD (J/mol·K)	54.2	48.1
> 50 K & < 250 K	45	788	AAPD (%)	12.8	22.6
			AAD (J/mol·K)	19.5	19.5
			SD (J/mol·K)	42.3	49.0
> 250 K	22	160	AAPD (%)	14.3	11.5
			AAD (J/mol·K)	53.2	33.3
			SD (J/mol·K)	92.9	43.8

801 database. There is no apparent bias for the PF method, and the correlation reproduces reasonably well the experimental data for the compounds in the training set. Example comparisons of C_p^s values predicted by the PL and PF methods to experimental data for three compounds are shown in Figure 2.7.

In order to test the extrapolative capability of the correlations developed here, the PL and PF methods were used to predict C_p^s values for 45 compounds available in the TRC handbooks[16] but not in the DIPPR 801 database from which the training set was developed. The results of this test are shown in Table 2.4. While the AAPD is approximately 8% for correlation of the training set (Table 2.3), the test set results suggest an expected average accuracy for new predictions of about 13%. Unfortunately, we do not know the accuracy of the experimental values compiled in the TRC handbook for the test set compounds, but we do not expect experimental error to be a large

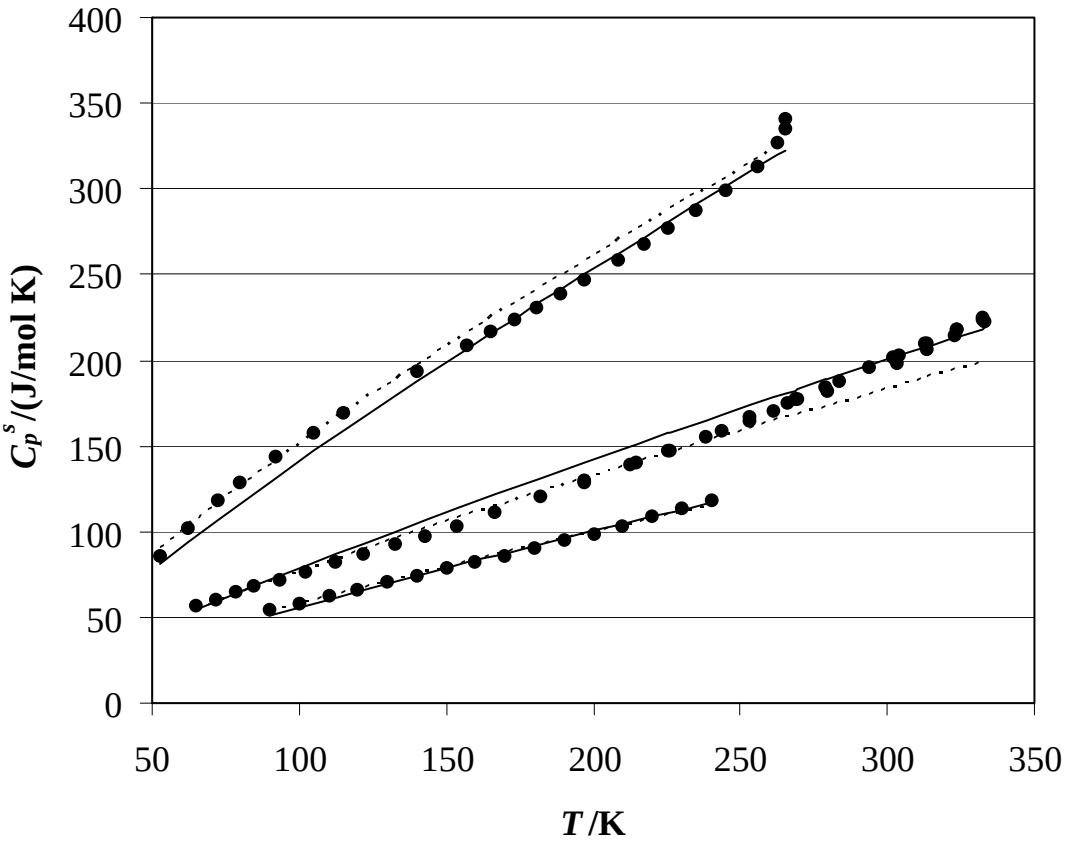

Figure 2.7. Comparison of experimental (●) C_p^s data to values predicted using the PL (– – – –) and PF (———) methods for *n*-decylcyclohexane (top series), biphenyl (middle series), and iodobenzene (bottom series).

component of this error. We have also divided the comparison in Table 2.4 into two different temperature ranges. In the low-temperature region, defined as 50 K < T < 250 K, the PL method is a better predictor of C_p^s in terms of percent error than the PF method with AAPD values of 12.8 % and 22.6 %, respectively. However, in terms of absolute deviations, the PF method is equivalent to the PL method in the low temperature range, but superior overall and particularly at higher temperatures. The PF method is a better

Table 2.5. Comparison of C_p^s predictive methods at 298 K

	# Compounds	AAPD (%)	AAD (J/mol·K)	SD (J/mol·K)
A. Common Test Set				
Domalski & Hearing	83	8.95	20.5	38.2
Modified Kopp	83	7.13	20.6	32.3
PL method	83	9.53	26.9	43.6
PF method	83	7.72	23.8	43.9
B. Larger Test Set				
Modified Kopp	127	9.52	23.4	37.5
PL method	127	10.7	26.4	41.6
PF method	127	8.47	22.3	40.3

predictor of C_p^s above 250 K with an AAPD of 11.5 % compared to 14.3% for the PL method, and we recommend this method for predictions above 250 K.

While there are not currently any other estimation methods for organic solid heat capacities that can be applied to the wide range of temperatures to which the proposed methods can be compared, there are two commonly-used methods for predicting C_p^s at 298 K which are available for comparison purposes: the modified Kopp's rule[1] and the Domalski-Hearing[3] method. We have compared estimations from these two methods to those made at 298 K using the PL and PF methods in Table 2.5. The comparison test set included 127 compounds obtained from the DIPPR 801 database for compounds with experimental data within 10 K of 298 K. Although these compounds were in our training

Table 2.6. Comparison of C_p^s predictive methods between 50 - 150 K

	# Compounds	# Data Points	AAPD (%)	AAD (J/mol·K)	SD (J/mol·K)
A. Alkanes					
Kabo et al. #1	46	627	3.34	4.57	10.7
PL Method	46	627	5.14	6.01	11.2
PF Method	46	627	4.69	5.78	11.0
B. Alkanes, Alkenes, Alkanols. and Alkanones					
Kabo et al. #2	87	962	3.33	3.45	10.7
PL Method	87	962	5.54	5.51	9.89
PF Method	87	962	5.64	5.63	9.99

set, it is highly likely that they were also in the training set used to obtain the Domalski-Hearing group values owing to the limited C_p^s data available. The test set for part A of Table 2.5 is a subset of the available data which includes 83 compounds for which Domalski-Hearing group values are available; part B shows the results for all 127 compounds for the three methods that can be used for this entire test set. The PL and PF methods compare well with the methods developed exclusively for use at 298 K, but they have the added capability of predicting the temperature dependence of C_p^s over a large range. As noted above, the PF method is preferred at temperatures above 250 K. Additionally, Kabo et al.[6] have presented two simple additive correlations for estimating the heat capacities of some types of solid organics at 10 K increments from 10- 150 K. A comparison between these methods and the proposed methods for alkanes, monoalkenes, monoalkanols, and monoalkanones between 50 - 150 K from the DIPPR database is

presented in Table 2.6. As can be seen, the methods of Kabo et al. are more accurate within their limited scopes and their use is recommended where they apply.

Solid Heat Capacity Summary

Two group-contribution methods have been developed to predict heat capacities of organic solids at ambient pressure. The PL method utilizes an empirical temperature dependence based on a power-law expression observed for solid hydrocarbons and ionic crystals. The PF method is based on the Einstein partition function for crystals and the Debye idea of using a temperature-dependent vibration distribution function. Whereas Debye used a quadratic temperature dependence, the power to which the temperature in the frequency distribution is raised has been allowed to be optimized for the whole training set of organic compounds available to us from the DIPPR 801 database. Both methods then have a fixed or universal temperature functionality. At this time, this a necessary constraint because of the relatively small amount of C_p^s data available. Results suggest that some flexibility in this temperature dependence for families of compounds, perhaps correlation of it with molecular descriptors, might be an avenue for improvement of the methods when additional data are available. The compound-specific constants in the PL and PF methods have been correlated primarily in terms of first-order structural group contributions, but radius of gyration was also found to be a significant correlating property for the PF method.

The methods developed in this study fill an important gap in predictive capabilities for organic solid properties. Both methods correlate the training set within an

AAPD of about 8%. From the limited evaluations on other test data sets that have been performed, an average accuracy of approximately 13% is estimated for the two methods. However, the PL method is expected to be slightly more accurate than this at lower temperatures. The PF method, on the other hand, is expected to have this accuracy or better at temperatures above 250 K, with reduced accuracy below 250 K. Neither method is recommended below 50 K. The simpler PL method for temperatures is recommended between 50 and 250 K or when quick estimates are needed, but use of the PF method is recommended for temperatures above 250 K.

For ionic compounds, the use of Mostafa et al.[5] is recommended. This method is simple to use, has excellent temperature dependence built in, and has an expected accuracy of better than 10 % for most compounds.

References

(1) J. E. Hurst and B. K. Harrison, "Estimation of Liquid and Solid Heat Capacities using a Modified Kopp's Rule," *Chem. Eng. Comm.*, 1992, **112**, 21-30.

(2) F. H. Brock, "Estimation of Specific Heats at Normal Temperatures," *ARS J.*, 1961, **31**, 265-268.

(3) E. S. Domalski and E. D. Hearing, "Estimation of the Thermodynamic Properties of C-H-N-O-S-Halogen Compounds at 298.15 K," *J. Phys. Chem. Ref. Data*, 1993, **22**, 805-1159.

(4) O. Kubaschewski and H. Ünal, "An empirical estimation of the heat capacities of inorganic compounds," *High Temp. - High Press.*, 1977, **9**, 361-365.

(5) A. T. M. G. Mostafa, J. M. Eakman, M. M. Montoya, and S. L. Yarbo, "Prediction of Heat Capacities of Solid Inorganic Salts from Group Contributions," *Ind. Eng. Chem. Res.*, 1996, **35**, 343-348.

(6) G. J. Kabo, A. A. Kozyro, and V. V. Diky, "Additivity of Thermodynamic Properties of Organic Compounds in the Crystalline State. 1. Additive Calculations for Thermodynamic Properties of Alkanes, Alkenes, Alkanols, and Alkanones," *J. Chem. Eng. Data*, 1995, **40**, 160-166.

(7) G. J. Kabo, A. A. Kozyro, V. V. Diky, and V. V. Simirsky, "Additivity of Thermodynamic Properties of Organic Compounds in Crystalline State. 2. Heat Capacities and Enthalpies of Phase Transition of Alkyl Derivatives of Urea in Crystalline State," *J. Chem. Eng. Data*, 1995, **40**, 371-393.

(8) V. V. Diky, A. A. Kozyro, and G. J. Kabo, “Additivity of Thermodynamic Properties of Organic Compounds in the Crystalline State. 3. Heat Capacities and Related Properties of Urea Phenyl Derivatives,” *J. Chem. Eng. Data*, 2002, **47**, 239-244.

(9) L. P. Guilyazetdinov, “Structural Group Composition and Thermodynamic Properties of Petroleum and Coal Tar Fractions,” *Ind. Eng. Chem. Res.* 1995, **34**, 1352-1363.

(10) D. A. McQuarrie, *Statistical Mechanics*, Harper & Row: New York, 1973.

(11) A. Bondi, *Physical Properties of Molecular Crystals, Liquids, and Glasses*, John Wiley and Sons: New York, 1968.

(12) K. G. Joback and R. C. Reid, “Estimation of Pure-Component Properties from Group-Contributions,” *Chem. Eng. Comm.* 1987, **57**, 223-243.

(13) *TSAR Version 3.2*, Oxford Molecular Group, Oxford Molecular Limited: Oxford, 1998.

(14) D. Wininger, “SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules,” *J. Chem. Inf. Comput. Sci.* 1988, **28**, 31-36.

(15) “SMILES Tutorial,” www.daylight.com, 1998.

(16) *TRC Thermodynamic Tables-Hydrocarbons*, 1998, Volume XIII, Table vc, Thermodynamics Research Center, The Texas A & M University System, College Station, TX; *TRC Thermodynamic Tables-Non-Hydrocarbons*, ibid.

CHAPTER 3 - HEAT OF SUBLIMATION AND SOLID VAPOR PRESSURE

Relationship between Heat of Sublimation and Solid Vapor Pressure

Heat of sublimation and solid vapor pressure are closely related through the Clapeyron Equation

$$\frac{dP_v^s}{dT} = \frac{\Delta H_s}{T\Delta V_s} \tag{3.1}$$

where ΔV_s is the change in volume upon sublimation, and are therefore considered together in this chapter. Alternatively, Eq. 3.1 can be recast in terms of the change in compressibility factor upon sublimation, ΔZ_s, as

$$\frac{d\ln P_v^s}{d(1/T)} = -\frac{\Delta H_s}{R\Delta Z_s}\,. \tag{3.2}$$

The compressibility factor of the solid is much smaller than that of the gas, and because the vapor pressures of solids are so low, the compressibility factor for the saturated vapor is very close to unity. Equation 3.2 then simplifies to the Clausius-Clapeyron equation

$$\frac{d\ln P_v^s}{d(1/T)} = -\frac{\Delta H_s}{R}\,. \tag{3.3}$$

As liquid vapor pressure (P_v^l) is often better known than P_v^s and there is an equilibrium between the three phases at the triple point ($P_v^s = P_v^l$), the triple point is a

useful reference point for the integration of Eq. 3.3 to obtain

$$\ln \frac{P_v^s}{P_{TP}} = -\frac{\Delta H_s}{R}\left(\frac{1}{T} - \frac{1}{T_{TP}}\right) . \qquad \textbf{(3.4)}$$

Equation 3.4 gives the value of P_v^s at a given temperature as long as ΔH_s, the triple point temperature (T_{TP}), and pressure (P_{TP}) are known. It is assumed in the integration of Eq. 3.3 that ΔH_s is independent of temperature over the range T to T_{TP}. Examination of the difference between the solid and vapor heat capacities for a few compounds suggests that the change in ΔH_s over an 80 K range is small (less than the error introduced later in correlating ΔH_s), and to a very good approximation, ΔH_s can be taken as a constant over a fairly large temperature range. Prediction of P_v^s therefore becomes a matter of accurately predicting ΔH_s, given that P_{TP} is commonly available.

Previously Published Prediction Methods

Bondi developed a group-contribution method for estimating ΔH_s at the lowest condensed phase transition.[1,2] Where a molecule only has one solid phase, this is the triple point. Bondi uses this lowest phase transition because group contributions are more reliable there. However, if a method could be developed that was based on the triple point, this would be preferable for most engineering applications, as solid-liquid unit processes more often involve the solid phase stable at the triple point. Considerably more data are available now than were available when Bondi developed his method in 1963, making it possible to extend the available palette of functional groups and thus the number of compounds for which such a method can be used. The method developed in

this dissertation is similar to Bondi's method, but more extensive with regards to organic compounds.

Mackay et al. developed four methods, all similar, for predicting vapor pressures based on the normal boiling point (T_b) of a compound.[3] The four methods are the Trouton constant ΔH_v (TCH), the Kistiakowsky constant ΔH_v (KCH), the Trouton linear ΔH_v (TLH), and the Kistiakowsky ΔH_v (KLH). The Trouton methods are based on Trouton's rule that the ratio of the heat of liquid vaporization (ΔH_v) to T_b is a universal constant. The Kistiakowsky methods are similar, but include a factor based on T_b in the ratio. The constant methods assume that ΔH_v does not change with temperature, but the linear methods assume that ΔH_v has a linear temperature dependence. While these methods were primarily tested with liquids, they can be used with solids if an extra term based on the melting point (T_f) is added. These methods have the advantage of requiring only the normal boiling point and the melting point (or T_{TP}) as input, properties usually commonly available. Of these four methods, the KLH has the most robust design and is expected to be superior.

Neau et al. used the Peng-Robinson equation of state to estimate P_v^l as a function of temperature, from which they were able to calculate ΔH_v at the triple point.[4] ΔH_s at the triple point was then obtained from a known value of the heat of fusion (ΔH_f) and the estimated ΔH_v using

$$HSUB(TPT) = HVP(TPT) + HFUS(TPT)\,. \qquad \textbf{(3.5)}$$

The Peng-Robinson equation of state requires the critical temperature, critical pressure, and acentric factor. Because experimental values for the critical constants and the vapor

pressure curve, hence the acentric factor, may be unavailable for compounds that are solids at normal operating conditions, Neau et al. recommend using the group-contribution methods of Constantinou and Gani to estimate critical constants[5] and that of Constantinou et al. to estimate the acentric factor.[6] They also reported an alternative method of supplying the parameters used by the equation of state that requires the normal boiling point, group contributions, a "shape factor," and van der Waals volumes instead of critical constants and the acentric factor.[7] Neau et al. preferred this second method to the standard Peng-Robinson equation and claimed that it can be used with "hydrocarbon, ethylenic and sulphured compounds," but did not give any details, referring only to additional publications unavailable to this researcher.[8,9] This method, as explained by Coniglio et al.,[7] can be used with alkanes, aromatics, alkenes, esters, alcohols, and carboxylic acids. Neau et al. recommend the use of a method reported by Avaullée et al.[10] to estimate T_b if it is not known. The biggest drawback to this method is that the preferred method (using the "shape factor") is limited in the number of compounds to which it can be applied. The error greatly increases if the critical constants must be predicted.

While this brief summary of the methods available for predicting ΔH_s and P_v^s shows that there are rudimentary methods for coarse estimations of these two properties, it is difficult to assign an accuracy to the resultant estimations. In this chapter an attempt is made to improve upon these methods, and a more complete evaluation of their accuracy as a comparison standard for the methods developed here is made.

The Partition Function

Theoretically, once the partition function has been developed for the crystalline phase (Eq. 2.4), it can be used to obtain other properties in addition to the application to solid heat capacity as outlined in Chapter 2. Solid vapor pressure can be indirectly related to the partition function through chemical potential,

$$-\frac{\mu}{kT} = \left(\frac{\partial \ln Q}{\partial N}\right)_{T,V} . \tag{3.6}$$

Since the solid and vapor phases are in equilibrium, the chemical potentials must be equal. The vapor phase can be considered an ideal gas because the solid vapor pressure is usually very low. The ideal gas chemical potential for non-linear molecules is

$$-\frac{\mu}{kT} = \ln\frac{kT}{P} - \sum_{\nu} \ln\left(1 - e^{-\theta_\nu/T}\right) + \ln\left[\frac{1}{\sigma}\left(\frac{\pi T^3}{\theta_A\theta_B\theta_C}\right)^{1/2}\right] - \frac{3}{2}\ln\left(\frac{h^2}{2\pi mkT}\right) \tag{3.7}$$

where θ_ν are the ideal gas vibrations summed over all ν, θ_A, θ_B, and θ_C are moments of inertia, σ is the molecular symmetry number, and m is the mass of the molecule. By equating Eqs. 3.6 and 3.7 and substituting Eq. 2.4 for the partition function, one can solve for the pressure, P, to obtain for P_ν^s

$$\begin{aligned}\ln P_\nu^s = {} & \frac{N_a U_0}{2RT} + 2.55 N_a X_G^{-0.85}\int_0^{X_G}\left(\frac{\ln\left(1-e^{-x}\right)}{x^{0.15}} + \frac{x^{0.85}}{2}\right)dx - \ln\left(\frac{h^3}{(2\pi m)^{3/2}(kT)^{5/2}}\right) \\ & - \sum_{\nu}\ln\left(1-e^{-\theta_\nu/T}\right) + \ln\left[\frac{1}{\sigma}\left(\frac{\pi T^3}{\theta_A\theta_B\theta_C}\right)^{1/2}\right].\end{aligned} \tag{3.8}$$

Using the Clausius-Clapeyron equation (Eq. 3.3) with Eq. 3.8 gives an equation for heat of sublimation for non-linear molecules

$$\frac{\Delta H_s}{R} = -\frac{N_a U_0}{2R} + \frac{867}{400}\frac{N_a T}{X_G^{17/20}}\int_0^{X_G}\frac{\ln\left(1-e^{-x}\right)}{x^{3/20}}dx - \frac{51}{74}N_a\Theta_G - \frac{51}{20}N_a T\ln\left(1-e^{-X_G}\right) + \sum_\nu \frac{\theta_\nu}{e^{\theta_\nu/T}} + 4T\,. \tag{3.9}$$

Due to the assumption that Θ_G represents a hybridization of the molecular and atomic vibrations, U_0 does not have the same meaning that it does in the classical Einstein canonical partition function. Part of the zero point energy is contained in the third term of Eq. 3.9, i.e.,

$$U_{0,\text{classic}} = N_a U_0 + \frac{51}{37}N_a\Theta_G R\,. \tag{3.10}$$

It is to be expected that the biggest contribution to ΔH_s would come from the zero point energy and this is the case. A regression to find U_0 using Θ_G from C_p^s (Eq. 2.1) and ΔH_s values extracted from the DIPPR database using Eq. 3.5 showed that the terms in Eq. 3.10 had the largest influence on determining ΔH_s using Eq. 3.9. The temperature-dependent terms had comparatively little influence.

In applying the U_0 derived from ΔH_s (Eq. 3.9) and the Θ_G derived from C_p^s (Eq. 2.14) to P_v^s (Eq. 3.8), it was determined that Eq. 3.8 was too sensitive to U_0. Small changes in U_0 (within the uncertainty expected from Eq. 3.9) would cause unexpectedly large errors in P_v^s. A rearrangement of Eq. 3.8 shows that P_v^s is an exponential function of U_0 and Θ_G,

$$P_v^s = \left[\frac{h^3}{(2\pi m)^{3/2}(kT)^{5/2}}\right]\left[\prod_v \left(1 - e^{-\theta_v/T}\right)\right]\left[\sigma\left(\frac{\theta_A\theta_B\theta_C}{\pi T^3}\right)^{1/2}\right] \times \exp\left[\frac{N_a U_0}{2RT} + \frac{51}{74}\frac{N_a \Theta_G}{T} + 2.55 N_a X_G^{0.85}\int_0^{X_G}\frac{\ln\left(1 - e^{-x}\right)}{x^{0.15}}dx\right], \quad \textbf{(3.11)}$$

where the pre-exponential term has between 10 and 13 orders of magnitude. A 5% change in the $U_0/2R$ of *n*-hexane, which itself has four orders of magnitude, changes P_v^s by 6 orders of magnitude. The C_p^s correlation developed in Chapter 2 provided a method for prediciting Θ_G (and the related X_G), but P_v^s is insensitive to the last term within the exponent of Eq. 3.11. Since both U_0 and Θ_G must be obtained through group contributions and can be combined into one parameter through Eq. 3.10, it makes little sense to keep them separate and retain the complexity of Eq. 3.11 solely for the very small contribution of the integral term of that equation. The rest of this chapter instead presents a method similar to Bondi's that correlates ΔH_s directly.

Correlation of the Heat of Sublimation

Although the DIPPR database does not contain raw data specifically for ΔH_s, the value recommended in the database for the heat of fusion (ΔH_f) at the melting point and the evaluated correlation of the liquid heat of vaporization (ΔH_v), as a function of temperature (from the triple point to the critical point) can be used to obtain ΔH_s. For most compounds, the triple-point temperature (T_{TP}) and the normal melting point are very similar, and we can obtain ΔH_s using Eq. 3.5.

A training set of ΔH_s values at the triple point found using Eq. 3.5 for 218 organic compounds was used in a QSPR scheme similar to that used previously in developing the correlation for C_p^s. As in the heat-capacity work, the significant correlating factors for ΔH_s were functional groups and the radius of gyration, r_g. Additionally, quadratic terms for the two most common groups, methylene and aromatic carbon atoms, and correction terms for multiple halogens based on the fraction of carbon or silicon terminal valences

Table 3.1. Linear group values for ΔH_s

Group	**Description**	**Example**	**SMILES formula**	d_i **Eq. 3.12**
–CH_3	Methyl	n-butane	**C**CCC	736.5889
>CH_2	Methylene	1-heptanol	OC**C**CCCCC	561.3543
>CH–	Secondary carbon	2,3-dimethylpentane	C**C**(C)C(C)CC	111.0344
>C<	Tertiary carbon	2,2-dimethylbutane	C**C**(C)(C)CC	-800.517
CH_2=	Terminal alkene	1-octene	**C**=CCCCCCC	572.6245
–CH=	Alkene	1,3-butadiene	C=**C**C=C	541.2918
>C=	Substituted alkene	Isobutene	C=**C**(C)C	117.9504
Ar –CH=	Aromatic carbon	Biphenyl	c1**c**cccc1(c2ccccc2)	626.7621
Ar >C=	Substituted aromatic C	Toluene	c1ccccc1(C)	348.8092
Ar –O–	Furan oxygen	Furan	C1=C**O**C=C1	763.284
Ar –N=	Pyridine nitrogen	Quinoline	c1(ccc**n**2)c2cccc1	1317.056
Ar –S–	Thiophene sulfur	Thiophene	**S**1C=CC=C1	911.2903
–O–	Ether	Dimethyl ether	C**O**C	970.4474
–OH	Alcohol	1-pentanol	CCCCC**O**	3278.446
–COH	Aldehyde	1-butanal	CCC**C=O**	2402.093
>C=O	Ketone	3-hexanone	CC**C(=O)**CCC	1816.093
–COO–	Ester	Methyl methacrylate	C=C(C)**C(=O)O**C	2674.525
–COOH	Acid	*n*-butyric acid	CCC**C(=O)O**	5006.188
–NH_2	Primary amine	Methylamine	C**N**	2219.148
>NH	Secondary amine	Piperidine	C1CCCC**N**1	1561.222
>N–	Tertiary amine	Trimethylamine	C**N**(C)C	325.9442
–NO_2	Nitro	Nitrobenzene	c1(**N(=O)=O)**ccccc1	3661.233
–SH	Thiol/mercaptan	*n*-hexyl mercaptan	CCCCCC**S**	1921.097
–S–	Sulfide	Diethyl sulfide	CC**S**CC	1930.84
–SS–	Disulfide	Di-*n*-propyl disulfide	CCC**SS**CCC	2782.054
–F	Fluoride	Benzotrifluoride	c1(C(F)**(F)**F)ccccc1	626.4494
–Cl	Chloride	Ethyl chloride	CC**[Cl]**	1243.445
–Br	Bromide	Bromobenzene	c1**(Br)**ccccc1	669.9302
>Si<	Silane	Tetramethylsilane	C**[Si]**(C)(C)C	-83.7034
>Si(O–)–	Siloxane	Hexamethyldisiloxane	C[Si](C)(C)**O[Si]**(C)(C)C	-16.0597

Table 3.2. Nonlinear and halogen group values for ΔH_s

Group	Description	Eq. 3.12
A. Nonlinear terms		f_i
>CH_2	Methylene	9.5553
Ar =CH–	Aromatic carbon	-2.21614
B. Halogen fraction terms		g_i
–Cl	Cl fraction	-1543.66
–F	F fraction	-1397.4
–Br	Br fraction	5812.49

occupied by halogen atoms were also found to be significant in the correlation. Using the multiple regression package in Oxford Molecular Tsar 3.2,[11] the following was obtained

$$\frac{\Delta H_s(T_{TP})}{R} = 6.98.04 + 3.838 \times 10^{12} r_g + \sum_i^{NG} d_i n_i + \sum_i^{NG} f_i n_i^2 + \sum_i^{NG} g_i \frac{n_i}{n_X} \qquad \textbf{(3.12)}$$

where d_i, f_i, and g_i are values for group i regressed from the training set, n_i is the number of times that group i appears in the molecule for all NG number of groups, and n_X is the total number of all halogen and hydrogen atoms attached to C and Si atoms in the molecule. Values of the radius of gyration, r_g, should be entered in meters and are obtainable from several sources, including the DIPPR 801 database.

Tables 3.1 and 3.2 contain the values of the group contributions obtained from the regression. Linear groups are given in Table 3.1; the nonlinear terms for methylene and aromatic carbon groups and the correction terms for the halogen fractions are given in

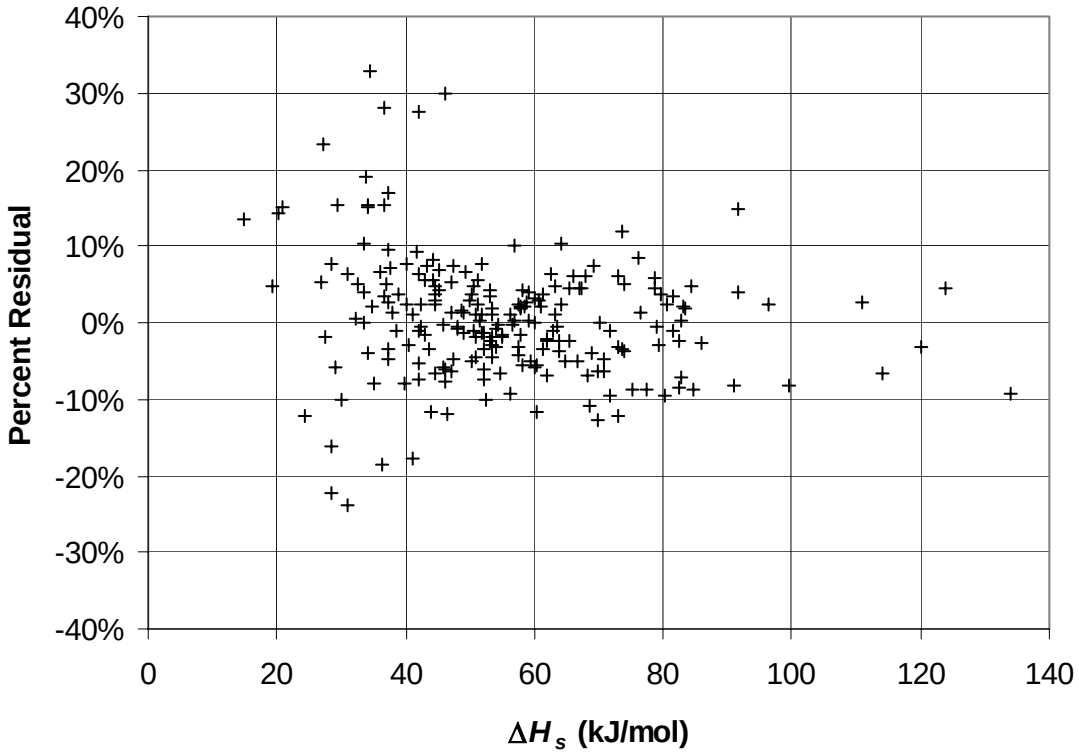

Figure 3.1. Percent residual of ΔH_s for the 218 compounds of the training set.

Table 3.2. Table 3.1 also illustrates group definitions using SMILES formulas in bold typeface in the same manner as in Table 2.1.

Equation 3.12 has an average absolute deviation (AAD) of 3.01 kJ/mol, an average absolute percent deviation (AAPD) of 5.89%, and an R^2 value of 95.8% with respect to the training set. A plot of the percent residuals versus the ΔH_s from the training set is shown in Figure 3.1. Due to the small quantity of ΔH_s data available, we chose to validate Eq. 3.12 with an independent P_v^s data set rather than holding back data from the training set for testing extrapolation of the correlation to additional compounds.

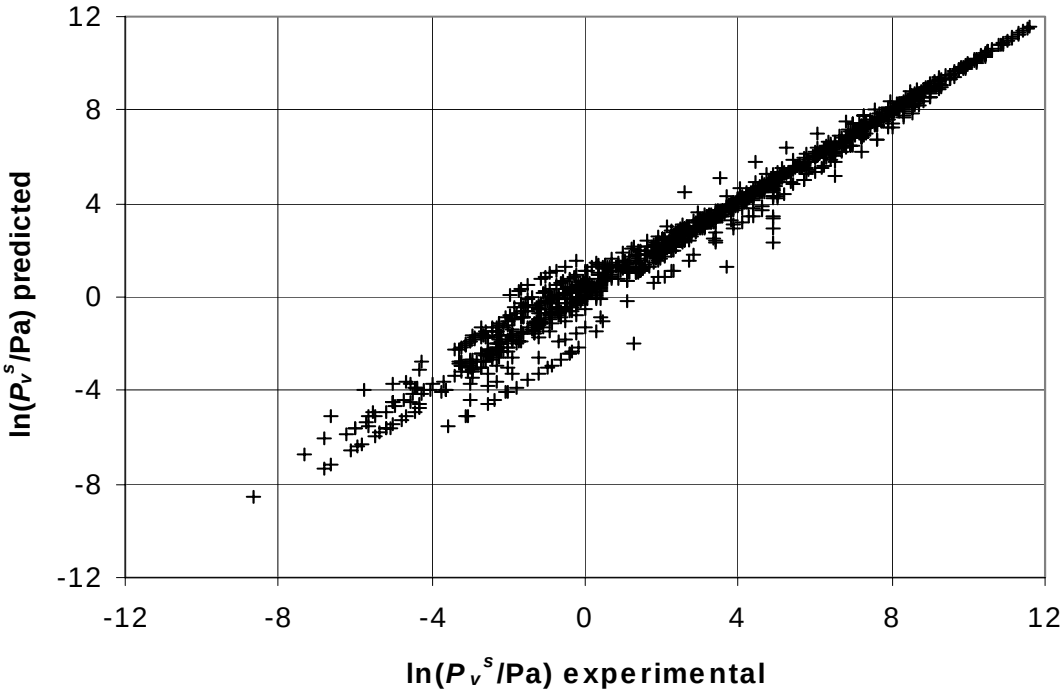

Figure 3.2. Predicted vs. experimental ln(P_v^s/Pa) for the 87 compounds (1103 data points) of the test set.

Estimating Solid Vapor Pressure

Solid vapor pressures for 87 compounds (1103 separate data points) were computed using Eq. 3.4, with ΔH_s calculated from Eq. 3.12, and the resultant values were compared to experimental values from the DIPPR database. Figure 3.2 shows the predicted values versus the experimental values of the test set on a logarithmic scale. As P_v^s data span several orders of magnitude, we used the average absolute logarithmic deviation (AALD), defined as

$$AALD = \frac{\sum_{i=1}^{n} \left| \ln\left(P^s_{v\ pred,i}\right) - \ln\left(P^s_{v\ exp,i}\right) \right|}{n} \qquad \textbf{(3.13)}$$

for a comparison statistic, where n is the number of data points, $P^s_{v\,pred,i}$ is the predicted value of P^s_v at a specific temperature, and $P^s_{v\,exp,i}$ is the experimental value at that temperature. The AALD for this test set was 0.371. This AALD corresponds to errors in the actual vapor pressure of 4.49×10^{-3} Pa, 0.449 Pa, and 44.9 Pa at nominal values of 0.01 Pa, 1 Pa, and 100 Pa, respectively. Figure 3.3 shows the performance of this prediction method for three compounds over a range of temperatures.

Comparison of the Correlations

To appraise the value of Eq. 3.12, we compared P^s_v values obtained from it with those calculated from the earlier methods discussed in the previous section of this chapter

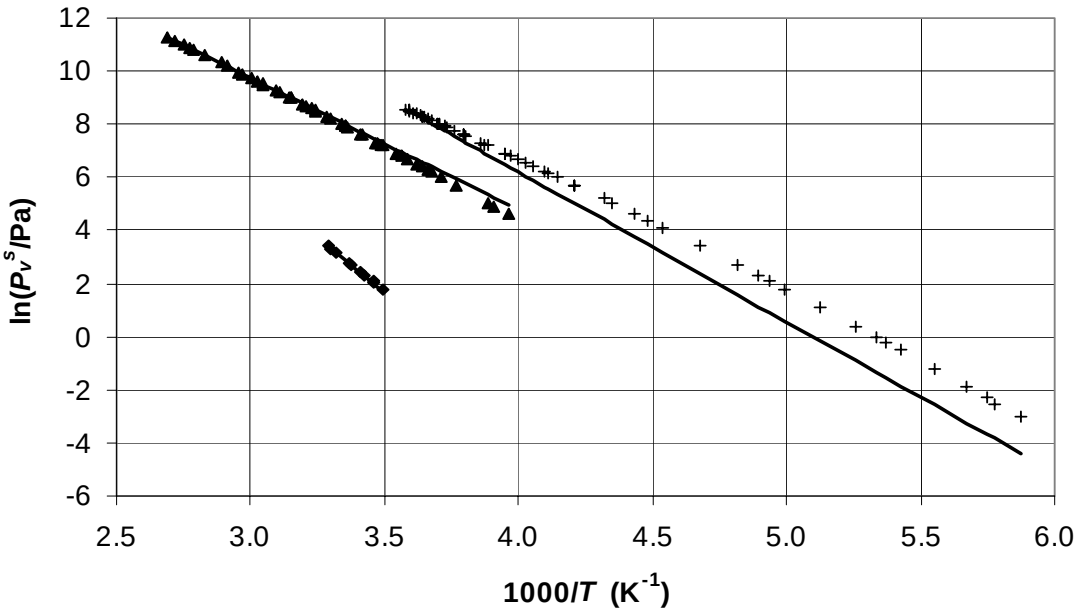

Figure 3.3. Experimental (–) and predicted P^s_v for 1,2,3-trichlorobenzene (◆), 2,2,3,3-tetramethylbenzene (▲), and cyclohexane (+).

using the same P_v^s test set, or a subset thereof as appropriate for the model examined. Table 3.3 contains a summary of this evaluation.

Solid vapor pressures were calculated from Bondi's ΔH_s correlation using Eq. 3.4 and the assumption that ΔH_s is independent of temperature. The entire test set could not be used because Bondi's correlation and Eq. 3.12 do not share the same functional group building blocks, so a subset of the 87 compounds was used. While both methods are targeted toward organic compounds, Bondi's method includes more inorganic groups while Eq. 3.12 has a more extensive palette of organic groups. As shown in Table 3.3, the

Table 3.3. Comparison of P_v^s prediction methods

Method	Compounds	Points	AALD	AALD Eqs. 3.12 & 3.4
A. Bondi				
Bondi	39	591	0.326	0.233
B. Mackay et al.				
TCH	74	947	1.76	0.360
KCH	74	947	1.72	0.360
TLH	74	947	1.01	0.360
KLH	74	947	0.888	0.360
C. Neau et al.				
Experimental critical constants	22	346	1.27	0.273
Group-contribution constants	47	755	3.26	0.340
Experimental boiling point	30	493	1.18	0.347
Group-contribution boiling point	30	490	0.691	0.350

AALD for the method developed here is 0.23 for 591 vapor pressure points for 39 compounds as compared to 0.33 for Bondi's method applied to the same test set.

The four methods Mackay et al. developed to predict vapor pressure were tested against that portion of the test set for which boiling point data were available in the DIPPR database. The results of the comparison with the test set for these methods are also given in Table 3.3. While the AALD of each of these methods is higher than that of Eq. 3.12 for the same subset of the test set, these methods have the advantage of requiring only the normal boiling point and the melting point (or T_{TP}), properties which are commonly available. Of these four methods, KLH is superior as expected.

As there are several ways to apply the methodology of Neau et al., the results of the comparison shown in Table 3.3 are for the four cases of (1) using experimental values of the critical constants and the recommended value of the acentric factor obtained from the DIPPR database, (2) using group contributions[5,6] to estimate these constants, (3) using the alternative method with experimental boiling points from the DIPPR database, and (4) using the alternative method with normal boiling points calculated from the primary group-contribution method[12] used in the DIPPR database. Using Eq. 3.12 to estimate ΔH_s gives values of P_v^s closer to the experimental values than any of these four cases, probably due to assumptions made in using the equation of state to predict ΔH_v.

Figure 3.4 compares experimental $\ln(P_v^s)$ values with those predicted using the method of Eq. 3.12, Bondi, McKay et al. (KLH), and Neau et al. (using both critical constants and T_b) for benzene over a range of inverse temperatures.

P_v^s and ΔH_s Summary

A group-contribution method was developed for estimating the heat of sublimation of organic compounds at the triple point. This method was been applied also to estimating solid vapor pressure through the Clausius-Clapeyron relationship. The accuracy of this method for ΔH_s is similar to Bondi's correlation, but it has additional functional groups. It is consistent with and uses the same form as the heat capacity correlation developed in the previous chapter for organic solids. The temperature

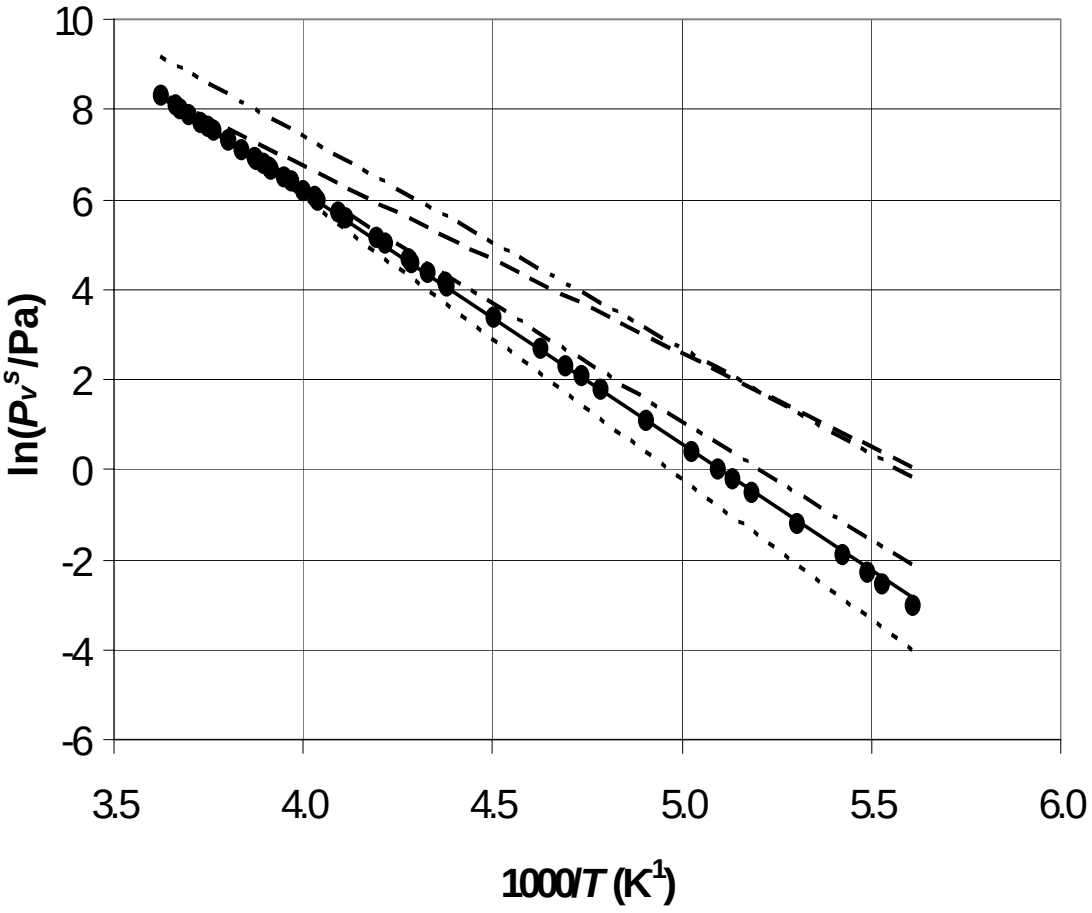

Figure 3.4. Comparison of P_v^s data of benzene for experimental (●), Eqs. 3.12 and 3.4 (–), KLH (- - -), Bondi (– - –), Neau et al. critical constants (– - - –) , and Neau et al. boiling point.

dependence of the solid vapor pressure is obtained from the integrated form of the Clausius-Clapeyron equation in conjunction with known triple-point conditions.

The method was tested against P_v^s data from the DIPPR database. These data were not used in development of the correlation for ΔH_s. The AALD for this comparison was 0.371, marking a substantial improvement over existing methods for predicting P_v^s. While there were no ΔH_s data available to test the correlation, it is assumed that the correlation has an uncertainty of 13% due to the fact that the AAPD for the training set (~6%) was similar to the AAPD of the training set for C_p^s.

References

(1) A. Bondi, "Heat of Sublimation of Molecular Crystals," *J. Chem. Eng. Data*, 1963, **8**, 371-381.

(2) A. Bondi, *Physical Properties of Molecular Crystals, Liquids and Glasses*; Wiley: New York, 1968.

(3) D. MacKay, A. Bobra, D. W. Chan, and W. Y. Shlu, "Vapor Pressures Correlations for Low-Volatility Environmental Chemicals," *Environ. Sci. Technol.* 1982, **16**, 645-649.

(4) E. Neau, S. Garnier, and L. Avaullée, "A consistent estimation of sublimation pressures using a cubic equation of state and fusion properties" *Fluid Phase Equil.* 1999, **164**, 173-186.

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

(6) L. Constantinou, R. Gani, and J. P. O'Connell, "Estimation of the acentric factor and the liquid molar volume at 298 K using a new group contribution method," *Fluid Phase Equil.* 1995, **103**, 11-22.

(7) L. Coniglio, L. Trassy, E. Rauzy, "Estimation of Thermophysical Properties of Heavy Hydrocarbons through a Group Contribution Based Equation of State," *Ind. Eng. Chem. Res.* 2000, **39**, 5037-5048.

(8) L. Coniglio, Thesis, Université d'Aix-Marseille III, 1993.

(9) L. Trassy, Thesis, Université de la Méditerranée, 1998.

(10) L. Avaullée, L. Trassy, E. Neau, and J. N. Jaubert, "Thermodynamic modeling for petroleum fluids - I. Equation of state and group contribution for the

estimation of thermodynamic parameters of heavy hydrocarbons," *Fluid Phase Equil.* 1997, **139**, 155-170.

(11) *TSAR Version 3.2*; Oxford Molecular Group, Oxford Molecular Limited: Oxford, 1998.

(12) D. Ericksen, W. V. Wilding, J. L. Oscarson, and R. L. Rowley, "Use of the DIPPR Database for the Development of QSPR Correlations: Normal Boiling Point," *J. Chem. Eng. Data* 2002, **47**, 1293-1302.

CHAPTER 4 - SOLID DENSITY

Previously Developed Solid Density Prediction Methods

There is a paucity of prediction methods for solid density. The biggest challenge in formulating a prediction method for solid density (ρ^s) is the lack of data. Many fewer data are archived in the DIPPR 801 database for ρ^s than for solid heat capacity or melting point. The paucity of data limits the types of methods that can be used to those with few independent variables. Two methods have been developed for estimating solid density, both of them simple.

Internally, the DIPPR 801 database has used the following relationship between ρ^s and liquid density (ρ^l) developed by the database staff at Penn State in 1983 when not as many experimental data were available,

$$\rho^s(T_{TP}) = 1.17\rho^l(T_{TP}) . \quad \textbf{(4.1)}$$

As the melting point and triple point are usually very similar, T_f may be used in place of T_{TP}.

Girolami[1] presented a very simple method for estimating liquid and solid densities. It was developed as a "back of the envelope" method, and as such does not give a specific temperature dependence or temperature at which it is valid, but the liquid method was tested by the author against liquid densities at room temperature, and so it is

assumed that the solid method is also designed to work best at that temperature. Most likely, the method was not expected to produce sufficiently accurate densities to warrant concern about the temperature dependence. This method is an elemental additive method where elements of the same period (row of the periodic table) have the same additive contribution. The liquid density estimation also includes correction factors to account for hydrogen bonding and ring structures.

Adding Temperature Dependence to the ρ^s-ρ^l Relationship

An examination of the ρ^s data showed that solid density varies linearly with respect to temperature for most compounds. Of the 54 compounds examined with more than 2 data points, 28 linear fits had R^2 values of greater than 99% and 43 had R^2 values greater than 95% with respect to temperature. This linear relationship between ρ^s and temperature suggested that a simple temperature dependence could easily be added to Eq. 4.1.

In addition to adding a temperature dependence to Eq. 4.1, efforts were made to develop a new predictive method using QSPR methods. A group of 40 descriptors (see

Table 4.1. Statistics for regression of ρ^s/ρ^l ratio with 22 descriptors

Statistic	$T/T_{TP} = 0.85$	$T/T_{TP} = 1.0$
R^2	82.9%	81.0%
Cross-validation R^2	41.8%	35.3%
Lowest absolute descriptor t-value	1.61	1.68
Highest descriptor probability of non-significance	11.7%	10.1%

Table 4.2. Descriptors used to generate ρ^s/ρ^l relationship

Descriptor	Description
Radius of Gyration	Distance where a point-mass would have equivalent moment of inertia (in meters)
Van der Waals Volume	Volume and surface area of molecule calculated by bond distances and non-bonded contact radii
Van der Waals Area	
Symmetry Number	Number of identical positions the backbone of the molecule can rotate to achieve - assume spherical molecules have 100, conical have 20, and cylindrical have 10
Number of Rigid Atoms	Number of atoms in the plane of the molecule
Number of Branch Atoms	Number of atoms initiating an alkyl branch
Number of Backbone Atoms	Number of atoms excluding H and F
Weighted Number of Rigid Atoms	Similar to above, except that S and Br have weight 2 and I has weight 4
Weighted Number of Branch Atoms	
Weighted Number of Backbone Atoms	
Molecular Mass	From Periodic Table
Molecular Surface Area	Connolly surface area - probe radius 1.4 Å
Molecular Volume	Van der Waals volume
Ellipsoidal Volume	Volume within ellipsoid defined by molecule
log P	$\log_{10}$ octanol-water partition coefficient
Total Lipole	Total lipophilic distribution - calculated from log P and analogous to dipole
Molecular Refractivity	Ratio of speed of light to that in a vacuum
Kier Chi 0 (atoms)	Connectivity indexes - order determines how many atoms are used as the connective base
Kier Chi 1 (bonds)	
Kier Chi 2 (path)	
Kier Chi 3 (cluster)	
Kier Chi 4 (cluster)	

Table 4.2 (continued)

Descriptor	Description
Kier Chi V 0 (atoms)	As above, except that it takes into account the atomic number, rather than just the valence electrons
Kier Chi V 1 (bonds)	
Kier Chi V 2 (path)	
Kier Chi V 3 (cluster)	
Kier Chi V 4 (cluster)	
Kier Chi V 4 (path/cluster)	
Kappa 1 Index	Shape indices - indicate degree of complexity of bonding pattern, degree of linearity, and degree of branching at center, respectively
Kappa 2 Index	
Kappa 3 Index	
Kappa Alpha 1 Index	As above, except that a comparison is made to C sp^3 atoms
Kappa Alpha 2 Index	
Kappa Alpha 3 Index	
Shape Flexibility Index (ϕ)	Flexibility - branching and rings decrease, backbone length increases
Rotatable Bonds	Single order, non-terminal, non-ring, non-amide bond
Randic Topological Index	Topological indices derived from molecular graphs - measure sum of distances, sum of weighted edges, and average-distance sum connectivity, respectively
Balaban Topological Index	
Wiener Topological Index	
Sum of E-State Indices	Electrotopological index - related to Chi indices
Number of Atoms	Number of atoms

Table 4.2) were examined as candidate independent variables for use in a QSPR correlation. The first 3 descriptors in Table 4.2 were taken from the DIPPR 801 database, the next 7 descriptors were counted by hand, and the last 30 descriptors were calculated by Tsar 3.3.[2,3] A correlation for the ratio of ρ^s/ρ^l was regressed at the triple point

temperature using a least-squares method for 59 compounds (303 data points evaluated to have an uncertainty of less than 5%). Tsar's statistical output was examined and the descriptor with the t-value closest to 0 was determined to be the least significant and discarded. The correlation was re-regressed using 39 parameters and the least significant descriptor was again discarded. This process was repeated until removing descriptors was judged to offer no benefit and the order of removal was noted. This process was repeated for T/T_{TP} = 0.0, 0.8, and 0.85. The order of removal of descriptors was different for each value of T/T_{TP}, but was similar for 1.0 and 0.85. The 22 most significant descriptors for T/T_{TP} = 0.85 were chosen as the final descriptor set. These descriptors are shown with their coefficients in Table 4.3. These 22 descriptors corresponded to the 22 most significant descriptors for T/T_{TP} = 1.0 except in 2 cases. Kappa 3 and Kier Chi V 0 ranked as 19 and 20, respectively for T/T_{TP} = 1.0, but were not as significant for T/T_{TP} = 0.85. They were replaced by Kappa Alpha 1 and Kappa Alpha 3 which were ranked 24 and 36, but fell within the top 22 for T/T_{TP} = 0.85. The statistics for the regression with these 22 descriptors for both T/T_{TP} = 0.85 and 1.0 are shown in Table 4.1.

The coefficients in Table 4.3, p_i and q_i, are used with the following equations

$$\rho^s(0.85T_{TP}) = \left(1.339 + \sum_i^{NG} n_i p_i\right)\rho^l(T_{TP}), \quad \textbf{(4.2)}$$

$$\rho^s(T_{TP}) = \left(1.301 + \sum_i^{NG} n_i q_i\right)\rho^l(T_{TP}). \quad \textbf{(4.3)}$$

Equations 4.2 and 4.3 can be used with a simple linear equation to provide full temperature dependence.

Table 4.3. Coefficients for ρ^s/ρ^l descriptor coefficients

Descriptor	p_i (Eq. 4.2)	q_i (Eq. 4.3)
Radius of Gyration	-2.345×10^{9}	-1.808×10^{9}
Number of Rigid Atoms	0.1494	0.1529
Number of Branch Atoms	0.1393	0.1410
Weighted Number of Rigid Atoms	-0.1191	-0.1256
Molecular Mass	1.491×10^{-3}	1.688×10^{-3}
Ellipsoidal Volume	2.625×10^{-4}	2.001×10^{-4}
log P	6.183×10^{-2}	6.689×10^{-2}
Total Lipole	8.120×10^{-3}	7.640×10^{-3}
Molecular Refractivity	-3.439×10^{-2}	-3.668×10^{-2}
Kier Chi 3 (cluster)	0.4378	0.3821
Kier Chi 4 (cluster)	-2.258	-2.055
Kier Chi 4 (path/cluster)	-6.596×10^{-2}	-5.089×10^{-2}
Kier Chi V 1 (bonds)	0.2334	0.2350
Kier Chi V 3 (cluster)	-0.1072	-0.1389
Kier Chi V 4 (cluster)	0.6167	0.6473
Kappa Alpha 1	-4.153×10^{-2}	-3.845×10^{-2}
Kappa Alpha 2	0.3372	0.3005
Kappa Alpha 3	-1.590×10^{-2}	-1.639×10^{-2}
Shape Flexibility (ϕ)	-0.2641	-0.2371
Balaban Topological Index	9.687×10^{-2}	7.155×10^{-2}
Wiener Topological Index	-2.205×10^{-4}	-1.533×10^{-4}
Number of Atoms	1.701×10^{-2}	1.513×10^{-2}

Because the cross-validation R^2 values in Table 4.1 were so low, suggesting that the prediction capability of this method was low, a simple linear extension of Eq. 4.1 was

tested. This extension was created by using Tsar 3.3 to create a least-squares regression using 65 compounds (350 data points) which included the same 59 compounds used above plus small compounds for which not all of the descriptors in Table 4.2 were available. The resultant equation is

$$\rho^s = \left(1.28 - 0.16\frac{T}{T_{TP}}\right)\rho^l(T_{TP}) . \qquad \textbf{(4.4)}$$

A quick examination of Eq. 4.4 shows that it does not reduce to Eq. 4.1 at the triple point (having a factor of 1.12 rather than 1.17). The new value of 1.12 for the ratio of ρ^s/ρ^l at the triple point is an improvement over the older form used by DIPPR. The regression of the new value is based on substantially more data points and includes many new compounds and revised values that were not available at the time that the value of 1.17 was obtained.

Comparison of the Correlations

Deviations from experimental values are compared for a linear fit of the descriptor method (Eqs. 4.2 and 4.3) and the results from the simple method (Eq. 4.4) in Table 4.4 for both the training set and a 117 compound (170 data point) test set. This test set was taken from the DIPPR database under the same conditions as the training set (5% uncertainty or less). These data were not included in the training set because they failed the linearity requirement, mostly because only one temperature was available for that compound. As can be seen from Table 4.4, the descriptor method fits the training set better than the simple extension of the old DIPPR method (Eq. 4.4). However, numerous

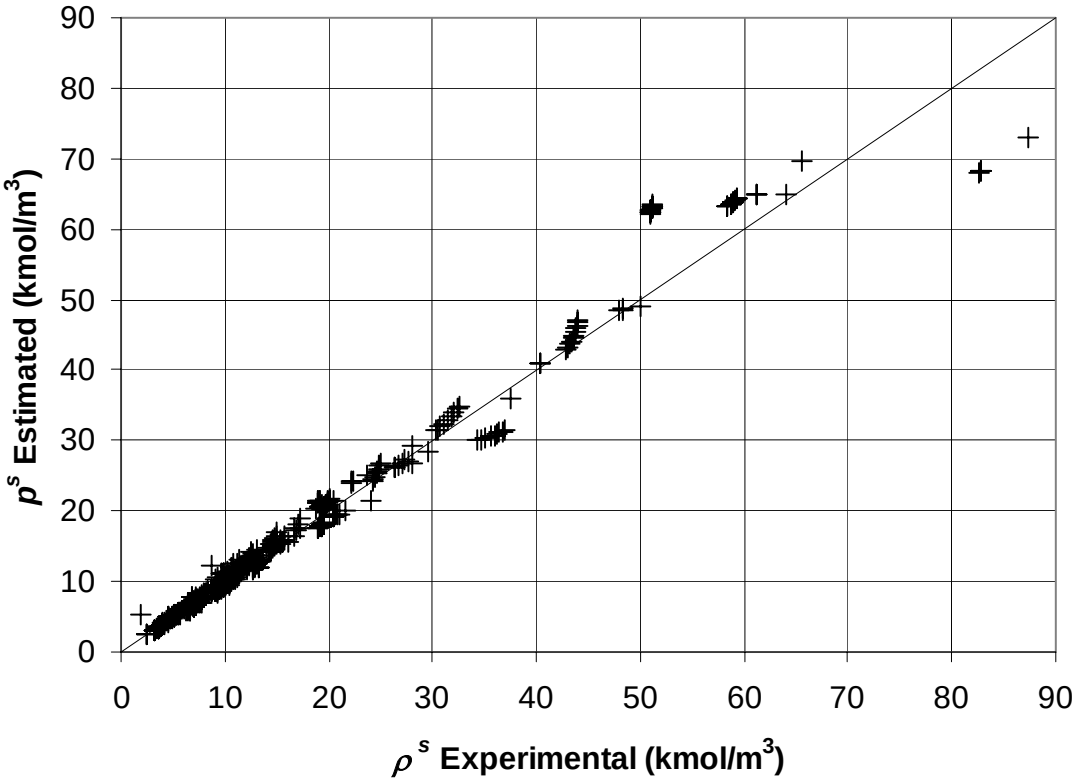

Figure 4.1. Comparison of experimental versus predicted ρ^s values using the simple method (Eq. 4.4)

Table 4.4. ρ^s deviations for training and test sets for descriptor and simple methods

	Descriptor Method (Eqs. 4.2 & 4.3)	Simple Method (Eq. 4.4)
Training Set		
AAD	0.234 kmol/m^3	0.656 kmol/m^3
AAPD	1.82%	4.49%
Test Set		
AAD	1.07 kmol/m^3	0.560 kmol/m^3
AAPD	11.8%	6.28%

independent descriptors were required in order to achieve a reasonable correlation. As the number of independent correlation variables increases for a relatively small training set, there is the danger that the resultant correlation becomes a unique fit to the specific training set and loses its ability to predict other systems. The premise in these equations is that the underlying physics to which the fit is most sensitive are captured by the appropriate descriptors such that extrapolation to other data sets is retained. This is often not the case when so many descriptors must be used, and this appears to be a problem with the correlations developed as Eqs. 4.2 and 4.3 because they do not have the capability of describing the test data set in a predictive mode even to the extent the extended simple DIPPR correlation does. No other method was found during this work that could provide a better extrapolation ability and so the new extended DIPPR method (Eq. 4.4) is proposed here as the preferred method for ρ^s prediction.

A comparison of experimental versus predicted ρ^s values for the simple method is shown in Fig. 4.1. The results of calculations using this same method for 3 compounds are shown with experimental values in Fig. 4.2.

The deviations from experimental values of both the training set and testing set for the simple method are compared with values obtained from the Girolami method and the original DIPPR relationship (Eq. 4.1) in Table 4.5. The comparison with Girolami is at room temperature, which for this purpose is considered 288 - 308 K, and covers 95 compounds (163 data points), 30 of which are from the training set. The comparison with Eq. 4.1 is for the triple point and covers 21 compounds, 18 of which are from the training set. It should be noted that Eq. 4.4 produces a lower deviation than Eq. 4.1 at the triple

Table 4.5. Deviation of ρ^s prediction methods from experimental values

	Multiple Temperatures		Triple Point		Room Temperature	
	AAD ($kmol/m^3$)	AAPD	AAD ($kmol/m^3$)	AAPD	AAD ($kmol/m^3$)	AAPD
Equation 4.4	1.10	5.61%	1.36	4.84%	0.723	6.23%
Equation 4.1			1.80	6.44%		
Girolami					1.17	11.1%

point. As mentioned, the increase in the amount of data in the DIPPR 801 database over the last 20 years is a likely explanation for this improvement.

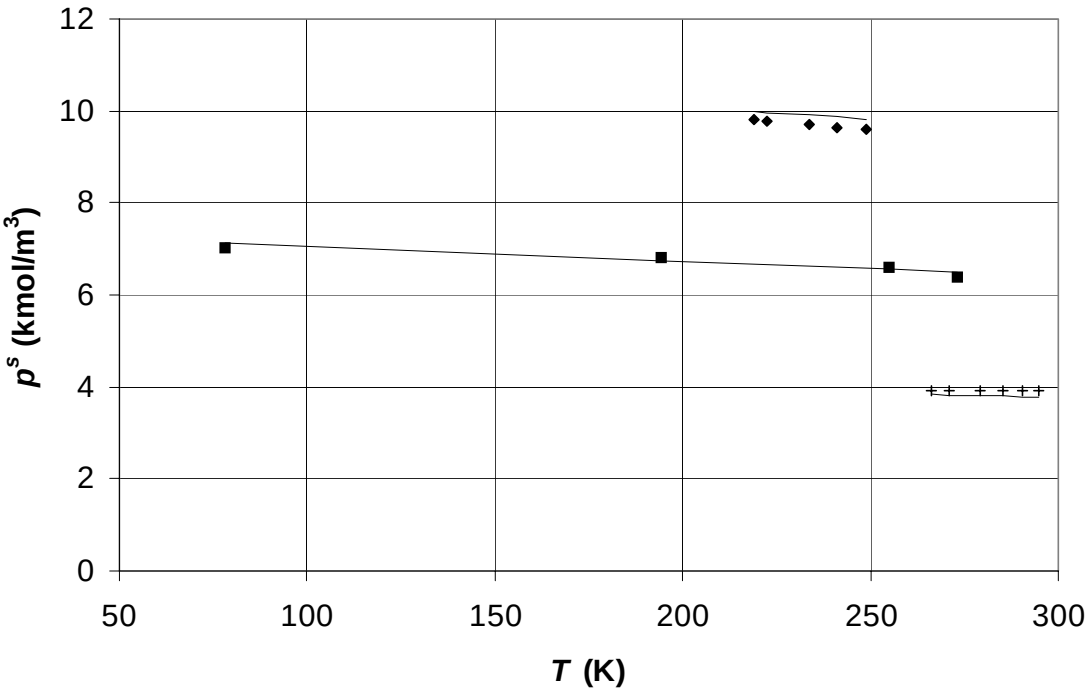

Figure 4.2. Experimental (—) and predicted values of ρ^s for neopentane (◆), *n*-nononoic acid (■), and *n*-hexadecanoic acid (+)

Solid Density Summary

Two new methods are presented for calculating solid density based on the liquid density at the triple point. These methods are an adaption of a method developed internally by the DIPPR 801 project staff to predict solid density at the triple point. The first of these uses 22 descriptors to calculate the density at two points in relation to the triple point. The second uses a simple linear relationship. While the descriptor method correlates very well to the original training set, its prediction power is limited and the simple method is preferred. The simple method adds temperature dependence to the original DIPPR relationship and correlates solid density at the triple point with a smaller deviation from experimental values. Overall, the simple method predicted test set values with an uncertainty of 6.28% (0.560 kmol/m^3).

References

(1) Girolami, G. S. A Simple “Back of the Envelop” Method for Estimating the Densities and Molecular Volumes of Liquids and Solids. *J. Chem. Edu.* **1994**, *71*, 962 – 964.

(2) *TSAR Version 3.3*, Oxford Molecular Group, Oxford Molecular Limited: Oxford, 2000.

(3) *TSAR 3.3 for Silicon Graphics Reference Guide*, Oxford Molecular Limited: Oxford, 2000.

CHAPTER 5 - MELTING POINT

Melting Point Prediction Methods

There are many correlations for determining melting points. These correlations fall into three classifications: classic group contributions, enthalpy/entropy methods, and non-group QSPR methods.

Group-contribution methods are the simplest methods for determining melting points. They are quick and easy to use and are applicable to many compounds, but the results may have greater error than the results of other methods. Joback and Reid[1] developed a first-order method with 40 groups for use with organic compounds that may contain halogens, oxygen, nitrogen, and sulfur. The authors reported an AAPD of 11.2% for this method.[1] Constantinou and Gani[2] reported a combination first-order group contribution with a second-order correction (or overlay) for organic compounds similar to those for which the Joback-Reid method is applicable. The Constantinou-Gani method uses 63 first-order groups and 40 second-order groups. The authors reported slightly smaller deviations than those reported for the Joback-Reid method, having an AAPD of 8.90% for the first-order base and 7.23% by including the second-order overlay (standard deviation: 22.51 K and 18.28 K, respectively).[2] Note that the second-order overlay allows for greater accuracy and the ability to distinguish between isomers, but requires more

time and effort to use. The second-order overlay does not and is not required to account for every atom in the molecule (as the first-order base is). Marrero and Gani[3] updated the Constantinou-Gani method to account for more groups. This method has 165 first-order groups, 115 second-order groups, and 64 "third-order" groups. The only difference between Marrero and Gani's second- and third-order groups is that the third-order overlay is reserved for long chains and fused rings. This method has the advantage of being widely applicable to a large number of compounds due to its large number of groups. However the large number of groups also makes it unwieldy. Marrero and Gani report that their method has an AAPD of 7.5% for the same compounds for which the Joback-Reid produced a 14.6% AAPD.[3]

The melting point can also be determined by dividing the enthalpy of fusion (or melting) by the entropy of fusion ($T_f = \Delta H_f / \Delta S_f$). Methods based on this procedure typically use group contributions to calculate the enthalpy. The entropy is typically correlated from QSPR descriptors for symmetry, flexibility, and sphericity. This type of correlation has been developed for alkanes,[4] other aliphatic compounds,[5,6] and aromatic[7] compounds. Some of these correlations are further limited to only rigid[7] or non-hydrogen-bonding compounds.[4,6] The aromatic correlation has a reported standard error of 37.7 K.[7] The most general of the aliphatic correlations has a reported AAPD of 20% (compared to 34% for the Joback and Reid group method for the same compounds) and root mean square error of 34.4 K.[5] There is also a correlation between melting and boiling points for non-hydrogen-bonding organic compounds, based upon the relationship between entropy and the descriptors representing flexibility and symmetry,

which has a reported root mean square error of 35.5 K.[8] This method is unique among the enthalpy/entropy methods in that it uses T_b rather than group contributions to estimate the enthalpy. One group of researchers[4] noticed that when using correlations to fit the values of enthalpy and entropy for alkanes, the melting point was systematically underestimated by 5.4 K which they corrected by fitting the coefficients directly for the melting point, illustrating the need to check the assumptions inherent in a method. Dearden[9] points out that the method of Simamora et al.[7] used with aromatic compounds is the only method that appears to adequately account for the impact of hydrogen bonding.

Several QSPR correlations have been developed for calculating melting points. These QSPRs are for specific families of compounds, and extrapolation outside of these specific families is not valid. Examples of families used in QSPRs for melting point are: alkanes,[10,11] aldehydes,[12] amines,[12,13] ketones,[12] and benzenes.[14] According to a summary by Katritzky et al.,[14] published QSPR predictive methods range in accuracy from very good (SD = 0.51 K for normal alkanes) to not much better than group contributions (SD = 36.1 K for pyridines). The main disadvantage with QSPR methods is that currently they need to be used within specific families. The reason for this is that there are many factors that influence the melting point temperature. By restricting the method to a specific family, many of those factors are the same and are therefore never correlated. This permits a smaller number of descriptors to be used, but restricts generality because extrapolation outside of the family, where the unknown factors are no longer the same, will produce poor results. The more similar the members of the families are, the better the correlation is, but the more limited extrapolation outside of the training set becomes. For

example, a six-parameter correlation for ortho-substituted benzenes (SD = 28.30 K) has a lower standard deviation than a nine-parameter correlation for all substituted benzenes (SD = 30.19K).[14]

In addition to the methods mentioned above, Chickos and Nichols[15] have developed a method for estimating the melting point of compounds with long alkyl chains based on a "parent compound." For instance, 1-undecene could be estimated based on a known T_f value for 1-pentene. This method distinguishes between even and odd numbered chains because they rarely correlate well with each other. The need for this distinction can readily be seen when T_f values for straight-chain hydrocarbons are plotted versus carbon number. A saw-toothed pattern results with a trend for the even-numbered compounds quite distinct from that for the odd-numbered. This occurs because of the different abilities of the even and odd chains to fit into the crystal lattice. The Chickos-Nichols method can also be applied to other homologous series such as *n*-perfluoroalkanes. The correlation of these series was very high, having an R^2 value of 99% for many of them. The authors have compiled values for many series so that the method may be used immediately in many cases.

Evaluating Melting Point Predictions

In general, the more specific the QSPR, the more accurate it is for the intended compounds, but the less accurate it is for compounds not involved in the training set. As part of this rule of thumb, a QSPR designed for a specific family is usually preferred for members of that family to more general group-contribution methods. However, general

methods are valuable for their range of applicability. Many of the compounds for which predicted data are required for compounds that either do not fit within well-defined families or belong to families for which specific methods are not available. General methods are also more applicable to screening processes.

The purpose of this work has been to create and evaluate general methods, applicable to a wide range of compounds, such that a database like DIPPR 801 can adopt them as part of its general procedures. Therefore, the more specific, family-derived QSPR methods are not dealt with in this dissertation. Instead, the broad group-contributions and enthalpy/entropy methods have been evaluated for use in the DIPPR repertoire of prediction methods. Even so, one should remember the caveat that T_f predictions sought within specific families, such as the *n*-alkanes, will be more accurate if correlations designed specifically for that family are used.

Deviations from experimental values of melting point are shown in Table 5.1 for 11 methods of the group-contribution and enthalpy/entropy types. Two of the methods use overlays to make them more accurate (2nd order for Constantinou and Gani,[2] 2nd and 3rd orders for Marrero and Gani[3]). The test set for these methods comes from the DIPPR 801 database. Melting points in the test set are those experimental values that have an uncertainty of 5% or less. As the different methods have different ranges of applicability, the actual subset used for each method ranges from 72 to 1175 melting points. The four methods described in Yalkowsky et al.[8] are described below

$$T_f = -97.2 + 0.833T_b + 96.7\log_{10}(\sigma) - 3.85RIGID1 - 7.23FLEX1 + 0.21BRANCH1 \quad \textbf{(5.1)}$$

Table 5.1. Deviations in general melting point methods

Method	Compounds	AAD (K)	AAPD	SD (K)	Type
Tsakanikas & Yalkowsky[4]	72	15.7	7.60%	26.4	Alkanes (non-ring, non-infinitely symmetrical)
Simamora et al.[7]	123	47.6	13.1%	86.6	Rigid aromatic compounds
Krzyzaniak et al.[6]	497	30.3	14.5%	43.3	Non-hydrogen-bonding aliphatic compounds
Zhao & Yalkowsky[5]	794	39.9	17.0%	54.5	Aliphatic compounds
Yalkowsky et al.[8] (Eq. 5.1)	918	30.9	14.0%	40.7	Organic compounds with known T_b
Yalkowsky et al.[8] (Eq. 5.2)	918	31.7	14.7%	41.3	Organic compounds with known T_b
Yalkowsky et al.[8] (Eq. 5.3)	918	31.8	14.5%	42.3	Organic compounds with known T_b
Yalkowsky et al.[8] (Eq. 5.4)	918	31.2	14.1%	42.2	Organic compounds with known T_b
Constantinou & Gani,[2] 1st Order	1014	37.6	14.3%	59.3	Organic compounds
Constantinou & Gani,[2] 1st & 2nd Order	1007	34.6	13.5%	52.9	Organic compounds
Joback & Reid[1]	1117	43.0	16.3%	69.5	Organic compounds
Marrero & Gani,[3] 1st Order	1175	34.3	14.1%	52.1	Organic compounds
Marrero & Gani,[3] 1st & 2nd Order	1170	32.4	13.3%	50.5	Organic compounds

$$T_f = -105 + 0.876T_b + 97.9\log_{10}(\sigma) - 5.00RIGID2 - 8.02FLEX2 + 8.03BRANCH2 \quad \textbf{(5.2)}$$

$$T_f = -107 + 0.893T_b + 103\log_{10}(\sigma) - 6.20TOTAL1 \quad \textbf{(5.3)}$$

$$T_f = -125 + 0.959T_b + 107\log_{10}(\sigma) - 7.70TOTAL2 \quad \textbf{(5.4)}$$

where σ is the symmetry number and *RIGID1*, *FLEX1*, *BRANCH1*, and *TOTAL1* are the number of rigid, flexible, branch, and total atoms in the molecule, respectively. The variables ending in "*2*" are analogous to the ones ending in "*1*" except that the former weight the larger atom (Br, S, and I) more heavily. The smallest atoms (H and F) are not included in any of the 4 methods. Of the 4 methods, Yalkowsky et al. report a slightly lower root mean square error for Eq. 5.2 (35.5 K) than for Eq. 5.1 (35.6 K), but they recommend Eq. 5.4 (36.1 K) because it gives a comparable accuracy with two fewer parameters.

Table 5.2. Deviations in organic melting point with a common test set

Method	AAD (K)	AAPD	SD (K)
Yalkowsky et al.[8] (Eq. 5.1)	30.4	13.9%	39.7
Yalkowsky et al.[8] (Eq. 5.2)	31.3	14.7%	40.6
Yalkowsky et al.[8] (Eq. 5.3)	31.4	14.4%	41.5
Yalkowsky et al.[8] (Eq. 5.4)	30.7	14.0%	41.2
Constantinou & Gani,[2] 1st Order	31.3	13.9%	46.9
Constantinou & Gani,[2] 1st & 2nd Order	29.8	13.4%	43.6
Joback & Reid[1]	31.0	13.7%	42.8
Marrero & Gani,[3] 1st Order	29.4	13.6%	43.5
Marrero & Gani,[3] 1st & 2nd Order	27.5	12.6%	41.7
Marrero & Gani,[3] 1st, 2nd, & 3rd Order	27.1	12.5%	40.5

Table 5.2 shows the seven most general methods of Table 5.1, those applicable to most organic compounds, with a common test set of 779 compounds. Table 5.1 is useful for showing the breadth of the methods while Table 5.2 is useful for showing how the methods compare when used for the same compounds.

The lowest AAD and AAPD of Table 5.2 for the common test set of compounds was produced by the Marrero-Gani method. Based on this comparative study, the Marrero-Gani method is given the recommendation as a primary predictive method. The methods of Yalkowsky et al. are given a secondary recommendation to be used when the Marrero-Gani method is too cumbersome. The 344 groups in the full Marrero-Gani method, and even the 165 groups used in the first-order Marrero-Gani method, make the method difficult to use, and one may justifiably be concerned about how many compounds were used in the regression of some of the less-common groups. It would be wise to compare such predictions to those obtained from one of the Yalkowsky methods since they only require the T_b, symmetry number, and 1-3 categories of atomic counts as input data, which ensures that egregious extrapolation errors are not present.

Development of New Melting Point Methods

The purpose of this study was to evaluate all available methods for prediction of solid properties and improve upon those methods or develop new methods where possible. As can be seen from Table 5.2, most of the predictive methods available for melting point produce similar relative errors. As part of this study, brief attempts at developing a new method or improving upon previous methods were also undertaken.

However, no new insights into the physics underpinning the melting point were developed and correlational efforts produced results similar to those shown in Table 5.2 for existing methods. Therefore, the final recommendations for T_f developed in this study are for retention of existing methods and a usage priority as explained in the previous section.

Melting Point Summary

There are many methods for predicting melting points. In general, methods specific to narrowly-defined families provide the most accurate predictions. These methods are typically of the non-group QSPR variety. More general methods are needed when no specific method is available for the compound or the compound does not fit the narrow description type for which the correlation was developed.

General methods for estimating melting points fall into the classical group-contributions or the enthalpy/entropy category with the group-contributions methods covering the broadest classifications. Of the broadest methods examined, those that cover most organic compounds, the group-contribution method of Marrero and Gani produced the smallest deviations from experimental values for a 779 compound test set (27.1 K AAD and 12.5% AAPD). The large number of distinct groups makes the method somewhat difficult to use and it may lead to inaccuracies in extrapolation when groups are involved that were regressed from only a few compounds. Based on this study, the Marrero-Gani method is recommended as a primary method with the methods of

Yalkowsky et al. as secondary and comparison methods (30.4 K AAD and 13.9% AAPD).

References

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

(2) L. Consantinou and R. Gani, "New Group Contribution Method for Estimating Properties of Pure Compounds," *AIChE J.* 1994, **40**, 1697-1710.

(3) J. Marrero and R. Gani, "Group-contribution based estimation of pure component properties," *Fluid Phase Equil.* 2001, **183**, 183-208.

(4) P. D. Tsakanikas and S. H. Yalkowsky, "Estimation of Melting Point of Flexible Molecules: Aliphatic Hydrocarbons," *Toxic. Environ. Chem.* 1988, **17**, 19-33.

(5) L. Zhou and S. H. Yalkowsky, "A Combined Group Contribution and Molecular Geometry Approach for Predicting Melting Points of Aliphatic Compounds," *Ind. Eng. Chem. Res.* 1999, **38**, 3581-3584.

(6) J. F. Krzyzaniak, P. B. Myrdal, P. Simamora, and S. H. Yalkowsky, "Boiling Point and Melting Point Prediction for Aliphatic, Non-Hydrogen-Bonding Compounds," *Ind. Eng. Chem. Res.* 1995, **34**, 2530-2535.

(7) P. Simamora, A. H. Miller, and S. H. Yalkowsky, "Melting Point and Normal Boiling Point Correlations: Applications to Rigid Aromatic Compounds," *J. Chem. Inf. Comput. Sci.* 1993, **33**, 437-440.

(8) S. H. Yalkowsky, J. F. Krzyzaniak, and P. B. Myrdal, "Relationships between Melting Point and Boiling Point of Organic Compounds," *Ind. Eng. Chem. Res.* 1994, **33**, 1872-1877.

(9) J. C. Dearden, "The Prediction of Melting Point," *Advances in Quantitative Structure Property Relationships, Volume 2*, JAI Press, 1999.

(10) M. Charton and B. Charton, “Quantitative Description of Structural Effects on Melting Points of Substituted Alkanes,” *J. Phys. Org. Chem.* 1994, **7**, 196-206.

(11) D. E. Needham, I. Wei, and P. G. Seybold, “Molecular Modeling of the Physical Properties of the Alkanes,” *J. Am. Chem. Soc.* 1988, **110**, 4186-4194.

(12) A. R. Katritzky and E. V. Gordeeva, “Traditional Topological Indices *vs* Electronic, Geometrical, and Combined Molecular Descriptors in QSAR/QSPR Research,” *J. Chem. Inf. Comput. Sci.* 1993, **33**, 835-857.

(13) J. C. Dearden, “The QSAR prediction of melting point, a property of environmental relevance,” *Sci. Total Environ.* 1991, **109/110**, 59-68.

(14) A. R. Katritzky, U. Maran, M. Karelson, and V. S. Lobanov, “Prediction of Melting Points for the Substituted Benzenes: A QSPR Approach,” *J. Chem. Inf. Comput. Sci.* 1997, **37**, 913-919.

(15) J. S. Chickos and G. Nichols, “Simple Relationships for the Estimation of Melting Temperatures of Homologous Series,” *J. Chem. Eng. Data* 2001, **46**, 562-573.

CHAPTER 6 - CONCLUSIONS AND RECOMMENDATIONS

In this study, the available methods for predicting some of the key properties for solid organic compounds have been evaluated. Existing methods have been improved or new methods where possible have been developed. Prediction methods are used when reliable experimental data are not available due to cost, scarcity, safety, or environment. Prediction methods are also used to fill holes in property databases so that the characteristics of a compound can be more fully understood, even in the absence of experimental data.

Where reliable prediction methods are unavailable, new correlations for estimating property data have been created. As prediction methods for solid properties are scarcer than for the liquid and vapor phases, this has been done for several properties. The creation of new prediction methods has been carried out by (1) extracting experimental data from the DIPPR 801 database, (2) selecting the class of equation to use for the correlation, (3) refining the form of the equation through the use of least squares regression of the experimental data (training set), (4) selecting groups and/or other indicators to use in the prediction, (5) calculating the values of the groups, adding groups where needed, and (6) testing the correlation against an independent set of experimental data (test set).

In Chapter 2, the heat capacity of solid organic compounds was examined. Prediction methods existed for compounds at room temperature (298 K) or very low temperature ($T \leq 150$ K), but these were viewed as rather crude and inadequate. Two new methods for predicting solid heat capacity were created. The first is a simple power law method (PL) that uses first-order functional groups. The second utilizes a semi-empirical modification of the Einstein-Debye canonical partition function (PF) that utilizes the same groups with the addition of a few more indicators to account for molecule size and multiple halogen groups. The PL method gives better results in the temperature range between 50 and 250 K. The PF method achieves lower deviations in the temperature range above 250 K. The PF method has comparable results to the room temperature prediction methods, but has the advantage of functioning at other temperatures. The low temperature method ($T \leq 150$ K) achieves lower uncertainties than the PL method and is recommended for that temperature range for the subset of organic compounds for which it is available (alkanes, alkenes, alkanols, and alkanones). Both the PL and PF methods have been assigned an uncertainty of 13% in their intended temperature range as specified above. Solid heat capacity prediction could be improved with a more flexible base equation (such as Eq. 2.1). The requirement to develop such a correlation is a wider range of experimental families. There are many data points in families such as the *n*-alkanes, but few for families like the anhydrides.

In Chapter 3, solid vapor pressure and heat of sublimation were examined. These properties are related through the Clausius-Clapeyron equation. The use of the partition function used to predict solid heat capacity was deemed unusable for predicting solid

vapor pressure due to the necessity to correlate the crystal lattice constant in addition to the characteristic vibrational cut-off temperature. This led to only adequate representation of the heat of sublimation with a resultant poor accuracy in solid vapor pressure. Instead, a method for estimating heat of sublimation at the triple point was created using the same groups as are used in the heat capacity PF method. This method and the Clausius-Clapeyron equation were used to predict solid vapor pressure. This method produced an average absolute logarithmic deviation of 0.371 which translates to a 44.9% error. As most sublimation pressures are very low, on the order of a Pascal or so, on an absolute scale, this uncertainty is acceptable. Compared to two previously-developed methods for predicting solid vapor pressure and one for predicting heat of sublimation, this method yields a lower average absolute logarithmic deviation in solid vapor pressure. Future work on heat of sublimation and solid vapor pressure should expand the functional groups in the method. This requires a broader set of experimental data to use in the training set. Another way to improve predictions would be to add temperature dependence to the heat of sublimation equation. While there is not a strong temperature dependence to heat of sublimation, the addition of temperature dependence to the heat of sublimation may improve the solid vapor pressure predictions for compounds with higher melting points.

Solid density is the focus of Chapter 4. A method for predicting solid density at the triple point from the liquid density at the triple point, previously used by the DIPPR 801 staff, was extended to include a temperature dependence. This modified relation also yielded a lower deviation from experimental data at the triple point than the previous

version which was only applicable at that temperature. This was achieved by utilizing data archived over the past 20 years. Adding topological and other types of descriptors to the relationship decreased deviations within the training set, but drastically decreased extrapolation capability as evidenced by the results on a test set of compounds. The modified DIPPR relation is therefore recommended over the QSPR descriptor method. The modified DIPPR correlation had a deviation from experimental data of 6.28%. The deviation at room temperature was lower than that of a simple “back of the envelop” method which had been published in the literature. Improvements to solid density prediction can be made with more experimental data on which to base the correlation. This would allow more independent variables, such as functional groups, to be used. More data would also produce equations with better predictive capabilities than Eqs. 4.2 and 4.3. It would also be desirable to use a independent variable other than liquid density. While liquid density for many organic compounds is better known than the solid density, not all compounds form a liquid at convenient temperatures and pressures. The most desirable prediction methods are those that use only data derived from a knowledge of the molecular structure.

In Chapter 5, the various ways to predict melting point were examined. Specific family-oriented QSPR methods provide the most accurate predictions, but are not always available or practical. Enthalpy/entropy methods provide ways to predict melting points for broad families, but the most generalized methods are group contributions methods. Of the methods for predicting melting points for wide ranges of organic compounds, the method of Marrero and Gani gave the lowest average absolute deviation from

experimental values (12.5%). This method can be unwieldy due to its very large number of groups, and so a secondary recommendation is extended to the method of Yalkowsky et al. (deviation of 13.9%) which depends on the normal boiling point and indicators of symmetry and flexibility rather than a large number of groups. Future work in melting point prediction can best be aided by a greater understanding of how to predict the entropy of fusion. While the enthalpy of fusion can be predicted well by group-contributions methods, the entropy cannot. While it is known that symmetry, flexibility, shape, and the amount of hydrogen bonding play roles, these are not enough to make predicting melting points as reliable as predicting boiling points.

One other solid property for which a prediction method would be desirable is solid thermal conductivity. While there are a lot of data available for the thermal conductivity of metals and oxides, very little has been recorded for organic compounds. When more experimental data become available, this topic needs to be revisited.

In general, more experimental data are needed to facilitate the development of better prediction methods. While some families of compounds are well represented in databases such as DIPPR 801, others are not. It is desirable that data used to develop prediction methods be well distributed across the types of compounds for which it is developed so that the reliability of the methods does not deteriorate for some compounds.

Another general, but key, issue in the development of solid property prediction methods is dealing with different crystalline structures. The methods developed in this work are designed for the crystalline phase stable at the triple point (heat of sublimation, solid vapor pressure, solid density, and melting point) or the phase stable at absolute zero

(solid heat capacity). If there was a way to characterize crystalline structures, it could be added to the correlations either as an individual independent variable (such as r_g in the C_p^s correlation) or as a set of descriptors. While monatomic crystals have easily characterized patterns, organic molecules are large, complex, and comparatively asymmetrical. This makes them hard to compartmentalize. If such a characterization were developed, it would greatly improve prediction for all the solid properties.

While improvements can be made when more data become available, this work has examined the state of property prediction methods for solid organic compounds and provided improved methods for four properties.