Use of the DIPPR Database for Development of QSPR Correlations: Normal Boiling Point

Daniel Ericksen, W. Vincent Wilding, John L. Oscarson, and Richard L. Rowley*

Department of Chemical Engineering, 350 CB, Brigham Young University, Provo, Utah 84602

Tabulation of evaluated physical property constants and of their estimated uncertainties makes the DIPPR database a valuable tool for developing correlations for physical properties of pure fluids. In this study, we have used the DIPPR database to develop a group-contribution method for the normal boiling point (NBP) of pure compounds. The resultant correlation uses the molecular descriptors of molecular weight and van der Waals volume in addition to Domalski−Hearing (DH) second-order group definitions. A training set of 1141 evaluated normal boiling points of >95% accuracy was selected from the database and correlated. The average absolute deviation (AAD) was 7.8 K (1.9%) with zero bias. Estimations of NBP for a test set of 384 compounds not used in the regression gave an AAD of 13.0 K (2.7%). The results suggest that the method is comparable in accuracy to the best methods available for estimating NBP, but it has the convenient feature of DH group designations that are immediately compatible with currently available DH algorithms and software.

Introduction

In conjunction with modern QSPR (quantitative structure−property relationship) software that facilitates rapid division of molecules into structural groups and calculation of molecular descriptors, the DIPPR database[1] provides a quantitative foundation for rapid development of physical property prediction methods. We recently reported use of the DIPPR database in development of a group-contribution method (GCM) for prediction of surface tension.[2] We follow up that work with a report here on development of a new GCM for the normal boiling point (NBP) of organic compounds.

Two key elements in developing accurate QSPR correlations are the ability to generate the molecular descriptors, structural groups in the case of GCMs, and the availability of accurate properties from which the correlation coefficients may be regressed. The focus here is on the chemical database from which accurate correlations for properties, here NBP, can be developed and tested. The DIPPR pure-component database, containing 44 properties for over 1800 compounds, is an *evaluated* database. The evaluated nature of the database is particularly important for the regression step because the accuracy of the final correlation is only as good as the data upon which it is based. Extensive data of lower quality tend to decrease the accuracy of a correlation relative to use of a more selective data set of higher quality data. Obviously, a large data set of high-quality values is the ideal. Property values in the DIPPR database underwent a comprehensive evaluation, and the single *accepted* value for NBP was based on this evaluation. The evaluation for NBP involved not only relative comparisons of experimental values but also an analysis of the values relative to related properties and related compounds. Thus, where possible, trends within chemical families were used to adjudicate between reported values from different laboratories as were thermodynamic relationships to other reported property values, such as vapor pressures and heats of vaporization.

The accepted values in the database are also assigned an accuracy level in terms of an uncertainty. The available uncertainties in the DIPPR database are <0.2%, <1%, <3%, <5%, <10%, <25%, and so forth. We have found that for most properties data of accuracy <3% or <5% provides a reasonable breadth of data across chemical constituency with sufficient accuracy. Specifically for NBP we found that the accuracy of the GCM itself appears limited to no better than the 5% level, and so we chose <5% as our quality criterion for the training set used in this work.

Current NBP Estimation Methods

A large number of methods for estimating boiling points have been devised. Excellent reviews of these methods are available,[3−7] and no attempt is made here to review these methods. However, it is appropriate to put the method developed in this work in context with current capabilities, both to understand the motivation for yet another NBP estimation method and to understand the relative merits of various techniques.

From statistical mechanics, the boiling point of a compound can be found by equating the gas- and liquid-phase chemical potentials, thereby obtaining an expression in terms of the ratio of the gas- and liquid-phase partition functions. To the extent that internal modes are independent of density, this ratio is equivalent to the ratio of the configurational partition functions, Z. As Z is unity for an ideal gas, the Z ratio is nearly numerically equivalent to Z for the liquid. Therefore, it seems reasonable to base the calculation of NBP strictly on the structure and interactions (or properties that characterize them) of the liquid phase. Structural and interaction-related molecular descriptors for this purpose are often computed within the general QSPR methodology using computational chemistry packages. Such descriptors generally relate to molecular structure and electron probability distributions within the

* To whom correspondence should be addressed. E-mail: rowley@byu.edu.

10.1021/je0255372 CCC: $22.00
Published on Web 07/26/2002

molecule. Here, we make a convenient distinction between descriptors based on bonds, atoms, and groups, which we designate as a general GCM, and those based on molecular descriptors, which we designate as MolD. Often, tabulated GCM values can be used directly with a known formula and molecular structure, but MolD values are usually specific to the molecule instead of individual groups and therefore must be calculated as part of the prediction methodology using computational chemistry software.

Within the GCM techniques are methods by Lydersen,[8] Joback and Reid,[9] and Constantinou and Gani[7] (CG). The Lydersen and Joback methods both employ first-order group contributions, where each group value is independent of neighboring groups to which it is connected. Thus, in the correlation

$$\mathrm{NBP} = \mathrm{Tb}_0 + \sum_k n_k \mathrm{Tb}_k \qquad (1)$$

where Tb_0 is a constant base contribution, n_k is the number of type k groups in the molecule, and Tb_k is the incremental value for the functional group. The Tb_k group values only need be tabulated for unique functional groups. We tested the Joback method on 1200 compounds in the DIPPR database and found an average absolute deviation (AAD) of 12 K (3.8%). The CG method is a second-order method. It utilizes additive group values based on UNIFAC-defined groups[10] to obtain the first-order contribution to NBP, but it also includes tables for adding second-order corrections to this value for more accurate estimations. The second-order corrections are based on combinations of the first-order groups. This has the effect of correcting the sum of first-order groups for the more extended environment within the molecule. The CG method is

$$\mathrm{NBP} = \mathrm{Tb}_0 \ln\left(\sum_k n_k \mathrm{Tb1}_k + W \sum_j m_j \mathrm{Tb2}_j\right) \qquad (2)$$

where Tb1 indicates a group value for a first-order term and Tb2 indicates a second-order correction term. W in this equation is either 0 or 1, depending on whether one wants to include second-order corrections, and m_j is the number of second-order corrections of type j. Constantinou and Gani[7] used the DIPPR database for their training set and reported an AAD for the regression of 10.5 K (2.0%).

While MolD methods may also include contributions from the sum of atoms or functional groups, their strength is that they go beyond summations of individual group contributions and attempt to correlate properties in terms of the actual, unique internal electronic environment of the molecule. Instead of correcting first-order groups for the influence of neighboring groups, molecular descriptors are calculated that are specific to the molecule itself. Properties are then correlated in terms of the most statistically significant descriptors. Tabulation of molecular (as opposed to group) descriptors is not practical; instead, users desiring to predict a thermophysical property must first calculate the descriptors in a way identical to that used by the developer. In addition to this limitation, MolD methods generally tend to be specific for certain classes of fluids, for example, alkanes, haloalkanes, pyridines. For instance, Katritzky et al.[11] summarize the results for the best MolD methods (they list 19 different methods) for NBP by saying that they "provide satisfactory predictions for various classes of organic compounds, but a more general model is desirable. The failure to devise a general QSPR model for the prediction of the boiling points of organic compounds is due to the inability of the available descriptors to reflect quantitatively variations in the intermolecular interactions in liquid media." Katritzky et al. successfully developed an eight-parameter correlation for the prediction of NBP for any organic compound containing C, H, O, N, S, F, Cl, Br, and I atoms, which seems to be the only general NBP method available based on MolD. However, the method was based on a training set of only 541 compounds for which the standard prediction error was reported as 15.5 K. The chemical domain and accuracy both appear to be less than that of the CG method.

These considerations of accuracy, broad applicability, and ease of use led sponsors of the DIPPR database project to select CG as the primary estimation method for compounds put into the database for which there is no experimental NBP. However, the application of second-order corrections in the CG method is not as convenient as the Domalski−Hearing (DH) method[12] for automated programs that use standard group recognition algorithms to do property predictions. The DH method extends the Benson group definitions[13] in which the central functional group with its attached neighbors defines a unique group. In this notation, $\mathrm{C-(H)_3(C)}$ represents a central carbon atom with bonds to three hydrogen and one carbon atoms. Group values then depend not only on the central atom but also upon the bonded environment; for example, $\mathrm{C-(H)_3(C)}$, $\mathrm{C-(H)_2(C)_2}$, $\mathrm{C-(H)(C)_3}$, and $\mathrm{C-(C)_4}$ would all have unique contributions to the summation in eq 1.

In this work, we develop from the DIPPR database a second-order NBP GCM based on DH groups and other molecular descriptors already tabulated in the DIPPR database. The advantage of this method over MolD methods is a broader range of applicability for a given accuracy level and the ease of use because a quantum mechanical package is not required to determine the descriptors. The goal is to develop a second-order method of comparable accuracy to the CG method, but one that is consistent with automated implementations of the DH method. This allows tables of DH group values to be inserted into the software for the new property without reprogramming. For example, CHETAH[14] utilizes DH groups, and DIADEM[15] has an automatic formula parser that works in conjunction with the DIPPR database. The parser in DIADEM is based on the compound's SMILES (Simplified Molecular Input Line Specification) formula,[16−18] a convenient in-line chemical notation included in the DIPPR database, and DH group designations are extremely compatible with SMILES formulas.

New GCM for NBP

A training set containing experimental NBP data for 1141 compounds was obtained from the DIPPR database. Each compound in the test set, with the exception of three, had a quality code of <5% uncertainty. The three exceptions (<10% quality code) were included to broaden the method to include important groups that otherwise would have been excluded. A commercial QSPR (quantitative structure−property relationship) software package (TSAR[19]) was used to facilitate the regression, statistical analysis, and group counts within each molecule. Constant properties from the DIPPR database were examined as possible molecular descriptors to include in the correlation in addition to the DH groups. As mentioned above, we restricted our use of molecular descriptors to those tabulated in the database to avoid the added requirement of molecular geometry optimization and descriptor calculation by users of the method. The DH groups were used to obtain a second-order method tied to group definitions that have

Table 1. Group Values for NBP Correlation

group	Tb/K	example
A. CH Groups		
C−(H)$_3$(C)	−10.30	**C**C [ethane]
C−(H)$_2$(C)$_2$	−0.04	C**C**C [propane]
C−(H)(C)$_3$	4.54	CC(**C**)C [isobutane]
C−(C)$_4$	9.73	C**C**(C)(C)C [neopentane]
Cd−(H)$_2$	−13.19	**C**=C [ethylene]
Cd−(H)(C)	41.78	C=**C**C [propylene]
Cd−(C)$_2$	91.28	C=**C**(C)C [isobutene]
Cd−(H)(Cd)	2.40	C=**C**C=C [1,3-butadiene]
Cd−(C)(Cd)	57.23	C=**C**(C)C=C [isoprene]
Cd−(H)(CB)	−7.19	C=**C**c1ccccc1 [styrene]
Cd−(C)(CB)	32.27	C=**C**(c1ccccc1)CC [2-phenylbutene-1]
Cd−(H)(Ct)	14.52	C=**C**C#C [vinylacetylene]
Cd−(C)(Ct)	26.82	C=**C**(C)C#C [2-methyl-1-butene-3-yne]
C−(H)$_4$	−64.87	**C** [methane]
C−(H)$_2$(C)(Cd)	−38.69	C=C**C**C [1-butene]
C−(H)(C)$_2$(Cd)	−42.77	C=C**C**(C)C [3-methyl-1-butene]
C−(C)$_3$(Cd)	−36.71	C=C**C**(C)(C)C [3,3-dimethyl-1-butene]
C−(H)$_2$(Cd)$_2$	−80.74	C=C**C**C=C [1,4-pentadiene]
C−(H)$_3$(Cd)	−49.27	C=C**C** [propylene]
C−(H)$_2$(Cd)(CB)	−37.90	c1(**C**C=C2)c2cccc1 [indene]
C−(H)(C)(Cd)(CB)	−34.60	C1=CC=C2C(=C1)C=C**C**2C [1-methylindene]
Ct-(H)	−9.38	**C**#C [acetylene]
Ct-(C)	2.25	C**C**#C[methylacetylene]
Ct-(Cd)	1.33	C=**C**C#C [vinylacetylene]
Ct-(CB)	26.82	c1(**C**#C)ccccc1 [phenylacetylene]
C−(H)$_3$(Ct)	0.82	**C**C#C[methylacetylene]
C−(H)$_2$(C)(Ct)	7.84	C#C**C**CC [1-pentyne]
C−(H)$_2$(Cd)(Ct)	−28.25	C#C**C**C=C [1-pentene-4-yne]
Ca	41.34	C=**C**=C [propadiene]
C−(H)$_2$(Ca)	−31.75	C=C=**C** [propadiene]
Ca2-(H)(C)(Ca)	−29.68	C**C**=C=C [1,2-butadiene]
C−(H)$_3$(Ca2)	6.68	**C**C=C=C [1,2-butadiene]
C−(H)$_2$(C)(Ca2)	20.77	C**C**C=C=C [1,2-pentadiene]
Ca2-(C)$_2$(Ca)	−36.11	C**C**(C)=C=C [3-methyl-1,2-butadiene]
CB−(H)(CB)$_2$	2.22	c1cc**c**cc1 [benzene]
CB−(C)(CB)$_2$	1.40	c1cccc**c**1(C) [toluene]
CB−(Cd)(CB)$_2$	30.17	C=C**c**1ccccc1[styrene]
CB−(Ct)(CB)$_2$	20.12	**c**1(C#C)ccccc1 [ethynylbenzene]
CB−(CB)$_3$	8.72	c1cccc**c**1(c2ccccc2) [biphenyl]
C−(C)$_2$(CB)$_2$	−14.53	c1ccccc1**C**(C)(C)c2ccc(O)cc2 [*p*-cumylphenol]
C−(H)$_2$(C)(CB)	6.28	c1ccccc1(**C**C) [ethylbenzene]
C−(H)(C)$_2$(CB)	9.70	c1(**C**(C)C)ccc(C)cc1 [*p*-cymene]
C−(CB)(C)$_3$	7.14	c1(**C**(C)(C)C)ccccc1 [*tert*-butylbenzene]
C−(H)$_3$(CB)	3.70	c1cccc**c**1(**C**) [toluene]
C−(H)$_2$(CB)$_2$	27.27	c1ccccc1**C**c2ccccc2 [diphenylmethane]
C−(H)(C)(CB)$_2$	−37.99	c1ccccc1**C**(C)c2ccccc2 [1,1-diphenylethane]
C−(H)(CB)$_3$	3.30	c1ccccc1**C**(c3ccccc3)c2ccccc2 [triphenylmethane]
CB−(O)(CB)(CBF)	−15.72	C2(=1)N=CC=CC1C=CC=**C**2O [8-hydroxyquinoline]
CB−(Cd)(CB)(CBF)	20.37	C1=C2**C**3=C(C=C1)C=CC=C3C=C2 [acenaphthalene]
CB−(C)(CB)(CBF)	5.06	**c**1(C)cccc2ccccc21 [1-methylnaphthalene]
CBF−(CBF)(CB)$_2$	8.77	c1(C)ccc**c**2ccccc21 [1-methylnaphthalene]
CBF−(CB)(CBF)$_2$	10.25	C1=CC2=C(C=C1)**C**=C1C(=**C**2)C2=C(C=C1)C=CC=C2 [benzanthracene]
CB−(CB)$_2$(CBF)	16.70	c1(**c**2ccccc2)cccc2ccccc21 [1-phenylnaphthalene]
CB−(H)(CB)(CBF)	7.85	c1(C)cc**c**c2ccccc21 [1-methylnaphthalene]
CB−(H)(CBF)$_2$	12.82	**c**1cc**c**c2oc(C)**c**c21 [2-methylbenzofuran]
B. CHO Groups		
CO−(H)$_2$	23.09	**C**=**O** [formaldehyde]
CO−(O)(CO)	−183.98	CCOC(=O)**C(=O)**OCC [diethyl oxalate]
CO−(Cd)(O)	−108.34	C=C**C(=O)**O [acrylic acid]
CO−(C)(O)	−34.61	C**C(=O)**O [acetic acid]
CO−(H)(O)	−195.34	**C(=O)**O [formic acid]
CO−(O)$_2$	−376.95	C1O**C(=O)**OC1 [ethylene carbonate]
CO−(H)(Cd)	111.13	C=C**C**=**O** [acrolein]
CO−(CB)$_2$	88.44	c1ccccc1(**C(=O)**c1ccccc1) [benzophenone]
CO−(C)(CB)	205.67	c1ccccc1(**C(=O)**C) [acetophenone]
CO−(H)(CB)	57.37	c1ccccc1(**C**=**O**) [benzaldehyde]
CO−(O)(CB)	−167.19	c1c(**C(=O)**OCC)cccc1 [ethyl benzoate]
CO−(C)$_2$	341.65	C**C(=O)**C [acetone]
CO−(H)(C)	184.15	C**C**=**O** [acetaldehyde]
CO−(C)(Cd)	273.41	C**C(=O)**C(=C)C [methyl isopropenyl ketone]
O−(CO)$_2$ (aliphatic)	422.82	CC(=O)**O**C(=O)C [acetic anhydride]
O−(CO)$_2$ (aromatic)	454.65	c1cc2C(=O)**O**C(=O)c2cc1 [phthalic anhydride]
O−(Cd)(CO)	230.33	CC(=O)**O**C=C [vinyl acetate]
O−(C)(CO)	125.04	C**O**C(=O) [methyl formate]
O−(H)(CO)	265.06	CC(=O)**O** [acetic acid]

Table 1 (Continued)

group	Tb/K	example
B. CHO Groups (continued)		
O−(C)(O)	−61.45	CC(C)(C)**O**O [*tert*-butyl hydroperoxide]
O−(H)(O)	55.80	CC(C)(C)O**O** [*tert*-butyl hydroperoxide]
O−(Cd)$_2$	62.54	C=C**O**C=C [divinyl ether]
O−(C)(Cd)	−38.64	CC**O**C=C [ethyl vinyl ether]
O−(CB)$_2$	−26.95	c1ccccc1(**O**c1ccccc1) [diphenyl ether]
O−(C)(CB)	−75.47	c1(O)c(**O**C)cc(C=O)cc1 [vanillin]
O−(H)(CB)	29.64	c1(**O**)c(OC)cc(C=O)cc1 [vanillin]
O−(C)$_2$	−145.40	C**O**C [dimethyl ether]
O−(H)(C)	−18.16	CO [methanol]
Cd−(H)(CO)	−75.47	C=**C**C(=O)OC [methyl acrylate]
Cd−(C)(CO)	−30.44	C=**C**(C)C(=O)OC [methyl methacrylate]
Cd−(O)(C)	36.15	**C**=C1CC(=O)O1 [diketene]
Cd−(O)(H)	−36.29	CC(=O)O**C**=C [vinyl acetate]
CB−(CO)(CB)$_2$	−23.71	c1cccc**c**1(C=O) [benzaldehyde]
CB−(O)(CB)$_2$	18.01	c1cccc**c**1(O) [phenol]
C−(H)$_2$(CO)$_2$	−321.04	CC(=O)**C**C(=O)OC [methyl acetoacetate]
C−(CO)(C)$_3$	−159.60	C**C**(C)(C)C(=O)O [neopentanoic acid]
C−(H)(CO)(C)$_2$	−160.69	C**C**(C)C(=O)O [isobutyric acid]
C−(H)$_2$(CO)(C)	−162.04	C**C**C(=O)C(C)C [ethyl isopropyl ketone]
C−(H)$_3$(CO)	−168.37	**C**C(=O)C(C)C [methyl isopropyl ketone]
C−(H)$_2$(CO)(Cd)	−176.28	C=C1**C**C(=O)O1 [diketene]
C−(H)(O)(CO)(C)	−112.06	C**C**(O)C(=O)OC [methyl lactate]
C−(H)(C)(O)(CB)	59.26	**C**(c1ccccc1)(O)CC [1-phenyl-1-propanol]
C−(H)(O)$_2$(C)	140.16	CCO**C**(C)OCC [acetal]
C−(H)$_2$(O)$_2$	150.51	CO**C**OC [methylal]
C−(H)$_2$(O)(CB)	73.18	c1ccccc1(**C**O) [benzyl alcohol]
C−(H)$_2$(O)(Cd)	39.08	**C**1O**C**C=C1 [2,5-dihydrofuran]
C−(O)(C)$_3$ ether	67.23	CO**C**(C)(C)C [methyl *tert*-butyl ether]
C−(O)(C)$_3$ ester	59.64	C**C**(C)(C)OC(=O)C [*tert*-butyl acetate]
C−(H)(O)(C)$_2$ ether	60.66	CO**C**(C)C [methyl isopropyl ether]
C−(H)(O)(C)$_2$ ester	65.42	CC(=O)O**C**(C)C [isopropyl acetate]
C−(O)(C)$_3$ alc./perox.	54.14	C**C**(C)(O)CC [2-methyl-2-butanol]
C−(H)(O)(C)$_2$ alc./perox.	58.01	C**C**(O)C [2-propanol]
C−(H)(O)(C)$_2$ epoxy	78.08	C**C**1OC1 [1,2-propylene oxide]
C−(O)(C)$_3$ epoxy	65.75	C**C**1(C)CO1 [1,2-epoxy-2-methylpropane]
O−(C)$_2$ epoxy	−124.86	C1**O**C1 [ethylene oxide]
C−(H)$_2$(O)(C)	67.74	**C**1OC1 [ethylene oxide]
C−(H)$_3$(O)	67.74	**C**O [methanol]
C−(C)$_2$(O)(CB)	49.57	c1ccccc1(**C**(C)(C)O) [2-phenyl-2-propanol]
C−(H)$_2$(O)(CO)	−70.28	OC(=O)**C**OC [methoxyacetic acid]
C−(C)$_2$(O)(CO)	−108.13	C**C**(C)(O)C(=O)O [α-hydroxyisobutyric acid]
C−(H)$_2$(O)(Ct)	92.85	O**C**C#C**C**O [2-butyne-1,4-diol]
CB−(H)(CB)(OA)	83.62	O1**C**=CC=C1(C=O) [furfural]
OA-(CB)$_2$	−209.55	**O**1C=CC=C1(C=O) [furfural]
OA-(CB)(CBF)	−301.16	c1cccc2**o**c(C)cc21 [2-methylbenzofuran]
CBF−(OA)(CB)(CBF)	110.97	c1cccc2c3cccc**c**3o**c**21 [dibenzofuran]
OA-(CBF)$_2$	−243.33	c1cccc2c3ccccc3**o**c21 [dibenzofuran]
CB−(C)(CB)(OA)	130.20	O1**C**(CO)=CC=C1 [furfuryl alcohol]
CB−(CO)(CB)(OA)	93.44	O1C=CC=**C**1(C=O) [furfural]
C. CHN and CHNO Groups		
C−(H)$_3$(N)	4.20	**C**N [methylamine]
C−(H)$_2$(C)(N)	9.84	C**C**N [ethylamine]
C−(H)(C)$_2$(N)	3.19	C**C**(C)N [isopropylamine]
C−(C)$_3$(N)	−5.02	C**C**(C)(C)N [*tert*-butylamine]
C−(H)$_2$(CB)(N)	20.40	c1(**C**N)ccccc1 [benzylamine]
C−(H)$_2$(Cd)(N)	−28.57	C=C**C**N [allylamine]
C−(H)(C)$_2$(N) cyclic imine	−9.87	N1**C**(C)C1 [propyleneimine]
N−(H)$_2$(C)	21.53	C**N** [methylamine]
N−(H)(C)$_2$	0.70	CC**N**CC [diethylamine]
N−(C)$_3$	−30.64	C**N**(C)C [trimethylamine]
N−(H)(C)$_2$ cyclic imine	40.56	C1C**N**1 [ethyleneimine]
N−(H)$_2$(CB)	−1.74	c1(**N**)ccccc1 [aniline]
N−(H)(C)(CB)	−17.57	c1(**N**C)ccccc1 [*N*-methylaniline]
N−(C)$_2$(CB)	−45.23	c1(**N**(C)C)ccccc1 [*N*,*N*-dimethylaniline]
N−(H)(CB)$_2$	−86.30	c1ccccc1**N**c2ccccc2 [diphenylamine]
CB−(H)(CBF)(NI)	86.01	C1=CC2=C(C=C1)C=CN=**C**2 [isoquinoline]
N−(H)(CB)(CBF)	−174.33	C1=CC2=C(C=C1)C=C**N**2 [indole]
CB−(C)(CB)(NI)	66.16	n1**c**(C)cccc1 [2-methylpyridine]
CB−(H)(NI)(OA)	179.63	O1**C**=NC=C1 [oxazole]
N−(H)(CBF)$_2$	−232.30	C1=CC2=C(C=C1)C1=C(C=CC=C1)**N**2 [dibenzopyrrole]
NI−(CBF)$_2$	184.08	**n**1c(cccc2)c2cc(cccc3)c31 [acridine]
NI−(CB)(CBF)	37.81	C2(=1)**N**=CC=CC1C=CC=C2 [quinoline]
NI−(CB)$_2$	−132.09	**n**1ccccc1 [pyridine]
NI−(NI)(CB)	−17.16	**N**1=NC=CC=C1 [pyridazine]
CBF−(CB)(CBF)(N)	116.25	C1=C**C**2=C(C=C1)C=CN2 [indole]

Table 1 (Continued)

group	Tb/K	example
		C. CHN and CHNO Groups (continued)
CBF−(CB)(CBF)(NI)	−86.87	**C**2(=1)N=CC=CC1C=CC=C2 [quinoline]
CB−(N)(CB)$_2$	60.15	**c**1(N)**c**cccc1 [aniline]
CB−(NO$_2$)(CB)$_2$	38.56	**c**1(N(=O)=O)ccccc1 [nitrobenzene]
CB−(CN)(CB)$_2$	52.67	**c**1(C#N)ccccc1 [benzonitrile]
CB−(H)(NI)$_2$	151.19	N1=**C**N=CC=C1 [pyrimidine]
CB−(H)(CB)(NI)	77.40	N1=CN=**C**C=**C**1 [pyrimidine]
CO−(H)(N)	−11.73	CN**C=O** [*N*-methylformamide]
CO−(C)(N)	148.95	C**C(=O)**N [acetamide]
N−(H)$_2$(CO)	205.23	CC(=O)**N** [acetamide]
N−(H)(C)(CO)	152.88	C**N**C=O [*N*-methylformamide]
N−(C)$_2$(CO)	84.86	C**N**(C)C=O [*N*,*N*-dimethylformamide]
N−(H)(CB)(CO)	64.84	c1ccccc1(**N**C(=O)C) [acetanilide]
C−(H)$_3$(CN)	91.71	**C**C#N [acetonitrile]
C−(H)$_2$(C)(CN)	83.64	C**C**C#N [propionitrile]
C−(H)(C)$_2$(CN)	62.99	C**C**(C)C#N [isobutyronitrile]
C−(H)$_2$(Cd)(CN)	37.27	C=C**C**C#N [vinylacetonitrile]
Cd−(H)(CN)	61.05	C=**C**C#N [acrylonitrile]
Cd−(C)(CN)	100.10	C=**C**(C)C#N [methacrylonitrile]
C−(H)$_2$(CO)(CN)	−72.67	N#C**C**C(=O)OC [methyl cyanoacetate]
C−(H)$_3$(NO$_2$)	61.08	**C**N(=O)=O [nitromethane]
C−(NO$_2$)$_4$	−138.10	N(=O)(=O)**C**(N(=O)=O)(N(=O)=O)N(=O)=O [tetranitromethane]
C−(H)$_2$(C)(NO$_2$)	36.36	C**C**N(=O)=O [nitroethane]
C−(H)(C)$_2$(NO$_2$)	40.18	C**C**(N(=O)=O)C [2-nitropropane]
O−(C)(NO$_2$)	−70.24	O=N(=O)**O**CC**O**N(=O)=O [ethylene glycol dinitrate]
C−(H)$_2$(C)(NCO)	9.90	CCC**C**N=C=O [*n*-butyl isocyanate]
C−(H)(C)$_2$(NCO)	9.27	**C**1(N=C=O)CCCCC1 [cyclohexyl isocyanate]
CB−(NCO)(CB)$_2$	4.33	**c**1(N=C=O)ccccc1 [phenyl isocyanate]
CN−(H)	78.27	**C#N** [hydrogen cyanide]
C−(H)$_2$(CN)$_2$	166.60	N#C**C**C#N [malononitrile]
(CN)$_2$	−40.81	**N#CC#N** [cyanogen]
		D. S Groups
C−(H)$_3$(S)	13.04	**C**SCC [methyl ethyl sulfide]
C−(H)$_2$(C)(S)	23.70	CS**C**C [methyl ethyl sulfide]
C−(H)(C)$_2$(S)	17.27	C**C**(S)C [isopropyl mercaptan]
C−(C)$_3$(S)	17.74	CS**C**(C)(C)C [methyl *tert*-butyl sulfide]
C−(H)$_2$(CB)(S)	38.49	c1(**C**S)ccccc1 [benzyl mercaptan]
CB−(S)(CB)$_2$	−5.89	**c**1(S)ccccc1 [phenyl mercaptan]
S−(C)(H)	−10.15	c1(C**S**)ccccc1 [benzyl mercaptan]
S−(CB)(H)	28.94	c1(**S**)ccccc1 [phenyl mercaptan]
S−(C)$_2$	−29.40	C**S**C [dimethyl sulfide]
S−(C)(S)	−13.36	CS**S**C [dimethyl disulfide]
CB−(C)(CB)(SA)	2.17	S1**C**(C)=CC=C1 [2-methylthiophene]
CB−(H)(CB)(SA)	−1.50	S1C(C)=CC=**C**1 [2-methylthiophene]
SA-(CB)$_2$	−19.21	**S**1C(C)=CC=C1 [2-methylthiophene]
SA-(CB)(CBF)	−240.37	c1(**S**C=C2)c2cccc1 [benzothiophene]
CBF−(SA)(CB)(CBF)	260.72	**c**1(SC=C2)c2cccc1 [benzothiophene]
SA-(CBF)$_2$	−501.57	C1=CC2=C(C=C1)**S**C1=C2C=CC=C1 [dibenzothiophene]
		E. Si Groups
Si−(C)$_3$(O)	22.84	C**[Si]**(C)(C)O[Si](C)(C)O**[Si]**(C)(C)C [octamethyltrisiloxane]
Si−(C)(Cd)(Cl)$_2$	18.19	C=C**[Si]**(C)(Cl)Cl [methyl vinyl dichlorosilane]
Si−(C)$_2$(O)$_2$	38.32	C[Si](C)(C)O**[Si]**(C)(C)O[Si](C)(C)C [octamethyltrisiloxane]
Si−(H)$_2$(C)(Cl)	−58.72	C**[Si]**([H])([H])Cl [methyl chlorosilane]
Si−(H)(C)(Cl)$_2$	−89.39	C**[Si]**([H])(Cl)Cl [methyl dichlorosilane]
Si−(C)(Cl)$_3$	−118.69	C**[Si]**(Cl)(Cl)Cl [methyl trichlorosilane]
Si−(Cd)(Cl)$_3$	−33.25	C=C**[Si]**(Cl)(Cl)Cl [vinyltrichlorosilane]
Si−(H)$_3$(C)	−45.20	C**[Si]**([H])([H])[H] [methylsilane]
Si−(H)$_2$(C)$_2$	−27.55	C**[Si]**(C)([H])[H] [dimethylsilane]
Si−(H)(C)$_3$	−18.42	C**[Si]**(C)(C)[H] [trimethylsilane]
Si−(H)(C)$_2$(Cl)	−44.01	C**[Si]**(C)(Cl)[H] [dimethylchlorosilane]
Si−(C)$_3$(Cl)	−32.72	C**[Si]**(C)(Cl)C [trimethylchlorosilane]
Si−(C)$_2$(Cl)$_2$	−67.67	C**[Si]**(C)(Cl)Cl [dimethyldichlorosilane]
Si−(H)(O)$_3$	46.54	CO**[Si]**([H])(OC)OC [trimethoxysilane]
C−(H)$_3$(Si)	−14.24	**C**[Si]([H])([H])[H] [methylsilane]
Cd−(H)(Si)	−85.64	C=**C**[Si](Cl)(Cl)Cl [vinyltrichlorosilane]
O−(Si)$_2$	−55.64	C[Si](C)(C)**O**[Si](C)(C)**O**[Si](C)(C)C [octamethyltrisiloxane]
O−(C)(Si)	−109.35	C**O**[Si]([H])(**O**C)**O**C [trimethoxysilane]
		F. Halogen Groups
C−(H)$_3$(F)	−49.23	**C**F [methyl fluoride]
C−(H)$_3$(Cl)	−39.88	**C**[Cl] [methyl chloride]
C−(H)$_3$(Br)	−109.52	**C**Br [methyl bromide]
C−(H)$_3$(I)	−149.11	**C**I [methyl iodide]
C−(C)(F)$_3$	−121.69	F**C**(F)(F)CF [1,1,1,2-tetrafluoroethane]
C−(H)$_2$(C)(F)	−26.06	FC(F)(F)**C**F [1,1,1,2-tetrafluoroethane]
C−(H)(C)$_2$(F)	−3.56	FC(F)(F)**C**(F)C(F)F [1,1,1,2,3,3-hexafluoropropane]

Table 1 (Continued)

group	Tb/K	example
F. Halogen Groups (continued)		
C−(H)(C)(F)$_2$	−72.65	FC(F)(F)C(F)**C**(F)F [1,1,1,2,3,3-hexafluoropropane]
C−(C)$_2$(F)$_2$	−61.13	**C**1(F)(F)**C**(F)(F)**C**(F)(F)**C**1(F)(F) [octafluorocyclobutane]
C−(C)(Cl)(F)$_2$	−109.43	Cl**C**(F)(F)C(Cl)F [1,2-dichloro-1,1,2-trifluoroethane]
C−(H)(C)(Cl)(F)	−70.09	ClC(F)(F)**C**(Cl)F [1,2-dichloro-1,1,2-trifluoroethane]
C−(C)(Cl)$_3$	−77.44	Cl**C**(Cl)(Cl)C(F)(F)F [1,1,1-trichlorotrifluoroethane]
C−(H)(C)(Cl)$_2$	−43.98	**C**([Cl])([Cl])C[Cl] [1,1,2-trichloroethane]
C−(H)$_2$(C)(Cl)	−11.65	C([Cl])([Cl])**C**[Cl] [1,1,2-trichloroethane]
C−(H)(C)$_2$(Cl)	−16.64	**C**([Cl])(C)C [isopropyl chloride]
C−(C)$_3$(Cl)	−22.71	**C**(C)([Cl])(C)C [*tert*-butyl chloride]
C−(H)(C)(Br)$_2$	−103.53	C**C**(Br)Br [1,1-dibromoethane]
C−(H)$_2$(C)(Br)	−67.07	C**C**Br [bromoethane]
C−(H)(C)$_2$(Br)	−74.51	C**C**(Br)C [2-bromopropane]
C−(H)$_2$(C)(I)	−111.50	C**C**I [ethyl iodide]
C−(H)(C)$_2$(I)	−120.30	C**C**(I)C [isopropyl iodide]
C−(H)(C)(Br)(Cl)	−95.86	FC(F)(F)**C**(Br)Cl [halothane]
C−(C)(Cl)$_2$(F)	−95.24	Cl**C**(Cl)(F)C [1,1-dichloro-1-fluoroethane]
C−(C)(Br)(F)$_2$	−148.51	Br**C**(F)(F)C(F)(F)Br [1,2-dibromotetrafluoroethane]
Cd−(H)(F)	−62.86	C=**C**F [vinyl fluoride]
Cd−(H)(Cl)	−31.76	**C**([Cl])=C([Cl])[Cl] [trichloroethylene]
Cd−(H)(Br)	−104.51	C=**C**Br [vinyl bromide]
Cd−(C)(Cl)	15.67	C**C**(Cl)=C [2-chloropropene]
Cd−(F)$_2$	−102.31	ClC=**C**(F)F [2-chloro-1,1-difluoroethylene]
Cd−(Cl)$_2$	−54.43	C([Cl])=**C**([Cl])[Cl] [trichloroethylene]
Cd−(Cl)(F)	−74.76	Cl**C**(F)=C(F)F [chlorotrifluoroethylene]
Cd−(Br)(F)	−118.82	Br**C**(F)=C(F)F [bromotrifluoroethylene]
CB−(F)(CB)$_2$	−28.78	**c**1(F)ccccc1 [fluorobenzene]
CB−(Cl)(CB)$_2$	−11.54	c1cccc**c**1([Cl]) [chlorobenzene]
CB−(Br)(CB)$_2$	−55.76	**c**1(Br)ccccc1 [bromobenzene]
CB−(I)(CB)$_2$	−97.91	**c**1(I)ccccc1 [iodobenzene]
C−(H)$_2$(CO)(Cl)	−163.75	Cl**C**C(=O)O [chloroacetic acid]
C−(H)(CO)(Cl)$_2$	−221.74	Cl**C**(Cl)C(=O)O [dichloroacetic acid]
C−(CB)(F)$_3$	−101.19	c1(**C**(F)(F)F)ccccc1 [benzotrifluoride]
C−(H)$_2$(CB)(Cl)	3.62	c1ccccc1(**C**[Cl]) [benzyl chloride]
CO−(C)(Cl)	134.96	C**C(=O)**Cl [acetyl chloride]
CO−(CB)(Cl)	21.34	c1(**C(=O)**Cl)ccccc1 [benzoyl chloride]
C−(H)$_2$(Cd)(Cl)	−45.84	C=C**C**([Cl]) [3-chloropropene]
C−(H)$_2$(Ct)(Cl)	−3.88	C#C**C**[Cl] [propargyl chloride]
CB−(CB)(CBF)(Br)	18.45	c12ccccc1ccc**c**2(Br) [1-bromonaphthalene]
CB−(CB)(CBF)(Cl)	−7.39	c12**c**(Cl)cccc1cccc2 [1-chloronaphthalene]
C−(CO)(Cl)$_3$	−263.55	**C**([Cl])([Cl])([Cl])C(=O)O [trichloroacetic acid]
C−(H)$_2$(O)(Cl)	51.40	Cl**C**OC [chloromethyl methyl ether]
C−(Cl)$_4$	−130.29	**C**([Cl])([Cl])([Cl])[Cl] [carbon tetrachloride]
Cd−(Cd)(Cl)	−10.60	C=**C**(Cl)C=C [chloroprene]
C−(CB)(Cl)$_3$	−61.10	c1ccccc1(**C**([Cl])([Cl])([Cl])) [benzotrichloride]
C−(Cd)$_2$(Cl)$_2$	−125.93	C1([Cl])=C([Cl])**C**([Cl])([Cl])C([Cl])=C1([Cl]) [hexachlorocyclopentadiene]
C−(H)(C)(Cd)(Cl)	−62.75	C=C**C**(Cl)CCl [3,4-dichloro-1-butene]
C−(H)(CB)(Cl)$_2$	−14.03	c1ccccc1**C**(Cl)Cl [benzyl dichloride]
Cd−(C)(F)	30.58	FC(F)(F)**C**(F)=C(F)F [hexafluoropropylene]
C−(Cd)(F)$_3$	−165.69	F**C**(F)(F)C(F)=C(F)F [hexafluoropropylene]
C−(CO)(F)$_3$	−302.34	F**C**(F)(F)C(=O)**C**(F)(F)F [hexafluoroacetone]
C−(C)$_2$(Cl)(F)	−64.56	C1(Cl)(F)**C**(Cl)(F)C(F)(F)C1(F)(F) [1,2-dichlorohexafluorocyclobutane]
C−(H)(O)(F)$_2$	25.61	F**C**(F)OCC(F)(F)F [2-(difluoromethoxy)−1,1,1-trifluoroethane]
CO−(O)(Cl)	−221.08	[Cl]**C(=O)**OCC [ethyl chloroformate]
C−(H)$_2$(Cl)$_2$	−52.17	[Cl]**C**[Cl] [dichloromethane]
C−(H)(Cl)$_3$	−92.47	**C**([Cl])([Cl])[Cl] [chloroform]
C−(H)$_2$(Cl)(F)	−67.94	Cl**C**F [chlorofluoromethane]
C−(Cl)$_2$(F)$_2$	−187.45	Cl**C**(Cl)(F)F [dichlorodifluoromethane]
C−(Cl)$_3$(F)	−159.16	Cl**C**(Cl)(Cl)F [trichlorofluoromethane]
C−(H)(Cl)(F)$_2$	−135.81	Cl**C**(F)F [chlorodifluoromethane]
C−(Cl)(F)$_3$	−210.42	Cl**C**(F)(F)F [chlorotrifluoromethane]
C−(H)$_2$(F)$_2$	−71.28	F**C**F [difluoromethane]
C−(H)(F)$_3$	−144.33	F**C**(F)F [trifluoromethane]
C−(F)$_4$	−227.76	F**C**(F)(F)F [carbon tetrafluoride]
C−(H)(Br)(F)$_2$	−189.61	Br**C**(F)F [bromodifluoromethane]
C−(H)$_2$(I)$_2$	−171.28	I**C**I [diiodomethane]
C−(H)(Cl)$_2$(F)	−117.34	Cl**C**(F)Cl [dichlorofluoromethane]
C−(H)(Br)$_3$	−188.04	Br**C**(Br)Br [tribromomethane]
CO−(F)$_2$	−137.86	F**C(=O)**F [carbonyl fluoride]
CO−(Cl)$_2$	−111.31	Cl**C(=O)**Cl [phosgene]
C−(H)$_2$(Br)$_2$	−141.52	Br**C**Br [dibromomethane]
C−(H)2(Br)(Cl)	−103.59	Br**C**Cl [bromochloromethane]
C−(Br)(Cl)$_3$	−164.01	Br**C**(Cl)(Cl)Cl [bromotrichloromethane]
C−(Br)(Cl)(F)$_2$	−229.06	F**C**(Br)(Cl)F [bromochlorodifluoromethane]
C−(Br)(F)$_3$	−260.12	F**C**(Br)(F)F [bromotrifluoromethane]
C−(Br)$_2$(F)$_2$	−262.29	F**C**(Br)(Br)F [dibromodifluoromethane]

Table 2. Extension of DH Groups to New Structural Configurations

notation	definition	example
Ca2	C adjacent to allenic carbon	C−$(H)_2$(C)(Ca2) in C=C=C**C**C
OA	O in aromatic ring	OA−$(CB)_2$ in c1c**o**cc1 (furan)
NI	N with double bond in aromatic ring	CB−(H)(CB)(NI) in n1**c**cccc1 (pyridine)
SA	S atom in aromatic ring	SA−(CB)(CBF) in c1(**S**C=C2)c2cccc1 (benzothiophene)
NCO	isocyanate group	CB−(NCO)$(CB)_2$ in c1cccc**c**1N=C=O (phenyl isocyanate)
CBF	common C in fused aromatic rings	CBF−(SA)(CB)(CBF) in **c**1(SC=C2)c2cccc1 (benzothiophene)

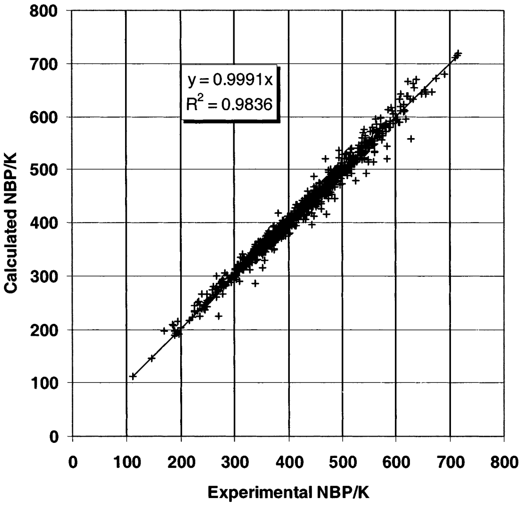

Figure 1. Regression results for NBP calculated from eq 3 compared to experimental data for the training set.

already found acceptance and use in some software applications. The best correlation obtained was

$$\mathrm{NBP} = 35.11 + 36.93\sqrt{\frac{M}{\mathrm{kg/mol}}} + 52.83\sqrt{\frac{V_{\mathrm{VDW}}}{\mathrm{m^3/kmol}}} + \sum_k n_k \mathrm{Tb}_k \quad (3)$$

where V_{VDW} is the van der Waals volume and M is the molecular weight. The values of the group contributions, Tb_k, obtained from the global least-squares regression of the entire training set are given in Table 1. For each group shown in Table 1, the SMILES formula is also provided for an example molecule in which the applicable group is shown in bold to help clarify the definition. Values for V_{VDW} can be obtained from several sources including the DIPPR database, computational chemistry packages such as PC SPARTAN[20] and from references such as that by Bondi.[21] A brief primer on SMILES nomenclature is provided in the Appendix that can be used as a key to the example molecules in Table 1.

Although the DH group definitions form the basis of the GCM used in this work, there were a few cases in which greater specificity was desired for groups treated equivalently in the DH definitions, even though one of the bonded atoms is chemically distinct. In many cases, this differentiation appears as a word after the group. For example, the key word *ether*, *ester*, *epoxy*, or *alcohol* is appended to the C−(H)(O)$(C)_2$ structural configuration to define four separate groups based on the chemical nature of the oxygen atom. In six cases, it was necessary to extend the DH definitions to new structural configurations. These cases with their definitions and illustrative molecules are listed in Table 2.

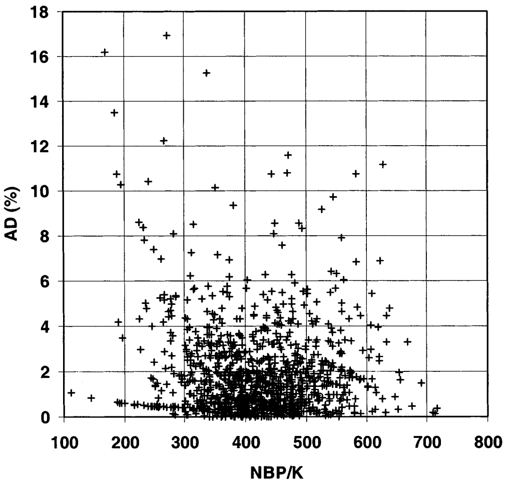

Figure 2. Percentage absolute deviation (AD) of eq 3 values from experimental NBP data for the training set.

Table 3. Overall NBP Correlation Results

	eq 3			CG−2nd order		
data set	AAD (K)	AAD (%)	σ/K	AAD (K)	AAD (%)	σ/K
training (1141 compds)	7.75	1.9	8.3	7.71	2.0	10.5
test (384 compds)	13.0	2.7	13.3	13.9	3.0	14.5

The regression results for the 1141-compound training set are shown in Figures 1 and 2. Figure 1 shows the agreement between experimental NBP values in the training set and values calculated using eq 3 and the regressed groups in Table 1. Figure 2 also applies to the training set but illustrates the percentage absolute deviation (AD) as a function of NBP. The AAD for the training set was 7.75 K (1.9%) with a standard deviation of 8.27 K. These results compare favorably with those reported for the CG training set, also from the DIPPR database, as shown in Table 3. The maximum deviation is seen from Figure 2 to be 17% (46 K); this is for decafluorobutane. Perfluorinated compounds are notoriously difficult to handle without secondary corrections because of the high electronegativity of each fluorine. Also, the first member of some organic families are the most difficult to correlate with the DH groups because of the different extended electronic environment experienced by a group within a larger molecule compared to that found in a very short molecule in which little electron delocalization can occur. Thus, besides decafluorobutane, the next three largest deviations shown in Figure 2 are for ethylene, methanol, and ethane, respectively. One seldom needs to predict the first member in a family, so estimations with this method are generally expected to agree with experimental values approximately within the 5% uncertainty of the selected training set. The results appear to be unbiased as evidenced by the 0.0 K average difference between predicted and experimental values and by a slope of 0.991 (R^2 = 0.984; see Figure 1) for a straight-line regression of the data, constrained through the origin. It is also of interest to examine the results in terms of chemical families, as defined in the DIPPR database. This is done in Table 4. As can be seen from this table, 1-alkenes, fluorohydrocarbons, cycloalkanes, and *n*-alcohols have

Table 4. NBP Correlation Results by Chemical Family

		abs. deviation/K		abs. % deviation	
family	no. of compds	mean	std dev	mean	std dev
n-alcohols	12	18.26	13.98	4.50	4.23
n-aliphatic acids	13	14.36	8.84	2.97	1.88
n-aliphatic primary amines	13	7.72	1.57	1.93	0.60
n-alkanes	20	9.94	5.70	2.81	3.04
1-alkenes	19	10.38	5.89	3.06	3.67
2,3,4-alkenes	18	4.26	3.48	1.33	1.25
n-alkylbenzenes	10	4.26	3.47	1.01	1.00
acetates	20	1.60	1.44	0.38	0.35
aldehydes	23	5.02	5.57	1.25	1.29
aliphatic ethers	31	6.32	5.01	1.81	1.61
alkylcyclohexanes	15	5.35	4.08	1.25	0.90
alkylcyclopentanes	11	6.72	4.90	1.83	1.38
alkynes	17	3.13	2.00	1.08	0.98
anhydrides	5	9.53	12.04	2.02	2.55
aromatic alcohols	30	9.31	7.43	1.83	1.36
aromatic amines	28	9.11	14.08	1.68	2.28
aromatic carboxylic acids	2	9.91	2.69	1.76	0.42
aromatic chlorides	15	4.86	3.76	1.05	0.83
aromatic esters	12	9.12	3.77	1.71	0.65
C, H, Br compounds	17	8.88	9.36	2.29	2.55
C, H, F compounds	27	8.35	10.64	3.45	4.26
C, H, I compounds	8	6.43	7.72	1.65	2.00
C, H, multihalogen compounds	36	5.34	5.70	1.81	2.01
C, H, NO_2 compounds	10	4.03	3.28	0.91	0.71
C1/C2 aliphatic chlorides	18	7.26	5.22	2.20	1.77
C3 and higher aliphatic chlorides	26	4.63	4.95	1.25	1.27
cycloaliphatic alcohols	10	6.90	5.15	1.46	1.01
cycloalkanes	6	12.80	7.10	4.27	3.31
cycloalkenes	9	9.55	7.71	2.78	2.37
dialkenes	25	5.44	4.73	1.57	1.38
dicarboxylic acids	1	8.06	NA	1.31	NA
dimethylalkanes	21	5.19	5.18	1.44	1.75
diphenyl/polyaromatics	11	16.26	16.23	2.80	2.89
epoxides	13	10.35	6.95	2.76	1.75
ethyl and higher alkenes	11	3.25	2.72	0.92	0.79
formates	12	2.49	1.53	0.68	0.47
isocyanates/diisocyanates	3	1.19	0.00	0.28	0.02
ketones	32	8.02	6.36	1.97	1.58
mercaptans	21	6.29	4.47	1.63	1.23
methylalkanes	17	6.09	4.39	1.66	1.57
methylalkenes	21	4.54	4.03	1.36	1.29
multi-ring cycloalkanes	3	6.07	4.02	1.24	0.78
naphthalenes	13	7.30	6.34	1.33	1.08
nitriles	21	7.21	6.98	1.65	1.58
organic salts	3	10.70	9.95	2.84	2.74
other aliphatic acids	12	6.53	5.15	1.42	1.11
other aliphatic alcohols	26	7.98	6.45	1.98	1.67
other aliphatic amines	18	8.82	7.98	2.21	1.65
other alkanes	23	7.92	5.41	1.95	1.28
other alkylbenzenes	43	4.48	4.41	0.95	0.88
other amines, imines	29	8.47	11.30	1.91	2.27
other condensed rings	10	5.98	6.81	0.95	1.05
other ethers/diethers	13	11.06	10.62	2.63	2.62
other hydrocarbon rings	11	5.28	5.82	1.29	1.44
other monoaromatics	13	3.70	3.69	0.81	0.82
other polyfunctional C, H, O	36	13.36	13.18	2.80	2.54
other saturated aliphatic esters	12	10.38	11.60	1.82	1.83
peroxides	2	4.32	1.48	1.12	0.39
polyfunctional acids	2	11.15	13.58	2.33	2.84
polyfunctional amides/amines	14	13.97	13.76	2.97	3.06
polyfunctional C, H, N, halide, (O)	5	8.51	4.26	1.68	0.84
polyfunctional C, H, O, halide	33	7.61	8.28	1.90	2.00
polyfunctional C, H, O, N	12	9.45	6.76	2.00	1.58
polyfunctional C, H, O, S	4	9.48	5.66	2.08	1.38
polyfunctional esters	11	13.05	10.70	2.76	2.15
polyfunctional nitriles	1	1.19	NA	0.23	NA
polyols	26	18.24	11.44	3.63	2.30
propionates and butyrates	13	2.13	1.99	0.56	0.54
silanes/siloxanes	23	7.98	10.14	1.74	1.91
sulfides/thiophenes	19	5.75	4.69	1.39	1.05
terpenes	5	6.03	5.39	1.33	1.18
unsaturated aliphatic esters	16	8.14	4.10	1.95	0.92
all compounds	**1141**	**7.75**	**8.27**	**1.88**	**2.02**

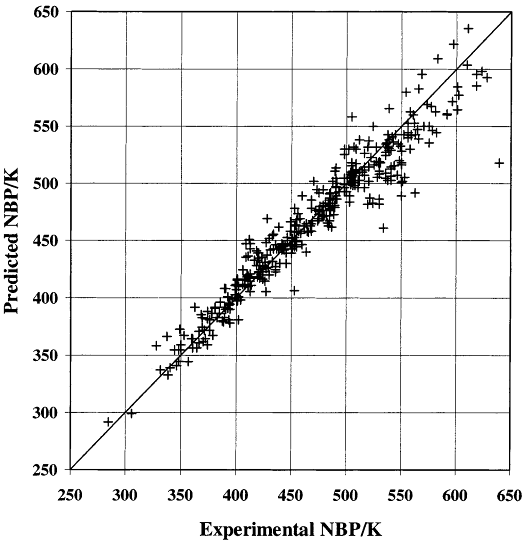

Figure 3. Prediction results for NBP calculated from eq 3 compared to experimental data for the test set.

absolute average percentage deviations above 3% and could be the focus of additional work. It was through this type of analysis that we decided to make the few extensions to the DH group designations mentioned previously.

Testing of the new NBP correlation in a "predictive" mode was done using a test set of 384 compounds that were not in the DIPPR database. These NBP values were obtained from the *Handbook of Chemistry and Physics*.[21] Because both CG and eq 3 were trained with data from the DIPPR database, this test set should provide a reasonable basis for testing their extrapolation capability to additional compounds. Results for this test set are shown in Figure 3 and overall statistics are included in Table 3. The predictive results are quite good and compare very favorably to those for the CG method. There is more scatter about the 45° line in Figure 3 than in Figure 1, as will always be the case when comparing predicted and regressed results. We have no way of separating predictive accuracy from possible experimental uncertainty for these data because, unlike the training set, no evaluation and accuracy designation is available for the test set. This reasserts the importance of evaluated data in a database used for correlation development. What we can say is that the DH group definitions in conjunction with eq 3 extrapolate at least as well as the CG second-order method and provide predicted NBP values of comparable accuracy. Advantages of eq 3 are the broad compound domain (including silanes, siloxanes, isocyanates, aromatic imines, epoxides, thiophenes, etc.), the computational simplicity of DH groups (no second-order corrections are required), and compatibility with current software implementations of DH thermodynamic properties. Linear regression of the data in Figure 3 gives a slope of 0.992 ($R^2 = 0.910$), again indicating no bias in predicted data with increasing NBP.

Example Calculations

We show here a few calculation examples to illustrate the use of the method. In each example, we construct a table that shows the number of groups of type k (or n_k), the contribution for group k (or Tb_k), and the product of these two terms to be summed in eq 3. Also shown in each tabular example is the value of the molecular weight in kg/kmol and the van der Waals volume in m^3/kmol as obtained from the DIPPR database as well as the product of the appropriate coefficient and the square root of these quantities as shown in eq 3. This is done to make the calculated NBP the sum of the entries in the last column in accordance with eq 3 and a spreadsheet implementation of the calculation.

Example 1: Estimate NBP for 4-Methylphenol (*p*-Cresol)

group k	$n(k)$	Tb(k)	$n(k)$·Tb(k)
$C-(H)_3(CB)$	1	3.7	3.7
$CB-(C)(CB)_2$	1	1.4	1.4
$CB-(CB)_2(H)$	4	2.22	8.88
$CB-(CB)_2(O)$	1	18.01	18.01
$O-(CB)(H)$	1	29.64	29.64

eq 3 term	value	coeff.	coeff·sqrt(value)
Tb_0	1	35.11	35.11
M	108.14	36.93	384.04
V(VDW)	0.06503	−52.83	−13.47
predicted			467.30
experimental			475.13
error			−1.6%

Example 2: Estimate NBP for Isopropyl Acetate

group k	$n(k)$	Tb(k)	$n(k)$·Tb(k)
$C-(H)_3(C)$	2	−10.3	−20.6
$C-(H)(O)(C)_2$ ester	1	65.42	65.42
$O-(CO)(C)$	1	125.04	125.04
$CO-(O)(C)$	1	−34.61	−34.61
$C-(H)_3(CO)$	1	−168.37	−168.37

eq 3 term	value	coeff.	coeff.·sqrt(value)
Tb_0	1	35.11	35.11
M	102.133	36.93	373.22
V(VDW)	0.06299	−52.83	−13.26
predicted			361.95
experimental			361.65
error			0.1%

Example 3: Estimate NBP for 2-Methyl-1-pentene

group k	$n(k)$	Tb(k)	$n(k)$·Tb(k)
$Cd-(H)_2$	1	−13.19	−13.19
$Cd-(C)_2$	1	91.28	91.28
$C-(H)_3(Cd)$	1	−49.27	−49.27
$C-(H)_2(C)(Cd)$	1	−38.69	−38.69
$C-(H)_2(C)_2$	1	−0.04	−0.04
$C-(H)_3(C)$	1	−10.3	−10.3

eq 3 term	value	coeff.	coeff.·sqrt(value)
Tb_0	1	35.11	35.11
M	84.161	36.93	338.79
V(VDW)	0.06475	−52.83	−13.44
predicted			340.25
experimental			335.25
error			1.5%

It is important that the correct (most precisely defined) groups are used when estimating NBP with this method. Substitution of a group for the correct one can lead to poor results, even though one might think the groups should be quite similar. This occurs for example when two groups are usually found in the same molecule and have large values of opposite sign. For example, a methyl ketone would contain both of the groups $CO-(C)_2$ and $C-(H)_3(CO)$. The values for these two groups are large and of opposite signs so that they offset each other in ketones. It would be incorrect to substitute $C-(H)_3C$ for $C-(H)_3(CO)$ because the compensation of two correctly adjoined groups is left out. This characteristic is inherent in the DH group definitions, but it is necessitated by the very different electronic environments that different neighboring atoms can produce on the same central group.

Summary

Coupled with commercial QSPR software, the DIPPR database provides a convenient base for development and testing of GCM- and MolD-style property correlations. Because of the evaluated nature of the database, training sets of known accuracy can be selected from it to optimize accuracy and breadth of the developed equation. In this work, we have used these features to develop a second-order GCM for NBP that has a large compound domain. Regression of the group constants produced an AAD of 7.75 K (1.9%); the AAD was 13.0 K (2.7%) when used in a predictive mode on a 384-compound test set. These results compare favorably in both accuracy and extrapolation capability to the CG second-order method, which has been arguably the best generalized correlation for NBP available. The new method is based on the DH groups and is therefore directly applicable to automated software already based on the DH method.

Appendix. Primer on SMILES Nomenclature

Though SMILES is a comprehensive chemical notation, five simple rules are adequate to represent the structure of a molecule for the illustrations used in Table 1.

Rule 1. Atoms are represented by atomic symbols: B, C, N, O, F, P, S, Cl, Br, and I. Hydrogen atoms may be omitted as they are inferred by the valance of the atom. Thus, C represents methane, CC represents ethane, and CCCl represents chloroethane.

Rule 2. Double bonds are represented with = and triple bonds are represented with #. Thus, C=C is ethylene and C#C is acetylene.

Rule 3. Branching is indicated by parentheses, and the original chain is continued after closure of the parentheses. Thus, CC(C)C represents isobutane, CC(C)(C)C represents neopentane, and CC(O)C represents 2-propanol.

Rule 4. Ring closures are indicated by pairs of matching digits. A digit appears after the first atom in the ring; the sequence of atoms in the SMILES formula subsequently shows the connected atoms in the ring, and finally the matching digit connects the atom before it to the first atom that opened the ring. Thus, C1CCCCC1 represents cyclohexane, C1CCCC1C(C)C represents isopropylcyclopentane, and C1CCCCC1C2CCCCC2 represents bicyclohexane.

Rule 5. Aromatic structures are represented with lower case letters. For example, c1ccccc1 represents benzene, Cc1ccc(Br)cc1 represents 4-bromotoluene, and n1ccccc1 represents pyridine.

Literature Cited

(1) Rowley, R. L.; Wilding, W. V.; Oscarson, J. L.; Zundel, N. A.; Marshall, T. L.; Daubert, T. E.; Danner, R. P. *DIPPR® Data Compilation of Pure Compound Properties*; Design Institute for Physical Properties, AIChE: New York, 2002.

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

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

(4) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods*; McGraw-Hill: New York, 1982.

(5) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally diverse quantitative structure−property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.

(6) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.

(7) Constantinou, L.; Gani, R. New group contribution method for estimating properties of pure compounds. *AIChE J.* **1994**, *40*, 1697−1710.

(8) Lydersen, A. L. *Estimation of critical properties of organic compounds*; Univ. Wisconsin Coll. Eng., Eng. Exp. Stn., Rept. 3; University of Wisconson, Madison, WI, 1955.

(9) Joback, K. G.; Reid, R. C. *Chem. Eng. Commun.* **1987**, *57*, 233.

(10) Fredenslund, Aa.; Gmehling, J.; Rasmussen, P. *Vapor Liquid Equilibria Using UNIFAC*; Elsevier: Amsterdam, 1977.

(11) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Normal boiling points for organic compounds: correlation and prediction by a quantitative structure−property relationship. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28−41.

(12) Domalski, E. S.; Hearing, E. D. 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.

(13) Benson, S. W. *Thermochemical Kinetics, 2nd Edition*; John Wiley & Sons: New York, 1976.

(14) *CHETAH Version 7.2: The ASTM Computer Program for Chemical Thermodynamic and Energy Release Evaluation* (NIST Special Database 16); NIST: Washington, DC, 1998; 4th ed.

(15) Rowley, J. R.; Wilding, W. V.; Oscarson, J. L.; Rowley, R. L. *DIADEM Version 2.0, DIPPR® Information And Data Evaluation Manager*; BYU-DIPPR TPL; Brigham Young University, Provo, UT, 2002.

(16) Wininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

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

(18) Rowley, R. J.; Oscarson, J. L.; Rowley, R. L.; Wilding, W. V. Development of an Automated SMILES Pattern Matching Program To Facilitate the Prediction of Thermophysical Properties by Group Contribution Methods. *J. Chem. Eng. Data* **2001**, *46*, 1110−1113.

(19) *TSAR Version 3.1*; Oxford Molecular Group, Oxford Molecular Limited: Oxford, 1997.

(20) *PC SPARTAN*; Wavefunction, Inc.: Irvine, CA, 1996.

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

(22) Lide, D. R., Ed.; *CRC Handbook of Chemistry and Physics, 81st Edition*; CRC Press: Boca Raton, FL, 2000.

Received for review April 15, 2002. Accepted May 29, 2002. Financial support of this work through the DIPPR 801 project is gratefully acknowledged.

JE0255372