# Modeling the Octanol−Water Partition Coefficients by an Optimized Molecular Connectivity Index†

Milan Šoškić*

Department of Chemistry, Faculty of Agronomy, University of Zagreb, HR-10000 Zagreb, The Republic of Croatia

Dejan Plavšić

The Ruđer Bošković Institute, P.O. Box 180, HR-10002 Zagreb, The Republic of Croatia

Received January 25, 2005

A procedure that makes it possible to generate a coherent model for prediction of the octanol−water partition coefficient within the molecular connectivity formalism was put forward. The method is based on the optimization of weights for corresponding skeletal atoms and is similar to the method for calculation of a variable connectivity index proposed by Randić. In contrast to Randić's method, we incorporate in the algorithm the possibility that the contribution of a term describing a carbon−heteroatom bond may be negative. When tested on a set of about 300 structurally diverse organic molecules, our procedure proved to be superior to the standard valence connectivity method. External validation on a smaller set of compounds confirmed the superiority of our procedure with respect to the standard one. Intramolecular interactions, which are operative in more complex compounds, are treated in a similar fashion to that in the Hansch−Leo or Rekker methods, by inclusion of empirical correction factors.

## INTRODUCTION

Lipophilicity plays an important role in processes such as the transport of bioactive molecules through biological membranes or their interactions with the corresponding receptors.[1] In addition, this molecular property also affects the environmental fate of organic molecules.[2] Therefore, it is not surprising that lipophilicity frequently appears as a key parameter in QSAR (quantitative structure−activity relationship) models.[3] It is most usually described by the logarithm of the partition coefficient, log *P*, of a compound determined in the *n*-octanol/water system. Experimental determination of log *P* is often complex and time-consuming and can be done only for already synthesized compounds. For this reason, a number of computational methods for the prediction of this parameter have been proposed.[4]

Molecular connectivity indices[5,6] are very popular graph-theoretical[7] descriptors that have proven useful in numerous QSAR[7−13,15] and QSPR[6−8,14,15] (quantitative structure−property relationship) studies. Derived directly from the structural formula, these indices encode important structural features such as size, branching, and cyclicity, or the presence of heteroatoms in an organic molecule. Molecular connectivity indices have been successfully used for describing lipophilicity in sets of closely related compounds[15] but are less suitable for groups of heterogeneous compounds.[16,17] To circumvent this obvious shortcoming, here we propose a method for improving the efficiency of molecular connectivity indices in estimating log *P*. The method is essentially based on the optimization of the weights for corresponding skeletal atoms and is similar to the method for calculation of a variable connectivity index recently proposed by Randić.[18,19]

## METHODS OF CALCULATION AND EXPERIMENTAL DATA

Molecular connectivity indices are calculated from hydrogen-depleted molecular graphs. The valence first-order molecular connectivity index is calculated from the following expression:

$$^1\chi^{\mathrm{v}} = \sum(\delta_i^{\mathrm{v}}\delta_j^{\mathrm{v}})^{-0.5} \qquad (1)$$

where $\delta^{\mathrm{v}}$ denotes valence delta values for the skeletal atoms making up a bond and the summation goes over all of the bonds.

The valence delta values are defined by the equation

$$\delta^{\mathrm{v}} = (Z^{\mathrm{v}} - h)/(Z - Z^{\mathrm{v}} - 1) \qquad (2)$$

where $Z^{\mathrm{v}}$ is the number of valence electrons in an atom, $Z$ is its atomic number, and $h$ is the number of hydrogen atoms attached to those atoms.

The optimized first-order molecular connectivity index is calculated in a similar way:

$$^1\chi^{\mathrm{opt}} = \sum(\delta_i^{\mathrm{opt}}\delta_j^{\mathrm{opt}})^{-0.5} \qquad (3)$$

where $\delta^{\mathrm{opt}}$ denotes the optimized weights for the adjacent skeletal atoms and the sums are over all of the bonds. The optimized weights are empirical values, which can be

* Corresponding author tel.: +385-1-2393-872; e-mail: msoskic@agr.hr; fax: +385-1-2315-300.

† Presented as a poster at the 15th European Symposium on Quantitative Structure−Activity Relationships & Molecular Modeling, September 5−10, 2004, Istanbul, Turkey.

10.1021/ci050024v CCC: $30.25 
Published on Web 06/03/2005

obtained by iterative regression analysis similar to the procedure proposed by Randić.[18,19] In contrast to Randić's optimization method, we built into the algorithm an option that the contributions of the terms $(\delta_i^{opt}\delta_j^{opt})^{-0.5}$ describing carbon−heteroatom bonds may be negative. The optimization procedure will be outlined using the example of three primary alcohols. The expressions for the calculation of the optimized molecular connectivity indices for ethanol (eq 4), *n*-propanol (eq 5), and *n*-buthanol (eq 6) are as follows:

$$^1\chi^{opt} = (1 \times 2)^{-0.5} - (2 \times X)^{-0.5} \quad (4)$$

$$^1\chi^{opt} = (1 \times 2)^{-0.5} + (2 \times 2)^{-0.5} - (2 \times X)^{-0.5} \quad (5)$$

$$^1\chi^{opt} = (1 \times 2)^{-0.5} + 2(2 \times 2)^{-0.5} - (2 \times X)^{-0.5} \quad (6)$$

The numbers in parentheses are valence delta values for the corresponding primary and secondary carbon atoms of the alcohols, and $X$ is the optimal weight of the oxygen for the OH group, which is to be determined. One starts by replacing $X$ with a definite numerical value, and the indices thus obtained are then regressed against the experimental log $P$ values for the alcohols. The process is continued until a minimal standard deviation for the regression is found.

The experimental log $P$ values used in this study are taken from refs 20 and 21.

The calculation of the molecular connectivity indices and multiple regression analysis were carried out by our own programs. The code of the program for calculation of the optimized first-order molecular connectivity index is available upon request.

In all regression equations, $n$ is the number of compounds used in the analysis, $r^2$ is the squared correlation coefficient, $s$ is the standard deviation of the estimates, and $F$ is the ratio of the regression and the residual variances. The degrees of freedom ($k$, $n-k-1$) associated with $F$ are specified in the superscript, where $k$ is the number of independent variables in the equation. PRESS is the sum of squared prediction errors from the leave-one-out cross-validation analysis, and $r^2_{cv}$ is the cross-validated coefficient of determination.

The 95% confidence intervals for the regression coefficients are given in parentheses. The significance of all of the derived models is above the 95% level or higher.

The predictive capacity of the models was additionally tested on an external validation data set, which was constructed in the following way: As a rule, depending on the availability of data, several representatives were selected from the literature[21] for the major classes of compounds considered in the paper, and then at least one compound from each of the preselected groups was randomly chosen for the validation set.

## RESULTS AND DISCUSSION

The primary goal of this study is to generate a model, based on the molecular connectivity formalism, which accurately predicts the log $P$ values for a set of structurally diverse compounds. The compounds used in this study together with their molecular connectivity indices and experimental and calculated log $P$ values are shown in Table 1. As expected, the statistics of a common model (eq 7) for the selected classes of simple organic compounds (compounds 1−232) is far from satisfactory when the valence first-order molecular connectivity index is used as a predictor variable (Figure 1):

$$\log P = -0.638\ (\pm 0.266) + 1.068\ (\pm 0.077) \times {}^1\chi^v \quad (7)$$

$$n = 232,\ r^2 = 0.762,\ s = 0.847,\ F^{1,230} = 736.3,\ r^2_{cv} = 0.758,\ \text{PRESS} = 168.1$$

From this and previous models,[16,17] it is obvious that the existing molecular connectivity indices are not appropriate for modeling log $P$ in the case of a mixed set of molecules. Clearly, certain modifications in the calculation procedure are necessary to make these indices more efficient in describing this physicochemical property. The octanol−water partition coefficient can generally be factorized into molecular bulk (size) and polarity terms. Because the molecular connectivity index has, in numerous studies, proven to be a good descriptor of molecular bulk, it appears that it is polarity that is not adequately parameterized by this molecular descriptor. This shortcoming can be overcome by replacing the standard valence delta values for describing the skeletal atoms in a molecule with an empirical weighting scheme, specifically derived for modeling lipophilicity. The optimization method for calculation of the variable connectivity index, recently proposed by Randić,[18,19] is, in principle, a suitable tool for this purpose.

However, because the presence of heteroatoms in molecules often lowers their partition coefficients, as compared to the corresponding hydrocarbon analogues, it appears reasonable to modify the algorithm by incorporating the option that the contributions of polar bonds (the bonds containing a heteroatom) may be negative. A similar line of reasoning was followed by Dearden et al.[17] in an attempt to improve the descriptor performances of the original valence first-order molecular connectivity index in predicting log $P$. Recently, Estrada introduced a generalized graph-theoretical matrix,[22−24] from which, in principle, a huge number of optimized topological indices could be deduced, including the proposed index.

To evaluate whether the existing weightings (delta values) for carbon atoms should also be optimized, we correlated the octanol−water partition coefficients for hydrocarbons against their valence first-order molecular connectivity indices. For apolar alkanes and alkenes (compounds 1−29), a model of satisfactory statistical performance was obtained (eq 8):

$$\log P = 0.873\ (\pm 0.108) + 1.043\ (\pm 0.032) \times {}^1\chi^v \quad (8)$$

$$n = 29,\ r^2 = 0.994,\ s = 0.144,\ F^{1,27} = 4379.7,\ r^2_{cv} = 0.993,\ \text{PRESS} = 0.6592$$

We failed, however, to obtain a model of similar statistical quality when we tried to incorporate the only slightly more polar alkynes (compounds 30−37) in the model (eq 9):

$$\log P = 0.603\ (\pm 0.189) + 1.092\ (\pm 0.060) \times {}^1\chi^v \quad (9)$$

$$n = 37,\ r^2 = 0.975,\ s = 0.296,\ F^{1,35} = 1358.8,\ r^2_{cv} = 0.972,\ \text{PRESS} = 3.4397$$

On the other hand, when we used, for alkynes, the numerical values of the index $^1\chi^{opt}$ obtained by the modified calculation

**Table 1.** Molecular Connectivity Indices and Observed and Calculated log *P* Values for the 306 Compounds Used in This Study

| | | | | log *P* | | | | | | log *P* | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| number | compounds | $^1\chi^{opt}$ | $^1\chi^v$ | exptl. | eq 18 | number | compounds | $^1\chi^{opt}$ | $^1\chi^v$ | exptl. | eq 18 |
| 1 | methane | 0.000 | 0.000 | 1.09 | 0.83 | 68 | 2,3-dimethylbutan-2-ol | 0.554 | 2.667 | 1.48 | 1.42 |
| 2 | ethane | 1.000 | 1.000 | 1.81 | 1.89 | 69 | diethyl ether | 0.123 | 1.992 | 0.89 | 0.96 |
| 3 | propane | 1.414 | 1.414 | 2.36 | 2.32 | 70 | di-*n*-propyl ether | 1.123 | 2.992 | 2.03 | 2.02 |
| 4 | *n*-butane | 1.914 | 1.914 | 2.89 | 2.85 | 71 | di-*n*-butyl ether | 2.123 | 3.992 | 3.21 | 3.07 |
| 5 | *n*-pentane | 2.414 | 2.414 | 3.39 | 3.38 | 72 | ethyl-*n*-butyl ether | 1.123 | 2.992 | 2.03 | 2.02 |
| 6 | *n*-hexane | 2.914 | 2.914 | 3.90 | 3.91 | 73 | propanone | −1.041 | 1.204 | −0.24 | −0.26 |
| 7 | *n*-heptane | 3.414 | 3.414 | 4.50 | 4.43 | 74 | butanone | −0.481 | 1.765 | 0.29 | 0.33 |
| 8 | *n*-octane | 3.914 | 3.914 | 5.15 | 4.96 | 75 | pentan-2-one | 0.019 | 2.265 | 0.91 | 0.86 |
| 9 | *n*-nonane | 4.414 | 4.414 | 5.65 | 5.49 | 76 | pentan-3-one | 0.080 | 2.325 | 0.82 | 0.92 |
| 10 | *n*-undecane | 5.414 | 5.414 | 6.54 | 6.54 | 77 | hexan-2-one | 0.519 | 2.765 | 1.38 | 1.38 |
| 11 | *n*-dodecane | 5.914 | 5.914 | 6.80 | 7.07 | 78 | heptan-2-one | 1.019 | 3.265 | 1.98 | 1.91 |
| 12 | *n*-tridecane | 6.414 | 6.414 | 7.56 | 7.60 | 79 | heptan-4-one | 1.080 | 3.325 | 2.04 | 1.97 |
| 13 | *n*-tetradecane | 6.914 | 6.914 | 8.00 | 8.12 | 80 | octan-2-one | 1.519 | 3.765 | 2.37 | 2.43 |
| 14 | 2-methylpropane | 1.732 | 1.732 | 2.76 | 2.66 | 81 | nonan-2-one | 2.019 | 4.265 | 3.14 | 2.96 |
| 15 | 3-methylpentane | 2.808 | 2.808 | 3.60 | 3.79 | 82 | nonan-5-one | 2.080 | 4.325 | 2.88 | 3.03 |
| 16 | 2,2-dimethylpropane | 2.000 | 2.000 | 3.11 | 2.94 | 83 | decan-2-one | 2.519 | 4.765 | 3.73 | 3.49 |
| 17 | 2,2-dimethylbutane | 2.561 | 2.561 | 3.82 | 3.54 | 84 | undecan-2-one | 3.019 | 5.265 | 4.09 | 4.02 |
| 18 | 2,3-dimethylbutane | 2.643 | 2.643 | 3.85 | 3.62 | 85 | 3-methylbutan-2-one | −0.098 | 2.148 | 0.84 | 0.73 |
| 19 | 3,3-dimethylheptane | 4.121 | 4.121 | 5.19 | 5.18 | 86 | 4-methylpentan-2-one | 0.375 | 2.621 | 1.31 | 1.23 |
| 20 | ethene | 0.500 | 0.500 | 1.13 | 1.36 | 87 | 5-methylhexan-2-one | 0.875 | 3.121 | 1.88 | 1.76 |
| 21 | propene | 0.986 | 0.986 | 1.77 | 1.87 | 88 | 5-methyloctan-2-one | 1.913 | 4.159 | 2.92 | 2.85 |
| 22 | but-1-ene | 1.524 | 1.524 | 2.40 | 2.44 | 89 | 3,3-dimethylbutan-2-one | 0.209 | 2.454 | 1.20 | 1.05 |
| 23 | pent-1-ene | 2.024 | 2.024 | 2.80 | 2.97 | 90 | acetic acid | −0.947 | 0.928 | −0.17 | −0.17 |
| 24 | hex-1-ene | 2.524 | 2.524 | 3.39 | 3.49 | 91 | propanoic acid | −0.387 | 1.488 | 0.33 | 0.43 |
| 25 | hept-1-ene | 3.024 | 3.024 | 3.99 | 4.02 | 92 | butanoic acid | 0.113 | 1.988 | 0.79 | 0.96 |
| 26 | oct-1-ene | 3.524 | 3.524 | 4.57 | 4.55 | 93 | pentanoic acid | 0.613 | 2.488 | 1.39 | 1.48 |
| 27 | non-1-ene | 4.024 | 4.024 | 5.15 | 5.08 | 94 | hexanoic acid | 1.113 | 2.988 | 1.92 | 2.01 |
| 28 | isobutene | 1.354 | 1.354 | 2.34 | 2.26 | 95 | heptanoic acid | 1.613 | 3.488 | 2.41 | 2.53 |
| 29 | 2-methylbut-2-ene | 1.866 | 1.866 | 2.67 | 2.80 | 96 | octanoic acid | 2.113 | 3.988 | 3.05 | 3.06 |
| 30 | ethyne | −0.333 | 0.333 | 0.37 | 0.48 | 97 | decanoic acid | 3.113 | 4.988 | 4.09 | 4.12 |
| 31 | propyne | 0.211 | 0.789 | 0.94 | 1.06 | 98 | tetradecanoic acid | 5.113 | 6.988 | 6.10 | 6.22 |
| 32 | but-1-yne | 0.772 | 1.349 | 1.46 | 1.65 | 99 | hexadecanoic acid | 6.113 | 7.988 | 7.17 | 7.28 |
| 33 | but-2-yne | 0.750 | 1.250 | 1.46 | 1.62 | 100 | octadecanoic cid | 7.113 | 8.988 | 8.23 | 8.33 |
| 34 | hex-1-yne | 1.772 | 2.349 | 2.73 | 2.71 | 101 | eicosanoic acid | 8.113 | 9.988 | 9.29 | 9.39 |
| 35 | hept-1-yne | 2.272 | 2.849 | 3.32 | 3.23 | 102 | 3-methylbutanoic acid | 0.469 | 2.344 | 1.16 | 1.33 |
| 36 | oct-1-yne | 2.772 | 3.349 | 3.92 | 3.76 | 103 | 2-ethylbutanoic acid | 1.072 | 2.947 | 1.68 | 1.96 |
| 37 | non-1-yne | 3.272 | 3.849 | 4.51 | 4.28 | 104 | 2-methylpentanoic acid | 1.034 | 2.909 | 1.80 | 1.92 |
| 38 | ethanol | −0.874 | 1.023 | −0.30 | −0.09 | 105 | 2-propylpentanoic acid | 2.072 | 3.947 | 2.75 | 3.02 |
| 39 | propan-1-ol | −0.374 | 1.523 | 0.25 | 0.44 | 106 | 2-ethylhexanoic acid | 2.072 | 3.947 | 2.64 | 3.02 |
| 40 | butan-1-ol | 0.126 | 2.023 | 0.88 | 0.96 | 107 | ethyl acetate | −0.136 | 1.904 | 0.73 | 0.69 |
| 41 | pentan-1-ol | 0.626 | 2.523 | 1.56 | 1.49 | 108 | *n*-propyl acetate | 0.364 | 2.404 | 1.24 | 1.22 |
| 42 | hexan-1-ol | 1.126 | 3.023 | 2.03 | 2.02 | 109 | *n*-butyl acetate | 0.864 | 2.904 | 1.78 | 1.74 |
| 43 | heptan-1-ol | 1.626 | 3.523 | 2.72 | 2.55 | 110 | isobutyl acetate | 0.720 | 2.760 | 1.78 | 1.59 |
| 44 | octan-1-ol | 2.126 | 4.023 | 3.07 | 3.07 | 111 | *n*-pentyl acetate | 1.364 | 3.404 | 2.30 | 2.27 |
| 45 | nonan-1-ol | 2.626 | 4.523 | 3.67 | 3.60 | 112 | isopentyl acetate | 1.220 | 3.260 | 2.17 | 2.12 |
| 46 | undecan-1-ol | 3.626 | 5.523 | 4.72 | 4.66 | 113 | *n*-hexyl acetate | 1.864 | 3.904 | 2.83 | 2.80 |
| 47 | dodecan-1-ol | 4.126 | 6.023 | 5.13 | 5.18 | 114 | ethyl propanoate | 0.425 | 2.465 | 1.21 | 1.28 |
| 48 | tridecan-1-ol | 4.626 | 6.523 | 5.82 | 5.71 | 115 | *n*-propyl propanoate | 0.925 | 2.965 | 1.71 | 1.80 |
| 49 | tetradecan-1-ol | 5.126 | 7.023 | 6.36 | 6.24 | 116 | *n*-pentyl propanoate | 1.925 | 3.965 | 2.67 | 2.86 |
| 50 | 2-methylpropan-1-ol | −0.018 | 1.879 | 0.76 | 0.81 | 117 | ethyl butanoate | 0.925 | 2.965 | 1.71 | 1.81 |
| 51 | 2-methylbutan-1-ol | 0.520 | 2.417 | 1.16 | 1.38 | 118 | *n*-propyl butanoate | 1.425 | 3.465 | 2.15 | 2.34 |
| 52 | 3-methylbutan-1-ol | 0.482 | 2.379 | 1.16 | 1.34 | 119 | ethyl pentanoate | 1.425 | 3.465 | 2.30 | 2.34 |
| 53 | propan-2-ol | −0.671 | 1.413 | 0.05 | 0.12 | 120 | *n*-butyl pentanoate | 2.425 | 4.465 | 3.36 | 3.39 |
| 54 | butan-2-ol | −0.133 | 1.951 | 0.61 | 0.69 | 121 | ethyl isobutanoate | 0.808 | 2.847 | 1.55 | 1.68 |
| 55 | pentan-2-ol | 0.367 | 2.451 | 1.19 | 1.22 | 122 | acetonitrile | −1.167 | 0.724 | −0.34 | −0.40 |
| 56 | pentan-3-ol | 0.405 | 2.489 | 1.21 | 1.26 | 123 | propionitrile | −0.606 | 1.284 | 0.16 | 0.19 |
| 57 | hexan-2-ol | 0.867 | 2.951 | 1.76 | 1.75 | 124 | 1-cyanopropane | −0.106 | 1.784 | 0.53 | 0.72 |
| 58 | hexan-3-ol | 0.905 | 2.989 | 1.65 | 1.79 | 125 | 1-cyanobutane | 0.394 | 2.284 | 1.12 | 1.25 |
| 59 | heptan-2-ol | 1.367 | 3.451 | 2.31 | 2.27 | 126 | 1-cyanopentane | 0.894 | 2.784 | 1.66 | 1.78 |
| 60 | heptan-3-ol | 1.405 | 3.489 | 2.24 | 2.31 | 127 | 2-cyanopropane | −0.223 | 1.667 | 0.46 | 0.60 |
| 61 | heptan-4-ol | 1.405 | 3.489 | 2.44 | 2.31 | 128 | 2-cyanobutane | 0.315 | 2.205 | 1.10 | 1.16 |
| 62 | octan-2-ol | 1.867 | 3.951 | 2.90 | 2.80 | 129 | 1-cyano-2-methylpropane | 0.250 | 2.140 | 1.10 | 1.10 |
| 63 | octan-4-ol | 1.905 | 3.989 | 2.68 | 2.84 | 130 | 2-cyano-2-methylpropane | 0.083 | 1.974 | 1.08 | 0.92 |
| 64 | 2-methylpropan-2-ol | −0.390 | 1.724 | 0.35 | 0.42 | 131 | ethylamin | −0.874 | 1.115 | −0.13 | −0.09 |
| 65 | 2-methylbutan-2-ol | 0.171 | 2.284 | 0.89 | 1.10 | 132 | *n*-propylamine | −0.374 | 1.615 | 0.48 | 0.44 |
| 66 | 3-methylbutan-2-ol | 0.240 | 2.324 | 1.28 | 1.08 | 133 | *n*-butylamine | 0.126 | 2.115 | 0.97 | 0.96 |
| 67 | 2-methylpentan-2-ol | 0.671 | 2.784 | 1.53 | 1.54 | 134 | *n*-pentylamine | 0.626 | 2.615 | 1.49 | 1.49 |

**Table 1** (Continued)

| number | compounds | $^1\chi^{opt}$ | $^1\chi^v$ | log *P* exptl. | log *P* eq 18 | number | compounds | $^1\chi^{opt}$ | $^1\chi^v$ | log *P* exptl. | log *P* eq 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 135 | *n*-hexylamine | 1.126 | 3.115 | 2.06 | 2.02 | 202 | *n*-decylbenzene | 6.363 | 6.971 | 7.40 | 7.54 |
| 136 | *n*-heptylamine | 1.626 | 3.615 | 2.57 | 2.55 | 203 | phenol | 0.615 | 2.134 | 1.46 | 1.48 |
| 137 | *n*-octylamine | 2.126 | 4.115 | 3.09 | 3.07 | 204 | *o*-cresol | 1.005 | 2.551 | 1.98 | 1.89 |
| 138 | *n*-nonylamine | 2.626 | 4.615 | 3.60 | 3.60 | 205 | *m*-cresol | 1.003 | 2.545 | 1.98 | 1.89 |
| 139 | isopropylamine | −0.586 | 1.488 | 0.26 | 0.21 | 206 | *p*-cresol | 1.003 | 2.545 | 1.97 | 1.89 |
| 140 | isobutylamine | −0.018 | 1.971 | 0.73 | 0.81 | 207 | 2,4-dimethylphenol | 1.394 | 2.962 | 2.30 | 2.30 |
| 141 | *sec*-butylamine | −0.048 | 2.026 | 0.74 | 0.78 | 208 | 2,5-dimethylphenol | 1.394 | 2.962 | 2.33 | 2.30 |
| 142 | *tert*-butylamine | −0.390 | 1.789 | 0.40 | 0.42 | 209 | 2,6-dimethylphenol | 1.396 | 2.968 | 2.36 | 2.30 |
| 143 | diethylamine | −0.276 | 2.121 | 0.58 | 0.54 | 210 | 3,4-dimethylphenol | 1.394 | 2.962 | 2.23 | 2.30 |
| 144 | di-*n*-propylamine | 0.724 | 3.121 | 1.53 | 1.60 | 211 | 3,5-dimethylphenol | 1.391 | 2.956 | 2.35 | 2.30 |
| 145 | di-*n*-butylamine | 1.724 | 4.121 | 2.83 | 2.65 | 212 | 2-ethylphenol | 1.585 | 3.112 | 2.47 | 2.50 |
| 146 | diisopropylamine | 0.329 | 2.887 | 1.16 | 1.18 | 213 | 3-ethylphenol | 1.583 | 3.106 | 2.40 | 2.50 |
| 147 | triethylamine | 0.621 | 3.070 | 1.45 | 1.27 | 214 | 4-ethylphenol | 1.583 | 3.106 | 2.58 | 2.50 |
| 148 | acetamide | −1.989 | 0.993 | −1.26 | −1.26 | 215 | 2,4,6-trimethylphenol | 1.784 | 3.378 | 2.97 | 2.71 |
| 149 | butanamide | −0.928 | 2.054 | −0.21 | −0.15 | 216 | 2,3,6-trimethylphenol | 1.786 | 3.384 | 2.67 | 2.72 |
| 150 | ethylthiol | 0.391 | 1.571 | 1.18 | 1.24 | 217 | chlorobenzene | 1.935 | 2.477 | 2.89 | 2.87 |
| 151 | *n*-propylthiol | 0.891 | 2.071 | 1.81 | 1.77 | 218 | 1,2-dichlorobenzene | 2.477 | 2.960 | 3.43 | 3.44 |
| 152 | *n*-butylthiol | 1.391 | 2.571 | 2.28 | 2.30 | 219 | 1,3-dichlorobenzene | 2.474 | 2.954 | 3.53 | 3.44 |
| 153 | 1-fluorobutane | 1.000 | 1.974 | 2.00 | 1.89 | 220 | 1,4-dichlorobenzene | 2.474 | 2.954 | 3.44 | 3.44 |
| 154 | 1-fluoropentane | 1.500 | 2.474 | 2.33 | 2.41 | 221 | 1,2,3,-trichlorobenzene | 3.018 | 3.442 | 4.05 | 4.02 |
| 155 | chloroethane | 0.636 | 1.508 | 1.43 | 1.49 | 222 | 1,2,4-trichlorobenzene | 3.016 | 3.436 | 4.02 | 4.01 |
| 156 | 1-chloropropane | 1.136 | 2.008 | 2.04 | 2.03 | 223 | 1,3,5-trichlorobenzene | 3.014 | 3.430 | 4.19 | 4.01 |
| 157 | 1-chlorobutane | 1.636 | 2.508 | 2.64 | 2.56 | 224 | 1,2,3,4-tetrachlorobenzene | 3.560 | 3.925 | 4.64 | 4.59 |
| 158 | 1-chlropentane | 2.136 | 3.008 | 3.11 | 3.08 | 225 | 1,2,3,5-tetrachlorobenzene | 3.558 | 3.919 | 4.65 | 4.58 |
| 159 | 1-chlorohexane | 2.636 | 3.508 | 3.66 | 3.61 | 226 | 1,2,4,5-tetrachlorobenzene | 3.558 | 3.919 | 4.60 | 4.58 |
| 160 | 1-chloroheptane | 3.136 | 4.008 | 4.15 | 4.14 | 227 | pentachlorobenzene | 4.102 | 4.408 | 5.18 | 5.16 |
| 161 | 1-chlorooctane | 3.636 | 4.508 | 4.73 | 4.67 | 228 | hexachlorobenzene | 4.646 | 4.897 | 5.73 | 5.73 |
| 162 | 2-chloropropane | 1.026 | 1.808 | 1.90 | 1.91 | 229 | bromobenzene | 2.028 | 2.891 | 2.99 | 2.97 |
| 163 | 2-chlorobutane | 1.564 | 2.346 | 2.52 | 2.48 | 230 | 1,2-dibromobenzene | 2.662 | 3.788 | 3.64 | 3.64 |
| 164 | bromoethane | 0.807 | 2.094 | 1.61 | 1.68 | 231 | 1,3-dibromobenzene | 2.660 | 3.782 | 3.75 | 3.64 |
| 165 | 1-bromopropane | 1.307 | 2.594 | 2.10 | 2.21 | 232 | 1,4-dibromobenzene | 2.660 | 3.782 | 3.79 | 3.64 |
| 166 | 1-bromobutane | 1.807 | 3.094 | 2.75 | 2.74 | 233 | methanol | −2.236 | 0.447 | −0.74 | −0.95 |
| 167 | 1-bromopentane | 2.307 | 3.594 | 3.37 | 3.26 | 234 | dimethyl ether | −1.826 | 0.816 | 0.10 | 0.06 |
| 168 | 1-bromohexane | 2.807 | 4.094 | 3.80 | 3.79 | 235 | methyl-*n*-propyl ether | −0.351 | 1.904 | 1.21 | 1.04 |
| 169 | 1-bromoheptane | 3.307 | 4.594 | 4.36 | 4.32 | 236 | methyl acetate | −1.090 | 1.31 | 0.18 | 0.26 |
| 170 | 1-bromooctane | 3.807 | 5.094 | 4.89 | 4.85 | 237 | methyl propanoate | −0.530 | 1.877 | 0.82 | 0.85 |
| 171 | 1-bromodecane | 4.807 | 6.094 | 6.00 | 5.90 | 238 | methyl butanoate | −0.030 | 2.377 | 1.29 | 1.38 |
| 172 | 1-bromo-2-methylpropane | 1.663 | 2.950 | 2.53 | 2.59 | 239 | methyl pentanoate | 0.470 | 2.877 | 1.96 | 1.91 |
| 173 | iodoethane | 1.115 | 2.475 | 2.00 | 2.01 | 240 | methyl hexanoate | 0.970 | 3.377 | 2.42 | 2.43 |
| 174 | iodopropane | 1.615 | 2.975 | 2.54 | 2.54 | 241 | methyl nonanoate | 2.470 | 4.877 | 3.87 | 4.01 |
| 175 | 1-iodobutane | 2.115 | 3.475 | 3.08 | 3.06 | 242 | methyl decanoate | 2.970 | 5.337 | 4.41 | 4.54 |
| 176 | 1-iodopentane | 2.615 | 3.975 | 3.62 | 3.59 | 243 | methylamine | −2.236 | 0.577 | −0.57 | −0.95 |
| 177 | 1-iodohexane | 3.115 | 4.475 | 4.16 | 4.12 | 244 | dimethylamine | −2.390 | 1.000 | −0.38 | −0.53 |
| 178 | 1-iodoheptane | 3.615 | 4.975 | 4.70 | 4.64 | 245 | trimethylamine | −2.121 | 1.342 | 0.22 | 0.33 |
| 179 | benzene | 1.395 | 2.000 | 2.13 | 2.30 | 246 | methylthiol | −0.447 | 1.222 | 0.65 | 0.36 |
| 180 | toluene | 1.784 | 2.411 | 2.73 | 2.71 | 247 | fluoromethane | −1.000 | 0.378 | 0.51 | 0.36 |
| 181 | ethylbenzene | 2.363 | 2.971 | 3.15 | 3.32 | 248 | chloromethane | −0.100 | 1.132 | 0.91 | 0.73 |
| 182 | *o*-xylene | 2.174 | 2.827 | 3.12 | 3.12 | 249 | bromomethane | 0.141 | 1.961 | 1.19 | 0.98 |
| 183 | *m*-xylene | 2.172 | 2.821 | 3.20 | 3.12 | 250 | iodomethane | 0.577 | 2.500 | 1.51 | 1.44 |
| 184 | *p*-xylene | 2.172 | 2.821 | 3.15 | 3.12 | 251 | cyclopropane | 1.500 | 1.500 | 1.72 | 1.79 |
| 185 | *n*-propylbenzene | 2.863 | 3.471 | 3.72 | 3.85 | 252 | cyclopentane | 2.500 | 2.500 | 3.00 | 2.84 |
| 186 | isopropylbenzene | 2.755 | 3.354 | 3.66 | 3.74 | 253 | cyclohexane | 3.000 | 3.000 | 3.44 | 3.37 |
| 187 | 1,2,3-trimethylbenzene | 2.564 | 3.244 | 3.66 | 3.54 | 254 | cyclooctane | 4.000 | 4.000 | 4.45 | 4.42 |
| 188 | 1,2,4-trimethylbenzene | 2.562 | 3.238 | 3.56 | 3.53 | 255 | methylcyclopentane | 2.894 | 2.894 | 3.37 | 3.26 |
| 189 | 1,3,5-trimethylbenzene | 2.560 | 3.232 | 3.59 | 3.53 | 256 | methylcyclohexane | 3.394 | 3.394 | 3.61 | 3.78 |
| 190 | 2-ethyltoluene | 2.754 | 3.388 | 3.53 | 3.74 | 257 | cyclopentanol | 0.491 | 2.575 | 0.71 | 0.72 |
| 191 | 4-ethyltoluene | 2.752 | 3.382 | 3.63 | 3.73 | 258 | cyclohexanol | 0.991 | 3.075 | 1.23 | 1.25 |
| 192 | *n*-butylbezene | 3.363 | 3.971 | 4.38 | 4.38 | 259 | cycloheptanol | 1.491 | 3.575 | 1.83 | 1.78 |
| 193 | *tert*-butylbenzene | 3.066 | 3.661 | 4.11 | 4.07 | 260 | cyclooctanol | 1.991 | 4.075 | 2.39 | 2.30 |
| 194 | 4-isopropyltoluene | 3.143 | 3.765 | 4.10 | 4.15 | 261 | cyclopentanone | 0.166 | 2.411 | 0.38 | 0.38 |
| 195 | 1,2,3,4-tetramethylbenzene | 2.955 | 3.661 | 3.98 | 3.95 | 262 | cyclohexanone | 0.666 | 2.911 | 0.81 | 0.91 |
| 196 | 1,2,3,5-tetramethylbenzene | 2.953 | 3.655 | 4.04 | 3.95 | 263 | cyclododecanone | 3.666 | 5.911 | 4.10 | 4.07 |
| 197 | 1,2,4,5-tetramethylbenzene | 2.953 | 3.655 | 4.00 | 3.95 | 264 | formic acid | −1.671 | 0.494 | −0.54 | −0.57 |
| 198 | *n*-pentylbezene | 3.863 | 4.471 | 4.90 | 4.91 | 265 | methyl formate | −1.706 | 0.880 | −0.26 | −0.02 |
| 199 | pentamethylbezene | 3.346 | 4.077 | 4.56 | 4.36 | 266 | ethyl formate | −0.751 | 1.467 | 0.27 | 0.40 |
| 200 | *n*-hexylbenzene | 4.363 | 4.971 | 5.52 | 5.43 | 267 | *n*-propyl formate | −0.251 | 1.967 | 0.83 | 0.93 |
| 201 | *n*-octylbenzene | 5.363 | 5.971 | 6.34 | 6.49 | 268 | formamide | −2.873 | 0.569 | −1.51 | −1.83 |

**Table 1** (Continued)

| number | compounds | $^1\chi^{opt}$ | $^1\chi^v$ | log *P* exptl. | log *P* eq 18 | number | compounds | $^1\chi^{opt}$ | $^1\chi^v$ | log *P* exptl. | log *P* eq 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 269 | *N*-methylformamide | −2.770 | 1.024 | −0.97 | −1.15 | 288 | 1,4-dichlorobutane | 1.358 | 3.101 | 2.24 | 2.26 |
| 270 | *N*-methylacetamide | −2.090 | 1.454 | −1.05 | −0.79 | 289 | 1,1,1-trichloroethane | 0.050 | 2.198 | 2.49 | 2.61 |
| 271 | *N*,*N*-dimethylformamide | −3.133 | 1.388 | −1.01 | −0.95 | 290 | 1,1,2-trichloroethane | 0.079 | 2.516 | 1.89 | 1.95 |
| 272 | *N*,*N*-dimethylacetamide | −2.486 | 1.822 | −0.77 | −0.63 | 291 | 1,1,2,2-tetrachloroethane | −0.183 | 2.948 | 2.39 | 2.33 |
| 273 | *N*,*N*-diethylacetamide | −0.477 | 2.974 | 0.34 | 0.33 | 292 | 1,1,1,2-tetrachloroethane | −0.167 | 2.853 | 2.66 | 2.76 |
| 274 | *N*,*N*-dimethylpropanamide | −1.925 | 2.383 | −0.11 | −0.04 | 293 | pentachloroethane | −0.420 | 3.294 | 3.22 | 3.15 |
| 275 | methyl acrylate | −0.894 | 1.513 | 0.80 | 0.92 | 294 | hexachloroethane | −0.650 | 3.647 | 4.14 | 3.97 |
| 276 | ethyl acrylate | 0.061 | 2.101 | 1.32 | 1.35 | 295 | 1,2-dichloropropane | 0.786 | 2.601 | 2.02 | 2.04 |
| 277 | *n*-butyl acrylate | 1.061 | 3.101 | 2.36 | 2.40 | 296 | dimethoxymethane | −3.117 | 1.394 | 0.18 | −0.03 |
| 278 | isobutyl acrylate | 0.917 | 2.957 | 2.22 | 2.25 | 297 | diethoxymethane | −1.168 | 2.569 | 0.84 | 0.87 |
| 279 | methyl methacrylate | −0.487 | 1.920 | 1.38 | 1.35 | 298 | cyclohexanediol | −1.072 | 3.154 | 0.16 | 0.34 |
| 280 | ethyl methacrylate | 0.468 | 2.508 | 1.94 | 1.78 | 299 | 1,2-ethandiol | −2.662 | 1.132 | −1.36 | −1.37 |
| 281 | *n*-butyl methacrylate | 1.468 | 3.508 | 2.88 | 2.83 | 300 | 2,3-butanediol | −2.163 | 2.004 | −0.92 | −0.85 |
| 282 | isobutyl methacrylate | 1.324 | 3.363 | 2.66 | 2.68 | 301 | 1,2-dimethoxyethane | −2.617 | 1.894 | −0.21 | −0.17 |
| 283 | dichloromethane | −0.141 | 1.601 | 1.25 | 1.34 | 302 | 1,2-diethoxyethane | −0.668 | 3.069 | 0.66 | 0.73 |
| 284 | 1,1-dichloroethane | 0.319 | 1.885 | 1.79 | 1.82 | 303 | 1,5-dimethoxydiethyl ether | −3.408 | 2.971 | −0.36 | −0.40 |
| 285 | trichloromethane | −0.387 | 1.961 | 1.97 | 2.15 | 304 | methoxyethanol | −2.634 | 1.513 | −0.77 | −0.76 |
| 286 | 1,2-dichloroethane | 0.359 | 2.101 | 1.48 | 1.59 | 305 | ethoxyethanol | −1.665 | 2.101 | −0.32 | −0.32 |
| 287 | 1,3-dichloropropane | 0.859 | 2.601 | 2.00 | 1.74 | 306 | butoxyethanol | −0.665 | 3.101 | 0.83 | 0.73 |

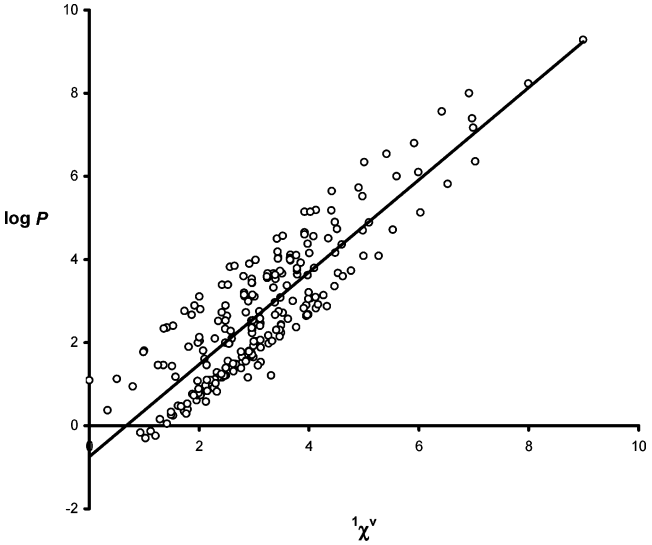


**Figure 1.** Correlation between experimental log *P* values and valence first-order molecular connectivity indices for compounds 1−232 listed in Table 1.

algorithm (in which the term describing the triple bond was subtracted from the rest of the terms), these compounds fitted the model well (eq 10):

$$\log P = 0.826\ (\pm 0.088) + 1.059\ (\pm 0.028) \times {}^1\chi^{opt} \quad (10)$$

$$n = 37,\ r^2 = 0.994,\ s = 0.146,\ F^{1,35} = 5685.7,\ r^2_{cv} = 0.993,\ \mathrm{PRESS} = 0.8532$$

Therefore, throughout the optimization process, we used the standard delta values for aliphatic (primary, secondary, tertiary, and quaternary) carbon atoms as their optimal weights.

Applying the above outlined procedure, we determined the optimal weights for characteristic skeletal atoms (Table 2) in several classes of simple aliphatic organic compounds (1−178), introducing one class of molecules at a time. When the analysis had been extended to benzene and its derivatives (compounds 179−232), it was found that a distinct set of optimized weights for aromatic carbon atoms and heteroatoms attached to them is required for inclusion of these compounds in the model (Table 2). This is not unexpected because, as a rule, bonding to an aromatic nucleus makes substituents more lipophilic.[1,25] In the spirit of the molecular connectivity formalism, we adopted the value 5.30 [optimal value for unsubstituted aromatic carbon (4.30) + 1] as optimal for the aromatic carbons to which substituents are attached.

The model generated for simple aliphatic and aromatic compounds (Figure 2) reads as follows (eq 11):

$$\log P = 0.816\ (\pm 0.021) + 1.061\ (\pm 0.009) \times {}^1\chi^{opt} \quad (11)$$

$$n = 232,\ r^2 = 0.996,\ s = 0.110,\ F^{1,230} = 57\,784.4,\ r^2_{cv} = 0.996,\ \mathrm{PRESS} = 2.8011$$

Its standard deviation is more than 7 times lower than that in the model based on the valence first-order molecular connectivity index (eq 7). We propose the term "optimized molecular connectivity indices" for the indices, which are calculated from optimal weights for the skeletal atoms rather than from corresponding valence delta values. In this paper, this is indicated by the abbreviation opt in the superscript of the index.

As evident from Table 2, the price for more accurately predicted log *P* values is a greater number of weights, relative to the standard procedure, that are required for adequate treatment of heteroatoms in different atomic environments. Thus, although there are only two valence delta values for oxygen in the valence connectivity formalism, depending on whether it is bonded to hydrogen or not, we found several optimal weights for oxygen just for compounds considered in our analysis. However, a similar situation with a great number of fragmental constants is also present in some of the most successful procedures for the calculation of log *P*, such as the Hansch−Leo,[1] the Rekker,[25] or the KOWWIN[26] methods. It is worth noting that the optimal weights for the heteroatoms depend on the type of carbon atom to which they are attached, which further increases the number of required weights. See, for example, the listed empiric weights

**Table 2.** Optimal Weights for Selected Skeletal Atoms (Highlighted in Bold) Obtained by the Described Optimization Procedure[a]

| Atom | Weight |
|---|---|
| —C≡**C**H | - 3.00 |
| —$CH_2$—**O**H | - 0.20 |
| —CH(—**O**H)— | - 0.10 |
| —C(—**O**H)— | - 0.07 |
| —$CH_2$—**O**—$CH_2$— | - 1.20 |
| —C(=**O**)— | - 0.06 |
| —C(=**O**)—OH | - 1.25 |
| —C(=**O**)—**O**H | - 0.25 |
| —$CH_2$—O—C(=**O**)— | - 2.40 |
| —$CH_2$—**O**—C(=**O**)— | - 1.40 |
| —$CH_2$—**N**$H_2$ | - 0.20 |
| —CH(—**N**$H_2$)— | - 0.11 |
| —C(—**N**$H_2$)— | - 0.07 |
| —$CH_2$—**N**H—$CH_2$— | - 0.70 |
| —CH(—**N**H—CH<)— | - 0.34 |
| —$CH_2$—**N**(—$CH_2$—)—$CH_2$— | - 2.00 |
| —C≡**N** | -0.09 |
| —C(=**O**)—$NH_2$ | - 1.25 |
| —C(=**O**)—**N**$H_2$ | - 0.06 |
| —C(=**O**)—**N**H—$CH_2$— | - 0.49 |
| —C(=**O**)—**N**(—$CH_2$—)—$CH_2$— | - 0.97 |
| —$CH_2$—**S**H | - 5.00 |
| —$CH_2$—**F** | - 1.00 |
| —$CH_2$—**Cl** | - 100.00 |
| —CH(—**Cl**)— | - 20.00 |
| —C(—**Cl**)— | - 11.10 |
| —$CH_2$—**Br** | 50.00 |
| —$CH_2$—**I** | 3.00 |
| —(Ar)**C**H | 4.30 |
| —(Ar)C—**O**H | - 0.35 |
| —(Ar)C—**Cl** | 0.55 |
| —(Ar)C—**Br** | 0.41 |

[a] A negative sign in front of the weight denotes that in the computation of the optimized descriptor, the bond terms $(\delta_i^{opt}\delta_j^{opt})^{-0.5}$ including the corresponding skeletal atom are to be subtracted.

for oxygen in primary, secondary, and tertiary alcohols. This is a consequence of the nature of the algorithm for calculation of the first-order molecular connectivity index. Namely, in searching for an optimal weight, we optimize the product of the weights for a particular carbon atom (a fixed value) and an adjacent heteroatom (a variable value). A simple way to treat this problem is to average the corresponding optimal weights, but this would certainly result in decreased accuracy for the predicted partition coefficients. Alternatively, a unique weight for a given heteroatom could be used in calculations of the optimized molecular connectivity index, no matter to which type of carbon atom it is attached, and according to need the log *P* thus obtained could be corrected by empirically found correction factors. If the optimal weight for the oxygen of primary alcohols is used in calculations of the indices for secondary and tertiary alcohols, their calculated partition coefficients are systematically overestimated. This can be corrected by subtracting empirical correction factors from the calculated log *P* values or, as we did, by inclusion of corresponding indicator variables in the model, where they play the same role. In general, an indicator variable shows the presence or absence of a structural element in a molecule. It takes a value of one if the element is present in the structure and zero if it is not. The question arises whether the corrections established for one class of compound can be extended to other classes of molecules. Our analysis showed that the same corrections can be applied to alcohols (38−68) and amines (131−142), whereas chloroalkanes (155−163) do not require such corrections. The following model is obtained when the above-discussed corrections are incorporated:

$$\log P = 0.818\ (\pm 0.023) + 1.060\ (\pm 0.009) \times {}^1\chi^{opt} - 0.553\ (\pm 0.060) \times I_{SC} - 0.837\ (\pm 0.099) \times I_{TC} \quad (12)$$

$$n = 232,\ r^2 = 0.996,\ s = 0.110,\ F^{3,228} = 18949.0,\ r^2_{cv} = 0.996,\ \mathrm{PRESS} = 2.8585$$

The indicator variables stand for compounds having, at the secondary ($I_{SC}$) or tertiary ($I_{TC}$) carbon, functional groups for which optimized indices were calculated from the corresponding weights for heteroatoms attached to a primary carbon. Clearly, a more thorough analysis is necessary for evaluating the real potential of the procedure.

Heteroatoms attached to a methyl group (233−250) also require a new set of optimized weights. To reduce the total number of such weights, we handle those compounds as described in the above example. In the calculation of their indices, we treat the heteroatoms as if they were bonded to primary carbon. It was found that the partition coefficients calculated for amino, flouro, and hydroxyl derivatives of methane as well as for various methyl ethers and methyl esters require such correction ($I_{MET}$), whereas the correction is not needed for chloro, bromo, iodo, and thiol derivatives of methane.

$$\log P = 0.832\ (\pm 0.021) + 1.053\ (\pm 0.009) \times {}^1\chi^{opt} + 0.594\ (\pm 0.045) \times I_{MET} \quad (13)$$

$$n = 250,\ r^2 = 0.996,\ s = 0.115,\ F^{2,247} = 28\ 996.4,\ r^2_{cv} = 0.996,\ \mathrm{PRESS} = 3.799$$

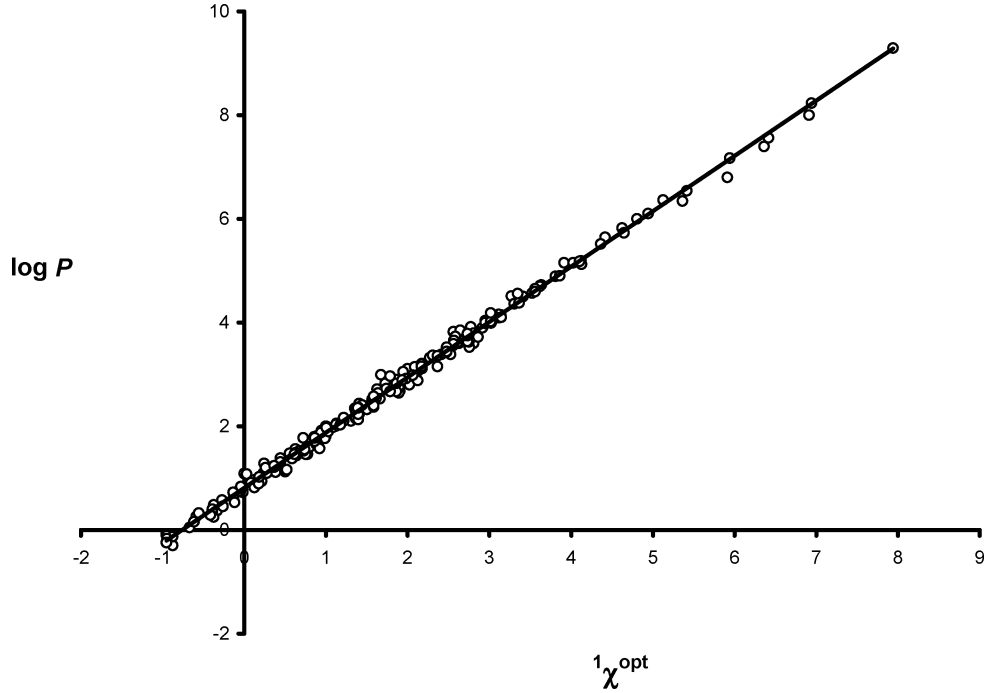


**Figure 2.** Correlation between experimental log *P* values and optimized first-order molecular connectivity indices for compounds 1−232 listed in Table 1.

It is known from early applications of molecular connectivity indices that certain physicochemical properties of cyclic compounds are not successfully modeled by this parameter. Kier et al.[27] proposed subtracting the quantity 0.5 from the index as a way to improve such correlations. Our analysis showed that the lipophilicity of alicyclic compounds (251−263) is another molecular property that cannot be modeled well by the index. As in the previous examples, this could be overcome by finding empirically determined optimal weights for carbon atoms forming rings. However, this does not appear to be an optimal solution here. Because the weights of heteroatoms are generally dependent on the type of carbon to which they are bonded (vide supra), the replacement of the delta values by the optimal ones for the carbon atoms in the rings would entail the recalculation of most of the previously determined optimal weights for heteroatoms. To avoid this, thus keeping the total number of optimal weights as low as possible, we introduced a correction factor ($I_{ALRIN}$) for the presence of an aliphatic ring in a molecule. Interestingly, the numerical value of the regression coefficient in front of the indicator variable is similar to the correction factor proposed by Kier et al.

$$\log P = 0.831\ (\pm 0.021) + 1.055\ (\pm 0.009) \times {}^1\chi^{opt} + 0.598\ (\pm 0.046) \times I_{MET} - 0.614\ (\pm 0.064) \times I_{ALRIN} \quad (14)$$

$$n = 263,\ r^2 = 0.996,\ s = 0.114,\ F^{3,259} = 20\ 272.8,\ r^2_{cv} = 0.996,\ \text{PRESS} = 3.4942$$

A hydrogen atom bonded to a strongly electronegative group is more lipophilic than usual.[25] Hence, compounds containing such hydrogen [for example, formic acid and its derivatives (264−269, 271)] require special treatment. As hydrogen atoms are usually neglected in the graph-theoretical approach, we solved this peculiarity by inclusion of indicator variable $I_{PHYD}$.

$$\log P = 0.831\ (\pm 0.021) + 1.054\ (\pm 0.009) \times {}^1\chi^{opt} + 0.592\ (\pm 0.045) \times I_{MET} + 0.614\ (\pm 0.064) \times I_{ALRIN} + 0.250\ (\pm 0.115) \times I_{PHYD} \quad (15)$$

$$n = 267,\ r^2 = 0.996,\ s = 0.114,\ F^{4,262} = 15\ 698.9,\ r^2_{cv} = 0.996,\ \text{PRESS} = 3.5586$$

Up to now we have not considered organic molecules with more than one functional group. The problem, which may arise in predicting the octanol−water partition coefficients of such compounds, is that their lipophilicity is affected by interactions among functional groups. As the optimized weights for the skeletal atoms are derived from simple organic molecules, they cannot be expected to cope with complexities of this kind. In acrylates (275−282), for instance, the carbon−carbon double bonds are in conjugation with the carboxylic group, which makes them more lipophilic than they would be if the functional groups were independent of each other. Hence, it is not surprising that the partition coefficients for acrylates calculated from eq 15 are systematically underestimated. In a similar fashion to that in the previous example, we compensate the effect of conjugation[1,25] by inclusion of an interaction factor (indicator variable $I_{CONJUG}$).

$$\log P = 0.832\ (\pm 0.021) + 1.054\ (\pm 0.009) \times {}^1\chi^{opt} + 0.568\ (\pm 0.039) \times I_{MET} + 0.368\ (\pm 0.092) \times I_{PHYD} - 0.614\ (\pm 0.065) \times I_{ALRIN} + 0.454\ (\pm 0.083) \times I_{CONJUG} \quad (16)$$

$$n = 282,\ r^2 = 0.996,\ s = 0.116,\ F^{5,276} = 13\ 258.9,\ r^2_{cv} = 0.996,\ \text{PRESS} = 3.9362$$

The proximity effect[1,25] is another intermolecular interaction, which will be considered here. In alkanes, the

accumulation of polar groups (such as halogen, alkoxy, ester, carbonyl, carboxyl, amino, etc.) at the same (geminal) or adjacent (vicinal) carbon atoms reduces their tendency toward hydration, which results in increased lipophilicity. Preliminary analysis showed that proximity effects for groups with a poor tendency to form hydrogen bonds (halogens) should be treated separately from those capable of forming such bonds. Adequate treatment of the analyzed multiply chlorinated compounds (283−295) requires the inclusion of three distinctive correction factors:

$$\log P = 0.831\ (\pm 0.021) + 1.055\ (\pm 0.009) \times {}^{1}\chi^{\mathrm{opt}} + 0.569\ (\pm 0.039) \times I_{\mathrm{MET}} + 0.370\ (\pm 0.093) \times I_{\mathrm{PHYD}} - 0.613\ (\pm 0.065) \times I_{\mathrm{ALRIN}} + 0.455\ (\pm 0.082) \times I_{\mathrm{CONJUG}} + 0.657\ (\pm 0.099) \times I_{\mathrm{HG2}} + 1.726\ (\pm 0.100) \times I_{\mathrm{HG3}} + 0.380\ (\pm 0.124) \times I_{\mathrm{HVIC}} \quad (17)$$

$$n = 295,\ r^2 = 0.996,\ s = 0.118,\ F^{8,286} = 8243.0,\ r^2_{\mathrm{cv}} = 0.995,\ \mathrm{PRESS} = 4.2725$$

The newly introduced indicator variables stand for compounds, with two ($I_{HG2}$) and three ($I_{HG3}$) geminal halogens on a carbon atom, and $I_{HVIC}$ is the correction factor taking into account interactions among halogens bonded to vicinal carbons.

To compensate for the proximity effects in the compounds (296−306) with polar groups capable of forming hydrogen bonds, two additional indicator variables are introduced in the model:

$$\log P = 0.829\ (\pm 0.020) + 1.055\ (\pm 0.008) \times {}^{1}\chi^{\mathrm{opt}} + 0.580\ (\pm 0.036) \times I_{\mathrm{MET}} + 0.367\ (\pm 0.093) \times I_{\mathrm{PHYD}} - 0.627\ (\pm 0.063) \times I_{\mathrm{ALRIN}} + 0.454\ (\pm 0.083) \times I_{\mathrm{CONJUG}} + 0.658\ (\pm 0.099) \times I_{\mathrm{HG2}} + 1.726\ (\pm 0.099) \times I_{\mathrm{HG3}} + 0.381\ (\pm 0.124) \times I_{\mathrm{HVIC}} + 1.271\ (\pm 0.138) \times I_{\mathrm{PG2}} + 0.605\ (\pm 0.074) \times I_{\mathrm{PVIC}} \quad (18)$$

$$n = 306,\ r^2 = 0.996,\ s = 0.117,\ F^{10,295} = 7165.6,\ r^2_{\mathrm{cv}} = 0.995,\ \mathrm{PRESS} = 4.5048$$

where $I_{PG2}$ and $I_{PVIC}$ are corrections that should be made for interactions between the polar groups separated by one and two carbon atoms, respectively. It should be mentioned that in the Hansch−Leo and the Rekker methods,[1,25] these effects are treated in a similar way. It appears realistic to expect that intramolecular interactions not considered here can be treated in a similar fashion as that outlined in the previous examples. As only a limited number of molecular interactions could be examined in this study, we did not probe to determine if some of the indicator variables appearing in the regression equations could be replaced by the (optimized) cluster or path-cluster molecular connectivity indices. We intend to analyze this issue extensively in one of our future studies.

In the closing paragraph, we present the results of testing the robustness of our model (Table 4). The 30 compounds included in the validation set (Table 3) are structurally similar but more complex than the molecules taken in the construction of the model. As a rule, they possess a combination of the functionalities that are present in the compounds of the training set. For the sake of comparison, the validation statistics for the model based on the valence connectivity index (eq 7) and for several other models based on the optimized connectivity index are also included in Table 4. As shown in the table, the average difference between the observed and predicted log *P* values for the best model (eq 18) amounts to 0.12 log units. This can be considered acceptable, taking into account that, in general, the standard deviation of experimental log *P* values ranges from ±0.04 to 0.10 log units.[1] On the other hand, even inclusion of the indicator variables appearing in eq 18 in the model based on the first-order valence connectivity index does not significantly improve its predictive ability (regression equation not shown). The average error of the log *P* values calculated from this model exceeds 1 log unit (1.1).

**Table 3.** Molecular Connectivity Indices and Experimental and Calculated log *P* Values for the Validation Set

| | | | | log *P* | |
|---|---|---|---|---|---|
| number | compounds | ${}^{1}\chi^{opt}$ | ${}^{1}\chi^{v}$ | exptl. | eq 18 |
| 1 | 1,4-butanediol | −1.662 | 2.132 | −0.83 | −0.92 |
| 2 | 1-methylcyclohexanol | 1.317 | 3.431 | 1.33 | 1.59 |
| 3 | tetrahydrofuran | 0.209 | 2.077 | 0.46 | 0.42 |
| 4 | tetrahydropyran | 0.709 | 2.577 | 0.95 | 0.95 |
| 5 | 4-methylcyclohexanone | 1.060 | 3.305 | 1.38 | 1.32 |
| 6 | 5-phenylpentan-2-one | 1.469 | 4.322 | 2.42 | 2.38 |
| 7 | glutaric acid | −1.187 | 2.563 | −0.29 | −0.42 |
| 8 | cyclohexanecarboxylic acid | 1.658 | 3.533 | 1.96 | 1.95 |
| 9 | tripropylamine | 2.121 | 4.570 | 2.79 | 3.07 |
| 10 | cyclohexylamine | 1.076 | 3.150 | 1.49 | 1.34 |
| 11 | methyl 4-phenylbutanoate | 1.419 | 4.434 | 2.77 | 2.91 |
| 12 | dimethyl adipate | −0.974 | 3.840 | 1.03 | 0.96 |
| 13 | 4-chlorobutyronitrile | −0.384 | 2.378 | 0.56 | 0.43 |
| 14 | octanenitrile | 1.894 | 3.784 | 2.73 | 2.83 |
| 15 | 4-phenylbutyramid | 0.521 | 4.111 | 1.41 | 1.38 |
| 16 | 4-chlorobutanol | −0.152 | 2.617 | 0.85 | 0.67 |
| 17 | cyclopropylamine | −0.424 | 1.650 | 0.07 | −0.24 |
| 18 | 4-*tert*-butylcyclohexanol | 2.596 | 4.680 | 3.06 | 2.94 |
| 19 | bromochloromethane | 0.029 | 2.187 | 1.41 | 1.52 |
| 20 | difluoromethane | −1.414 | 0.534 | 0.20 | 0.00 |
| 21 | 1,2-diiodoethane | 1.316 | 4.036 | 2.71 | 2.60 |
| 22 | thymol | 2.365 | 3.905 | 3.30 | 3.33 |
| 23 | bibenzyl | 3.813 | 5.028 | 4.79 | 4.85 |
| 24 | phenanthren | 3.451 | 4.815 | 4.46 | 4.47 |
| 25 | *p*-chlorotoluene | 2.323 | 2.888 | 3.33 | 3.28 |
| 26 | crotonic acid | −0.248 | 1.627 | 0.72 | 1.02 |
| 27 | acrylamide | −1.792 | 1.190 | −0.67 | −0.60 |
| 28 | 1,4-dioxane | −1.582 | 2.155 | −0.27 | −0.26 |
| 29 | 1,2-dibutoxyethane | 1.332 | 5.069 | 2.48 | 2.84 |
| 30 | 1-pentyne | 1.272 | 1.849 | 1.98 | 2.17 |

**Table 4.** Comparison of the Predictive Capacities of Selected Models[a]

| | $r^2_{\mathrm{test}}$ | MAE[b] |
|---|---|---|
| eq 7 | 0.621 | 1.25 |
| eq 11 | 0.913 | 0.38 |
| eq 13 | 0.935 | 0.31 |
| eq 14 | 0.978 | 0.18 |
| eq 16 | 0.973 | 0.18 |
| eq 17 | 0.987 | 0.13 |
| eq 18 | 0.989 | 0.12 |

[a] $r^2_{\mathrm{test}}$ is the squared correlation coefficient between the observed and predicted log *P* values (calculated from the indicated models) for the test set of compounds displayed in Table 3. [b] MAE is the mean absolute error of prediction for the respective model.

## CONCLUSION

In this paper, we put forward a molecular connectivity index specifically designed for estimating the octanol−water partition coefficients. Instead of the valence delta values, we used empirically determined optimized weights for characterization of the skeletal atoms in a molecule. In addition, we modified the algorithm for calculating the molecular connectivity indices, assuming that, in the case of log *P*, the contribution of a bond containing a heteroatom may be negative. The index proved to be superior to the valence connectivity index in modeling the log *P* values when tested on a set of simple, structurally diverse organic molecules. Additional correction factors are needed for the prediction of the partition coefficients of polyfunctional compounds for which intramolecular interactions are effective. It appears that the proposed molecular connectivity index offers a possibility for developing a global model for predicting this physicochemical property.

## ACKNOWLEDGMENT

This study was supported by Grants 0178067 (M.Š.) and 0098035 (D.P.), awarded by the Ministry of Science and Technology of the Republic of Croatia.

## REFERENCES AND NOTES

(1) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Application in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(2) *Handbook of Chemical Property Estimation Methods*; Lyman, W. J., Reehl, W. F., Rosenblatt, D. H., Eds.; American Chemical Society: Washington, DC, 1990.

(3) Hansch, C.; Leo, A.; Mekapati, S. B.; Kurup, A. QSAR and ADME. *Biorg. Med. Chem.* **2004**, *12*, 3391−3400.

(4) Manhold, R.; van de Waterbeemd, H. Substructure and Whole Molecule Approaches for Calculating log *P*. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 337−354.

(5) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(6) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press Ltd: Letchworth, England, 1986.

(7) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.

(8) *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Sci. Pub.: Amsterdam, The Netherlands, 1999.

(9) Estrada, E.; Uriarte, E. Recent Advances on the Role of Topological Indices in Drug Discovery Research. *Curr. Med. Chem.* **2001**, *8*, 1573−1588.

(10) Šoškić, M.; Klaić, B.; Magnus, V.; Sabljić, A. Quantitative Structure−Activity Relationships for N-(indol-3-ylacetyl)amino Acids Used as Sources of Auxin in Plant Tissue Culture. *Plant Growth Regul.* **1995**, *16*, 141−152.

(11) Šoškić, M.; Sabljić, A. QSAR Study of 4-Hydroxypyridine Derivatives as Inhibitors of the Hill Reaction. *Pestic. Sci.* **1995**, *45*, 133−141.

(12) Šoškić, M.; Plavšić, D.; Trinajstić, N. 2-Difluoromethylthio-4,6-bis(monoalkylamino)-1,3,5,-triazines as Inhibitors of Hill Reaction: A QSAR Study with Orthogonalized Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 146−150.

(13) Šoškić, M.; Plavšić, D. QSAR Study of 1,8-Naphthyridin-4-ones as Inhibitors of Photosystem II. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1316−1321.

(14) Plavšić, D.; Trinajstić, N.; Amić, D.; Šoškić, M. Comparison between the Structure-Boiling Point Relationships with Different Descriptors for Condensed Benzenoids. *New. J. Chem.* **1998**, 1075−1078.

(15) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(16) Saha, A.; Roy, K. Comparative QSPR Studies with Molecular Connectivity, Molecular Negentropy and TAU Indices. Part 2. Lipid-Water Partition Coefficient of Diverse Functional Acyclic Compounds. *Internet Electron. J. Mol. Des.* **2003**, *2*, 288−305.

(17) Dearden, J. C.; Wong, E. H. Y.; Walker, J. D. Molecular Connectivity: Polarity Correction to Improve Correlation with Hydrophobicity. *QSAR Comb. Sci.* **2004**, *23*, 75−79.

(18) Randić, M. Novel Graph Theoretical Approach to Heteroatoms in QSAR. *Chemom. Intell. Lab. Syst.* **1991**, *10*, 213−227.

(19) Randić M.; Mills, D.; Basak, S. C. On Characterization of Physical Properties of Amino Acids. *Int. J. Quantum Chem.* **2000**, *80*, 1199−1209.

(20) Abraham, M. H.; Chada, S. H.; Whiting, G. S.; Mitchell, R. C. Hydrogen Bonding. An Analyis of Water-Octanol and Water-Alkane Partitioning and Δlog P Parameter of Seiler. *J. Pharm. Sci.* **1994**, *83*, 1085−1100.

(21) Hansch, C.; Leo, A.; Hoekman, D. A. *Exploring QSAR. Hydrophobic, Electronic and Steric Constants*; American Chemical Society: Washington, DC, 1995.

(22) Estrada E., Generalization of topological indices. *Chem. Phys. Lett.* **2001**, *336*, 248−252.

(23) Estrada, E. Generalized Graph Matrix, Graph Geometry, Quantum Chemistry and Optimal Description of Physicochemical Properties. *J. Phys. Chem. A* **2003**, *107*, 7482−7489.

(24) Estrada, E. Three-Dimensional Generalized Graph Matrix, Harary Descriptors and a Generalized Interatomic Lennard-Jones Potential. *J. Phys. Chem. A* **2004**, *108*, 5468−5473.

(25) Rekker, R. F.; Mannhold, R. *Calculation of Drug Lipophilicity*; VCH: Weinheim, Germany, 1992.

(26) Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Methods for Estimating Octanol−Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83−93.

(27) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randić, M. Molecular Connectivity I. Relationship to Nonspecific Local Anesthesia. *J. Pharm. Sci.* **1976**, *64*, 1971−1974.

CI050024V