Group-Contribution Method for Predicting Properties of Biochemical and Safety Interest and Temperature-Dependent Properties of Pure Components

Emmanuel Stefanis[a], Leonidas Constantinou[b], Costas Panayiotou[a]

[a] *Department of Chemical Engineering, Aristotle University of Thessaloniki, GR 54124, Thessaloniki, Greece and*
[b] *Frederick Research Center, P.O. Box 24729, Nicosia, Cyprus*

Abstract

A new group-contribution method for predicting properties of pure organic compounds is presented. It uses two kinds of groups: First-order groups for the description of the basic molecular structure of the compounds and second-order groups, which are based on the conjugation theory and improve the accuracy of the predictions. New first-order groups have been defined to ensure that the molecular structures of any compound of biochemical interest, including complex aromatic, multiring, and heterocyclic compounds, can be easily described. In addition, new second-order groups have been defined to enhance the reliability of the predictions and the applicability of the method. The temperature-independent properties that have been predicted by the new method are the octanol-water partition coefficient (logKow), total (Hildebrand) solubility parameters at 25^0 C, and the flash point. These properties have many applications in the chemical, pharmaceutical, and nutraceutical industries, as well as in the environmental protection. The method was extended for the first time to the prediction of temperature-dependent properties, such as vapor pressure, liquid density, heat of vaporization, surface tension and more, at a wide range of temperatures through its combination with the NRHB lattice-fluid theory.

The Proposed Method

The basic equation that gives the value of each property according to the molecular structure is[1]:

$$f(p) = \sum_i n_i F_i + \sum_j m_j S_j \qquad (1)$$

, where F_i is the contribution of the first-order group of type i that appears n_i times in the compound and S_i is the contribution of the second-order group of type j that appears m_j times in the compound. $f(p)$ is a single equation of the property, p, and is selected after a thorough study of the physicochemical and thermodynamic behaviour of the property. C is the universal constant of the model. The quantity $\Sigma m_j S_j$ is considered to be zero for compounds that do not have second-order groups.

The determination of the contributions is done by a two-step regression analysis:
- First step: determination of first-order group contributions only (F_i's).

-    Second step: using the F_i's contributions, the second-order groups are activated and the second-order group contributions (S_j's) are calculated through regression. These contributions act as a correction to the first-order approximation.

The modified Levenberg approach was used to minimize the total sum of squared errors between the experimental and predicted values of the properties. This was the criterion for the selection of the most appropriate equation to fit the experimental data. The model is applicable to organic compounds with three or more carbon atoms excluding the atom of the characteristic group (e.g., –COOH or -CHO).

A.   Prediction of Temperature-Independent Properties of Chemical, Biochemical, and Safety Interest

Up to 94 first-order group contributions and 49 second-order group contributions to Octanol-Water Partition Coefficients, Total Solubility Parameters (at 25^0 C) , and the Flash Point of pure organic compounds have been estimated.

The selected equations of estimation for each property are the following:

Octanol-water partition coefficients (logKow):
$$\log Kow = \sum_i n_i F_i + \sum_j m_j S_j + 0.097 \qquad (2)$$

Total Solubility Parameters (25^0 C) ($(kJ/m^3)^{(1/2)}$):

$$So\,\text{lub}\,ility\ Parameter = (\sum_i n_i F_i + \sum_j m_j S_j + 75954.1)^{0.383837} - 56.14 \quad (3)$$

Flash Point (K):
$$Flash \quad Po\,\text{int} = \sum_i n_i F_i + \sum_j m_j S_j + 216 \qquad (4)$$

Table 1 illustrates the overall statistical values of the estimation of the previously-mentioned temperature-independent properties. Table 2 shows a comparison between the same values and the corresponding statistical values for logKow estimation using other similar methods of estimation.

Table 1: Statistical Values of First- and Second-Order Approximations for the Prediction of Properties of Biochemical and Safety Interest

Property	Data Points	Standard Deviation 1st-order	Standard Deviation 2nd-order	AAE 1st-order	AAE 2nd-order	AAPE (%) 1st-order	AAPE (%) 2nd-order
logKow	422	0.315	0.267	0.226	0 .188	-	-
Total Solubility Parameters	1017	1.468	1.308	0.996	0.901	5.15	4.67
Flash Point	418	16.530	14.733	11.898	10.689	3.66	3.27

, where AAE is the average absolute error and AAPE the average absolute percentage error.

Table 2: Comparison of the accuracy of existing methods for logKow estimation

	KowWin Software Version 1.66[2]	Marrero, Gani[3]	New Method
Standard Deviation (SD)	0.44	0.34	**0.27**
Absolute Average Error (AAE)	0.32	0.24	**0.19**
Correlation coefficient (r²)	0.95	0.97	**0.99**

B. Prediction of Temperature-Dependent Properties

The prediction of temperature-dependent thermophysical properties is based on the NRHB lattice-fluid theory[4]. The NRHB (Non-Random Hydrogen-Bonding) theory is a novel theory which is based on the estimation of non-random distribution of free volume in a system. The only parameters needed for the estimation of properties such as vapor pressure, liquid density, heat of vaporization, and surface tension at a wide range of temperatures are the following characteristic scaling constants:

ε^* which is the average interaction energy per molecular segment,
v^* which is the average segmental volume and
v_{sp}^* which is the hard core volume

For the estimation of surface tension the influence parameter, κ, is also needed.

The steps of our methodology were the following:

(1) Calculation of experimental values of vapor pressure, liquid density and heat of vaporization at a selected temperature range for 469 compounds using the equations of the DIPPR experimental database[5]. The temperature range was equal to $T_m+I^*((T_c-T_m-10)/k_p)$, where $I=1, \ldots, 18$ the increasing number of data points, $k_p = 18$ the total number of data points, T_m the melting point and T_c the critical temperature of each compound.

(2) Calculation of surface-to-volume ratio, s, a geometric feature of each molecule. This was done by using the 78 UNIFAC group contributions[6] to the surface area parameter, Q_i and the volume parameter, R_i for each group i. The appropriate equation is: $s = \Sigma Q_i / \Sigma R_i$.

(3) Fitting of DIPPR experimental data and "s" data to the ε^*, v^*, v_{sp}^* characteristic scaling constants with a Visual FORTRAN program which uses a BOTM subroutine. The experimental values of the characteristic scaling constants for 469 organic compounds were defined by following this methodology. The compounds were especially selected so as to cover a large amount of homologous series of organic compounds.

(4) Prediction of scaling constants with the Constantinou-Gani group-contribution method[1]. For the estimation of first-order and second-order group contributions to the ε^*, v^*, v_{sp}^* scaling constants, and to the influence parameter, κ, the same methodology as that for the temperature-independent properties was followed.

(5) Test estimation of temperature-dependent properties at any temperature using the estimated scaling constants.

In the case of surface tension the following methodology was followed:

(1) Definition of experimental values of influence parameters, κ, for 467 selected compounds after fitting experimental values of surface tension calculated by the DIPPR database equations. The fitting has been done at a temperature equal to $0.6T_c$ for each compound, where T_c is the critical temperature. The defined influence parameters were considered to validate at any other temperature.

(2) Prediction of influence parameters with the Constantinou-Gani group-contribution method.

(3) Test estimation of surface tension at any temperature using the estimated scaling constants and influence parameters.

The best model found after the least-square analysis for all scaling constants and the influence parameter, κ, was the linear one. The equations that give the characteristic scaling constants and the influence parameter through the first-order and second-order group contributions are the following:

$$\varepsilon^* = \sum_i n_i F_i + \sum_j m_j S_j + 4455.4 \text{ J/mol} \tag{5}$$

$$v^* = \sum_i n_i F_i + \sum_j m_j S_j + 9.1748 \, cm^3 / mol \tag{6}$$

$$v_{sp}^* = \sum_i n_i F_i + \sum_j m_j S_j + 1.18081 \, cm^3 / g \tag{7}$$

$$\kappa = \sum_i n_i F_i + \sum_j m_j S_j + 0.16451 \tag{8}$$

The group contributions for the 78 first-order (UNIFAC) groups and 31 second-order groups to the characteristic scaling constants and the influence parameter are available.

Table 3 shows the overall improvement of the estimation of the three scaling constants and the influence parameter, κ that has been achieved after the introduction of 2nd order groups in the regression.

Table 3: Statistical values of first-order and second-order approximation for the estimation of the scaling constants.

Scaling constant	Data Points	Standard Deviation 1st order	Standard Deviation 2nd order	AAE 1st order	AAE 2nd order	AAPE (%) 1st order	AAPE (%) 2nd order
ε^*	411	201	160	140	116	2.60	2.16
v^*	411	0.851	0.787	0.576	0.528	4.30	3.88
v_{sp}^*	411	0.0319	0.0295	0.0226	0.0206	2.14	1.97
Influence parameter, κ	406	0.0167	0.0144	0.0108	0.0099	6.65	5.87

Comments and Conclusions

First of all, the method is simple and relatively accurate compared to similar existing group-contribution methods. The twenty-six newly introduced first-order groups and twelve new second-order groups ensure that any complex molecular structure (multiring, aromatic, and heterocyclic) can be easily described. There are only two levels of approximation in contrary to other methods that use a third-order approximation[3, 7]. The second level of approximation is based on the theory of conjugation operators and is sufficient to give a physical meaning to the method and to improve the overall accuracy of it. The predicted temperature-independent properties are of significant importance for the design of pharmaceutical and nutraceutical products, for the environmental protection and for the safety of the personnel in both chemical industries and laboratories. The new method's main advantage is that it is exclusively based on the molecular structure of each organic compound. No other experimental data are needed.

The present group-contribution method combined with the NRHB lattice-fluid theory is able to predict a series of temperature-dependent properties at a wide range of temperatures. The predictions can be expanded further to a range of different pressures. Only a set of three scaling constants, ε^*, v^* and v_{sp}^*, and an influence parameter, κ, in the case of surface tension estimation, are needed. All constants are temperature-independent and can be easily predicted by the group contributions and the equations 5-8 found in the present work. Finally, it should be emphasized that the lack of need for other experimental data is of crucial importance in the design of new, unknown products. This is a very important advantage compared to other temperature-dependent properties estimation methods.

Acknowledgments

References

(1) Constantinou, L.; Gani, R. New Group Contribution Method for Estimating Properties of Pure Compounds. *AIChe J.* **1994**, *40*, 1697.

(2) *KowWin Software*, U.S. Environmental Protection Agency, free distribution at: www.epa.gov/oppt/exposure/docs/episuite.htm

(3) Marrero, J.; Gani, R. Group-Contribution-Based Estimation of Octanol/Water Partition Coefficient and Aqueous Solubility. *Ind. Eng. Chem. Res.* **2002**, *41*, 6623.

(4) Panayiotou, C; Pantoula, M.; Stefanis, E.; Tsivintzelis, I.; Economou, I. Nonrandom Hydrogen-Bonding Model of Fluids and Their Mixtures. 1.Pure Fluids. *Ind. Eng. Chem. Res.* **2004** (available on http://pubs.acs.org/cgi-bin/asap.cgi/iecred/asap/pdf/ie040114+.pdf).

(5) Daubert, T. E.; Danner, R. P. Physical and Thermodynamic Properties of Pure Compounds: Data Compilation; Hemisphere: New York, 1989.

(6) Fredeslund, A., Sørensen, J. M. Group Contribution Estimation Method in *Models for Thermodynamic and Phase Equilibria Calculations*, Sandler, S. I. Ed., Marcel Dekker Inc., New York, 1994

(7) Marrero, J.; Gani, R. Group Contribution Based Estimation of Pure Component Properties. *Fluid Phase Equil.* **2001**, *183-184*, 183-208.

APPENDIX: Example of the application of the method to the estimation of temperature-dependent properties: Ethyl Benzoate
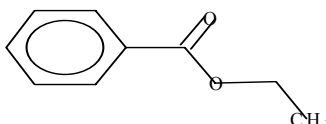
A) ε^* characteristic scaling constant.

1st Order approximation:

1st Order Groups	Occurences, n_i	Contributions, F_i	n_iF_i
-CH3	1	-140.9	-140.9
-CH2	1	122.2	122.2
ACH	5	124.3	621.5
AC	1	215.6	215.6
-COO	1	515.5	515.5
Σn_iF_i			1333.9
Universal Constant, C			4455.4

1st order approximation value:

$$\varepsilon^* = \sum_i n_i F_i + 4455.4 = 5789 \text{ J/mol}$$

1st order approximation error: (5789-5910)/5910 = -2 %.

2nd Order Approximation:

2nd Order Groups	Occurences, m_j	Contributions, S_j	m_jS_j
ACCOO	1	116.5	116.5
Σm_jS_j			116.5

$$\varepsilon^* = \sum_i n_i F_i + \sum_j m_j S_j + 4455.4 = 5906 \text{ J/mol}$$

Estimated characteristic scaling constant ε^*=5906 J / mol.
Experimental characteristic scaling constant ε^*=5910 J /mol.
2nd order approximation error = (5906-5910)/5910 = 0.07 %.

B) v^* characteristic scaling constant.

The same methodology was followed for the estimation of the other two scaling constants, v^*, and v_{sp}^*.

1st order approximation value:

$$v^* = \sum_i n_i F_i + 9.1748 = 11.354 \; cm^3 / mol$$

1st order approximation error: (11.354-12.049)/12.049 = -5.8 %.

2nd order approximation value:

$$v^* = \sum_i n_i F_i + \sum_j m_j S_j + 9.1748 = 12.056 \; cm^3 / mol$$

2nd order approximation error = (12.056-12.049)/12.049 = -0.06 %.

C) v_{sp}^* characteristic scaling constant.

1st order approximation value:

$$v_{sp}^* = \sum_i n_i F_i + 1.18081 = 0.9279 \; cm^3 / g$$

1st order approximation error: (0.9279-0.9380)/0.9380 = -1.1 %.

2nd order approximation value:

$$v_{sp}^* = \sum_i n_i F_i + \sum_j m_j S_j + 1.18081 = 0.9432 \; cm^3 / g$$

2nd order approximation error = (0.9432-0.9380)/0.9380 = 0.55 %.

In Figures 1, 2 and 3 the charts of vapor pressure, liquid density, and the heat of vaporization at the whole range of temperatures are presented. Triangular symbols represent the experimental values calculated by the DIPPR equations. The properties estimated by the experimental values of scaling constants ε^*, v^* and v_{sp}^* are depicted with circular symbols. The solid line represents the values of the properties estimated by the scaling constants predicted by the group-contribution method (Examples A, B, C).
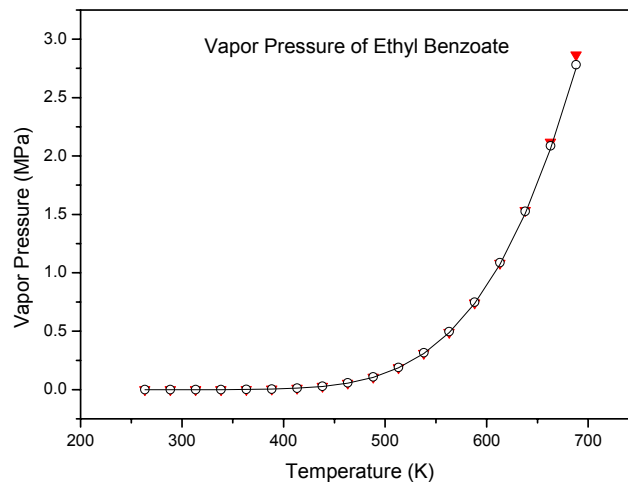

Fig. 1. Vapor pressure calculated by DIPPR equations (triangular symbols), by experimental scaling constants (circular symbols) and by estimated scaling constants (solid line)
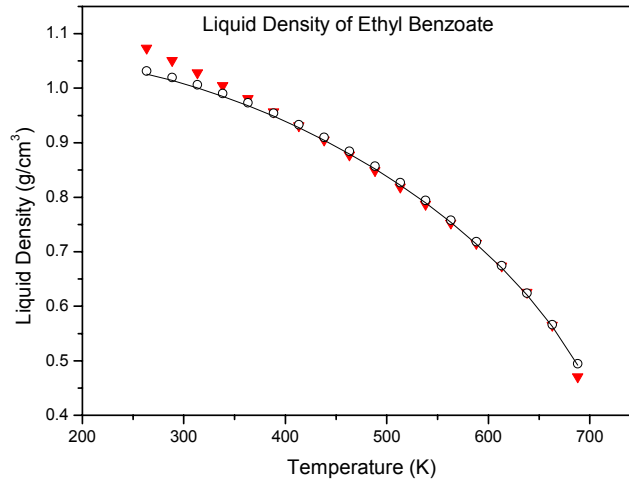
Fig. 2. Liquid density calculated by DIPPR equations (triangular symbols), by experimental scaling constants (circular symbols) and by estimated scaling constants (solid line)
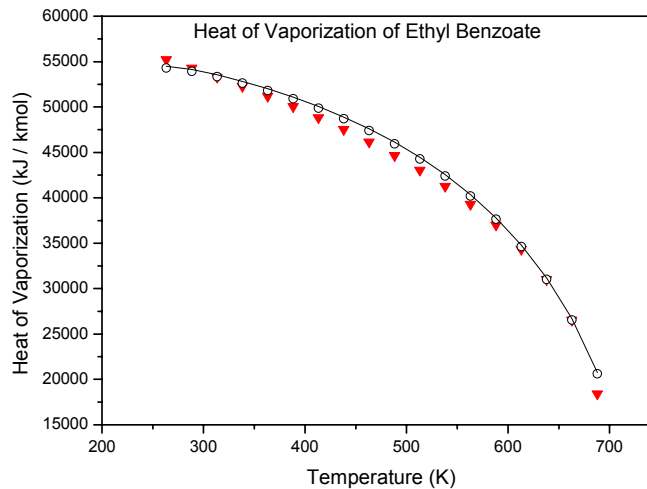

Fig. 3. Heat of vaporization calculated by DIPPR equations (triangular symbols), by experimental scaling constants (circular symbols) and by estimated scaling constants (solid line)

D) Influence parameter, κ.

For the estimation of surface tension at any given temperature the knowledge of the influence parameter of each compound is necessary. The methodology for the estimation of this parameter is same as that followed for the estimation of the three scaling constants.

1st order approximation value:

$$\kappa = \sum_i n_i F_i + 0.16452 = 0.15695$$

1st order approximation error: (0.15695-0.1652)/0.1652 = -5 %.

2nd order approximation value:

$$\kappa = \sum_i n_i \, F_i + \sum_j m_j \, S_j + 0.16452 = 0.1603$$

2nd order approximation error: = (0.1603-0.1652)/0.1652 = -2.97 %.

      In Figure 4 the chart of surface tension at any temperature is presented. Triangular symbols represent the surface tension values calculated by the DIPPR equations. The surface tension estimated by the experimental scaling constants ε^*, v^* and v_{sp}^* and of the experimental influence parameter, κ, is depicted with circular symbols. The solid line represents the values of surface tension estimated by the scaling constants and the influence parameter predicted by the group-contribution method (Examples A, B, C and D).
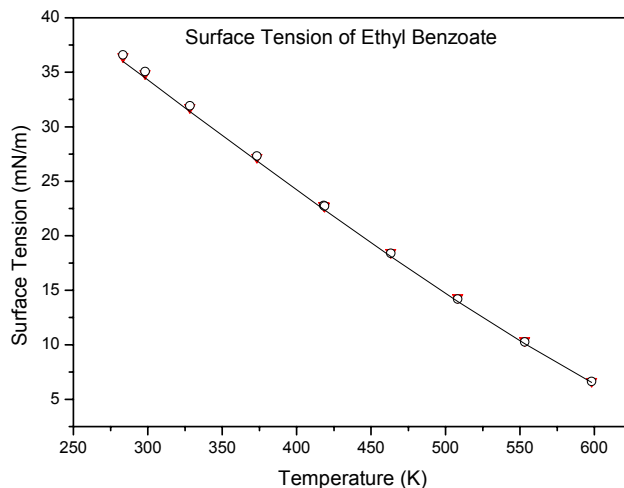

Fig. 9. Surface tension calculated by DIPPR equations (triangular symbols), by experimental scaling constants (circular symbols) and by estimated scaling constants (solid line)