JOURNAL OF PHYSICAL ORGANIC CHEMISTRY, VOL. 6, 634–636 (1993)

THEORETICAL SOLUTE PARAMETERS AND CORRELATIONS OF LOG P AND pK_a. A COMMENT ON A RECENT DEVELOPMENT AND STATISTICAL PROTOCOL

WILLIAM C. HERNDON

Department of Chemistry, University of Texas at El Paso, El Paso, Texas 79968, U.S.A.

Physical–chemical properties can be correlated and predicted using linear free energy or linear solvation energy relationships (LSER). Procedures to obtain theoretical parameters from calculated molecular structures that can be used in LSER have recently been evaluated [G R. Famini, C. A. Penski and L. Y. Wilson, *J. Phys. Org. Chem.* 5, 395–408 (1992)]. Among other applications, these procedures were applied to correlate octanol–water partition coefficients and pK_a values for two groups of structurally diverse solutes with very good reported results. In this paper, the statistical results of those studies are re-examined. The corrected statistical parameters do not provide tenability for the appropriateness of the methodology in these applications. However, valid multilinear relationships of the theoretical LSER parameters with the experimental properties are found which do substantiate the original conclusions of Famini *et al.*

Partition coefficients and acidities are key parameters for assessing biological activities, toxicity and persistence of chemicals in the environment. In this regard, the octanol–water partition coefficient (K_{ow} or $\log K_{ow} = \log P$) is the most widely used descriptor of partitioning behavior, and the Brønsted–Lowry pK_a is the common measure of acidity. Procedures to correlate and/or predict these properties based on molecular structure or on physical–chemical principles are, therefore, of substantial interest. Practical procedures for estimating such properties and reviews of the voluminous literature can be found in several recent publications.[1–3] Of particular relevance to this paper, Taft and co-workers[4] have demonstrated that solvatochromic parameters used in linear solvation energy relationships (LSER) provide useful rectifications of acidity data and of partitioning behavior in a variety of solvents for a large number of solute structural types.

Famini *et al.*[5] have recently outlined an important development in the LSER area, viz. the use of computationally derived independent variables rather than the experimentally based set of solvatochromic parameters. In principle, the theoretical LSER (TLSER) descriptors permit *a priori* prediction of solvatochromic properties based solely on the (calculated) molecular structure. There are numerous examples of the use of theoretical (quantum mechanical) molecular descriptors in structure–activity and structure–property studies.[6] The theoretical descriptors reported by Famini *et al.* are obtained using an additive procedure to estimate molecular volumes[7] (V_{mc}) and MNDO semi-empirical calculations[8] to obtain electronic terms. The electronic parameters are comprised of π_1 (polarizability term), hydrogen bond acidity terms ε_a and q_+ and hydrogen bond basicity terms ε_b and q_-. The original paper[5] should be consulted for the definitions of each term and explanations. The TLSER parameters are used to correlate several properties, among which are 67 log P values for diverse structures and 42 pK_a values for a variety of weak acids. The reported linear regression results for log P and pK_a are summarized in Table 1, entries 1–3.

During an attempt to fit non-linear models to the log P and pK_a data, the opportunity arose to examine plots of both sets of Famini *et al.*'s experimental and calculated values. The scatter of the points in these graphs was substantial, and it seemed unlikely that these plots could have correlation coefficients (R) as high as 0·971 and 0·942. Indeed, after reanalysis, the actual R values for equations (1) and (2) (Table 1) turn out to be 0·929 and 0·794, respectively. These values are confirmed by simple first-order linear regressions of Famini *et al.*'s calculated log P and pK_a values with the experimental values. In addition, it should be noted that the regression coefficient for the independent variable q_+ in equation (2) is presumably the result of a typographical error in Famini *et al.*'s paper[5] and should be replaced by the value −48·1.

On the basis of the actual value of R, the quality of

0894–3230/93/110634–03$06.50

Received 22 March 1993
Accepted 29 June 1993

Table 1. TLSER correlations of log P and pK_a data from ref. 5

Equation	Property	Regression coefficients (parameters) N	R	SD[a]
1	$\log P = 3{\cdot}14V_{mc} - 5{\cdot}92q_-$	67	0·974[b]	0·45
2	$pK_a = 127\varepsilon_a - 41{\cdot}1q_+$	42	0·941[c]	2·82
3[d]	$pK_a = 29{\cdot}4 - 155\varepsilon_b - 7{\cdot}33q_- + 55{\cdot}4\varepsilon_a - 14{\cdot}2q_+$	39[d]	0·950	1·06
4	$\log P = 3{\cdot}20V_{mc} - 5{\cdot}72q_- - 1{\cdot}60q_+$	67	0·933	0·44
5	$\log P = 0{\cdot}871 + 3{\cdot}25V_{mc} - 5{\cdot}83q_- - 2{\cdot}23q_+ - 7{\cdot}44\pi_1$	67	0·941	0·42
6	$pK_a = 17{\cdot}6 - 1{\cdot}20V_{mc} - 173{\cdot}7\varepsilon_b + 209{\cdot}9\varepsilon_a - 40{\cdot}6q_+$	42	0·923	1·85

Equation (4) (coefficient t-statistics, 29, 15, 1·9; f-ratio, 214·7)
Equation (5) (coefficient t-statistics, 2·6, 18, 15, 2·7, 2·9; f-ratio, 120·4)
Equation (6) (coefficient t-statistics, 2·9, 2·2, 5·3, 12, 8·4; f-ratio, 53·5)

[a] Standard deviation.
[b] Corrected $R = 0{\cdot}929$.
[c] Corrected $R = 0{\cdot}794$.
[d] Stated to be for 39 compounds with p$K_a < 13$. The data in ref. 5 are listed for 36 compounds of this type. Equation (3) was not re-examined because of this discrepancy.

the regression model for equation (1) can be characterized as passable, but the R for equation (2) is low enough to be considered worthless for use in support of a physical–chemical argument. Hence the results do not sustain an optimistic view[5] of the appropriateness of TLSER descriptors in applications of this type. However, standardized stepwise regression analyses of the log P and pK_a data showed that there do exist several statistically acceptable, multilinear relationships with the TLSER parameters which were not identified by Famini *et al.* Examples are given in equations (4)–(6) in Table 1.

Equations (4) and (5) are fair to good rectifications of the log P data, moderately better than equation (1) with its corrected $R = 0{\cdot}929$. The larger number of degrees of freedom in the statistical analysis accounts for the increased goodness of fit. The t-statistics for individual regression coefficients and the f-ratios satisfy criteria for valid multilinear regression equations (the values of the statistical options controlling inclusion of independent variables are those recommended on the basis of Monte Carlo studies[9]), and there is no statistical reason to exclude a constant term from the optimum equation for correlation of the experimental data. A constant term is also included in the best correlation of the pK_a data [equation (6)]. The relevant statistical results, summarized in Table 1, denote a fair correlation of all 42 pK_a values, much better than the fit of the data according to equation (2), with its corrected $R = 0{\cdot}794$.

The reasons for the discrepancies in the calculated R values may be due to statistical protocols, discussed in detail in the operating manual of one of the commercial statistical computer programs chosen by Famini *et al.* to analyze their data.[10] In that program, correct statistical significance-of-fit parameters (R, R^2 and f-ratio) for linear regressions without a constant term can only be derived by using an optional procedure for analysis, referred to in the manual as the mixture model. Otherwise, inappropriately high values are calculated unless the data are centred around the origin of the data coordinate system. In general, it is not necessary to consider adjustments of this type to obtain correct regression statistics in other available statistical packages.

In the final analysis, the corrections and the results of the re-examination of the log P and pK_a data actually lend support to the original contentions of Famini *et al.*[5] It does appear to be possible to use the computationally based TLSER parameters in the same way as the experimentally based LSER descriptors, and reasonable results have now been obtained for all properties examined so far. However, Famini *et al.*[5] pointed out that the LSER solvatochromic parameters generally give better correlations where common properties have been studied. This could indicate a need for more accurate or, possibly, different theoretical descriptors. In any event, additional investigations of this promising approach to the LSER type of analysis should be encouraged.

ACKNOWLEDGEMENTS

The financial support of the Welch Foundation and the Materials Research Center of Excellence at the University of Texas at El Paso (sponsored by the National Science Foundation) is gratefully acknowledged. The author also thanks Jose L. Canales (El Paso High School) for assistance with the data analysis.

REFERENCES

1. W. J. Lyman, W. F. Reehl and D. H. Rosenblatt, *Handbook of Chemical Property Estimation Methods*, McGraw-Hill, New York (1982).
2. D. Mackay, W. Y. Shiu and K. C. Ma, *Illustrated Handbook of Physical–Chemical Properties and Environmental Fate for Organic Chemicals, Vol. 1, Monoaromatic Hydrocarbons, Chlorobenzenes, and PCBs*. Lewis, Boca Raton, FL (1991); *Vol. 2, Polynuclear Aromatic Hydrocarbons, Polychlorinated Dioxins, and Dibenzofurans*, Lewis, Boca Raton, FL (1992).
3. P. J. Taylor, in *Comprehensive Medicinal Chemistry*, edited by C. A. Ramsden, Vol. 4, Chapt. 18.6. Pergamon Press, New York (1990).
4. For leading references, see: for log *P*, M. J. Kamlet, R. M. Doherty, M. H. Abraham, Y. Marcus and R. W. Taft, *J. Phys. Chem.* **92**, 5244–5255 (1988); for pK_a, R. W. Taft, I. A. Koppel, R. D. Topsom and F. Aniva, *J. Am. Chem. Soc.* **112**, 2047–2052 (1990).
5. G. R. Famini, C. A. Penski and L. Y. Wilson, *J. Phys. Org. Chem.* **5**, 395–408 (1992).
6. For a review, see G. H. Lowe and S. K. Burt, in *Comprehensive Medicinal Chemistry*, edited by C. A., Ramsden, Vol. 4, Chapt. 18.1. Pergamon Press, New York (1990).
7. A. J. Hopfinger, *J. Am. Chem. Soc.* **102**, 7196–7206 (1980).
8. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.* **99**, 4899–4907, 4907–4917 (1977).
9. R. B. Bendel and A. A. Afifi, *J. Am. Stat. Assoc.* **72**, 46–53 (1977).
10. L. Wilson, *SYSTAT: the System for Statistics*, pp. 176–180. SYSTAT, Evanston, IL (1990).