Use of the DIPPR Database for the Development of QSPR Correlations: Solid Vapor Pressure and Heat of Sublimation of Organic Compounds[1]

B. T. Goodman,[2] W. V. Wilding,[2] J. L. Oscarson,[2] and R. L. Rowley[2, 3]

A group-contribution method has been developed to estimate the heat of sublimation (ΔH_{sub}) at the triple point for organic solids. The correlation was developed using ΔH_{sub} values from a training set of 218 compounds, and it was tested by comparing predicted and solid vapor pressures for 87 compounds (1103 data points). The predicted solid vapor pressures were obtained from the ΔH_{sub} correlation using the Clausius–Clapeyron equation. The absolute average deviation in the logarithm of vapor pressure was 0.371. The new method compares favorably with Bondi's method for prediction of ΔH_{sub} and represents an improvement over other available methods for predicting solid vapor pressures.

KEY WORDS: group contributions; heat of sublimation; organic compounds; prediction; QSPR; solid vapor pressure.

1. INTRODUCTION

Evaluated databases such as the DIPPR 801 database [1] are useful tools in developing new property prediction methods. Because the data have been evaluated and an uncertainty assigned, training sets of specific accuracy can be used to develop correlations and/or weight the regression of parameters or constants in the correlating equations. Data in the database not used in developing the correlation can be used to test the applicability of the correlation to additional compounds. Additionally, a database serves as an important source of fundamental and molecular-based

---

[1] Paper presented at the Fifteenth Symposium on Thermophysical Properties, June 22–27, 2003, Boulder, Colorado, U.S.A.

[2] Department of Chemical Engineering, 350 CB, Brigham Young University, Provo, Utah 84602, U.S.A.

[3] To whom correspondence should be addressed. E-mail: rowley@byu.edu

constants that can be used to develop new correlations akin to the quantitative structure-property relationship (QSPR) methodology. In particular, we have found it convenient to cross correlate properties for which experimental data may be missing in terms of constants, molecular properties, and structural information contained in the DIPPR database.

In the more common use of QSPR, a computational chemistry package is used to compute molecular descriptors based on the optimized geometry of the molecule and the resultant electron distribution. The DIPPR database contains some structural constants obtained from quantum calculations on the optimized geometry as well as other fundamental constants, often obtained from experiment, that are either directly or indirectly related to the structure of the molecule. While the number of descriptors available in the database is currently very small relative to the more general QSPR techniques, many of the more commonly used properties are available. For example, group contributions are readily available through an automatic parsing of the SMILES formula stored in the database. Group definitions used by the automatic parser can be defined by the user in DIADEM, companion software available for the database. Additionally, stored molecular descriptors (e.g., dipole moment, radius of gyration, molecular weight, van der Waals volume, van der Waals area) can be used in conjunction with fundamental constants and defined descriptors (e.g., refractive index, critical properties, boiling point, solubility parameter, parachor) to obtain useful estimation techniques. We have used this methodology to develop new predictive techniques for normal boiling point [2], surface tension [3], and the heat capacity [4] of solid organic compounds. Here, we extend that previous work to heats of sublimation, ΔH_{sub}, and vapor pressures of solid organic compounds (SVP).

While estimation methods are available for many vapor and liquid properties, fewer techniques have been developed for solid properties, and there are no good methods for a few solid properties. One of the hallmarks of the DIPPR database is its completeness, meaning that if experimental data are not available for a particular property, quality estimation methods are used for that property when possible in order to provide a complete set of property values. Nevertheless, only 13% of the more than 1800 compounds in the DIPPR database list any SVP data, experimental or predicted.

2. CORRELATIONS

2.1. Correlation of the Heat of Sublimation

Although the DIPPR database does not contain raw data specifically for ΔH_{sub}, the value recommended in the database for the heat of fusion,

ΔH_{fus}, at the melting point and the evaluated correlation of the heat of vaporization, ΔH_{vap}, as a function of temperature (from the triple point to the critical point) can be used to obtain ΔH_{sub} values for development of the correlation. For most compounds, the triple point temperature, T_{TP}, and the normal melting point are very similar and we can obtain ΔH_{sub} from

$$\Delta H_{\text{sub}}(T_{\text{TP}}) = \Delta H_{\text{fus}} + \Delta H_{\text{vap}}(T_{\text{TP}}). \tag{1}$$

A training set of 218 organic compounds with ΔH_{sub} values at the triple point found using Eq. (1) was used in a QSPR scheme similar to that used previously with the correlation for the heat capacity of solids [4]. As with the heat capacity work, the significant correlating factors for ΔH_{sub} were functional groups and the radius of gyration, RG. Additionally, quadratic terms for the two most common groups, methylene and aromatic carbon, and correction terms for multiple halogen atoms, based on the fraction of carbon or silicon terminal valences occupied by halogen atoms, were also found to be significant in the correlation. Using the multiple regression package in Oxford Molecular Tsar 3.2 [5], we obtained

$$\frac{\Delta H_{\text{sub}}(T_{\text{TP}})}{R} = 698.04 + (3.83798 \times 10^{12}\,\text{m}^{-1})\,RG$$

$$+ \sum_i^{NG} n_i a_i + \sum_i^{NG} b_i n_i^2 + \sum_i^{NG} c_i\,\frac{n_i}{n_x} \tag{2}$$

where a_i, b_i, and c_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, 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 are obtainable from several sources including the DIPPR 801 database.

Tables I and II contain the values of the group contributions obtained from the regression. Linear groups are given in Table I; the nonlinear terms for methylene and aromatic carbon groups and the correction terms for the halogen fractions are given in Table II. Table I also illustrates group definitions. The designated group is highlighted with bold typeface in the SMILES formula [6] for the compound. SMILES (Simplified Molecular Line 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 software, such as Tsar, that automates the parsing of molecular structures into groups. Simple SMILES tutorials can be found in our previous work [2] or on the world wide web [7].

Table I. Linear Group Values for ΔH_{sub}

Group	Description	Example	SMILES formula	a Eq. (2)
$-CH_3$	Methyl	n-butane	CCCC	736.5889
$>CH_2$	Methylene	1-heptanol	OCCCCCCC	561.3543
$>CH-$	Secondary carbon	2,3-dimethylpentane	CC(C)C(C)CC	111.0344
$>C<$	Tertiary carbon	2,2-dimethylbutane	CC(C)(C)CC	−800.517
$CH_2=$	Terminal alkene	1-octene	C=CCCCCCC	572.6245
$-CH=$	Alkene	1,3-butadiene	C=CC=C	541.2918
$>C=$	Substituted alkene	Isobutene	C=C(C)C	117.9504
$Ar-CH=$	Aromatic carbon	Biphenyl	c1ccccc1(c2ccccc2)	626.7621
$Ar>C=$	Substituted aromatic C	Toluene	c1ccccc1(C)	348.8092
$Ar-O-$	Furan oxygen	Furan	C1=COC=C1	763.284
$Ar-N=$	Pyridine nitrogen	Quinoline	c1(ccn2)c2cccc1	1317.056
$Ar-S-$	Thiophene sulfur	Thiophene	S1C=CC=C1	911.2903
$-O-$	Ether	Dimethyl ether	COC	970.4474
$-OH$	Alcohol	1-pentanol	CCCCCO	3278.446
$-COH$	Aldehyde	1-butanal	CCCC=O	2402.093
$>C=O$	Ketone	3-hexanone	CCC(=O)CCC	1816.093
$-COO-$	Ester	Methyl methacrylate	C=C(C)C(=O)OC	2674.525
$-COOH$	Acid	n-butyric acid	CCCC(=O)O	5006.188
$-NH_2$	Primary amine	Methylamine	CN	2219.148
$>NH$	Secondary amine	Piperidine	C1CCCCN1	1561.222
$>N-$	Tertiary amine	Trimethylamine	CN(C)C	325.9442
$-NO_2$	Nitro	Nitrobenzene	c1(N(=O)=O)ccccc1	3661.233
$-SH$	Thiol/mercaptan	n-hexyl mercaptan	CCCCCCS	1921.097
$-S-$	Sulfide	Diethyl sulfide	CCSCC	1930.84
$-SS-$	Disulfide	Di-n-propyl disulfide	CCCSSCCC	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 II. Nonlinear and Halogen Group Values for ΔH_{sub}

Group	Description	Eq. (2)
	A. Nonlinear terms	
		b
>CH$_2$	Methylene	9.5553
Ar=CH–	Aromatic carbon	−2.21614
	B. Halogen fraction terms	
		c
–Cl	Cl fraction	−1543.66
–F	F fraction	−1397.4
–Br	Br fraction	5812.49

Equation (2) has an average absolute deviation (AAD) of 3.01 kJ·mol^{-1}, 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 AAPD versus the ΔH_{sub} from the training set is shown in Fig. 1. Due to the small quantity of ΔH_{sub} data available, we chose to validate Eq. (2) with an independent SVP data set rather than holding back data from the training set for testing extrapolation of the correlation to additional compounds.
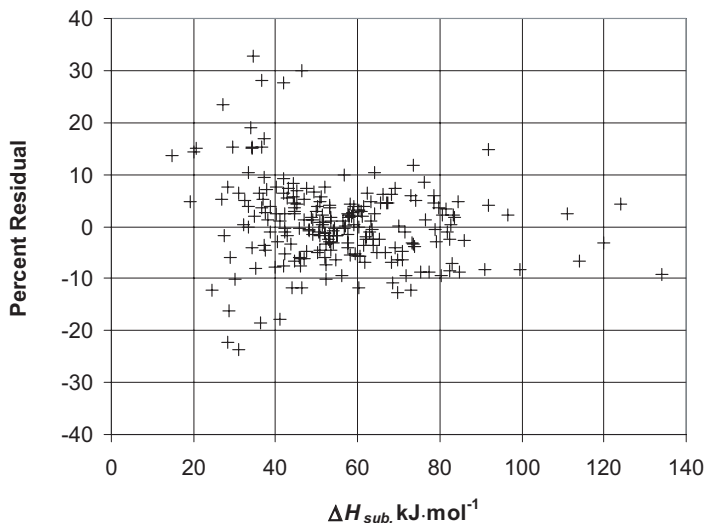
Fig. 1. Percent residual of ΔH_{sub} for the 218 compounds of the training set.

2.2. Relationship between SVP and HSUB

The vapor pressure of a solid and its heat of sublimation (HSUB) are related through the Clapeyron equation,

$$\frac{\mathrm{d}P}{\mathrm{d}T} = \frac{\Delta H_{\mathrm{sub}}}{T \Delta V_{\mathrm{sub}}} \tag{3}$$

where P is vapor pressure, T is temperature, and ΔV_{sub} is the change in molar volume upon sublimation. Alternatively, this can be written in terms of the change in compressibility factor upon sublimation, ΔZ_{sub}, as

$$\frac{\mathrm{d} \ln P}{\mathrm{d}(1/T)} = -\frac{\Delta H_{\mathrm{sub}}}{R \, \Delta Z_{\mathrm{sub}}} \tag{4}$$

where R is the universal gas constant. Often the compressibility factor of the solid is much smaller than that of the solid, and because the vapor pressure is so low, the compressibility factor for the saturated vapor is very close to unity, and Eq. (4) simplifies to the Clausius–Clapeyron equation:

$$\frac{\mathrm{d} \ln P}{\mathrm{d}(1/T)} = -\frac{\Delta H_{\mathrm{sub}}}{R}. \tag{5}$$

As the triple point pressure, P_{TP}, and temperature are also tabulated constants in the DIPPR database, we integrate Eq. (5) from the triple point temperature down to an arbitrary temperature to obtain the form of the correlation used here as

$$\ln \frac{P}{P_{\mathrm{TP}}} = -\frac{\Delta H_{\mathrm{sub}}}{R} \left(\frac{1}{T} - \frac{1}{T_{\mathrm{TP}}} \right). \tag{6}$$

We have assumed in the integration of Eq. (5) that ΔH_{sub} is independent of temperature over the range T to T_{TP}. Examination of heat capacity data for a few compounds suggests that the change in ΔH_{sub} over an 80 K range is less than the error in Eq. (2), and the constancy of ΔH_{sub} is therefore an excellent approximation.

2.3. Estimation of Solid Vapor Pressure

Solid vapor pressures for 87 compounds (1103 separate data points) were computed using Eq. (6), with ΔH_{sub} calculated from Eq. (2), and compared to experimental values from the DIPPR database. Experimental values of SVP are compared to predicted values on a logarithmic scale in
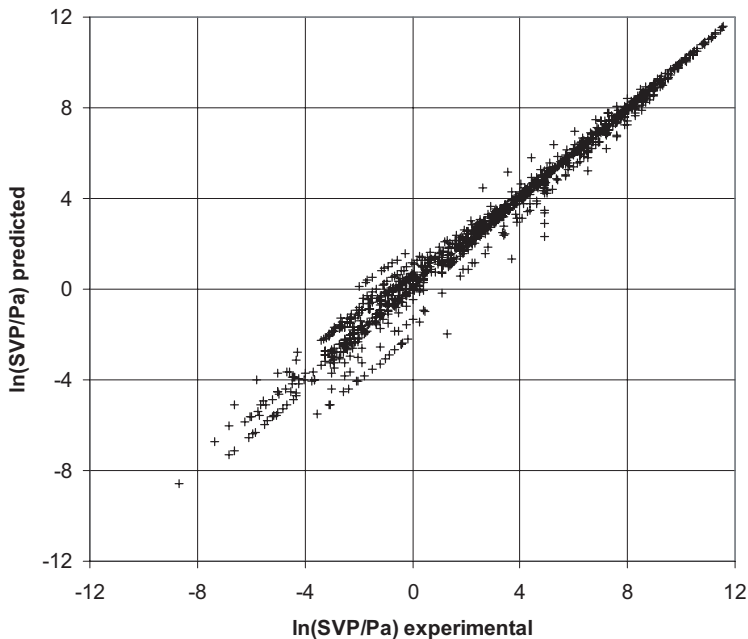
Fig. 2. Experimental vs. predicted ln(SVP/Pa) for the 87 compounds (1103 data points) of the test set.

Fig. 2. As SVP data span several orders of magnitude, we used as our statistical metric the average absolute logarithmic deviation (AALD), defined as

$$\text{AALD} = \frac{1}{n}\left[\sum_{i=1}^{n} |\ln(\text{SVP}_{\text{pred},i} - \ln(\text{SVP}_{\text{exp},i})| \right] \tag{7}$$

where n is the number of data points, $\text{SVP}_{\text{pred},i}$ is the predicted value of SVP at a specific temperature, and $\text{SVP}_{\text{exp},i}$ 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 shows the performance of this prediction method for three compounds over a range of temperatures. Included as an appendix to this paper are computational examples that illustrate the group definitions and the use of Eqs. (2) and (6) to obtain ΔH_{sub} and SVP for the same compounds as shown in Fig. 3.
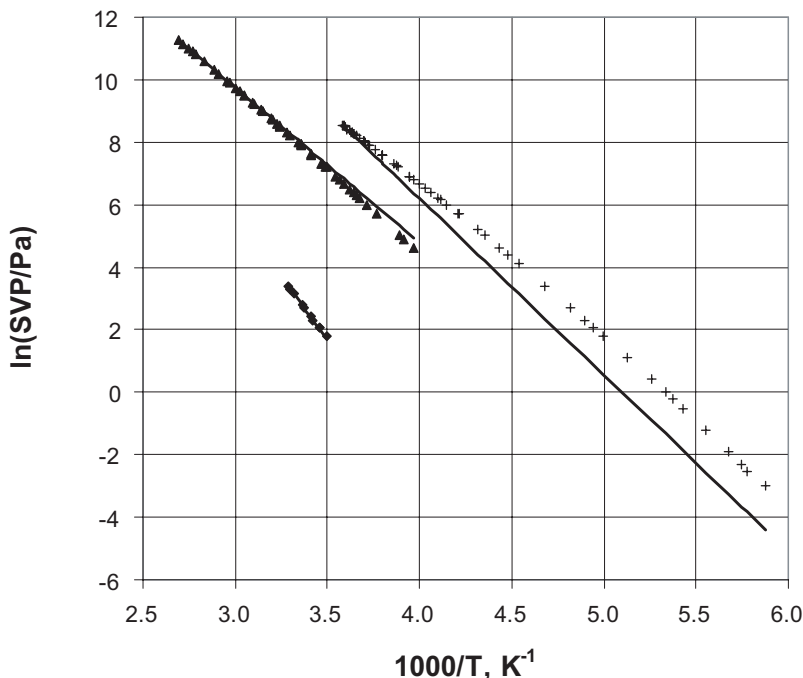
Fig. 3. Experimental (—) and predicted ln(SVP/Pa) for 1,2,3-trichlorobenzene (♦), 2,2,3,3-tetramethylbutane (▲), and cyclohexane (+).

3. COMPARISON TO OTHER PREDICTION METHODS

Bondi provided a group-contribution method for estimating ΔH_{sub} at the lowest phase transition [8, 9]. Where a molecule only has one solid phase, this is the triple point. Again using the assumption that ΔH_{sub} is independent of T, we have calculated SVP for a subset of our test set using Bondi's correlation for ΔH_{sub} and Eq. (6). The entire test set could not be used because Bondi's correlation and Eq. (2) do not share the same functional group building blocks. While both sets of groups are targeted toward organic compounds, Bondi's groups can be used with more molecules that contain inorganic components while Eq. (2) has a more extensive palette of organic groups. As shown in Table III, the AALD for the method developed here is 0.23 for 591 vapor pressure points for 39 compounds as compared to 0.33 for the Bondi method applied to the same test set.

Mackay et al. [10] developed four similar methods for predicting vapor pressures based on the normal boiling point (NBP) of a molecule.

Table III. Comparison of SVP Prediction Methods

Method	Compounds	Points	AALD	AALD Eqs. (2) & (6)
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

While these methods were primarily tested with liquid compounds, they can be used with solid compounds if an extra term based on the melting point (MP) is added. Using a subset of the previous test set for which these methods could be applied, we compared these methods (labeled TCH, KCH, TLH, and KLH) to ours; the results are also given in Table III. While the AALD of each of these methods is higher than that of our method, these methods have the advantage of requiring only the normal boiling point and the melting point (or T_{TP}), commonly available properties. Of these four methods, KLH is superior.

Neau et al. [11] used the Peng–Robinson equation of state to estimate P_{vap} and ΔH_{vap} at the triple point. Along with ΔH_{fus} and T_{TP}, this information was used to estimate ΔH_{sub} in the manner shown in Eq. (1). The Peng–Robinson equation of state requires the critical temperature, critical pressure, and acentric factor. As experimental values for the critical constants and the vapor pressure curve, hence the acentric factor, may be unavailable for compounds that are solids for normal operating conditions, Neau et al. recommend using the group-contribution methods of Constantinou and Gani [12] to estimate critical constants and that of Constantinou et al. [13] to estimate the acentric factor. 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 [14]. 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 do not give any details, referring only to additional publications unavailable to us [15, 16]. This method, as explained by Coniglio et al. [14], can be used with alkanes, aromatics, alkenes, esters, alcohols, and carboxylic acids. Neau et al. recommend the use of Avaullee et al. [17] to estimate the NBP if this is not known. The results of a comparison of our correlation with that of Neau et al. are also shown in Table III, for the four cases of (1) using experimental values of the critical constants and the acentric factor obtained from the DIPPR database, (2) using group contributions 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 with our previously developed group-contribution method [2].

Figure 4 compares experimental ln(SVP) values with those predicted using the method of Eq. (2), Bondi, McKay et al. (KLH), and Neau et al. (using both critical constants and NBP) for benzene over a range of inverse temperatures.
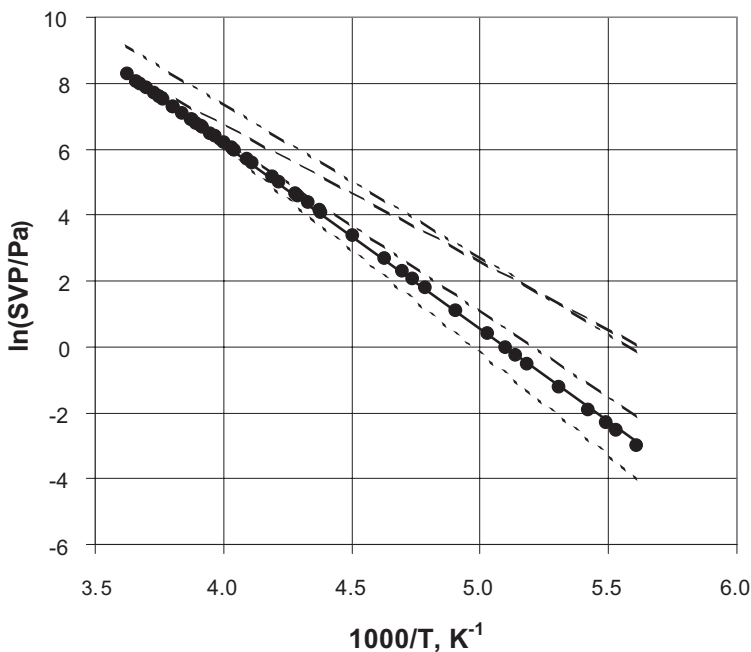

Fig. 4. Comparison of ln(SVP/Pa) data of benzene for experimental (●), Eqs. (2) and (6) (solid line), KLH (short dash), Bondi (dash-dot-dash), Neau et al. critical constants (dash-dot-dot-dash), and Neau et al. boiling point (long dash).

4. SUMMARY

A group-contributions method was developed for estimating the heat of sublimation of organic compounds at the triple point. This method has also been applied to estimating solid vapor pressure through the Clausius–Clapeyron relationship. The accuracy of the method for ΔH_{sub} is similar to Bondi's correlation, but has additional functional groups and is consistent with and uses the same form as the heat capacity correlation developed earlier for organic solids. The temperature 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 SVP data from the DIPPR database not used in development of the correlation for ΔH_{sub}. The AALD of this comparison was 0.371, marking a substantial improvement over existing methods.

APPENDIX-SAMPLE CALCULATIONS

Example 1. Calculation of ΔH_{sub} and SVP for 2,2,3,3-tetramethylbutane.

SMILES formula: CC(C)(C)C(C)(C)C

$RG = 3.785 \times 10^{-10}$ m $TPT = 373.96$ K $P_{vp}(TPT) = 86930.2$ Pa

Linear Groups (Table I)			Nonlinear Groups/ Corrections (Table II)			
Group	n_i	a_i	Group	n_i	b_i	c_i
$-CH_3$	6	736.5889	none			
$>C<$	2	-800.517				

Eq. (2):

$$\frac{\Delta H_{sub}}{R} = 698.04 + (3.83798 \times 10^{12})(3.785 \times 10^{-10})$$

$$+ (6)(736.5889) + (2)(-800.517)$$

$$= 4969.215 \text{ K}$$

$$\Delta H_{sub} = (4969.215 \text{ K})(8.314 \times 10^{-3} \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}) = 41.314 \text{ kJ} \cdot \text{mol}^{-1}$$

SVP Results:

T (K)	exp (Pa)	pred (Pa)
298.15	2780	2963 (6.6%)
353.15	40000	39729 (-0.7%)

Example 2. Calculation of ΔH_{sub} and SVP for 1,2,3-trichlorobenzene.

SMILES formula: c1(Cl)c(Cl)c(Cl)ccc1

RG $= 4.455 \times 10^{-10}$ m   TPT $= 325.65$ K P_{vp}(TPT) $= 182.957$ Pa

Linear Groups (Table I)			Nonlinear Groups/ Corrections (Table II)			
Group	n_i	a_i	Group	n_i	b_i	c_i
Ar $=$ CH$-$	3	626.7621	Ar $=$ CH$-$	3	-2.21614	
Ar $=$ C $<$	3	348.8092	$-$Cl	3		-1543.66
$-$Cl	3	1243.445	n_X	6		

Eq. (2):

$$\frac{\Delta H_{sub}}{R} = 698.04 + (3.83798 \times 10^{12})(4.455 \times 10^{-10})$$

$$+ (3)(626.7621) + (3)(348.8092)$$

$$+ (3)(1243.445) + (3)^2(-2.21614) + (\tfrac{3}{6})(-1543.66) = 8273.134 \text{ K}$$

$$\Delta H_{sub} = (8273.134 \text{ K})(8.314 \times 10^{-3} \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}) = 68.783 \text{ kJ} \cdot \text{mol}^{-1}$$

SVP Results:

T (K)	exp (Pa)	pred (Pa)
286.15	6	5.487 (-8.5%)
301.15	23.087	23.16 (0.3%)

Example 3. Calculation of ΔH_{sub} and SVP for cyclohexane

SMILES formula: C1CCCCC1

$RG = 3.216 \times 10^{-10}$ m $TPT = 279.69$ K $P_{vp}(TPT) = 5362.51$ Pa

Linear Groups (Table I)			Nonlinear Groups/Corrections (Table II)			
Group	n_i	a_i	Group	n_i	b_i	c_i
$> CH_2$	6	561.3543	$> CH_2$	6	9.5553	

Eq. (2):

$$\frac{\Delta H_{sub}}{R} = 698.04 + (3.83798 \times 10^{12})(3.216 \times 10^{-10})$$

$$+ (6)(561.3543) + (6)^2 (9.5553)$$

$$= 5644.451 \text{ K}$$

$$\Delta H_{sub} = (5644.451 \text{ K})(8.314 \times 10^{-3} \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{K}^{-1}) = 46.928 \text{ kJ} \cdot \text{mol}^{-1}$$

SVP Results:

T (K)	exp (Pa)	pred (Pa)
173.15	0.08	0.022 (-73%)
273.16	3637.8	3310.2 (-9.0%)

REFERENCES

1. R. L. Rowley, W. V. Wilding, J. L. Oscarson, N. A. Zundel, T. L. Marshall, T. E. Daubert, and R. P. Danner, *DIPPR^{&registered>} Data Compilation of Pure Compound Properties* (Design Institute for Physical Properties, AIChE, New York, 2002).

2. D. Ericksen, W. V. Wilding, J. L. Oscarson, and R. L. Rowley, *J. Chem. Eng. Data* **47**:1293 (2002).
3. T. A. Knotts, W. V. Wilding, J. L. Oscarson, and R. L. Rowley, *J. Chem. Eng. Data* **46**:1007 (2001).
4. B. T. Goodman, W. V. Wilding, J. L. Oscarson, and R. L. Rowley, *J. Chem. Eng. Data* **49**:24 (2004).
5. *TSAR Version 3.2* (Oxford Molecular Group, Oxford Molecular Limited, Oxford, 1998).
6. D. Wininger, *Chem. Inf. Comput. Sci.* **28**:31 (1998).
7. *SMILES Tutorial* (1998), www.daylight.com/smiles/f_smiles.html.
8. A. Bondi, *J. Chem. Eng. Data* **8**:371 (1963).
9. A. Bondi, *Physical Properties of Molecular Crystals, Liquids and Glasses* (Wiley, New York, 1968), pp. 453–479.
10. D. MacKay, A. Bobra, D. W. Chan, and W. Y. Shlu, *Environ. Sci. Technol.* **16**:645 (1982).
11. E. Neau, S. Gamier, and L. Avaullee, *Fluid Phase Equilib.* **164**:173 (1999).
12. L. Constantinou and R. Gani, *AIChE J.* **40**:1697 (1994).
13. L. Constantinou, R. Gani, and J. P. O'Connell, *Fluid Phase Equilib.* **103**:11 (1995).
14. L. Coniglio, L. Trassy, and E. Rauzy, *Ind. Eng. Chem. Res.* **39**:5037 (2000).
15. L. Coniglio, Thesis (Université d'Aix-Marseille III, 1993).
16. L. Trassy, Thesis (Université de la Méditerranée, 1998).
17. L. Avaullée, L. Trasy, E. Neau, and J. N. Jaubert, *Fluid Phase Equilib.* **139**:155 (1997).