Modeling of Gibbs Energy of Formation of Organic Compounds by Linear and Nonlinear Methods[†]

Aixia Yan*

State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering,
P.O. Box 53, Beijing University of Chemical Technology, 15 BeiSanHuan East Road,
Beijing 100029, P. R. China

Received January 8, 2006

Two quantitative models for the prediction of the Gibbs energy of formation ($\Delta G_f°$) of 177 organic compounds were developed. These molecules contain elements such as H, C, N, O, F, S, Cl, and Br, with the molecular weight in the range of 16.04−202.25. The molecules were represented by six selected 2D-structure descriptors. At first, the complex relationship between $\Delta G_f°$ and the six selected input descriptors was depicted by a two-dimensional Kohonen's self-organizing neural network (KohNN) map; on the basis of the KohNN map, the whole data set was split into a training set consisting of 130 compounds and a test set (or a validation set and a test set) including 47 compounds. Then, $\Delta G_f°$ was predicted using a multilinear regression (MLR) analysis and a back-propagation (BPG) neural network. For 177 organic compounds, root-mean-square deviations of 17.8 and 15.4 kcal mol^{-1} were achieved by MLR and the BPG neural network, respectively.

INTRODUCTION

The Gibbs energy of formation ($\Delta G_f°$) is a fundamentally important physicochemical property of organic compounds, which plays an important role in all kinds of chemical and biochemical reactions. The $\Delta G_f°$ of organic compounds is extensively applied in chemistry and chemical engineering,[1] environmental sciences,[2] molecular biology,[3] and so forth.

Several models have been developed for the calculation of the $\Delta G_f°$ of organic compounds with computational chemistry tools.[4−7] The Gibbs energy of formation of alkanes has been evaluated using quantitative structure−property relationship (QSPR) models with the topological descriptors based on the molecular graph[4,5] or on graphs of atomic orbitals.[6] These models based on the alkanes give good prediction results.[4−6] More recently, on the basis of a relatively more diverse set of organic compounds (the molecules include other elements such as N, O, F, S, Cl, and Br), Wang et al.[7] used a "DFT-NEURON" method to estimate some properties of organic compounds, including $\Delta G_f°$. In their method, $\Delta G_f°$ was first calculated by density-functional theory (DFT) methods and then followed by neural-networks-based and multiple-linear-regression- (MLR-)based correction approaches.[7] For three models [B3LYP/6-31G(d), B3LYP/6-311+G(3df,2p), and B3LYP/6-311+G-(d,p)] computed with DFT methods, the calculated Gibbs energies of formation for 180 organic molecules are 12.5, 13.8, and 22.3 kcal mol^{-1}, respectively; after the neural networks correction (multiple linear regression correction), the corresponding RMS deviations are reduced to 4.7 (5.4), 3.2 (3.5), and 3.0 (3.2) kcal mol^{-1}, respectively.[7]

This work aims at building suitable models for predicting the $\Delta G_f°$ of organic compound on the basis of their structures using QSPR methods. The descriptors representing the molecules are derived from the constitution of a molecule and incorporate important physicochemical effects. They can be obtained quickly by calculation from their molecular structures. First, a list of 54 descriptors for 177 molecules was computed; using statistical analysis, from these 54 descriptors, the six descriptors were then selected. Second, the relationship between $\Delta G_f°$ and the six selected descriptors was examined using a Kohonen's self-organizing neural network (KohNN); on the basis of the KohNN map distribution, the data set was split into a training set consisting of 130 compounds and a test set (or a validation set and a test set) including 47 compounds. Afterward, two quantitative models were developed by a MLR analysis and a back-propagation (BPG) neural network.[8]

DATA SETS

The experimental standard $\Delta G_f°$ values (kcal mol^{-1}) at 298 K of 180 small- or medium-sized organic molecules were derived from the former reference[7] and were originally taken from the Chemical Properties Handbook.[9]

However, in this work, three compounds (ethyl nitrate, propylnitrate, and isopropylnitrate) were excluded as they could not be converted by the β version (initial test version) of the ADRIANA.Code program.[10] This left a set of 177 molecules. These molecules contain elements such as H, C, N, O, F, S, Cl, and Br, with the molecular weight in the range of 16.04−202.25. $\Delta G_f°$ has a minimum value of −212.3 kcal mol^{-1}, a maximum value of 71.0 kcal mol^{-1}, and a mean value of −13.3 kcal mol^{-1}.

METHODS

In this work, the following programs and software packages were applied. The CACTVS system was used for

[†] Dedicated to Professor Johann Gasteiger.

* Author phone: +86-10-64421335; fax: +86-10-64416428; e-mail: yanax@mail.buct.edu.cn Or aixia_yan@yahoo.com.

10.1021/ci0600105 CCC: $33.50
Published on Web 06/03/2006

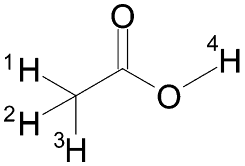

Figure 1. Example for autocorrelation coefficients calculation.

structure management, editing and comparing, and data extracting.[11] The ADRIANA.Code program was applied for the calculation of physicochemical properties of organic molecules.[10] SONNIA was utilized for building Kohonen's self-organizing neural networks.[12] SNNS was used for constructing the BPG neural network.[13]

Structure Representation and Descriptors Selection. All of the descriptors were calculated with the program ADRIANA.Code (Algorithms for the Encoding of Molecular Structures).[10] ADRIANA.Code comprises a series of methods for the generation of 3D structures, the calculation of physicochemical descriptors on the basis of empirical models for the influences of atoms in molecules, and a set of mathematical transformation techniques.[10] In particular, partial atomic charges are calculated by the PEOE partial equalization method[14,15] and its extension to π systems.[16] This method also provides electronegativity values for atoms in molecules that provide quantitative values for the inductive effect.[17] A damping method based on the construction of atoms in the hybridization state provides a quantitative measure of the polarizability effect.[18]

At first, five 2D whole-molecule descriptors were computed by using the program ADRIANA.Code. These include molecular weight, topological polar surface area (TPSA), mean molecular polarizability (MMP), the number of hydrogen-bonding acceptors, and the number of hydrogen-bonding donors. In ADRIANA.Code, the calculation of TPSA is implemented according to the approach developed by Ertl et al.[19] The calculation of MMP is based on a damping model, which uses a parametrization of the contribution of each atom in a compound.[20]

Simultaneously, using ADRIANA.CODE, the 2D molecular autocorrelation vectors[21] were calculated on the basis of the following seven atomic properties: σ charge (SigChg),[14,15] π charge (PiChg),[16] total charges (TotChg), σ electronegativity (SigEN), π electronegativity (PiEN), lone-pair electronegativity (LpEN), and atomic polarizability (Apolariz).[22]

In the autocorrelation vectors calculation, the hydrogen atoms were included. Topological autocorrelation vectors for each one of the above seven physicochemical atomic properties were calculated for each molecule by using the following equation:

$$A(d) = \sum_{ij} p_i p_j \tag{1}$$

A(d) is the topological autocorrelation coefficient referring to atom pairs *i* and *j*, which are separated by *d* bonds. p_i is an atomic property, for example, the σ charge on atom *i*. Thus, for each compound, a series of coefficients for different topological distances *d*, a so-called autocorrelation vector, is obtained; seven distances from a distance of $d = 0-6$ were considered. For example, acetic acid (Figure 1) has three pairs of atoms that are separated by four bonds: H_1-H_4, H_2-H_4, and H_3-H_4. Thus, the corresponding autocorrelation for the topological distance four computes to

$$A(4) = p_1p_4 + p_2p_4 + p_3p_4 \tag{2}$$

In statistical analyses, it was found that the seven 2D autocorrelation coefficients are highly correlated for the properties σ electronegativity, π electronegativity, and atomic polarizability; the first component of the 2D autocorrelation coefficients has the highest standard deviation for the properties σ charge, π charge, total charge, and lone-pair electronegativity. Thus, the first component ($d = 0$) of the autocorrelation coefficients for each property was selected for the following analysis. This value corresponds to the sum of the squares of each atomic property for a molecule. In addition, the third components ($d = 2$) of the 2D autocorrelation coefficients for the properties of σ charge and total charge were also selected as they are significantly correlated to $\Delta G_f°$ [the second components ($d = 1$) of the 2D autocorrelation coefficients for the properties of σ charge and total charge were removed because they are highly correlated to those of the corresponding first components]. All together, nine autocorrelation coefficients were selected for the following analysis.

The nine selected autocorrelation coefficients of the seven physicochemical properties were put together with the other five 2D descriptors. A pairwise correlation analysis was then done. A descriptor was eliminated if the correlation coefficient was equal to or higher than 0.85. A further statistical analysis was carried out, and several nonsignificant (from the linear model standpoint) descriptors were also omitted. This left six descriptors, as shown in Table 1.

The six selected descriptors include TPSA, MMP, the first and third components of the 2D autocorrelation coefficients for σ charge (Acorr_Sigchg_1 and Acorr_Sigchg_3), and the first components of the 2D autocorrelation coefficients for σ electronegativity (Acorr_SigEN_1) and lone-pair electronegativity (Acorr_LpEN_1). The intercorrelations between the six descriptors and $\Delta G_f°$ are given in Table 1.

Training/Test Set Selection by Kohonen's Self-Organizing Neural Network. KohNN has the special property of effectively creating a spatially organized internal representation of various features of input signals and their abstractions.[8] The perception of the similarity of objects is an essential feature. In a self-organizing neural network, the neurons are arranged in a two-dimensional array to generate a two-dimensional feature map such that similarity in the data is preserved. In other words, if two input data vectors are similar, they will be mapped into the same neuron or closely together in the two-dimensional map.

A Kohonen self-organizing neural network was applied to split the data set into a training set and a test set (or a validation set and a test set for the BPG model). The division based on a KohNN map is superior to random selection. The advantage of such a procedure was shown in previous work.[23−26] This method for splitting a data set into training and test sets ensures that both sets cover the information space as well as possible. Because the test set was not used during the training of the MLR or BPG model, it can still be considered as an external data set.

Table 1. Intercorrelations between the Six Descriptors and ΔG_f°[a]

descriptors	TPSA	MMP	Acorr_Sigchg_1	Acorr_Sigchg_3	Acorr_SigEN_1	Acorr_LpEN_1
TPSA	1					
MMP	−0.035	1				
Acorr_Sigchg_1	0.731	−0.040	1			
Acorr_Sigchg_3	0.308	0.044	0.332	1		
Acorr_SigEN_1	0.326	0.835	0.321	0.201	1	
Acorr_LpEN_1	0.224	−0.214	0.450	0.257	0.006	1
ΔG_f°	−0.284	0.231	−0.733	−0.366	−0.147	−0.616

[a] TPSA: topological polar surface area. MMP: mean molecular polarizability. Acorr_Sigchg_1: $\sum q_\sigma^2$ (q_σ: σ charge). Acorr_Sigchg_3: The third components of 2D autocorrelation coefficients for σ charge (where $d = 2$). Acorr_SigEN_1: $\sum \chi_\sigma^2$ (χ_σ: σ electronegativity). Acorr_LpEN_1: $\sum \chi_{LP}^2$ (χ_{LP}: lone-pair electronegativity). ΔG_f°: Gibbs energy of formation.

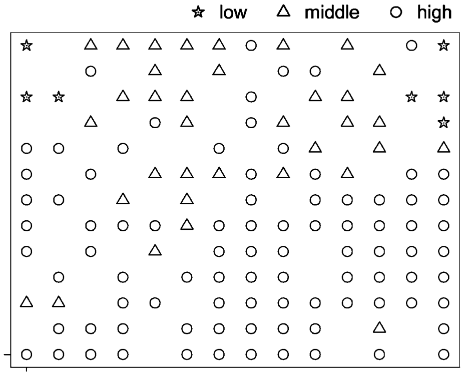

Figure 2. Rectangular KohNN map for 177 compounds obtained by using the six selected 2D input descriptors. "High" means compounds with a high ΔG_f° in the range of 71.0∼−23.42 kcal mol^{-1}; "middle" means compounds with a ΔG_f° in the range of −23.43∼−117.87 kcal mol^{-1}, and "low" refers to compounds with a ΔG_f° in the range of −117.88∼−212.30 kcal mol^{-1}.

RESULTS AND DISCUSSION

A rectangular KohNN with 14 × 13 neurons is utilized with the six descriptors used as input vectors. The initial learning spans are 7 and 6.5, with an initial learning rate of 0.7 and a rate factor of 0.95. The initial weights are randomly initialized, and training was performed for a period of 1600 epochs in an unsupervised manner. A map was formed according to the ranges of ΔG_f° of the most frequently occupied neuron. From Figure 2, one can see that compounds with a different range of ΔG_f° are projected into different areas.

In the Kohonen map, 130 of a total of 182 neurons are occupied. Afterward, one object of each neuron was taken for the training set; the other objects represented the test set (or validation and test sets). Thus, the 177 compounds were divided into a training set of 130 compounds and a test set of 47 compounds (or a validation set of 20 compounds and a test set of 27 compounds for the BPG model) after the KohNN classification.

Build a Model by MLR Analysis. A multilinear regression analysis was performed using six descriptors as input variables. The 130 compounds in the training set were used to build a model, and the 47 compounds were used for the prediction of ΔG_f°.

The following equation has been obtained:

$$\Delta G_f^{o} = \sum(c_i D_i) + 13.036 \tag{3}$$

In the equation, D_i is a descriptor and c_i is its corresponding regression coefficient in a MLR model. The corresponding regression coefficients are shown in Table 2. For the training set, $r = 0.93$, $s = 19.1$, MAE = 14.6 kcal mol^{-1}, $n = 130$, and $F = 135.5$, and for the test set, $r = 0.96$, $s = 14.1$, MAE = 11.7 kcal mol^{-1}, and $n = 47$ (r is the correlation coefficient; s is the standard deviation; MAE is the mean absolute error, which equals the mean value of the absolute errors, and n is the number of compounds). The root-mean-square (RMS) deviation of the calculated Gibbs energy of formation for the 177 organic compounds is 17.8 kcal mol^{-1}. The results are shown in Figure 3 and Table 3.

Build a Model by BPG. The SNNS program[13] was used for the back-propagation neural network. A standard back-propagation network was applied to estimate ΔG_f°. An input layer with six neurons, an output layer with one neuron representing ΔG_f°, and a hidden layer of several neurons were used. All layers were completely connected. The initial weights were randomly initialized between −0.1 and +0.1. Each input and output value was scaled between 0 and 1. The net was trained following the "standard back-propagation" algorithm as implemented in SNNS, employing a learning rate of 0.2.

The 130 compounds in the training set were used to train a neural network; 20 compounds in the validation set were used to determine the performance of a neural network, and the 27 compounds in the test set were used for the prediction of ΔG_f°. In the process, the architecture of the neural network was optimized. The number of hidden-layer neurons was varied from one to nine. The optimized neural network architecture was 6−2−1. The best number of training epochs was selected by the early stopping method[27] in order to avoid overtraining, and the training stops in the minimum of the validation set error; the best number of training epochs was 6000. For the training set, $r = 0.95$, $s = 14.5$, MAE = 11.8 kcal mol^{-1}, and $n = 130$; for the validation set, $r = 0.97$, $s = 11.2$, MAE = 11.9 kcal mol^{-1}, and $n = 20$, and for the test set, $r = 0.97$, $s = 8.7$, MAE = 11.2 kcal mol^{-1}, and $n = 27$. The RMS deviations of the calculated Gibbs energy of formation for the 177 organic compounds is 15.4 kcal mol^{-1}. The results are shown in Figure 4 and Table 3.

Compared with the Sets Split by Random Selection. In the above work, the models built on the sets were split into training, validation, and test sets on the basis of a Kohonen self-organizing map. To see how the model's performance depends on the different ways of splitting the data set, the MLR and BPG models on the other sets were split by random selection and then models were built.

In the random selection sets, the data set was randomly split into a training set of 130 compounds and a test set of 47 compounds for the MLR model (a validation set of 20 compounds and a test set of 27 compounds for the BPG

Table 2. Six Selected Descriptors and Their Corresponding Regression Coefficients in the Multilinear Regression Model

descriptors	coefficients	*t* score	partial *F*	*p* value
topological polar surface area (TPSA)	1.999	13.82	191.1	<0.0001
mean molecular polarizability (MMP)	10.934	10.67	113.8	<0.0001
Acorr_Sigchg_1= $\sum q_\sigma^2$ (q_σ: σ charge)	−295.306	−15.38	236.5	<0.0001
Acorr_Sigchg_3[a]	−113.289	−4.07	16.6	<0.0001
Acorr_SigEN_1 = $\sum \chi_\sigma^2$ (χ_σ: σ electronegativity)	−0.120	−9.78	95.7	<0.0001
Acorr_LpEN_1= $\sum \chi_{LP}^2$ (χ_{LP}: lone-pair electronegativity)	−0.267	−4.36	19.0	<0.0001

[a] Acorr_Sigchg_3: The third components of 2D autocorrelation coefficients for σ charge (where $d = 2$).

Table 3. Prediction Performances of the Obtained Models in This Work by Multilinear Regression (MLR) and Back-Propagation Neural Networks (NN)[a,b]

	training set				validation set				test set				
model	*n*	*r*	*s*	MAE	*n*	*r*	*s*	MAE	*n*	*r*	*s*	MAE	RMS
MLR1	130	0.93	19.1	14.6					47	0.96	14.1	11.7	17.8
MLR2	130	0.94	19.3	14.3					47	0.95	13.8	12.2	17.7
NN1	130	0.95	14.5	11.8	20	0.97	11.2	11.9	27	0.97	8.7	11.2	15.4
NN2	130	0.96	14.7	11.5	20	0.96	12.8	10.9	27	0.96	13.0	10.4	15.0

[a] The sets used for the models MLR1 and NN1 were selected on the basis of Kohonen's Self-organizing neural network; the sets used for the models MLR2 and NN2 were split by one random selection. [b] *n*: Number of compounds. *r*: Correlation coefficient. *s*: Standard deviation. MAE: Mean absolute error. RMS: Root-mean-square deviation for the whole model.

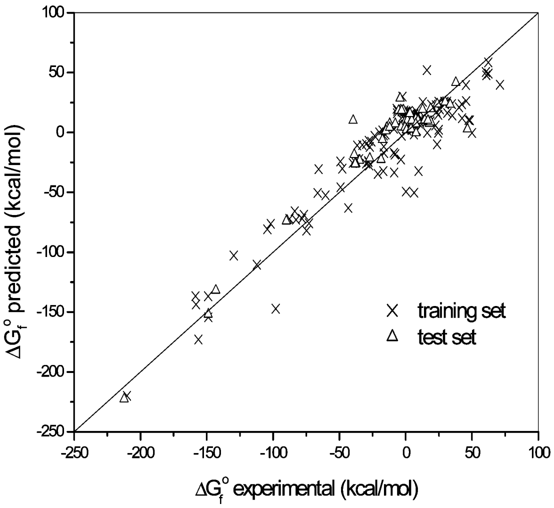

Figure 3. Predicted vs experimental values of ΔG_f° for the 177 compounds by multilinear regression analysis.

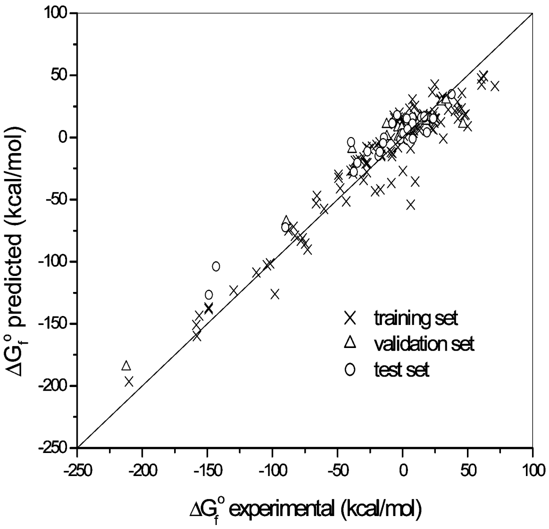

Figure 4. Predicted vs experimental values of ΔG_f° for the 177 compounds by a back-propagation neural network.

model). For the two MLR models, MLR1 was built on the sets based on a Kohonen self-organizing map and MLR2 was built on the sets split by random selection. The results are shown in Table 3. For the training set, for model MLR1, $r = 0.93$, $s = 19.1$, MAE = 14.6 kcal mol^{-1}, $n = 130$, and $F = 135.5$, and for model MLR2, $r = 0.94$, $s = 19.3$, MAE = 14.3 kcal mol^{-1}, $n = 130$, and $F = 143.6$; for the test set, for model MLR1, $r = 0.96$, $s = 14.1$, MAE = 11.7 kcal mol^{-1}, and $n = 47$, and for model MLR2, $r = 0.95$, $s = 13.8$, MAE = 12.2 kcal mol^{-1}, and $n = 47$. The RMS deviations for the two models are 17.8 and 17.7 kcal mol^{-1}, respectively. From these, it can be observed that the results of the two MLR models are similar.

Table 3 also gives the results for two back-propagation neural network models from two different sets-selection methods. Model NN1 was built on the sets based on a Kohonen self-organizing map; NN2 was built on the sets split by random selection. Similar to the case above, the structures of the neural networks based on the two random selection sets were also optimized. For model NN2, the optimum neural network architecture was 6−5−1; the best number of training epochs was 5500. For the training set, for model NN1, $r = 0.95$, $s = 14.5$, MAE = 11.8 kcal mol^{-1}, and $n = 130$, and for model NN2, $r = 0.96$, $s = 14.7$, MAE = 11.5 kcal mol^{-1}, and $n = 130$; for the validation set, for model NN1, $r = 0.97$, $s = 11.2$, MAE = 11.9 kcal mol^{-1}, and $n = 20$, and for model NN2, $r = 0.96$, $s = 12.8$, MAE = 10.9 kcal mol^{-1}, and $n = 20$; for the test set, for model NN1, $r = 0.97$, $s = 8.7$, MAE = 11.2 kcal mol^{-1}, and $n = 27$, and for model NN2, $r = 0.96$, $s = 13.0$, MAE = 10.4 kcal mol^{-1}, and $n = 27$. For the test set, the standard deviation of model NN1 is lower than that of model NN2. The RMS deviations for the two models are 15.4 and 15.0 kcal mol^{-1}, respectively. From these, it can be observed that the results of the two back-propagation neural network models are similar.

Basically, selecting sets on the basis of a Kohonen self-organizing map can ensure that the training set covers a larger chemical space than the validation and the test sets do, which may mean that the models would have good prediction results for the training set, validation set, and test set. One reason for the similar results from the different sets-selection methods may be that the sets split by the random selection are also casually good enough for building a model and for prediction; the other reason may be that the input descriptors used in this work are good for building the stable models.

CONCLUSIONS

The six descriptors used in this work represent fundamental characteristics of the molecules. The data set investigated in this study comprised a wide selection of different functional groups containing a variety of heteroatoms. Apparently, these four simple 2D autocorrelation coefficients, in particular, the first components for σ charge (Acorr_Sigchg_1) and lone-pair electronegativity (Acorr_LpEN_1), which have a substantial correlation with $\Delta G_f°$, (see Table 1), incorporate the influence of these heteroatoms and their bonding environment to a large extent. This alerts to the power of the PEOP method in taking account of the influence of heteroatoms and the network of bonds in its computational scheme and condensing it into charge and electronegativity values.

The descriptors applied in this work can be generated rapidly by calculation from the constitution of the molecules as reflected by a connection table. When a Windows XP (PM 790 MHZ) computer is used, for the 177 compounds, their five whole-molecule properties can be calculated by the ADRIANA.Code program in 1 s, and the seven 2D molecular autocorrelation vectors for all seven atomic properties can be calculated by the ADRIANA.Code program in about 4 s. The approach needs no experimental data for the description of compounds, and thus, the method is suitable to be extended to a larger data set.

In this work, a method for the modeling of $\Delta G_f°$ was developed on the basis of the 2D descriptors calculated from its structures. A nonlinear method such as the back-propagation neural network approach provides better models than multilinear regression analysis; the nonlinear methods are preferable in the modeling of the Gibbs energy of examined compounds. The performances of the models are acceptable, although the RMS values of the data set for $\Delta G_f°$ are larger than those in the work of Wang et al.[7] Further work on improving the prediction results and on using larger data sets will be undertaken.

ACKNOWLEDGMENT

This work was supported by the “Special Funding for the Talent Enrollment” of Beijing University of Chemical Technology. The author thanks Prof. J. Gasteiger, Dr. C. H. Schwab, and the Molecular Networks for the use of the programs ADRIANA.Code and SONNIA in our scientific work.

Supporting Information Available: The names of compounds used in this study, the six selected descriptors, the experimental and predicted values (by MLR analysis and a back-propagation neural network) of Gibbs energies of formation ($\Delta G_f°$) of all of the compounds, and the sets split by random selection. The material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Chirico, R. D.; Steele, W. V. High-Energy Components of “Designer Gasoline and Designer Diesel Fuel” I. Heat Capacities, Enthalpy Increments, Vapor Pressure, Critical Properties, and Derived Thermodynamic Functions for Biocyclopentyl between the T=(10 and 600) K. *J. Chem. Thermodyn.* **2004**, *36*, 633−643.

(2) Genoni, G. P. Influence of the Energy Relationships of Organic Compounds on Toxicity to the Cladoceran Daphnia Magna and the Fish Pimephales Promelas. *Ecotoxicol. Environ. Saf.* **1997**, *36*, 27−37.

(3) Alberty, R. A. Standard Transformed Gibbs Energies of Coenzyme A Derivatives as Functions of PH and Ionic Strength. *Biophys. Chem.* **2003**, *104*, 327−334.

(4) Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D.; Balaban, A. T. Evaluation in Quantitative Structure−Property Relationship Models of Structural Descriptors Derived from Information-Theory Operates. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 631−643.

(5) Ivanciuc, O.; Ivanciuc, T.; Klein, D. J.; Seitz, W. A.; Balaban, A. T. Wiener Index Extension by Counting Even/Odd Graph Distances. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 536−549.

(6) Toropov, A. A.; Toropova, A. P. QSPR Modeling of Alkanes Properties based on Graph of Atomic Orbitals. *THEOCHEM* **2003**, *637*, 1−10.

(7) Wang, X. J.; Wong, L. H.; Hu, L. H.; Chan, C. Y.; Su, Z.; Chen, G. H. Improving the Accuracy of Density-Function Theory Calculation: The Statistical Correction Approach. *J. Phys. Chem. A* **2004**, *108*, 8514−8525.

(8) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 1999.

(9) Yaws, C. L. *Chemical Properties Handbook*; McGraw-Hill: New York, 1999.

(10) (a) ADRIANA.Code can be obtained from Molecular Networks [http://www.mol-net.de (accessed Apr 2006)]. (b) The problem with the nitrate groups has been fixed in the latest version of ADRIANA.Code.

(11) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Flexibility. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 109−116. http://www2.chemie.uni-erlangen.de/software/cactvs/index.html (accessed Apr 2006).

(12) Terfloth, L.; Gasteiger, J. Self-organizing Neural Networks in Drug Design. *Screening—Trends Drug Discovery* **2001**, *2*, 49−51. http://www.mol-net.de (accessed Apr 2006).

(13) *SNNS: Stuttgart Neural Network Simulator*, version 4.2, Developed at University of Stuttgart, Maintained at University of Tübingen, 1995. http://www-ra.informatik.uni-tuebingen.de/SNNS/ (accessed Apr 2006).

(14) Gasteiger, J.; Marsili, M. A New Method for Calculating Atomic Charges in Molecules. *Tetrahedron Lett.* **1978**, *34*, 3181−3184.

(15) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(16) Kleinoeder, T. Prediction of Properties of Organic Compounds. Ph.D. Thesis, University of Erlangen-Nuernberg: Erlangen, Germany, 2005.

(17) Hutchings, M. G.; Gasteiger, J. Residual Electronegativity − An Empirical Quantification of Polar Influences and its Application to the Proton Affinity of Amines. *Tetrahedron Lett.* **1983**, *24*, 2541−2544.

(18) Gasteiger, J.; Hutchings, M. G. Quantification of Effective Polarisability. Applications to Studies of X-Ray Photoelectron Spectroscopy and Alkylamine Protonation. *J. Chem. Soc., Perkin Trans. 2*, **1984**, 559−564.

(19) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(20) Miller, K. J. Additivity Methods in Molecular Polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533−8542.

(21) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.

(22) Gasteiger, J.; Hutchings, M. G. Quantitative Models of Gas-Phase Proton-Transfer Reaction Involving Alcohols, Ethers and Their Thio Analogs. Correlation Analyses Based On Residual Electronegativity and Effective Polarizability. *J. Am. Chem. Soc.* **1984**, *106*, 6489−6495.

(23) Simon, V.; Gasteiger, J.; Zupan, J. A Combined Application of Two Different Neural Network Types for the Prediction of Chemical Reactivity. *J. Am. Chem. Soc.* **1993**, *115*, 9148−9159.

(24) Yan, A. X.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429−434.

(25) Yan, A. X.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds by Topological Descriptors. *QSAR Comb. Sci.* **2003**, *22*, 821−829.

(26) Yan, A. X.; Gasteiger, J.; Krug, M.; Anzali, S. Linear and Nonlinear Functions on Modeling of Aqueous Solubility of Organic Compounds by Two Structure Representation Methods. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 75−87.

(27) 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.

CI0600105