

# Neural Network Modeling for Estimation of Partition Coefficient Based on Atom-Type Electrotological State Indices

Jarmo J. Huuskonen,<sup>†</sup> David J. Livingstone,<sup>‡</sup> and Igor V. Tetko<sup>\*,§</sup>

Division of Pharmaceutical Chemistry, Department of Pharmacy, POB 56, University of Helsinki, Helsinki FIN-00014, Finland, ChemQuest, Delamere House, 1, Royal Crescent, Sandown, Isle of Wight PO36 8LZ, U.K., Centre for Molecular Design, University of Portsmouth, Portsmouth, Hants PO1 2EG, U.K., Laboratoire de Neuro-Heuristique, Institut de Physiologie, Université de Lausanne, Rue du Bugnon 7, Lausanne CH-1005, Switzerland, and Biomedical Department, Institute of Bioorganic & Petroleum Chemistry, Murmanskaya 1, Kiev-660, 253660 Ukraine

Received May 14, 1999

A method for predicting log *P* values for a diverse set of 1870 organic molecules has been developed based on atom-type electrotological-state (E-state) indices and neural network modeling. An extended set of E-state indices, which included specific indices with a more detailed description of amino, carbonyl, and hydroxy groups, was used in the current study. For the training set of 1754 molecules the squared correlation coefficient and root-mean-squared error were  $r^2 = 0.90$  and  $\text{RMS}_{\text{LOO}} = 0.46$ , respectively. Structural parameters which included molecular weight and 38 atom-type E-state indices were used as the inputs in 39-5-1 artificial neural networks. The results from multilinear regression analysis were  $r^2 = 0.87$  and  $\text{RMS}_{\text{LOO}} = 0.55$ , respectively. For a test set of 35 nucleosides, 12 nucleoside bases, 19 drug compounds, and 50 general organic compounds ( $n = 116$ ) not included in the training set, a predictive  $r^2 = 0.94$  and  $\text{RMS} = 0.41$  were calculated by artificial neural networks. The results for the same set by multilinear regression were  $r^2 = 0.86$  and  $\text{RMS} = 0.72$ . The improved prediction ability of artificial neural networks can be attributed to the nonlinear properties of this method that allowed the detection of high-order relationships between E-state indices and the *n*-octanol/water partition coefficient. The present approach was found to be an accurate and fast method that can be used for the reliable estimation of log *P* values for even the most complex structures.

## INTRODUCTION

The *n*-octanol/water partition coefficient is the ratio of the concentration of a chemical in *n*-octanol to that in water in a two-phase system at equilibrium. The measured values of this coefficient range from  $<10^{-4}$  to  $>10^8$ . The logarithm of this coefficient, log *P*, has been shown to be one of the key parameters in quantitative structure–activity relationship (QSAR) studies for simulating the lipophilicity of organic molecules and can be used to provide invaluable information for the overall understanding of the uptake, distribution, biotransformation, and elimination of a wide variety of chemicals. This coefficient is very important in the process of drug discovery and development from molecular design to pharmaceutical formulation and biopharmacy. A high lipophilicity is likely to hamper the bioavailability of the drugs and causes other difficulties with the use of such compounds. One strategy to find new lead compounds, which has proved valuable in recent years, is that of high-throughput screening, where collections of thousands of compounds are screened with the intention of finding relevant biological

activity.<sup>1</sup> It has been noticed that the synthesis of combinatorial libraries tends to result in compounds with higher molecular weights and higher lipophilicity than with conventional synthetic strategies. For this reason computational screening has been used to select sublibraries with physicochemical properties relevant to the range of known values, such as log *P*, of orally active drugs.<sup>2,34</sup> Hence, there is considerable interest in fast and accurate structure-based methods for estimation of log *P* for the rational development of new drugs before a promising drug candidate has been synthesized.

Several approaches have been developed for the prediction of log *P* based on nonexperimental structural parameters. Most of these methods use substructures (fragments/atom fragments)<sup>5–10</sup> or quantum chemical parameters (charges, electronic potentials, molecular volumes, and shape, etc)<sup>11–16</sup> and multilinear regression analysis to fit models to experimental data. The fragment-based methods work well for a large number of compounds; however, difficulties can arise in decomposing some structures into appropriate fragments whose constants are available. Several correction factors are also needed for some molecular interactions. Quantum chemical parameters are properties of the entire solute molecule. Although these parameters give promising results for many molecules, their merit is not so far sufficient for use as a general estimation method. In addition, the quantum

\* To whom correspondence should be addressed at the Université de Lausanne. Phone: ++41-21-692-5534. FAX: ++41-21-692-5505. E-mail: itetko@eliot.unil.ch.

<sup>†</sup> University of Helsinki.

<sup>‡</sup> ChemQuest and University of Portsmouth.

<sup>§</sup> Université de Lausanne and Institute of Bioorganic & Petroleum Chemistry.

chemical calculations are too time-consuming if their estimation for a large number of compounds is required. One other feature of most of these reported models is that they are based on linear combinations of the molecular descriptors. The possibility that nonlinear effects may improve their predictive ability needs to be examined.

Thus, there is a need to develop a model that can be reliably applied to estimate log *P* coefficients for a diverse set of molecules, the descriptors used in such estimations should be easily calculated for all analyzed molecules, and the overall estimation should be fast.

The electrotopological-state (E-state) indices were recently introduced by Hall and Kier<sup>17,18</sup> for the description of molecules. These indices combine together both electronic and topological characteristics of the analyzed molecules. For each atom type in a molecule the E-state indices are summed and can be used in a group contribution manner. Thus, the calculation of such indices is very straightforward and simple. The atom-type E-state indices have been found to be useful in quantitative structure–property relationship studies such as the prediction of boiling point, critical temperature,<sup>19</sup> and aqueous solubility (log *S*).<sup>20</sup> In our previous study with a heterogeneous set of 345 organic compounds we found that the E-state indices can be successfully used to estimate the partition coefficient.<sup>21</sup> We suggested that an extension of E-state indices by the inclusion of specific indices with a more detailed description of amino, carbonyl, and hydroxy groups could improve the calculated results. The current study reports the new results with an extended set of the E-state indices using a much larger database composed of 1754 compounds.

#### DATA SETS

The applicability and accuracy of a log *P* estimation model is directly affected by the size and quality of the training set. We have extended our previously used training set of 326 compounds<sup>21</sup> by including the Klopman learning database of 1663 molecules.<sup>8</sup> There were some duplicate compounds, and after their exclusion a training set of 1754 molecules was compiled. The partition coefficients expressed as logarithm values, log *P*, were in the range −4.20 to +5.90, corresponding to L-ornithine and thioridazine, respectively.

Three different test sets were used to validate the performance of multilinear regression and artificial neural network (ANN) models and to compare them with other proposed estimation methods. The first test set of 35 nucleosides and 12 nucleoside bases was introduced by Viswanadhan et al.<sup>22</sup> The second test set of 19 drug compounds<sup>23</sup> was analyzed in our preliminary study reported elsewhere.<sup>21</sup> The last test set of 50 general organic compounds was used to provide a comparison of the elaborated algorithm with neural network-based approaches used in modeling log *P*.<sup>24,25</sup> Thus, there were in total 116 compounds in all three test sets with a range of log *P* from −2.51 to +5.25 corresponding to cytosine and flufenamic acid, respectively.

The experimental log *P* values reported by different authors as well as commercial organizations, e.g. such as KOWWIN or CLOGP (available on-line at <http://esc-plaza.syrres.com/interkow/kowdemo.htm> and <http://www.daylight.com/daycgi/clogp>), can differ by as much as 0.1–0.3 log *P*

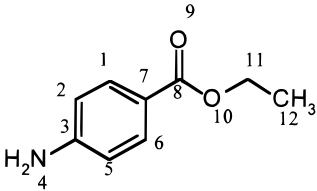
units. For example, CLOGP supplies experimental values 1.88 and 0.8 for methylene bromide and 2-nitropropane, while KOWWIN gives 1.70 and 0.93. Each commercial source provides homogeneous data sets and verifies if the reported experimental values were recorded in the most similar conditions using as much as possible the same methods. Thus, to have homogeneous training and test sets, we have used for all compounds reported in this study the experimental results as provided by the KOWWIN database. In the case of the test set compounds, we used both KOWWIN and the experimental values reported by the authors in their original studies.

#### METHODS

Molecular weights and atom-type E-state indices were calculated using a program developed in-house (checked against the Molconn-Z software, Hall Associated Consulting, Quincy, MA) with structure input for each analyzed compound using the SMILES line notation code. A set of 36 E-state indices, each of which contained values for at least 7 molecules, was selected for the analysis. In addition to the 36 atom-type E-state indices (basic set) we took into account the binding environment of amino, hydroxyl, and carbonyl groups. This extension of E-state indices was suggested in our preliminary study.<sup>21</sup> Compounds containing an amino group, i.e., parameters for SsNH<sub>2</sub>, SssN, and SsssN, were divided into three groups according to the neighborhood atom(s). These groups were aliphatic, aromatic attached, or the others (i.e., amides, ureas, etc.), and E-state values for these separate groups were used as contributors. In the same way the compounds containing a hydroxyl group, i.e., parameter for SsOH, were divided into alcohols, phenols, and carboxylic acids. The compounds containing a carbonyl group, i.e., parameter for SdO, were divided into ketones, carboxylic acids, esters, amides and ureas, nitro and nitroso compounds, sulfones, and sulfoxides. The scheme of calculation of atom-type E-state indices<sup>18</sup> is shown in Table 1. For example in anilides an amino group is attached with an aromatic ring and a carbonyl group. In this case parameter SssNH is considered as an aromatic attached. In the same way with amino groups in 1-phenyl-3-methyl urea, amino group 1 is considered as an aromatic attached and amino group 3 as an amide type, respectively. A total of 50 parameters were calculated for the extended set of atom type E-state indices and were used as contributors in the regression and neural network models.

Two separate multilinear regression analyses were performed. The first regression model calculated log *P* values using the basic set of 36 atom-type E-state indices, i.e., without an extension of the indices. The second regression was made with the extended set of 50 atom-type E-state indices. The multilinear regressions were performed with the SPSS software (v.8.0, SPSS Inc., Chicago, IL) running on a Pentium PC.

The ANNs employed in this study were fully connected feed-forward back-propagation networks with one hidden layer and bias neurons. The training of neural networks was carried out using the SuperSAB algorithm<sup>26</sup> programmed in ANSI C++. The logistic  $f(x) = 1/(1 + e^{-x})$  activation function was used both for hidden and output nodes. Both sets of E-state indices were used for neural network training.

**Table 1.** E-State Indices Calculated for Benzocaine along with the Atom-Type E-State Indices<sup>a</sup>


| atom ID | atom-type          | symbol    | E-state index |
|---------|--------------------|-----------|---------------|
| 1       | aCHa               | aaCH      | 1.646         |
| 2       | aCHa               | aaCH      | 1.673         |
| 3       | aCa                | aasC      | 0.642         |
| 4       | -NH <sub>2</sub>   | sNH2(ar)  | 5.449         |
| 5       | aCHa               | aaCH      | 1.673         |
| 6       | aCHa               | aaCH      | 1.646         |
| 7       | aCa                | aasC      | 0.533         |
| 8       | =C<                | dssC      | -0.308        |
| 9       | =O                 | dO(ester) | 11.093        |
| 10      | -O-                | ssO       | 4.788         |
| 11      | -CH <sub>2</sub> - | ssCH2     | 0.392         |
| 12      | -CH <sub>3</sub>   | sCH3      | 1.773         |

| atom-type | atom-type     | atom-type | Atom-type     |
|-----------|---------------|-----------|---------------|
|           | E-state value |           | E-state value |
| SsCH3     | 1.773         | SaasC     | 0.533         |
| SssCH2    | 0.392         | SsNH2(ar) | 5.449         |
| SaaCH     | 6.638         | SdO(es)   | 11.093        |
| SdssC     | -0.308        | SssO      | 4.788         |

<sup>a</sup> According to Kier and Hall.<sup>18</sup>

The pruning methods described elsewhere<sup>27</sup> were used to optimize the number of input parameters for neural network training with the extended set of indices. The number of neurons in the hidden layer was optimized as indicated in Results and Discussion. One single output node was used to code log *P* values.

Avoidance of overfitting/overtraining has been shown to be an important factor for the improvement of generalization ability and correct selection of variables in neural network studies.<sup>28,29</sup> The early stopping over ensemble technique was used in the current study to accomplish this problem. A detailed description of this approach can be found elsewhere.<sup>26,28</sup> In brief, each analyzed artificial neural network ensemble was composed of *M* = 200 networks. The values calculated for analyzed cases were averaged over all *M* neural networks and their means were used for computing statistical coefficients with targets. We used a subdivision of the initial training set into two equal learning/validation subsets. The first set was used to train the neural network, while the second one was used to monitor the training process as measured by root-mean-square error. An early stopping point determined as a best fit of a network to the validation set was used to stop the neural network learning. It has been shown that a neural network trained up to an early stopping point provides better prediction ability than a network trained to the error minimum for the learning set.<sup>26</sup> Thus, statistical parameters calculated at the early stopping point were used.

The calculations using ANNs were performed on the HP Workstation Cluster at the Swiss Center for Scientific Computing and at the computer server of the University of Lausanne. The developed software was programmed with a user-friendly Java interface and is available on-line at the World Wide Web address <http://www.lnh.unil.ch>.

An analysis of the generalization ability of the models was achieved using the leave-one-out method for the training set and actual prediction of the test sets. Predictive *q*<sup>2</sup> in LOO cross-validation was calculated for the training set as

$$\begin{cases} q^2 = \frac{SSY - \text{PRESS}}{SSY} \\ SSY = \sum (Y_{\text{exp}} - \bar{Y})^2, \quad \bar{Y} = \frac{1}{n} \sum Y_{\text{exp}} \\ \text{PRESS} = \sum (Y_{\text{calc}} - Y_{\text{exp}})^2 \end{cases} \quad (1)$$

where *Y*<sub>calc</sub> and *Y*<sub>exp</sub> are LOO values and experimental log *P* values, respectively, and summation is over the number of molecules, *n*, in the set. Other coefficients used to assess the predictive ability of the methods included the root-mean-square error (RMS)

$$\text{RMS} = \left[ \frac{\sum (Y_{\text{calc}} - Y_{\text{exp}})^2}{n} \right]^{1/2} \quad (2)$$

and the square of the correlation coefficient *r*<sup>2</sup>.

## RESULTS AND DISCUSSION

**Multilinear Regression.** Stepwise and backward methods were employed in the regression analysis of the basic and the extended sets of parameters. The following regression equations

$$\log P = \sum (a_i S_i) - 0.765 \quad (3)$$

$$n = 1754, \quad r^2 = 0.82, \quad \text{RMS} = 0.62, \quad F = 233, \\ q^2 = 0.81, \quad \text{RMS}_{\text{LOO}} = 0.64$$

$$\log P = \sum (a_i S_i) - 0.646 \quad (4)$$

$$n = 1754, \quad r^2 = 0.87, \quad \text{RMS} = 0.53, \quad F = 259, \\ q^2 = 0.85, \quad \text{RMS}_{\text{LOO}} = 0.55$$

were calculated for the basic and the extended sets of parameters, respectively. In both equations *n* is the number of compounds used in the fit, *F* is the overall *F*-statistic for the addition of each successive term, and *a<sub>i</sub>* and *S<sub>i</sub>* are the regression coefficients and the corresponding E-state indices. The regression coefficients in eqs 3 and 4 are indicated in Table 2 with the 95% confidence limits of the significant parameters. The correlation matrix of the 34 parameters in eq 3 and 41 parameters in eq 4 showed that all pairwise correlations had an *r*<sup>2</sup> < 0.50. The prediction of the MLR model, RMS<sub>LOO</sub> = 0.55, is only 0.02 units higher than for the fitting model. Such a small increase indicates robustness of the model.

The extension of the E-state indices significantly improved the statistical parameters of the regression equation for the training set. All extended parameters except three, SssNH2-(ar), SsOH(zwit), and SdO(amid), were found to be significant in an MLR regression equation. The extension of the parametrization made possible a clear interpretation of the factors influencing the hydrophobicity. In all cases the aliphatic amino group decreased hydrophobicity more compared to the case if it was in an aromatic environment. In the same way phenolic hydroxyls are more hydrophobic than

**Table 2.** Atom-Type E-State Indices<sup>a</sup> Used in Multilinear Regression and Neural Network Models

| no. <sup>b</sup> | symbol <sup>c</sup> | atom type <sup>d</sup> | remark           | basic <sup>e</sup>          | extend <sup>e</sup> | ANN            |
|------------------|---------------------|------------------------|------------------|-----------------------------|---------------------|----------------|
| 1                | MW                  |                        | molecular weight | -0.015 (0.001) <sup>f</sup> | -0.004 (0.001)      | × <sup>g</sup> |
| 2                | SsCH3               | -CH <sub>3</sub>       |                  | 0.432 (0.013)               | 0.343 (0.011)       | ×              |
| 3                | SdCH2               | =CH <sub>2</sub>       |                  | 0.318 (0.028)               | 0.255 (0.024)       | ×              |
| 4                | SssCH2              | -CH <sub>2</sub> -     |                  | 0.482 (0.016)               | 0.408 (0.012)       | ×              |
| 5                | StCH                | ≡CH                    |                  | 0.123 (0.048)               |                     |                |
| 6                | SdsCH               | =CH-                   |                  | 0.280 (0.019)               | 0.224 (0.016)       | ×              |
| 7                | SaaCH               | aCHa                   |                  | 0.299 (0.009)               | 0.217 (0.006)       | ×              |
| 8                | SsssCH              | >CH-                   |                  | 0.425 (0.032)               | 0.167 (0.031)       |                |
| 9                | SddC                | =C=                    |                  | 0.333 (0.088)               | 0.329 (0.079)       | ×              |
| 10               | StsC                | -C≡                    |                  | 0.125 (0.069)               | 0.155 (0.052)       |                |
| 11               | SdssC               | =C<                    |                  | -0.141 (0.038)              | -0.289 (0.042)      | ×              |
| 12               | SaasC               | saCa                   |                  | 0.200 (0.020)               | 0.116 (0.017)       |                |
| 13               | SaaaC               | aaCa                   |                  | 0.334 (0.028)               | 0.247 (0.022)       | ×              |
| 14               | SssssC              | >C<                    |                  | 0.245 (0.029)               | 0.134 (0.026)       |                |
| 15 <sup>#</sup>  | SsNH2               | -NH <sub>2</sub>       |                  | -0.028 (0.009)              |                     |                |
| 16 <sup>*</sup>  | SsNH2(al)           | -NH <sub>2</sub>       | aliphatic        |                             | -0.158 (0.015)      | ×              |
| 17 <sup>*</sup>  | SsNH2(ar)           |                        | aromatic         |                             |                     | ×              |
| 18 <sup>*</sup>  | SsNH2(oth)          |                        | the others       |                             | -0.132 (0.013)      | ×              |
| 19 <sup>#</sup>  | SssNH               | -NH-                   |                  | -0.106 (0.016)              |                     |                |
| 20 <sup>*</sup>  | SssNH(al)           | -NH-                   | aliphatic        |                             | -0.305 (0.029)      | ×              |
| 21 <sup>*</sup>  | SssNH(ar)           |                        | aromatic         |                             | 0.053 (0.028)       | ×              |
| 22 <sup>*</sup>  | SssNH(oth)          |                        | the others       |                             | -0.181 (0.020)      | ×              |
| 23               | StN                 | ≡N                     |                  | 0.128 (0.018)               | 0.048 (0.014)       | ×              |
| 24               | SdsN                | =N-                    |                  | 0.139 (0.016)               | 0.045 (0.014)       | ×              |
| 25               | SaaN                | aNa                    |                  |                             | -0.083 (0.008)      | ×              |
| 26 <sup>#</sup>  | SsssN               | >N-                    |                  | -0.357 (0.024)              |                     |                |
| 27 <sup>*</sup>  | SsssN(al)           | >N-                    | aliphatic        |                             | -0.559 (0.028)      | ×              |
| 28 <sup>*</sup>  | SsssN(ar)           |                        | aromatic         |                             | -0.215 (0.043)      | ×              |
| 29 <sup>*</sup>  | SsssN(oth)          |                        | the others       |                             | -0.413 (0.041)      | ×              |
| 30               | SddsN               | -N<                    |                  | -0.670 (0.099)              |                     |                |
| 31               | SaasN               | aaNs                   |                  |                             |                     | ×              |
| 32               | SaaNH               | aaNH                   |                  |                             |                     |                |
| 33               | SaadN               | aaNd                   |                  | -1.691 (0.230)              | -1.644 (0.204)      | ×              |
| 34 <sup>#</sup>  | SsOH                | -OH                    |                  | 0.074 (0.006)               |                     |                |
| 35 <sup>*</sup>  | SsOH(alc)           | -OH                    | alcohols         |                             | -0.011 (0.006)      | ×              |
| 36 <sup>*</sup>  | SsOH(phen)          |                        | phenols          |                             | 0.052 (0.005)       | ×              |
| 37 <sup>*</sup>  | SsOH(acid)          |                        | acids            |                             | 0.179 (0.020)       | ×              |
| 38 <sup>*</sup>  | SsOH(zwit)          |                        | amino acids      |                             |                     | ×              |
| 39 <sup>#</sup>  | SdO                 | =O                     |                  | 0.066 (0.005)               |                     |                |
| 40 <sup>*</sup>  | SdO(keto)           | =O                     | ketones          |                             | 0.019 (0.005)       | ×              |
| 41 <sup>*</sup>  | SdO(acid)           |                        | acids            |                             | -0.115 (0.017)      | ×              |
| 42 <sup>*</sup>  | SdO(ester)          |                        | esters           |                             | 0.030 (0.007)       | ×              |
| 43 <sup>*</sup>  | SdO(amid)           |                        | amides           |                             | 0.014 (0.006)       | ×              |
| 44 <sup>*</sup>  | SdO(nitro)          |                        | nitros           |                             | 0.046 (0.003)       | ×              |
| 45 <sup>*</sup>  | SdO(sulfo)          |                        | sulfones         |                             | -0.075 (0.018)      | ×              |
| 46               | SssO                | -O-                    |                  | 0.067 (0.008)               |                     | ×              |
| 47               | SaaO                | aOa                    |                  | 0.082 (0.035)               |                     | ×              |
| 48               | SsF                 | -F                     |                  | 0.137 (0.006)               | 0.086 (0.005)       | ×              |
| 49               | SsSH                | -SH                    |                  | 0.309 (0.055)               | 0.228 (0.046)       |                |
| 50               | SdS                 | =S                     |                  | 0.314 (0.033)               | 0.227 (0.030)       | ×              |
| 51               | SssS                | -S-                    |                  | 0.561 (0.055)               | 0.357 (0.046)       |                |
| 52               | SaaS                | aSa                    |                  | 0.821 (0.133)               | 0.611 (0.113)       |                |
| 53               | SddssS              | >>S<                   |                  |                             | -0.469 (0.108)      |                |
| 54               | SsCl                | -Cl                    |                  | 0.307 (0.010)               | 0.202 (0.007)       | ×              |
| 55               | SsBr                | -Br                    |                  | 0.755 (0.034)               | 0.451 (0.023)       | ×              |
| 56               | SsI                 | -I                     |                  | 1.616 (0.085)               | 0.939 (0.063)       |                |

<sup>a</sup> According to Hall and Kier. <sup>b</sup> Key: “#” marks parameters that were extended, and key “\*” marks the extended parameters. <sup>c</sup> S-states for the sum of the E-state values for a certain atom type or group. <sup>d</sup> The formula of the atom type or group; the bond types between heavy atoms are s = single (-), d = double (=), and a = aromatic (a). <sup>e</sup> The contribution values in regression equations. <sup>f</sup> Standard errors of the regression coefficients are indicated in parentheses. <sup>g</sup> Significant parameters in ANNs after pruning.

the hydroxyl groups in alcohols. These findings are in agreement with the conclusions by Klopman et al.<sup>8</sup> and Meylan and Howard<sup>9</sup> and, of course, have correspondence to the fragment constants used in the CLOGP method.<sup>5,6</sup> The calculated log *P* values for the test sets are reported in Tables 3–5. The prediction ability of the MLR equation for the third test set, RMS = 0.56, was similar to that estimated for the training set, RMS<sub>LOO</sub> = 0.55. The prediction results for the first two sets, RMS = 0.80 and RMS = 0.88, were much

lower. However, there were four outliers (ADO, GUO, dDAPR, and guanine, 8Aza) in test set 1 and one outlier (verapamil) in test set 2. When these molecules were excluded, multilinear regression was able to predict the log *P* values for 111 out of 116 compounds in the test set with a RMS = 0.59, which is in agreement with the result for the training set.

It was possible that there are some nonlinear dependencies between the electrotopological-state indices and the partition



**Table 3.** Comparison of the Predictive Ability of Multilinear Regression and Neural Network Models Using the Extended Set of Parameters

| model <sup>a</sup> | param     | no. <sup>b</sup> | training set |                    | test set 1 |      | test set 2 |      | test set 3 |      |
|--------------------|-----------|------------------|--------------|--------------------|------------|------|------------|------|------------|------|
|                    |           |                  | $q^2$        | RMS <sub>LOO</sub> | $r^2$      | RMS  | $r^2$      | RMS  | $r^2$      | RMS  |
| MLR                | regressed | 41               | 0.87         | 0.55               | 0.51       | 0.80 | 0.75       | 0.88 | 0.86       | 0.56 |
| ANN1               | all       | 51               | 0.90         | 0.46               | 0.75       | 0.39 | 0.93       | 0.47 | 0.92       | 0.42 |
| ANN2               | regressed | 41               | 0.89         | 0.48               | 0.72       | 0.42 | 0.89       | 0.54 | 0.90       | 0.46 |
| ANN3               | pruned    | 39               | 0.90         | 0.46               | 0.73       | 0.41 | 0.94       | 0.42 | 0.92       | 0.41 |

<sup>a</sup> MLR, multilinear regression; ANN, artificial neural networks. <sup>b</sup> Number of input parameters in the model.

coefficients. Thus, an application of nonlinear methods of data analysis could provide a better modeling of the data. Back-propagation artificial neural networks were used to detect the presence of nonlinear dependencies in the analyzed data set as described in the next section.

**Artificial Neural Network Analysis.** The number of iterations required for training the neural network was determined using the extended set of parameters. It was found that about 3000 iterations were required to determine a minimum error of neural network. Thus, the training was terminated by limiting the network run to 15 000 epochs (total number of epochs) or after 3000 epochs (local number of epochs) following the last improvement of RMS error at the early stopping point. The number of hidden neurons required for data analysis was investigated by the examination of neural networks with 1, 2, 3, 4, 5, 7, and 10 hidden neurons. It was found that the RMS<sub>LOO</sub> error decreased (i.e., RMS<sub>LOO</sub> = 0.736 ± 0.035, 0.502 ± 0.002, 0.475 ± 0.002, 0.466 ± 0.003, 0.456 ± 0.004) when the number of neurons in the hidden layer was changed from 1 to 5. However, further increase in the number of hidden neurons from 5 to 7 and 10 did not influence significantly the prediction ability of the neural networks (i.e., RMS<sub>LOO</sub> = 0.455 ± 0.003, 0.453 ± 0.003). Thus, the number of neurons in the hidden layer was selected to be 5. This number also provided a reasonable speed in neural network calculations, which of course is very important when training ensembles of neural networks.

The prediction ability of ANNs calculated for the training set using the fully extended set of parameters was improved compared to the MLR results giving LOO values of  $q^2$  = 0.898 ± 0.002, RMS<sub>LOO</sub> = 0.456 ± 0.004. The neural networks performed significantly better in prediction ability for the analyzed test sets (Table 3). An important fact was that both LOO and prediction results calculated for the test data sets are characterized by approximately the same error. This suggests that the models are robust.

Pruning algorithms, reported elsewhere,<sup>27</sup> were used to optimize the number of input parameters and to select only the most significant descriptors for ANN regression. These algorithms operated in a manner similar to stepwise multiple regression analysis and excluded on each step one parameter that was estimated to be nonsignificant. The set of parameters that calculated a minimum LOO error for the training set was selected as the optimized set. This was composed of the 39 parameters listed in Table 2 and also included the molecular weight of the compounds. The prediction ability calculated with this set was, in fact, similar to that calculated with the full set (Table 3). However, use of a smaller number of parameters provides a more robust model. The distribution of the errors calculated for the training set by ANNs and plots of the experimental versus calculated log *P* values for the compounds in all three test sets are shown in Figures 1

and 2, respectively. These figures and the statistics quoted above reveal the high quality of this neural network approach for modeling the lipophilicity of organic compounds.

The calculated results for test sets 1 and 2 were similar or superior to those calculated with other methods applied to these sets of organic compounds. In fact, ANNs provided the lowest standard deviation for test sets 1 and 2 compared to similar results calculated with such methods as KLOGP, ALOGP, CLOGP, and BLOGP and similar to that of KOWWIN. The calculated results for the second test set with the current method were significantly improved compared to our results reported previously<sup>21</sup> (RMS = 0.58,  $r^2$  = 0.87). The prediction results for the data set of 50 molecules calculated with the current approach were superior to those reported in the original work of Schaper and Samitier<sup>24</sup> (see Table 5). The AUTOLOGP program calculated slightly better results for the current set of molecules RMS = 0.37,  $r^2$  = 0.92, while the distribution of residual errors calculated with our program was very similar to that of AUTOLOGP (Table 5). It should be noted that Devillers and co-workers<sup>25</sup> used a much larger training data set of 7200 molecules, while our training set contained only 1754. KOWWIN provided the smallest prediction error for this set. However, it is possible, that some of these molecules were used in the training set to select the atom/fragment contribution coefficients used in this method.

The analysis of residuals showed that there were several compounds with a large estimation error in the training set. The compounds with a residual > 1.40, that is, equivalent to three times the RMS, are listed in Table 6. Some of these compounds, namely, piroxicam, timolol, methotrexate, phenformin, and minoxidil, were also outliers in our previous calculations. However, the number of outliers was reduced from 24 to 19 in the present study, despite the number of molecules in the training set being increased from 345 to 1754. The reason for large estimation errors can be explained by the fact that these compound types were represented by only a few members in the training set and were from the structural classes of N-epoxides (minoxidil and librium), sulfones and sulfoxides (piroxicam and meloxicam), pyridine carboxylic acid and amides, amino pteridine analogues (methotrexate), imides (glutarimide), morpholine analogues (timolol), guanidines (phenformin),  $\beta$ -lactones (pilocarpine), and amines with a highly basic amino group (labetalol and scopolamine). Thus, an extension of the training set with compounds from these structural classes might be expected to improve the present model for estimation of log *P* values.

The overall performance of this study suggests a high prediction ability of the developed approach. It is very important to note that we did not provide the fitting but, instead, leave-one-out results for the training set. The fitting results, i.e., results when neural networks were used to predict

**Table 4.** Experimental and Estimated log *P* Values for the Test Sets<sup>a</sup>

| (A) Test Set 1        |                 |                             |                          |                   |                    |                        |                     |                    |                     |
|-----------------------|-----------------|-----------------------------|--------------------------|-------------------|--------------------|------------------------|---------------------|--------------------|---------------------|
| no.                   | compd           | log <i>P</i> <sub>exp</sub> | MLR <sup>b</sup>         | ANN3 <sup>b</sup> | KLOGP <sup>c</sup> | ALOGP <sup>c</sup>     | CLOGP <sup>c</sup>  | BLOGP <sup>c</sup> | KOWWIN <sup>c</sup> |
| 1                     | ADO             | -1.05 (-1.23)               | -2.96                    | -1.4              | -2.13              | -1.24                  | -2.91               | 0.18               | -1.38               |
| 2                     | dADO            | -0.55 (-0.54)               | -2.06                    | -0.92             | -1.29              | -0.63                  | -2.53               | 0.36               | -0.71               |
| 3                     | ddADO           | -0.25 (-0.21)               | -0.93                    | -0.24             | -0.46              | -0.19                  | -1.12               | 0.41               | -0.65               |
| 4                     | ddeADO          | -0.5 (-0.35)                | -0.92                    | -0.21             | -0.42              | 0.04                   | -1.60               | 0.38               | -0.86               |
| 5                     | FddADO          | 0.08                        | -1.11                    | -0.14             | -0.23              | -0.02                  | -1.27               | 0.52               | -0.78               |
| 6                     | GUO             | -1.9 (-1.89)                | -3.56                    | -2.25             | -3.09              | -1.63                  | -3.92               | -0.21              | -1.71               |
| 7                     | dGUO            | -1.3                        | -2.79                    | -1.82             | -2.26              | -1.01                  | -3.34               | -0.01              | -1.04               |
| 8                     | ddGUO           | -1.01 (-1)                  | -1.09                    | -0.32             | -1.43              | -0.38                  | -1.92               | 0.18               | -0.88               |
| 9                     | ddeGUO          | -1.21                       | -1.23                    | -1.2              | -1.39              | -0.35                  | -2.41               | 0.12               | -0.72               |
| 10                    | dDAPR           | -0.52                       | -2.51                    | -1.1              | -1.38              | -0.91                  | -2.59               | -0.10              | -0.77               |
| 11                    | ddDAPR          | -0.46                       | -1.4                     | -0.51             | -0.55              | -0.47                  | 0.07                | 0.07               | -0.71               |
| 12                    | FddDAPR         | 0.05                        | -1.56                    | -0.4              | -0.32              | -0.30                  | -1.33               | 0.25               | -0.84               |
| 13                    | URD             | -1.98 (-1.71)               | -2.38                    | -1.91             | -2.28              | -1.59                  | -2.56               | 0.09               | -1.86               |
| 14                    | dURD            | -1.51 (-1.50)               | -1.75                    | -1.47             | -1.44              | -0.98                  | -2.09               | 0.46               | -1.19               |
| 15                    | ddURD           | -1 (-0.88)                  | -0.83                    | -0.74             | -0.61              | -0.54                  | -0.68               | 0.67               | -1.12               |
| 16                    | ddeURD          | -1.07                       | -0.69                    | -0.72             | -0.57              | -0.32                  | -1.16               | 0.65               | -1.34               |
| 17                    | FddURD          | -0.49 (-0.48)               | -0.76                    | -0.52             | -0.38              | -0.37                  | -0.83               | 0.91               | -1.26               |
| 18                    | dTHD            | -0.93 (-1.17)               | -1.67                    | -1.25             | -1.03              | -0.82                  | -1.59               | 0.82               | -0.64               |
| 19                    | ddTHD           | -0.6 (-0.63)                | -0.76                    | -0.49             | -0.20              | -0.39                  | -0.18               | 1.04               | -0.58               |
| 20                    | ddeTHD          | -0.72 (-0.81)               | -0.61                    | -0.45             | -0.16              | -0.16                  | -0.66               | 1.00               | -0.79               |
| 21                    | FddTHD          | -0.28 (-0.27)               | -0.65                    | -0.26             | 0.03               | -0.22                  | -0.33               | 1.27               | -0.71               |
| 22                    | CYD             | -2.51                       | -2.8                     | -2.03             | -2.86              | -1.42                  | -3.11               | 0.11               | -2.46               |
| 23                    | dCYD            | -1.77                       | -2.15                    | -1.58             | -2.03              | -0.81                  | -2.55               | 0.48               | -1.79               |
| 24                    | ddCYD           | -1.3                        | -1.21                    | -0.87             | -1.20              | -0.37                  | -1.13               | 0.62               | -1.72               |
| 25                    | ddeCYD          | -1.55 (-1.42)               | -1.09                    | -0.86             | -1.16              | -0.14                  | -1.62               | 0.57               | -1.94               |
| 26                    | FddCYD          | -0.89 (-0.91)               | -1.16                    | -0.68             | -0.97              | -0.20                  | -1.29               | 0.78               | -1.86               |
| 27                    | F6ddP           | 0                           | -0.22                    | 0.05              | 0.28               | 0.49                   | -0.91               | 0.25               | -0.54               |
| 28                    | F62AddP         | -0.05                       | -0.64                    | -0.02             | -0.11              | 0.21                   | -0.97               | 0.38               | -0.6                |
| 29                    | Br6ddP          | 0.35                        | 0.47                     | 0.67              | 0.81               | 1.08                   | -0.36               | 0.80               | 0.15                |
| 30                    | Br62AddP        | 0.34 (0.33)                 | -0.02                    | 0.5               | 0.42               | 0.80                   | -0.42               | 0.77               | 0.09                |
| 31                    | Cl6ddP          | 0.24 (0.23)                 | 0.18                     | 0.45              | 0.50               | 0.78                   | -0.39               | 0.49               | -0.09               |
| 32                    | Cl62AddP        | 0.21                        | -0.28                    | 0.33              | 0.11               | 0.50                   | -0.45               | 0.55               | -0.15               |
| 33                    | I6ddP           | 0.53 (0.52)                 | 0.87                     | 0.83              | 1.00               | 1.08                   | 0.07                | 1.01               | 0.43                |
| 34                    | I62AddP         | 0.52                        | 0.34                     | 0.69              | 0.61               | 0.80                   | 0.01                | 1.00               | 0.37                |
| 35                    | ddI             | -1.24                       | -0.65                    | -0.29             | -1.55              | -0.95                  | -1.76               | -0.34              | -0.47               |
| 36                    | uracil          | -1.07                       | -0.72                    | -1.08             | -0.86              | -0.72                  | -1.06               | -0.59              | -0.87               |
| 37                    | adenine, 8Aza   | -0.96                       | -1.76                    | -0.79             | -0.11              | -0.67                  | -0.06               | -0.03              | -0.79               |
| 38                    | guanine, 8Aza   | -0.71                       | -2.39                    | -2.1              | -1.14              | -1.06                  | -1.00               | -0.84              | -1.12               |
| 39                    | cytosine        | -1.73                       | -1.13                    | -1.29             | -1.45              | -0.55                  | -1.85               | -0.77              | -1.47               |
| 40                    | adenine         | -0.09                       | -1.09                    | -0.38             | 0.23               | -0.36                  | -0.43               | 0.04               | -0.73               |
| 41                    | thioguanine     | -0.07                       | -0.57                    | -0.19             | -0.37              | -0.02                  | -1.82               | 0.31               | -0.19               |
| 42                    | 9-propyladenine | 0.74                        | 0.35                     | 0.74              | 0.50               | 0.69                   | 0.44                | 1.33               | 0.8                 |
| 43                    | uracil, 6Aza    | -0.59                       | -0.8                     | -1.24             | -0.91              | -0.49                  | -0.59               | -1.01              | -1.27               |
| 44                    | quanine         | -0.91                       | -1.84                    | -1.66             | -1.17              | -0.75                  | -1.26               | -0.89              | -1.05               |
| 45                    | thymine         | -0.62                       | -0.65                    | -0.8              | -0.45              | -0.56                  | -0.56               | -0.13              | -0.32               |
| 46                    | hypoxanthine    | -1.11                       | -0.77                    | -0.79             | -0.89              | -0.88                  | -1.26               | -0.87              | -0.55               |
| 47                    | purine          | -0.37                       | -0.61                    | -0.23             | 0.53               | -0.28                  | -0.29               | -0.14              | -0.82               |
| <i>r</i> <sup>2</sup> |                 |                             | 0.51 (0.52) <sup>d</sup> | 0.73 (0.74)       | 0.78               | 0.70 (0.71)            | 0.50 (0.51)         | 0.16               | 0.72 (0.70)         |
| RMS                   |                 |                             | 0.80 (0.79)              | 0.41 (0.40)       | 0.46 (0.45)        | 0.52 (0.51)            | 0.92                | 1.19               | 0.42 (0.43)         |
| (B) Test Set 2        |                 |                             |                          |                   |                    |                        |                     |                    |                     |
| no.                   | compd           | log <i>P</i> <sub>exp</sub> | MLR <sup>b</sup>         | ANN3 <sup>b</sup> | XLOGP <sup>e</sup> | Moriguchi <sup>e</sup> | Rekker <sup>e</sup> | CLOGP <sup>e</sup> | KOWWIN <sup>c</sup> |
| 1                     | chlorothiazide  | -0.24                       | 0.09                     | 0.11              | -0.58              | -0.36                  | -0.68               | -1.24              | -0.23               |
| 2                     | cimetidine      | 0.4                         | 1.33                     | 0.91              | 0.20               | 0.82                   | 0.63                | 0.21               | 0.57                |
| 3                     | procainamide    | 0.88                        | 1.47                     | 1.45              | 1.27               | 1.72                   | 1.11                | 1.11               | 0.97                |
| 4                     | trimethoprim    | 0.91                        | 1.1                      | 1.12              | 0.72               | 1.26                   | -0.07               | 0.66               | 0.73                |
| 5                     | chloramphenicol | 1.14                        | 1.38                     | 1.52              | 1.46               | 1.23                   | 0.32                | 0.69               | 0.92                |
| 6                     | phenobarbital   | 1.47                        | 1.29                     | 1.44              | 1.77               | 0.78                   | 1.23                | 1.37               | 1.33                |
| 7                     | atropine        | 1.83                        | 1.92                     | 1.77              | 2.29               | 2.21                   | 1.88                | 1.32               | 1.91                |
| 8                     | lidocaine       | 2.44 (2.26)                 | 2.35                     | 2.44              | 2.47               | 2.52                   | 2.3                 | 1.36               | 1.66                |
| 9                     | phenytoin       | 2.47                        | 2.14                     | 2.34              | 2.23               | 1.8                    | 2.76                | 2.09               | 2.16                |
| 10                    | diltiazem       | 2.7                         | 3.69                     | 2.88              | 3.14               | 2.67                   | 4.53                | 3.55               | 2.79                |
| 11                    | propranolol     | 3.48 (2.98)                 | 2.77                     | 2.66              | 2.98               | 2.53                   | 3.46                | 2.75               | 2.6                 |
| 12                    | diazepam        | 2.82 (2.99)                 | 3.43                     | 3.03              | 2.98               | 3.36                   | 3.18                | 3.32               | 2.7                 |
| 13                    | diphenhydramine | 3.27                        | 4.13                     | 4.1               | 3.74               | 3.26                   | 3.41                | 2.93               | 3.11                |
| 14                    | tetracaine      | 3.51 (3.73)                 | 3.59                     | 3.67              | 2.73               | 2.64                   | 3.55                | 3.65               | 3.02                |
| 15                    | verapamil       | 3.79                        | 6.68                     | 4.3               | 5.29               | 3.23                   | 6.15                | 3.53               | 4.8                 |
| 16                    | haloperidol     | 3.36 (4.3)                  | 4.1                      | 3.55              | 4.35               | 4.01                   | 3.57                | 3.52               | 4.2                 |
| 17                    | imipramine      | 4.8                         | 4.84                     | 4.96              | 4.26               | 3.88                   | 4.43                | 4.41               | 5.01                |
| 18                    | chlorpromazine  | 5.41 (5.19)                 | 5.01                     | 4.98              | 4.91               | 3.77                   | 5.10                | 5.20               | 5.2                 |
| 19                    | flufenamic acid | 5.25                        | 4.05                     | 4.7               | 4.45               | 3.86                   | 5.81                | 5.58               | 5.15                |
| <i>r</i> <sup>2</sup> |                 |                             | 0.75 (0.77) <sup>f</sup> | 0.94 (0.95)       | 0.87 (0.89)        | 0.82 (0.87)            | 0.85 (0.84)         | 0.93 (0.94)        | 0.95 (0.93)         |
| RMS                   |                 |                             | 0.88 (0.84)              | 0.42 (0.40)       | 0.59 (0.54)        | 0.73 (0.68)            | 0.79 (0.80)         | 0.51 (0.49)        | 0.36 (0.44)         |

Table 4 (Continued)

| (C) Test Set 3        |  |                             |                  |                    |                   |                    |                       |       |         |       |
|-----------------------|--|-----------------------------|------------------|--------------------|-------------------|--------------------|-----------------------|-------|---------|-------|
| no.                   | compd  | log <i>P</i> <sub>exp</sub> | MLR <sup>b</sup> | resid <sup>b</sup> | ANN3 <sup>b</sup> | resid <sup>b</sup> | Schaper <sup>24</sup> | resid | KOW WIN | resid |
| 1                     | <i>n</i> -pentane                            | 3.39                        | 2.28             | 1.11               | 2.66              | 0.73               | 2.64                  | 0.75  | 2.8     | 0.59  |
| 2                     | 2,2-dimethylbutane                           | 3.82                        | 2.7              | 1.12               | 3.24              | 0.58               | 3.53                  | 0.29  | 3.18    | 0.64  |
| 3                     | 1-methylcyclopentane                         | 3.37                        | 2.45             | 0.92               | 2.73              | 0.64               | 2.47                  | 0.90  | 3.1     | 0.27  |
| 4                     | cycloocta-1,5-diene                          | 3.16                        | 3                | 0.16               | 3.23              | -0.07              | 3.79                  | -0.63 | 3.73    | -0.57 |
| 5                     | 1-butene                                     | 2.4                         | 1.61             | 0.79               | 1.86              | 0.54               | 1.86                  | 0.54  | 2.17    | 0.23  |
| 6                     | 1-pentyne                                    | 1.98                        | 1.03             | 0.95               | 1.25              | 0.73               | 1.96                  | 0.02  | 2.03    | -0.05 |
| 7                     | isoamyl alcohol                              | 1.16                        | 1                | 0.16               | 0.99              | 0.17               | 1.85                  | -0.69 | 1.26    | -0.1  |
| 8                     | cyclohexanol                                 | 1.23                        | 1.3              | -0.07              | 1.18              | 0.05               | 0.40                  | 0.83  | 1.64    | -0.41 |
| 9                     | 2,3-butanediol                               | -0.92                       | -0.29            | -0.63              | -0.49             | -0.43              | 0.07                  | -0.99 | -0.36   | -0.56 |
| 10                    | 2-cyclohexenone                              | 0.61                        | 1.12             | -0.51              | 1                 | -0.39              | 0.53                  | 0.09  | 1.2     | -0.59 |
| 11                    | 2-ethylpropenal                              | 1.24                        | 1.06             | 0.18               | 1.03              | 0.21               | 0.79                  | 0.46  | 1.23    | 0.01  |
| 12                    | 2-hexanone                                   | 1.38                        | 1.58             | -0.2               | 1.57              | -0.19              | 0.78                  | 0.60  | 1.24    | 0.14  |
| 13                    | glutaric acid                                | -0.29                       | 0.09             | -0.38              | -0.06             | -0.23              | -0.16                 | -0.13 | -0.26   | -0.03 |
| 14                    | 3-mercaptopropionic acid                     | 0.43                        | 0.58             | -0.15              | 0.27              | 0.16               | -0.05                 | 0.48  | 0.52    | -0.09 |
| 15                    | γ-butyrolactone                              | -0.64                       | 0.24             | -0.88              | 0.13              | -0.77              | 0.37                  | -1.01 | -0.31   | -0.33 |
| 16                    | hydroxyethyl acrylate                        | -0.21                       | 0.29             | -0.5               | -0.19             | -0.02              | 0.25                  | -0.46 | -0.25   | 0.04  |
| 17                    | 2,2,2-trifluoroethylamine                    | 0.24                        | 0.01             | 0.23               | 0.3               | -0.06              | 0.06                  | 0.18  | 0.27    | -0.03 |
| 18                    | <i>n</i> -propylamine                        | 0.48                        | -0.16            | 0.64               | 0.13              | 0.35               | 0.11                  | 0.37  | 0.34    | 0.14  |
| 19                    | methylethylamine                             | 0.15                        | 0.06             | 0.09               | 0.06              | 0.09               | 0.00                  | 0.15  | 0.32    | -0.17 |
| 20                    | 1,4-diethylenediamine                        | -1.5                        | -1.06            | -0.44              | -1.4              | -0.1               | -0.54                 | -0.96 | -0.8    | -0.7  |
| 21                    | <i>n,o</i> -dimethylcarbamate                | -0.06                       | 0.01             | -0.07              | -0.24             | 0.18               | -0.42                 | 0.36  | -0.04   | -0.02 |
| 22                    | 4-ethylsemicarbazide <sup>c</sup>            | -1.62                       | -0.69            | -0.93              | -0.89             | -0.73              | -1.22                 | 0.48  | -1.69   | 0.07  |
| 23                    | 2-(acetyl amino)- <i>n</i> -methylethanamide | -1.56                       | -0.55            | -1.01              | -0.69             | -0.87              | 0.14                  | -1.70 | -1.52   | -0.04 |
| 24                    | methylene bromide                            | 1.7 (1.88)                  | 1.84             | -0.14              | 1.94              | -0.24              | 1.27                  | 0.43  | 1.52    | 0.18  |
| 25                    | vinyl bromide                                | 1.57                        | 1.46             | 0.11               | 1.61              | -0.04              | 1.18                  | 0.39  | 1.52    | 0.05  |
| 26                    | 1,3-dichloropropane                          | 2                           | 1.99             | 0.01               | 2.31              | -0.31              | 3.36                  | -1.36 | 2.32    | -0.32 |
| 27                    | methyl acrylonitrile                         | 0.68                        | 1.03             | -0.35              | 0.76              | -0.08              | 0.37                  | 0.31  | 0.76    | -0.08 |
| 28                    | trifluoroacetamide                           | 0.12                        | 1.32             | -1.2               | 0.78              | -0.66              | -1.11                 | 1.23  | -0.11   | 0.23  |
| 29                    | 1-nitrobutane                                | 1.47                        | 1.23             | 0.24               | 1.41              | 0.06               | 0.05                  | 1.42  | 1.21    | 0.26  |
| 30                    | <i>o</i> -xylene                             | 3.12                        | 2.56             | 0.56               | 2.91              | 0.21               | 2.69                  | 0.43  | 3.09    | 0.03  |
| 31                    | <i>p</i> -xylene                             | 3.15                        | 2.56             | 0.59               | 2.91              | 0.24               | 2.84                  | 0.31  | 3.09    | 0.06  |
| 32                    | phloroglucinol                               | 0.16                        | 0.95             | -0.79              | 0.12              | 0.04               | 0.60                  | -0.44 | 0.55    | -0.39 |
| 33                    | <i>m</i> -toluic acid                        | 2.37                        | 1.73             | 0.64               | 1.93              | 0.44               | 1.80                  | 0.57  | 2.42    | -0.05 |
| 34                    | 2,4-dichlorophenol                           | 3.06                        | 2.53             | 0.53               | 3.09              | -0.03              | 3.12                  | -0.06 | 2.8     | 0.26  |
| 35                    | chlorohydroquinone                           | 1.4                         | 1.73             | -0.33              | 1.65              | -0.25              | 2.35                  | -0.95 | 1.68    | -0.28 |
| 36                    | 4-chloronitrobenzene                         | 2.39 <sup>g</sup>           | 2.13             | 0.26               | 2.26              | 0.13               | 2.06                  | 0.33  | 1.59    | 0.8   |
| 37                    | 5-methylfurfural                             | 0.67                        | 0.79             | -0.12              | 1.02              | -0.35              | 0.27                  | 0.40  | 1.38    | -0.71 |
| 38                    | 2-cyanomethylfuran                           | 0.85                        | 1.01             | -0.16              | 1.03              | -0.18              | 0.87                  | -0.02 | 0.93    | -0.08 |
| 39                    | 3-( <i>b</i> -nitrovinyl)furan               | 1.41                        | 1.32             | 0.09               | 1.62              | -0.21              | 0.53                  | 0.88  | 1.09    | 0.32  |
| 40                    | methyl pyrrole-2-carboxylate                 | 1.17                        | 0.94             | 0.23               | 1.12              | 0.05               | 0.54                  | 0.63  | 0.71    | 0.46  |
| 41                    | 2-cyanothiophene                             | 1.27                        | 1.86             | -0.59              | 1.07              | 0.2                | 1.74                  | -0.47 | 1.36    | -0.09 |
| 42                    | 2,3-dibromothiophene                         | 3.53                        | 3.65             | -0.12              | 2.98              | 0.55               | 3.13                  | 0.40  | 3.59    | -0.06 |
| 43                    | 3-picoline                                   | 1.2                         | 1.17             | 0.03               | 1.19              | 0.01               | 1.15                  | 0.05  | 1.35    | -0.15 |
| 44                    | 2-pyridinemethanol                           | 0.06                        | 0.19             | -0.13              | -0.09             | 0.15               | 0.28                  | -0.22 | -0.11   | 0.17  |
| 45                    | 3-pyridinemethanol                           | -0.02                       | 0.19             | -0.21              | -0.09             | 0.07               | 0.26                  | -0.28 | -0.11   | 0.09  |
| 46                    | 3-hydroxypyridine                            | 0.48                        | 0.54             | -0.06              | 0.21              | 0.27               | 0.16                  | 0.32  | 0.32    | 0.16  |
| 47                    | 4-cyanopyridine                              | 0.46                        | 0.87             | -0.41              | 0.51              | -0.05              | 1.09                  | -0.63 | 0.35    | 0.11  |
| 48                    | 3-hydroxy-2-pyridinecarboxamide              | 0.65                        | -0.36            | 1.01               | -0.67             | 1.32               | 0.12                  | 0.53  | 0.49    | 0.16  |
| 49                    | 2,6-dichloropyridine                         | 2.15                        | 1.93             | 0.22               | 2.15              | 0                  | 2.39                  | -0.24 | 2.09    | 0.06  |
| 50                    | 3,5-dichloropyridine                         | 2.56                        | 1.89             | 0.67               | 2.11              | 0.45               | 2.71                  | -0.15 | 2.09    | 0.47  |
| <i>r</i> <sup>2</sup> |  |                             | 0.86             |                    | 0.92              |                    | 0.78                  |       | 0.95    |       |
| RMS                   |  |                             | 0.56             |                    | 0.41              |                    | 0.65                  |       | 0.32    |       |

<sup>a</sup> ALOGP, atomic constant approach of Viswanadhan and co-workers;<sup>22</sup> BLOGP, the molecular orbital approach of Bodor and co-workers;<sup>12</sup> CLOGP, the fragmental constant approach;<sup>6</sup> KLOGP, extended group contribution approach of Klopman et al.;<sup>8</sup> XLOGP, atom-additive method of Wang et al.<sup>10</sup> KOWWIN, atom/fragment contribution method developed by Meylan and Howard.<sup>9</sup> The experimental log *P* values are given according to the KOWWIN database. The experimental values indicated in parentheses correspond to those reported in the previous studies, if they are different from KOWWIN values. <sup>b</sup> The results calculated with the extended set of parameters. The neural network results are reported for the pruned set of parameters. <sup>c</sup> The experimental (in parentheses) and calculated results are from ref 8. <sup>d</sup> The statistical coefficients calculated using the experimental values of compounds from ref 8 are indicated in parentheses. <sup>e</sup> The calculated values are cited from ref 10. <sup>f</sup> The coefficients calculated using the experimental values of the test set molecules from ref 10 are indicated in parentheses. <sup>g</sup> Compound 22 is 4-ethylsemicarbazide, H<sub>2</sub>N-NH-CO-NH-Et, instead of H<sub>2</sub>N-CO-NH-NH-Et; log *P* of compound 36 is 2.39 instead of 2.41, (Schaper, K.-J. Personal communication).

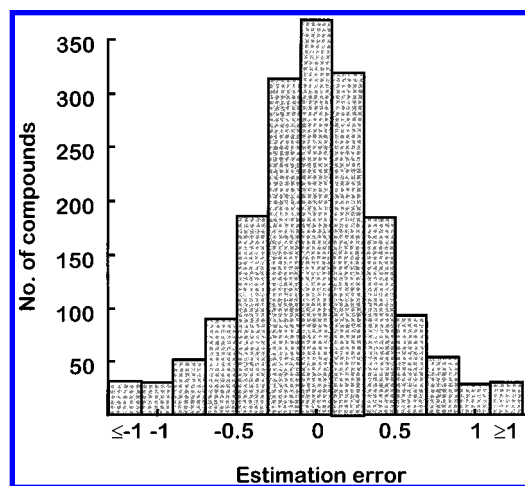
molecules from their training sets, were much better. For example, the ANN3 model calculated  $r^2 = 0.946 \pm 0.001$ ,  $s = 0.337 \pm 0.003$  which is much better than the LOO results ( $r^2 = 0.900 \pm 0.002$ ,  $\text{RMS}_{\text{LOO}} = 0.459 \pm 0.005$ ). These results could be further improved using a larger number of hidden neurons in the neural network. Indeed, the results calculated using 7 and 10 hidden neurons were  $r^2 = 0.947 \pm 0.001$ ,  $s = 0.332 \pm 0.003$  and  $r^2 = 0.948 \pm 0.001$ ,  $s =$

$0.319 \pm 0.003$ , respectively. However, as mentioned above, an increase in the number of hidden neurons did not provide any improvement of the LOO results, and the neural network with 5 hidden neurons was selected as a final model. It is clear that the fitting ANNs results alone are of no practical interest, since they do not estimate correctly the predictive performance of this method. It should be noted that, in many studies, the authors report only the fitting results for the

**Table 5.** Distribution of Residuals (Absolute Values in percent) between the Experimental and Calculated Values for the Test Set of Schaper and Samitier<sup>24</sup>

| range     | BNN <sup>a</sup> | AUTOLOGP <sup>b</sup> | ANN3    |
|-----------|------------------|-----------------------|---------|
| 0.00–0.20 | 18 (18)          | 54 (54)               | 52 (52) |
| 0.21–0.40 | 28 (46)          | 20 (74)               | 20 (72) |
| 0.41–0.60 | 20 (66)          | 16 (90)               | 12 (84) |
| 0.61–0.80 | 12 (78)          | 4 (94)                | 12 (96) |
| 0.81–1.00 | 12 (90)          | 4 (98)                | 2 (98)  |
| 1.01–1.20 | 2 (92)           | 2 (100)               | 0 (98)  |
| 1.21–1.40 | 4 (96)           |                       | 2 (100) |
| 1.41–1.60 | 2 (98)           |                       |         |
| 1.61–1.80 | 2 (100)          |                       |         |

<sup>a</sup> BNN, back-propagation neural network from the model of Schaper and Samitier.<sup>24</sup> <sup>b</sup> AUTOLOGP, the results reported by Devillers et al.<sup>25</sup>

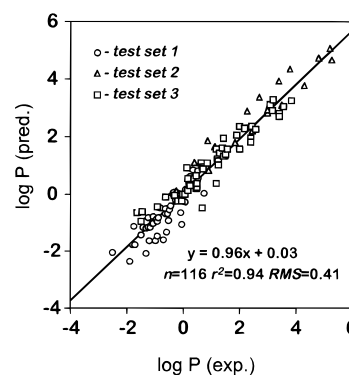
**Figure 1.** Distribution of the leave-one-out errors for the training set ( $n = 1754$ ,  $r^2 = 0.90$ ,  $RMS = 0.46$ ) calculated by the artificial neural network with pruned parameters (ANN3).

training set<sup>20,24,25,30</sup> since the standard implementation of neural network systems does not allow the rapid calculation of LOO results. Such fitting results, however, have a danger of overfitting/overtraining, a problem that has been discussed elsewhere.<sup>26</sup> The fast calculation of leave-one-out results is possible using an original approach reported previously.<sup>26–28</sup>

**Table 6.** Compounds with Large Prediction Errors in the Training Set

| no   | compound <sup>a</sup>               | ANN3                  |                        |       | MLR                    |       |
|------|-------------------------------------|-----------------------|------------------------|-------|------------------------|-------|
|      |                                     | $\log P_{\text{exp}}$ | $\log P_{\text{calc}}$ | resid | $\log P_{\text{calc}}$ | resid |
| 228  | methotrexate                        | -1.85                 | 0.27                   | 2.12  | -0.31                  | -1.54 |
| 1356 | fusaric acid                        | -1.29                 | 2.4                    | 3.69  | 2.28                   | -3.57 |
| 648  | 3-hydroxy-2-pyridinecarboxylic acid | -1.27                 | 0.45                   | 1.72  | 0.51                   | -1.78 |
| 240  | phenformin                          | -0.83                 | 1.11                   | 1.94  | 0.62                   | -1.45 |
| 180  | pilocarpine                         | 0.12                  | 1.64                   | 1.52  | 1.92                   | -1.8  |
| 689  | 6-aminonicotinamide                 | 0.7                   | -0.99                  | -1.69 | -0.76                  | 1.46  |
| 1645 | scopolamine                         | 0.98                  | 3.13                   | 2.15  | 2.34                   | -1.36 |
| 247  | minoxidil                           | 1.24                  | -0.78                  | -2.02 | -0.26                  | 1.5   |
| 1581 | timolol                             | 1.83                  | -0.23                  | -2.06 | 0.29                   | 1.54  |
| 116  | glutarimide                         | 1.9                   | -0.54                  | -2.44 | -0.38                  | 2.28  |
| 80   | librium                             | 2.44                  | 4.07                   | 1.63  | 3.09                   | -0.65 |
| 165  | piroxicam                           | 3.06                  | 1.65                   | -1.41 | 0.78                   | 2.28  |
| 1655 | labetalol                           | 3.09                  | 1.56                   | -1.53 | 2.17                   | 0.92  |
| 1622 | 3,6-dimethyl-4-aminosalicylic acid  | 3.38                  | 1.47                   | -1.91 | 1.4                    | 1.98  |
| 158  | meloxicam                           | 3.43                  | 1.63                   | -1.8  | 1.14                   | 2.29  |
| 222  | disulfiram                          | 3.88                  | 2.3                    | -1.58 | 4.00                   | -0.12 |
| 98   | domperidone                         | 3.9                   | 2.44                   | -1.46 | 1.48                   | 2.42  |
| 1546 | 1-(3,4-dichlorophenyl)-3-phenylurea | 4.7                   | 3.23                   | -1.47 | 3.29                   | 1.41  |
| 213  | loratadine                          | 5.2                   | 3.66                   | -1.54 | 3.72                   | 1.48  |

<sup>a</sup> Compounds with an absolute value of residuals > 1.4 log units for artificial neural networks are shown.

**Figure 2.** Correlation of calculated  $\log P_{\text{pred}}$  versus observed values  $\log P_{\text{exp}}$  for the three analyzed test sets calculated by the artificial neural network with pruned parameters (ANN3).

This method, polynomial neural networks<sup>31</sup> as well as an on-line version of the ANN  $\log P$  program, are available at <http://www.lnh.unil.ch>, Virtual Laboratory link.

## CONCLUSIONS

The atom-type electrotopological-state indices represent valuable tools in QSAR since they can be computed for any arbitrary molecule and the calculations are made in a clearly described and reproducible way. In addition, these parameters are weakly redundant, as shown in the low pairwise interparameter correlations in this large training set. This explains the growing number of successful applications of these indices in different fields of chemistry.<sup>17–21,29,32–36</sup> There are disadvantages, however, in the use of topological parameters in terms of their chemical interpretability, among other things, as discussed in a recent review.<sup>37</sup> It has been shown that the importance of these indices could be significantly increased if they are combined with artificial neural networks.<sup>19–21,29</sup> The results reported here, the prediction of  $\log P$  values for a large database of compounds, provides new evidence about the importance of atom-type electrotopological-state descriptors and artificial neural networks in QSAR studies.



## ACKNOWLEDGMENT

This study was partially supported by the Technology Development Center in Finland (TEKES) and INTAS-Ukraine Grant 95-0060. The authors thank Vsevolod Tanchuk for the development of the program for the calculation of the E-state indices. We are also grateful to Tamara Kasheva and Dmitry Shakhnin for their help in the verification of SMILES codes.

## REFERENCES AND NOTES

- Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Active Profiles Using Substructural Analysis and Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165–179.
- Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Underinger, T. L. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screen.* **1996**, *1*, 65–73.
- Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
- Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55–68.
- Hansch, L.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
- Leo, A. Calculating log *P*<sub>oct</sub> from Structures. *Chem. Rev.* **1993**, *93*, 1281–1306.
- Rekker, R. E. *Hydrophobic Fragment Constant*; Elsevier: New York, 1977.
- Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated log *P* Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752–781.
- Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol–Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83–92.
- Wang, R.; Fu, Y.; Lai, L. A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615–621.
- Klopman, G.; Iroff, L. Calculation of Partition Coefficients by the Charge Density Methodol. *J. Comput. Chem.* **1981**, *2*, 157.
- Bodor, N.; Huang, M.-J. An Extended Version of a Novel Method for Estimation of Partition Coefficients. *J. Pharm. Sci.* **1992**, *81*, 272–281.
- Haeberlin, M.; Brinck, T. Prediction of Water-Octanol Partition Coefficients Using Theoretical Descriptors Derived from the Molecular Surface Area and the Electrostatic Potential. *J. Chem. Soc., Perkin Trans. 2* **1997**, 289–294.
- Bodor, N.; Buchwald, P. Molecular Size Based Approach to Estimate Partition Properties for Organic Solutes. *J. Phys. Chem.* **1997**, *101*, 3404–3412.
- Breindl, A.; Beck, N.; Clark, T.; Glen, R. C. Prediction of the *n*-Octanol/Water Partition Coefficient, log*P*, Using a Combination of Semiempirical MO-Calculations and a Neural Network. *J. Mol. Model.* **1997**, *3*, 142–155.
- Buchwald, P.; Bodor, N. Octanol–Water Partitioning: Searching for Predictive Models. *Curr. Med. Chem.* **1998**, *5*, 353–380.
- Kier, L. B.; Hall, L. H. An Electrotopological State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801–807.
- Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.
- Hall, L. H.; Story, C. T. Boiling Point and Critical Temperature of a Heterogeneous Data Set: QSAR with Atom Type Electrotopological State Indices Using Artificial Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1004–1014.
- Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450–456.
- Huuskonen, J. J.; Villa, A. E. P.; Tetko, I. V. Prediction of Partition Coefficient Based on Atom-Type Electrotopological State Indices. *J. Pharm. Sci.* **1999**, *88*, 229–233.
- Viswanadhan, V. N.; Reddy, M. R.; Bacquet, R. J.; Erion, M. D. Assessment of Methods Used for Predicting Lipophilicity: Application to Nucleosides and Nucleoside Bases. *J. Comput. Chem.* **1993**, *14*, 1019–1025.
- Moriguchi, I.; Hirono, S.; Nakagome, I.; Hirano, H. Comparison of Reliability of log*P* Values for Drugs Calculated by Several Methods. *Chem. Pharm. Bull.* **1994**, *42* (2), 976–978.
- Schaper, K.-J.; Samitier, M. L. R. Calculation of Octanol/Water Partition Coefficients (log*P*) using Artificial Neural Networks and Connection Matrixes. *Quant. Struct.–Act. Relat.* **1997**, *16*, 224–230.
- Devillers, J.; Domine, D.; Guillon, C. Autocorrelation Modeling of Lipophilicity with a Back-Propagation Neural Network. *Eur. J. Med. Chem.* **1998**, *33*, 659–664.
- Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural Network Studies. 1. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826–833.
- Tetko, I. V.; Villa, A. E. P.; Livingstone, D. J. Neural Network Studies. 2. Variable Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794–803.
- Tetko, I. V.; Villa, A. E. P. Efficient Partition of Learning Data Sets for Neural Network Training. *Neural Networks* **1997**, *10*, 1361–1374.
- Huuskonen, J.; Salo, M.; Taskinen, J. Neural Network Modeling for Estimation of the Aqueous Solubility of Structurally Related Drugs. *J. Pharm. Sci.* **1997**, *86*, 450–454.
- Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.*, in press.
- Tetko, I. V.; Aksenova, T. I.; Volkovich, V. V.; Kasheva, T. N.; Filipov, D. V.; Villa, A. E. P.; Welsh, W. J.; Livingstone, D. J. Polynomial Neural Network for Linear and Nonlinear Model Selection in Quantitative-Structure Activity Relationship Studies on WWW. *SAR QSAR Environ. Res.*, in press.
- Abou-Shaaban, R. R.; al-Khamees, H. A.; Abou-Auda, H. S.; Simonelli, A. P. Atom Level Electrotopological State Indices in QSAR: Designing and Testing Antithyroid Agents. *Pharm. Res.* **1996**, *13*, 129–136.
- Buolamwini, J. K.; Raghavan, K.; Fesen, M. R.; Pommier, Y.; Kohn, K. W.; Weinstein, J. N. Application of the Electrotopological State Index to QSAR Analysis of Flavone Derivatives as HIV-1 Integrase Inhibitors. *Pharm. Res.* **1996**, *13*, 1892–1895.
- de Gregorio, C.; Kier, L. B.; Hall, L. H. QSAR Modeling with the Electrotopological State Indices: Corticosteroids. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 557–561.
- Liang, H. R.; Vuorela, H.; Vuorela, P.; Riekkola, M. L.; Hiltunen, R. Prediction of migration behavior of flavonoids in capillary zone electrophoresis by means of topological indices. *J. Chromatogr.* **1998**, *798*, 233–242.
- Gough, J. D.; Hall, L. H. Modeling Antileukemic Activity of Carboquinones with Electrotopological State and Chi Indices. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 356–361.
- Livingstone, D. J. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195–209.

CI9904261