# Improving the Accuracy of Density-Functional Theory Calculation: The Statistical Correction Approach

**XiuJung Wang, LaiHo Wong, LiHong Hu, ChakYu Chan, Zhongmin Su,[†] and GuanHua Chen***

*Department of Chemistry, The University of Hong Kong, Pokfulam Road, Hong Kong, China*

*Received: June 23, 2004*

Recently, a novel, neural-networks-based method, the DFT-NEURON method, was developed to improve the accuracy of first-principles calculations and was applied to correct the systematic deviations of the calculated heats of formation for small-to-medium-sized organic molecules (Hu, L. H.; Wang, X. J.; Wong, L. H.; Chen, G. H. *J. Chem. Phys.* **2003**, *119*, 11501). In this work, we examine its theoretical foundation and generalize it to adopt any other statistical correction approaches, in particular, the multiple linear regression method. Both neural-networks-based and multiple-linear-regression-based correction approaches are applied to calculate the Gibbs energies of formation, ionization energies, electron affinities, and absorption energies of small-to-medium-sized molecules and lead to greatly improved calculation results as compared to the conventional first-principles methods. For instance, after the neural networks correction (multiple linear regression correction), the root-mean-square (RMS) deviations of the calculated standard Gibbs energy of formation for 180 organic molecules are reduced from 12.5, 13.8, and 22.3 kcal·mol$^{-1}$ to 4.7 (5.4), 3.2 (3.5), and 3.0 (3.2) kcal·mol$^{-1}$ for B3LYP/6-31G(d), B3LYP/6-311+G(3df,2p), and B3LYP/6-311+G(d,p) calculations, respectively, and the RMS deviation of the calculated absorption energies of 60 organic molecules is reduced from 0.33 eV to 0.09 (0.14) eV for the TDDFT/B3LYP/6-31G(d) calculation. In general, the neural networks correction approach leads to better results than the multiple linear regression correction approach. All these demonstrate that the statistical-correction-based first-principles calculations yield excellent results and may be employed routinely as predictive tools in materials research and design.

## I. Introduction

First-principles quantum mechanical methods have become indispensable research tools in chemistry, condensed-matter physics, materials science, and molecular biology.[1,2] Experimentalists rely increasingly on these methods to interpret their experimental findings. Despite their successes, first-principles quantum mechanical methods are often not quantitatively accurate enough to predict the results of experimental measurements, in particular, on large systems. This is caused by the inherent approximations adopted in first-principles methods. Because of the computational costs, electron correlation has always been a difficult obstacle for ab initio molecular orbital calculations. For instance, highly accurate full configuration interaction (FCI) calculations have been limited to very small molecules.[3] Basis sets cannot cover an entire physical space, and this introduces inherent computational errors.[4] In practice, limited by the computational resources, we often adopt inadequate basis sets for medium-to-large molecules. Effective core potential (ECP) is frequently used to approximate the relativistic effects, which leads inevitably to approximated results for heavy-element-containing systems. Accuracy of a density-functional theory (DFT) is determined by the exchange-correlation (XC) functional.[2] The exact XC functional is, however, unknown. All DFT calculations employ the approximated XC functional, which lead to further calculation errors. Much less is understood about the XC functional of time-dependent density-functional theory (TDDFT). It is a common practice to employ the standard XC functional of DFT such as gradient-corrected BP86, BLYP, or B3LYP for TDDFT calculations. This often results in poor calculated excited-state properties.[5] All of these contribute to the discrepancies between calculated and measured results. One of the Holy Grails in computational science is to predict the properties of matter prior to the experiments. To achieve this, we must eliminate the systematic deviations of the calculation results and reduce the numerical uncertainties to the limit of chemical accuracy (i.e., 1−2 kcal·mol$^{-1}$ for energies). G2 and G3 methods produce root-mean-square (RMS) deviations of less than 2 kcal·mol$^{-1}$ for various thermochemical properties of small molecules.[6,7] For medium-to-large-sized molecules, the deviations from the experimental data remain quite significant and often substantially beyond the limit of chemical accuracy. Alternatives must be sought.

Despite the various approximations that first-principles quantum mechanical calculations adopt, the calculated results capture the essence of physics. For instance, although their absolute values may not agree well with the experimental data, the calculated results of different molecules often have the same relative tendency as their experimental counterparts. To predict a physical property of a material, it may thus be sufficient to correct the corresponding raw value from the first-principles calculation. The discrepancy between the calculated and measured results depends on the characteristic physical or chemical properties of the material. These properties include predominantly the calculated property of interest, and to a lesser degree, other related properties of the material. These related properties

* Corresponding Author. Phone: (852)-28592164. Fax: (852)-28571586. E-mail: ghc@everest.hku.hk.

† Current address: Institute of Functional Material Chemistry, Faculty of Chemistry, Northeast Normal University, Changchun 130024, China.
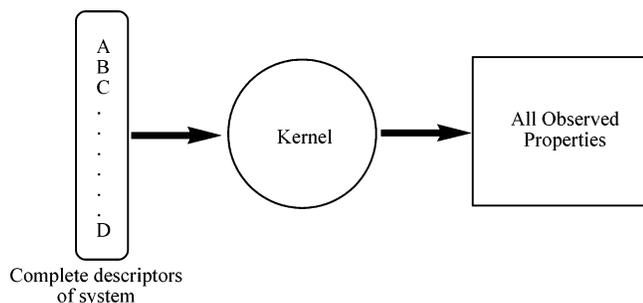
**Figure 1.** The universal computing machine.



**Figure 2.** The simplest universal computing machine.

may be evaluated via conventional first-principles methods. In other words, a quantitative relationship exists between the experimental and calculated properties. Although it is exceedingly difficult to be determined from the first-principles, the quantitative relationship can be obtained empirically. Statistical methods such as linear regression and neural networks may be employed to determine this relationship. Recently, we developed a neural-networks-based approach, DFT-NEURON, to determine the quantitative relationship between the experimental standard heat of formation and a set of physical descriptors for small-to-medium-sized organic molecules.[8] The resulting RMS deviation was reduced from 21.4 to 3.1 kcal·mol$^{-1}$ for the B3LYP/6-311+G(d,p) calculation and from 12.0 to 3.3 kcal·mol$^{-1}$ for the B3LYP/6-311+G(3df,2p) calculation after neural networks corrections.

In this work, we examine the theoretical foundation of the DFT-NEURON method and subsequently generalize it to encompass all statistical methodologies. Both multiple linear regression (MLR) and neural networks are used to improve the DFT calculation results on Gibbs energy of formation ($\Delta G_f^\circ$), ionization energy (IP), electron affinity (EA), and optical absorption energy. In section II, a general theoretical framework is established for the statistical correction approach to improve first-principles calculation results. In section III, the standard Gibbs energies of formation $\Delta G_f^\circ$'s at 298 K of 180 small- or medium-sized organic molecules used in ref 8, the IPs for 85 molecules in the G2 test set, and the EAs for 58 molecules in the G2 test set are evaluated via B3LYP calculations, and the absorption energies of 60 selected heterocyclic conjugated organic molecules[5] are calculated via the B3LYP/TDDFT method. All of these calculated values are then corrected by the MLR- and neural-networks-based correction approaches. The resulting linear regression expansions and neural networks are examined and analyzed in section IV. Discussion and conclusions are given in section V.

## II. Methodology

As stated in ref 8, the basic assumptions of the DFT-NEURON method are the following: (1) There is a quantitative relationship between the experimental measured property and the characteristic physical descriptors of the system, and (2) the primary descriptor is the calculated value of the property of interest. We will discuss and derive the theoretical foundation for the DFT-NEURON method and its basic assumptions here.[8]

All numerical computations of physical properties can be represented by the universal computing model (UCM) depicted in Figure 1, where the input is a description of the system, the kernel performs the calculations, and the output is the properties of interest. For instance, the description can be the number of electrons, the number of nuclei, the charge, and the position of an individual nucleus; the kernel can be the Schrödinger equation or the Kohn−Sham equation; and the output can be any proper-

ties of the system. Once a complete description of the system is given, the output is determined uniquely. The complete description is not unique, however. For example, it can be the ground-state electron density or the multipole moments of the system. According to the Hohenberg and Kohn theorem,[9] once the electron density is given, all physical properties of the system can thus be determined uniquely. There are different ways to describe the system of interest. Different sets of physical descriptors may be adopted to specify the same molecule or system. The kernel may vary depending on the choice of physical descriptors. With the proper selection of descriptors, the kernel can be rather simple and its computational cost trivial. Depicted in Figure 2 is the simplest kernel. The objective is to evaluate the exact value $A_{ex}$ of the property $\hat{A}$ of the system. If $A_{ex}$ is one of the physical descriptors, the kernel is simply
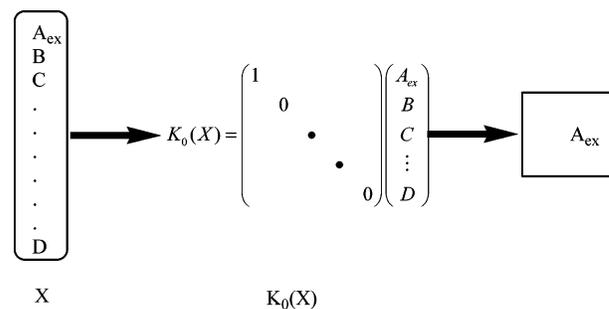
$$K_0 = \begin{pmatrix} 1 & & & & \\ & 0 & & & \\ & & 0 & & \\ & & & \ddots & \\ & & & & 0 \end{pmatrix}$$

as shown in Figure 2.

Because $A_{ex}$ is usually unknown, a different kernel $K$ other than $K_0$ is needed to evaluate the property $\hat{A}$. Although its exact value is difficult to compute, the existing quantum mechanical methods can be employed to obtain its approximate value $A_{cal}$ (i.e., $A_{ex} = A_{cal} + \delta A$). We adopt $A_{cal}$ as one of the physical descriptors, and the resulting kernel $K$ should thus be slightly different from $K_0$ if $|\delta A|$ is sufficiently small. We may express $K$ as follows:

$$A_{ex} = K(A_{cal}) = K_0(A_{cal}) + \delta K(A_{cal})$$

$$\delta K(A_{cal}) = -\delta A$$

where $\delta K(A, B, C, ..., D)$ is the functional deviation from $K_0(A, B, C, ..., D)$, and its functional form is to be determined. As long as $|\delta A|$ is small enough and $K(A, B, C, ..., D)$ is a well-behaved function of physical descriptors $(A, B, C, ..., D)$ around $\langle \hat{A} \rangle = A_{ex}$, $\delta K$ may be determined accurately via statistical methods. In other words, if the $A_{ex}$'s of a sufficient number of molecules are determined (by experiments), then $\delta K(A, B, C, ..., D)$ can be derived accurately by statistical methods, such as linear regression and neural networks. The analysis presented here validates the basic assumptions of the DFT-NEURON method[8] and expands it to include any other statistical correction approaches.

Having examined the theoretical foundation of the DFT-NEURON method or any other statistical correction approach to improve first-principles calculation results, we will generalize it to calculate $\Delta G_f^\circ$, IP, EA, and absorption energy. Besides

**8516** *J. Phys. Chem. A, Vol. 108, No. 40, 2004*

Wang et al.
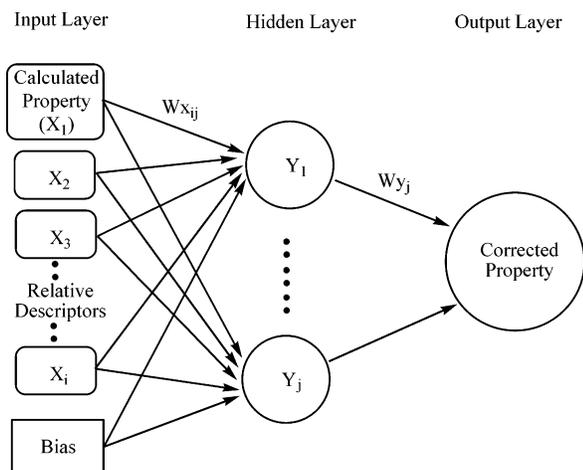


**Figure 3.** The structure of our neural network.

the neural networks correction, we will employ the MLR correction approach to correct the systematic errors of the DFT calculations.

Similar to ref 8, we adopt the three-layer architecture for our neural networks (see Figure 3). This architecture includes an input layer consisting of input from the physical descriptors ($X_1$, $X_2$, ..., $X_m$) and a bias, a hidden layer containing a number of hidden neurons ($Y_1$, ..., $Y_n$), and an output layer that outputs the corrected value for the property of interest (see Figure 3). The numbers of descriptors and hidden neurons are to be determined. The most important issue is to select the proper physical descriptors, which are to be used as the input for the neural network. If we are interested in determining the experimental $A_{ex}$ of the property $\hat{A}$, the first-principles calculated value $A_{cal}$ of $\hat{A}$ is set as the primary descriptor, as we have discussed already. Other physical descriptors are selected according to their correlations to $\hat{A}$. If it is related closely to $\hat{A}$, a property is chosen as a physical descriptor; otherwise, it is not. The physical properties, such as the number of atoms, the number of hydrogen atoms, the number of electrons, the number of valence electrons, total energy, zero point energy (ZPE), the highest occupied molecular orbital (HOMO) energy, the lowest unoccupied molecular orbital (LUMO) energy, the HOMO−LUMO energy gap, mass, the number of double bonds, the number of triple bonds, dipole moment, quadrupole moment, or the number of conjugated rings, have been chosen as the other physical descriptors depending on the property of interest. The bias is set to 1. The synaptic weights ($Wx_{ij}$'s) connect the input descriptors ($X_i$'s) and the hidden neurons ($Y_j$'s), while $Wy_j$'s connect the hidden neurons and the output $Z$. The output $Z$ is related to the input ($X_i$) as follows:

$$Z = \sum_{j=1}^{n} Wy_j \, \text{Sig}(\sum_{i=1}^{m} Wx_{ij} X_i)$$

where $\text{Sig}(v) = [1 + \exp(-\alpha v)]^{-1}$ and $\alpha$ is a parameter that controls the switch steepness of sigmoidal function $\text{Sig}(v)$. In our neural network, we adopt $\alpha = 4$. The error back-propagation learning procedure[10] is used to optimize the values of $Wx_{ij}$ and $Wy_j$ ($i = 1,..., m$; and $j = 1,..., n$). The output value is scaled to lie between 0 and 1, and all input values are scaled to lie between 0.1 and 0.9 except for the bias.

All experimental data for a particular property of interest are randomly divided into a training set and a testing set. To ensure the reliability of a neural network, a 5-fold cross-validation procedure is adopted.[11−14] The training set is divided further
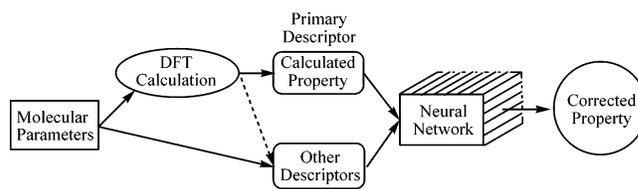


**Figure 4.** The flowchart of a DFT-NEURON calculation.

randomly into five subsets of equal size. Four of them are used to train the neural network, and the remaining one is used to evaluate its prediction. This procedure is repeated five times in rotation. The number of hidden neurons is varied to yield the optimal training results. Once the structure of the neural network and synaptic weights are determined, the neural network can be used to correct the raw first-principles calculation results. This is illustrated in Figure 4. The calculated property of interest and other closely related molecular properties are used as physical descriptors and are input to the neural network. The corrected property of interest is given at the output. $Z$ is very close to the rescaled $A_{ex}$ for a successful DFT-NEURON calculation.

For the MLR correction, the corrected physical property of interest, $Z$, is expressed in terms of the physical descriptors $X_1$, $X_2$, ..., and $X_m$ as

$$Z = C_0 + \sum_{i=1}^{m} C_i X_i$$

where $C_i$ ($i = 0, 1, ..., m$) are the coefficients to be determined.

To determine the values of $C_0$, $C_1$, $C_2$, ..., and $C_m$, we adopt the similar 5-fold cross-validation procedure employed to train the neural networks.

### III. Application for Calculating Gibbs Energy of Formation, Ionization Potential, Electron Affinity, and Absorption Energy

**A. Gibbs Energy of Formation.** We used the same 180 molecules[15−17] as in ref 8 to train, validate, and test a neural network to calculate the standard Gibbs energy of formation $\Delta G_f^\circ$ for small-to-medium-sized molecules. These molecules contain elements such as H, C, N, O, F, S, Cl, and Br. The heaviest molecule contains 14 heavy atoms, and the largest contains 32 atoms. The detailed experimental $\Delta G_f^\circ(298 \text{ K})$ and the differences between the calculated and experimental values for 180 compounds are collected in the Supporting Information. As in ref 8, we divide these molecules randomly into a training set containing 150 molecules and a testing set containing 30 molecules. To calculate $\Delta G_f^\circ$ of a molecule $A_x B_y$, we need to calculate its $\Delta H_f^\circ$ and its entropy $S^\circ$ at 298 K. $\Delta G_f^\circ$ at 298 K can be thus evaluated as

$$\Delta G_f^\circ(A_x B_y, 298 \text{ K}) = \Delta H_f^\circ(A_x B_y, 298 \text{ K}) - 298.15 \times$$
$$\{S^\circ(A_x B_y, 298 \text{ K}) - [x S^\circ(A, 298 \text{ K}) + y S^\circ(B, 298 \text{ K})]\}$$

where $\Delta H_f^\circ(A_x B_y, 298 \text{ K})$ is the standard enthalpy of formation of $A_x B_y$ at 298 K, $S^\circ(A_x B_y, 298 \text{ K})$ is the entropy of $A_x B_y$ at 298 K, and $S^\circ(A, 298 \text{ K})$ and $S^\circ(B, 298 \text{ K})$ are, respectively, the standard entropies of species $A$ and $B$ whose values are given in ref 18. Three slightly different approaches are used in the calculations. In approaches A and B, the geometry optimization and ZPE calculation are carried out at the B3LYP/6-31G(d) level[19] while the total energy is calculated at the B3LYP/6-31G(d) and B3LYP/6-311+G(3df,2p) levels, respectively. In

Density-Functional Theory Calculation

*J. Phys. Chem. A, Vol. 108, No. 40, 2004* **8517**
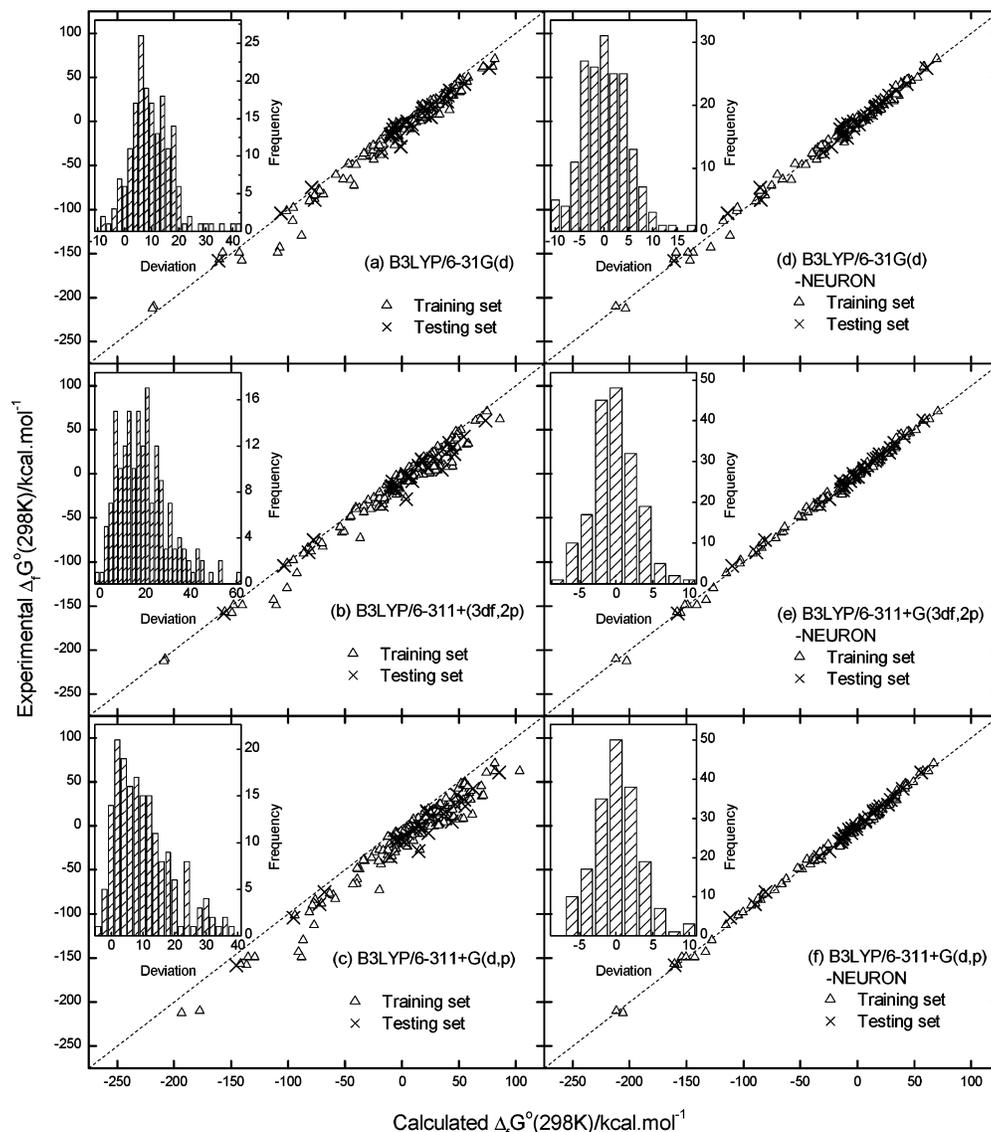


**Figure 5.** Calculated $\Delta G_f^\circ$ versus experimental $\Delta G_f^\circ$ values for all 180 compounds. Parts a, b, and c are for raw calculated $\Delta G_f^\circ$ values from approaches A, B, and C, respectively. Parts d, e, and f are for neural-networks-corrected $\Delta G_f^\circ$'s for approaches A, B, and C, respectively. Triangles ($\triangle$) are for the training set, and crosses ($\times$) are for the testing set. Insets are the histograms for the differences between the experimental and calculated $\Delta G_f^\circ$'s. All values are in the units of kcal·mol$^{-1}$.

approach C, the geometry optimization, total energy, and ZPE calculation are all carried out at the B3LYP/6-311+G(d,p) level.

The raw calculated $\Delta G_f^\circ$ values for approaches A, B, and C are compared to their experimental data in Figure 5a–c, respectively. The vertical coordinates are the experimental $\Delta G_f^\circ$'s, and the horizontal coordinates are the raw calculated values. The dashed line is where the vertical and horizontal coordinates are equal (i.e., where the B3LYP calculations and experiments would have the perfect match). The raw calculated values are mostly below the dashed line (i.e., most raw $\Delta G_f^\circ$ values are larger than the experimental data). Compared to the experimental measurements, the RMS deviations of the three approaches A, B, and C are 12.5, 13.8, and 22.3 kcal·mol$^{-1}$, respectively. Approach C has a larger deviation than the other two approaches. The RMS deviation of approach B is slightly larger than that of approach A. This is because of the unscaled ZPE values used in the raw DFT calculation.

There are clearly systematic deviations between the calculated and measured $\Delta G_f^\circ$'s. Pople and co-workers[6,7] used the scaled ZPE values to compute the total energy. They found that the best agreement to the experimental enthalpies for the G2 test

set is obtained with a scaling factor of 0.94 for B3LYP/ 6-311+G(3df,2p) calculations.[7] Adopting their strategy, we can reduce the RMS deviation for the 180 organic molecules to 9.3 kcal·mol$^{-1}$ for approach B. To reduce further the RMS deviation between the experiment and the calculation, we employ the MLR and the neural networks correction approaches.

The key issue is to determine the appropriate physical descriptors. As discussed in section II, the raw calculated $\Delta G_f^\circ$ contains the essence of the exact value and is thus the obvious choice for the primary descriptor. We observe that the molecular size affects the accuracy of the calculation. The more atoms a molecule has, the worse the calculated $\Delta G_f^\circ$ is. This is consistent with general observation in the field.[7] The total number of atoms of the molecule, $N_t$, is thus chosen as the second possible descriptor. ZPE is an important parameter in calculating $\Delta G_f^\circ$, whose raw calculated value is often scaled.[7] It is thus taken as the third physical possible descriptor. Finally, the hydrogen atom is much lighter than the heavy atoms. The number of hydrogen atoms in a molecule, $N_H$, is selected as the fourth and last possible descriptor to account for the distinctive contribution from hydrogen atoms.

**8518** *J. Phys. Chem. A, Vol. 108, No. 40, 2004*

Wang et al.

**TABLE 1: RMS Deviations of MLR and Neural Networks Corrections**[a]

| | $A$[b] | | $B$[b] | | $C$[b] | |
| | MLR | DFT-NEURON | MLR | DFT-NEURON | MLR | DFT-NEURON |
|---|---|---|---|---|---|---|
| I[c] | 5.8 | 5.0 | 4.1 | 3.7 | 6.0 | 5.0 |
| II[d] | 5.6 | 4.9 | 3.7 | 3.6 | 4.9 | 4.3 |
| III[e] | 5.5 | 4.8 | 3.7 | 3.5 | 3.2 | 3.0 |
| IV[f] | 5.4 | 4.7 | 3.5 | 3.2 | 3.2 | 3.0 |

[a] All data are in the units of kcal·mol$^{-1}$. [b] A, B, and C denote approaches A, B, and C, respectively. [c] I: DFT calculated $\Delta G_f°$ and ZPE as descriptors. [d] II: DFT calculated $\Delta G_f°$ and $N_t$ as descriptors. [e] III: DFT calculated $\Delta G_f°$, $N_t$, and ZPE as descriptors. [f] IV: DFT calculated $\Delta G_f°$, $N_t$, ZPE, and $N_H$ as descriptors.

The 150 molecules in the training set are divided randomly into five subsets of equal size. $\Delta G_f°$ is the primary physical descriptor and is thus always adopted. $N_t$, ZPE, and $N_H$ are selected for different statistical correction procedures. For the DFT-NEURON approach, we find that a hidden layer containing two neurons yields the best overall results.

The RMS deviations of MLR and neural networks corrections are listed in Table 1 for different procedures (I, II, III, and IV) with different descriptors. Both the MLR and neural networks corrections yield greatly improved $\Delta G_f°$'s over the raw DFT calculation results, while the neural networks correction approach gives slightly better results than the MLR correction approach. Procedure IV results in the smallest RMS deviations compared to the other three procedures I, II, and III. This is expected, because it has all four physical descriptors. Approach A with either the MLR or neural networks correction gives relatively large RMS deviations, because its basis set is the least sophisticated.

In the following, we will examine the DFT-NEURON results with procedure IV in detail. The neural-networks-corrected $\Delta G_f°$'s versus the experimental measured values are plotted in Figure 5d–f for the three approaches A, B, and C, respectively. The vertical coordinates are the experimental $\Delta G_f°$'s, and the horizontal coordinates are the neural-networks-corrected $\Delta G_f°$'s. The dashed line is again where the vertical and horizontal coordinates are equal. After the neural networks correction, the RMS deviations are reduced from 12.5, 13.8, and 22.3 kcal·mol$^{-1}$ to 4.7, 3.2, and 3.0 kcal·mol$^{-1}$ for approaches A, B, and C, respectively. Although the raw B3LYP/6-31G(d) results have RMS values of the same magnitude as the raw B3LYP/6-311+G(3df,2p) results, the neural-networks-corrected values for B3LYP/6-311+G(3df,2p) agree much better with the measured values than those of the B3LYP/6-31G(d). This implies that the 6-31G(d) basis set is not appropriate. In the insets of Figure 5a–f, we plot the histograms for the deviations of various approaches. Obviously, the raw calculated $\Delta G_f°$'s have large systematic deviations, while the neural-networks-corrected $\Delta G_f°$'s have virtually no systematic deviations. Moreover, the remaining numerical deviations are greatly reduced. This can be further demonstrated by the error analysis performed for the B3LYP/6-311+G(3df,2p) $\Delta G_f°$'s of all 180 molecules. For the training set, the RMS deviations before and after the neural networks correction are 14.0 and 3.2 kcal·mol$^{-1}$, respectively, while for the testing set, they are 13.1 and 3.1 kcal·mol$^{-1}$, respectively. For the MLR correction, the RMS deviations of the training and testing sets are reduced to 3.5 and 3.0 kcal·mol$^{-1}$, respectively. The consistency between testing and training sets implied that the neural network results can indeed predict the measured $\Delta G_f°$ with good accuracy. We

have performed the same error analysis for B3LYP/6-31+G(d) and B3LYP/6-311+G(d,p) $\Delta G_f°$'s and have reached a similar conclusion. Moreover, the deviations for large molecules are of the same magnitude as those for small molecules. The DFT-NEURON method does not discriminate against large molecules, unlike most other calculations that yield worse results for large molecules than for small ones.

**B. Ionization Potential.** We apply the MLR and neural networks correction approaches to improve the calculated IPs of small molecules or atoms. We took 85 atoms and molecules from the G2-1 and G2-2 test sets,[6] because their measured IPs are well-documented. They consisted of 18 atoms and 67 small molecules. B3LYP/6-311+G(3df,2p) is employed to calculate their IPs. In the calculation, the unscaled ZPEs are used. The calculated and experimental values are listed in Table 2. The RMS deviation of raw calculated IPs from their experimental counterparts is 4.9 kcal·mol$^{-1}$.

We divide these species randomly into a training set containing 70 species and a testing set containing 15 species. The calculated IP is the primary descriptor. The IP is mainly determined by the interaction among valence electrons, and the core electrons change little before and after the ionization. The number of valence electrons, $N_{ve}$, is thus set as the second physical descriptor. Because radicals are encountered, the multiplicity $g_s$ is selected as the third physical descriptor. The HOMO–LUMO energy gap $E_g$ is chosen as the fourth and last physical descriptor. Once again, we find that two hidden neurons yield the best overall results for the DFT-NEURON method.

The RMS deviations for the MLR and neural networks corrections are reduced to 3.6 and 3.0 kcal·mol$^{-1}$, respectively. The deviations from the experimental values for the MLR- and neural-networks-corrected IPs for all 85 atoms or molecules are tabulated in Table 2. Those belonging to the testing set are identified. The RMS deviations for the training and testing sets for the MLR correction are 3.6 and 3.5 kcal·mol$^{-1}$, respectively, while they are all 3.0 kcal·mol$^{-1}$ for the DFT-NEURON method. These validate the resulting MLR and neural networks correction approaches. The histograms for the deviations of raw calculated and the MLR- and neural-networks-corrected results are plotted in Figure 6a–c, respectively. The maximum positive and negative deviations for the raw calculated IPs are 12.7 and −13.0 kcal·mol$^{-1}$, respectively. In Figure 6b, the maximum positive and negative deviations are reduced to 8.4 and −9.8 kcal·mol$^{-1}$, respectively, while in Figure 6c, they are 7.4 and −6.6 kcal·mol$^{-1}$, respectively. The DFT-NEURON method results in slightly better IPs than the MLR correction approach.

**C. Electron Affinity.** We apply the MLR and neural networks correction approaches to improve the raw calculated EA. We employed 11 atoms and 47 molecules in the calculation, which we took from the G2-1 and G2-2 test sets.[6] Their experimental EAs are available and are well-documented in ref 6. We employ B3LYP/6-311+G(3df,2p) to compute their EAs. The raw calculated and experimental EAs are listed in Table 3. The RMS deviation between the raw calculation and measured EAs is 3.5 kcal·mol$^{-1}$.

We divide these species randomly into a training set containing 50 species and a testing set containing 8 species. Besides the primary physical descriptor, the raw calculated EA, the number of valence electrons $N_{ve}$, spin multiplicity $g_s$, and the HOMO–LUMO energy gap $E_g$ are chosen as the other physical descriptors just as we chose for the IPs. Two hidden neurons are adopted for the neural network. After the MLR and neural networks correction, the RMS deviations are reduced to 3.1 and 1.7 kcal·mol$^{-1}$, respectively. The DFT-NEURON approach

Density-Functional Theory Calculation

*J. Phys. Chem. A, Vol. 108, No. 40, 2004* **8519**

**TABLE 2: Experimental Ionization Potentials and the Differences between the Calculated and Experimental Values[a]**

| species | expt[b] | deviation[c] | deviation[d] | deviation[e] | species | expt[b] | deviation[c] | deviation[d] | deviation[e] |
|---|---|---|---|---|---|---|---|---|---|
| Li | 124.3 | 5.2 | 2.3 | 3.0 | $BCl_3$ | 267.5 | −9.6 | −4.6 | −4.6 |
| Be | 214.9 | −4.7 | −6.0 | −5.8 | $B_2F_4$ | 278.3 | −13.0 | −5.0 | −3.6 |
| B[f] | 191.4 | 10.1 | 7.1 | 3.3 | $CO_2$ | 317.6 | −0.8 | 1.0 | 3.2 |
| C | 259.7 | 6.6 | 1.1 | 0.1 | $CF_2$ | 263.3 | −1.5 | 1.6 | 0.9 |
| N | 335.3 | 3.1 | −5.3 | −4.1 | COS | 257.7 | 0.3 | 2.7 | 3.3 |
| O | 313.9 | 12.7 | 6.8 | 4.1 | $CS_2$ | 232.2 | −0.8 | 2.0 | 1.5 |
| F | 401.7 | 7.9 | 3.6 | −1.7 | $CH_2$ | 239.7 | 0.2 | −4.6 | −4.2 |
| Na | 118.5 | 6.5 | 3.6 | 4.7 | $CH_3$ | 227.0 | 2.4 | 0.2 | −2.6 |
| Mg | 176.3 | 1.9 | 0.9 | 2.3 | $C_2H_5$ ($^2A'$) | 187.2 | 2.3 | 2.4 | −0.1 |
| Al | 138.0 | 0.9 | −1.5 | −2.4 | $C_3H_4$(cyclopropene)[f] | 223.0 | −6.6 | −3.7 | −2.5 |
| Si | 187.9 | −0.8 | −5.5 | −3.5 | $CH_2=C=CH_2$[f] | 223.5 | −5.5 | −2.6 | −0.9 |
| P | 241.9 | −2.5 | −9.8 | 1.0 | sec-$C_3H_7$ | 170.0 | 0.0 | 2.3 | −0.1 |
| S | 238.9 | 4.3 | −0.6 | −0.1 | $C_6H_6$ | 213.2 | −4.4 | 3.0 | 4.0 |
| Cl | 299.1 | 2.3 | −0.7 | −3.2 | CN | 313.6 | 9.8 | 7.2 | 3.6 |
| $CH_4$ | 291.0 | −3.6 | −4.1 | −0.4 | CHO | 187.7 | 8.0 | 7.4 | 4.2 |
| $NH_3$ | 234.8 | 0.3 | 0.5 | 2.2 | $H_2COH$ ($^2A$) | 174.2 | 3.2 | 3.5 | 1.0 |
| OH | 300.0 | 5.2 | 2.1 | −0.5 | $CH_3O$ ($^2A'$) | 247.3 | −3.3 | −3.8 | −6.6 |
| $OH_2$ | 291.0 | 0.0 | −0.4 | 1.8 | $CH_3OH$ | 250.2 | −5.9 | −4.0 | −2.7 |
| FH[f] | 369.9 | 1.4 | 0.1 | 2.4 | $CH_3F$ | 287.6 | −4.0 | −2.5 | 0.1 |
| $SiH_4$ | 253.7 | −2.1 | −2.2 | 0.9 | $CH_2S$ | 216.2 | −2.1 | −0.3 | −1.6 |
| PH | 234.1 | 0.3 | −4.4 | −3.9 | $CH_2SH$ | 173.8 | 3.1 | 3.4 | 0.9 |
| $PH_2$[f] | 226.5 | 2.3 | 0.1 | −2.7 | $CH_3SH$ | 217.7 | −2.6 | −0.3 | 0.6 |
| $PH_3$[f] | 227.6 | −0.9 | −0.6 | 1.1 | $CH_3Cl$ | 258.7 | −3.1 | −1.3 | 0.0 |
| SH | 239.1 | 2.1 | −0.2 | −3.0 | $C_2H_5OH$[f] | 241.4 | −8.5 | −4.5 | −3.2 |
| $SH_2$ ($^2B_1$)[f] | 241.4 | −1.4 | −1.2 | −0.2 | $CH_3CHO$ | 235.9 | −3.5 | −0.1 | 0.1 |
| $SH_2$ ($^2A_1$) | 294.7 | −3.0 | −3.4 | −3.2 | $CH_3OF$ | 261.5 | 4.6 | −0.9 | −0.9 |
| ClH | 294.0 | −0.2 | −0.6 | 1.1 | $C_2H_4S$ (thiirane) | 208.7 | −2.8 | 0.9 | 1.7 |
| $C_2H_2$[f] | 262.9 | −3.5 | −2.9 | −1.1 | NCCN | 308.3 | −7.6 | −5.0 | −4.8 |
| $C_2H_4$ | 242.4 | −5.0 | −3.6 | −2.4 | $C_4H_4O$ (furan) | 203.6 | −2.3 | 3.9 | 5.0 |
| CO | 323.1 | 3.9 | 3.7 | 5.3 | $C_4H_5N$ (pyrrole)[f] | 189.3 | −2.6 | 3.8 | 5.3 |
| $N_2$ ($^2\Sigma$ cation) | 359.3 | 6.1 | 5.4 | 7.4 | $B_2H_4$ | 223.7 | −4.6 | −3.6 | −2.0 |
| $N_2$ ($^2\Pi$ cation) | 385.1 | −0.8 | −1.7 | −0.6 | NH* | 311.1 | 4.4 | −1.3 | −4.1 |
| $O_2$ | 278.3 | 11.9 | 8.4 | 6.8 | $NH_2$ | 256.9 | 4.5 | 1.9 | −0.3 |
| $P_2$ | 242.8 | 5.3 | 6.1 | 5.1 | $N_2H_2$ | 221.1 | −1.0 | 0.7 | 0.3 |
| $S_2$ | 215.8 | 4.4 | 1.8 | 1.9 | $N_2H_3$[f] | 175.5 | 6.6 | 6.8 | 4.5 |
| $Cl_2$ | 265.2 | −2.7 | −0.8 | −2.3 | HOF | 293.1 | −1.2 | 0.3 | 0.3 |
| ClF | 291.9 | −0.9 | 0.7 | −0.6 | $SiH_2$ ($^1A_1$) | 211.0 | −2.4 | −2.5 | −3.6 |
| SC[f] | 261.3 | 2.4 | 2.9 | 2.5 | $SiH_3$[f] | 187.6 | 1.0 | −0.7 | −2.6 |
| H | 313.6 | 1.5 | −3.7 | −2.9 | $Si_2H_2$[f] | 189.1 | −4.0 | −2.5 | −2.3 |
| He | 567.0 | 7.8 | 1.2 | 0.3 | $Si_2H_4$ | 186.6 | −4.5 | −2.4 | −2.2 |
| Ne | 497.2 | 4.8 | −0.6 | −3.4 | $Si_2H_5$ | 175.5 | 1.2 | 1.5 | −0.6 |
| Ar | 363.4 | 1.0 | −2.8 | 1.1 | $Si_2H_6$ | 224.6 | −5.3 | −3.0 | −0.8 |
| $BF_3$ | 358.8 | −9.8 | −5.9 | −3.0 | | | | | |

[a] All data are in the units of kcal·mol$^{-1}$. [b] Experimental data are taken from ref 6. [c] Differences between the raw calculated and experimental IPs. [d] Differences between the calculated and experimental IPs for DFT−MLR calculation. [e] Differences between the calculated and experimental IPs for DFT-NEURON calculation. [f] Molecules belong to the testing set.

generated better results than the MLR correction approach. The detailed results are tabulated in Table 3. The RMS deviations for the DFT-NEURON training and testing sets are 1.7 and 1.8 kcal·mol$^{-1}$, respectively, which justifies the validity of the resulting neural network. For MLR correction, the RMS deviations of the training and testing sets are 2.9 and 4.3 kcal·mol$^{-1}$, respectively. The histograms of the deviations for the raw DFT, MLR, and neural networks correction approaches are plotted in Figure 7a−c, respectively.

**D. Absorption Energy.** So far, the MLR and neural networks correction approaches have only been employed to calculate the ground-state properties of molecules or atoms. In principle, the method can be applied to compute the excited-state properties as well. As the first application for the excited-state properties, we generalize the MLR and neural networks correction approaches to calculate the optical absorption energies of organic molecules.

Marks et al.[5] evaluated the absorption energies of 60 heterocyclic organic molecules using ZINDO/CIS, ZINDO/RPA, HF/CIS, HF/RPA, TDDFT/TDA, and TDDFT/RPA calculations. The largest molecule contains 82 atoms, of which 48 are heavy atoms and the rest are H atoms. They concluded that

TDDFT/CIS and TDDFT/RPA methods yield relatively accurate results upon linear regression fit. The BLYP functional approach was employed in their TDDFT calculations. We employ the TDDFT/B3LYP calculation to evaluate the absorption energies of the same 60 organic molecules. The raw calculated absorption energies are subsequently corrected by the DFT-NEURON method and the MLR correction approaches.

Geometry optimization is performed at the B3LYP/6-31G(d) level for each of the 60 organic molecules. The TDDFT/B3LYP/6-31G(d) calculation is employed to determine the excited-state energies and their oscillator strengths. The calculated excitation energy with the largest oscillator strength is chosen as the optical gap of the molecule and compared to the experimental counterpart.[5] Results and comparisons are listed in Table 4. For most of our calculations, the lowest-energy transition possesses the largest oscillator strength. In some rare cases where the oligomers have only one repeat unit, the lowest transitions may not have the largest oscillator strengths. In such a case, we choose a low-lying transition with the largest oscillator strength whose energy is within 1 eV from the lowest excited-state energy. The calculated and measured absorption energies and their differences are listed in Table 4. The RMS

**8520** *J. Phys. Chem. A, Vol. 108, No. 40, 2004*

Wang et al.
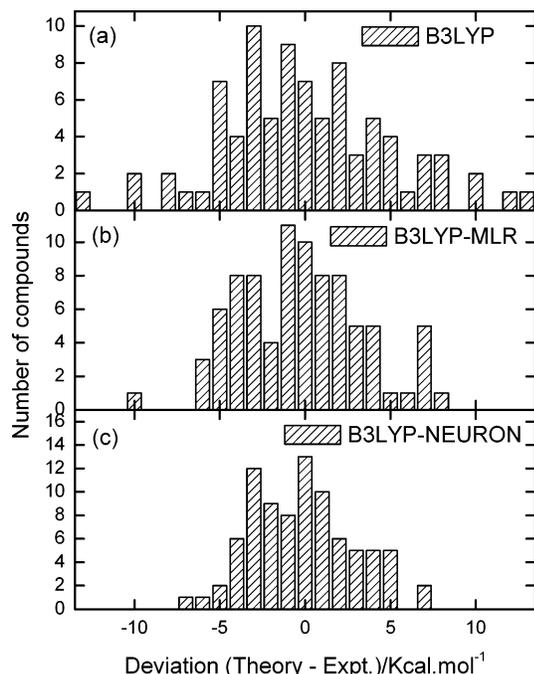


**Figure 6.** Histograms for the deviations of the B3LYP/6-311+G(3df,2p)-calculated IP for all 85 species. Parts a, b, and c are for the raw calculated, the MLR-corrected, and the neural-networks-corrected IPs, respectively.
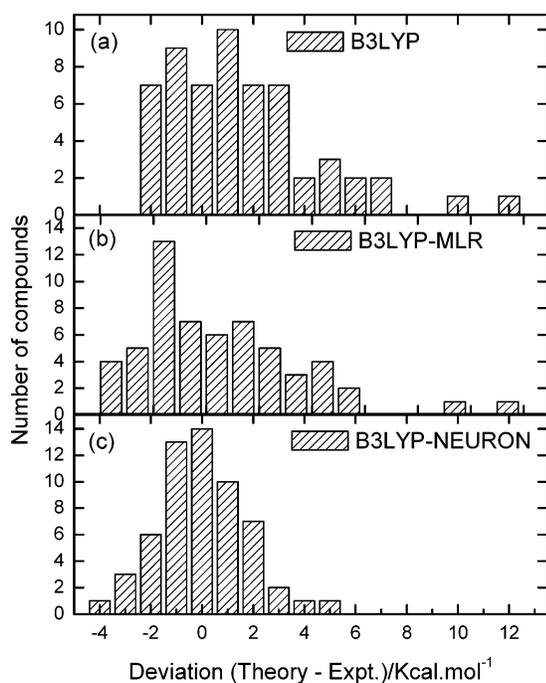


**Figure 7.** Histograms for the deviations of the B3LYP/6-311+G(3df,2p)-calculated EA values for all 58 species. Parts a, b, and c are for the raw calculated, the MLR-corrected, and the neural-networks-corrected EAs, respectively.

deviation between our calculated and measured absorption energies is 0.33 eV. It is less than that of ref 5, because we employed the B3LYP functional approach rather than BLYP. The B3LYP calculation reproduces the experimental data better than the BLYP calculation for heterocyclic organic molecules. This is different from the calculations on aromatic hydrocarbon radical cations, pyrene, perylene, and polyene.[20,21] There are clearly systematic deviations between calculated and measured absorption energies. The calculated deviation increases with an

**TABLE 3: Experimental Electron Affinities (EAs) and the Differences between Calculated and Experimental Values[a]**

| species | expt[b] | deviation[c] | deviation[d] | deviation[e] |
|---|---|---|---|---|
| C | 29.1 | 2.2 | 0.9 | 1.2 |
| O | 33.7 | 3.3 | 3.0 | −1.2 |
| F | 78.4 | 1.3 | 1.8 | 0.3 |
| Si | 31.9 | −1.5 | −2.8 | −2.4 |
| P | 17.2 | 4.4 | 3.7 | −1.1 |
| S | 47.9 | 2.6 | 2.2 | 0.7 |
| Cl | 83.4 | 1.5 | 1.8 | −0.3 |
| CH | 28.6 | 2.4 | 0.8 | 2.5 |
| $CH_2$ | 15.0 | 2.8 | 1.6 | 1.6 |
| $CH_3$ | 1.8 | 0.1 | −1.8 | 0.2 |
| NH | 8.8 | 1.2 | 0.0 | 0.9 |
| $NH_2$ | 17.8 | −1.8 | −3.2 | −1.6 |
| OH | 42.2 | −1.7 | −2.5 | −0.4 |
| SiH | 29.4 | −0.6 | −2.3 | −1.1 |
| $SiH_2$[f] | 25.9 | 0.9 | −1.1 | −2.2 |
| $SiH_3$[f] | 32.5 | 0.1 | −1.1 | 0.8 |
| PH | 23.8 | 1.3 | 0.3 | −0.4 |
| $PH_2$ | 29.3 | −0.6 | −1.9 | −0.1 |
| HS | 54.4 | −0.9 | −1.6 | 0.9 |
| $O_2$ | 10.1 | 1.9 | 1.1 | 1.3 |
| NO | 0.5 | 7.1 | 5.2 | 4.5 |
| CN | 89.0 | 4.4 | 4.9 | −0.3 |
| PO | 25.1 | 3.0 | 1.5 | 1.6 |
| $S_2$ | 38.3 | 0.1 | −0.4 | −0.8 |
| $Cl_2$[f] | 55.1 | 10.5 | 9.8 | 3.2 |
| Li[f] | 14.3 | −1.4 | −3.6 | −2.2 |
| B | 6.4 | 2.9 | 0.8 | 1.6 |
| Na | 12.6 | 0.9 | −1.3 | 0.2 |
| Al | 10.2 | −1.3 | −3.6 | −3.1 |
| $C_2$[f] | 75.5 | 2.1 | 2.4 | 0.1 |
| $C_2O$ | 52.8 | 0.5 | 0.5 | −0.2 |
| $CF_2$ | 4.1 | 6.2 | 4.5 | 0.2 |
| NCO | 83.2 | −2.0 | −1.6 | −4.2 |
| $NO_2$ | 52.4 | −1.0 | −1.5 | −0.8 |
| $O_3$ | 48.5 | 12.4 | 11.7 | 3.8 |
| OF | 52.4 | 0.1 | −0.3 | 1.7 |
| $SO_2$ | 25.5 | 6.5 | 5.2 | 2.1 |
| $S_2O$[f] | 43.3 | 5.4 | 4.4 | −1.5 |
| $C_2H$ | 68.5 | 2.5 | 2.4 | 3.1 |
| $C_2H_3$ | 15.4 | 0.2 | −1.3 | −0.6 |
| $H_2C=C=C$[f] | 41.4 | 3.3 | 2.0 | −1.9 |
| $H_3C=C=CH$ | 20.6 | 2.6 | 1.5 | 2.0 |
| $CH_2CHCH_2$ | 10.9 | 1.0 | −0.4 | −1.0 |
| HCO | 7.2 | 0.6 | −1.2 | −1.2 |
| HCF | 12.5 | 5.2 | 3.3 | 0.1 |
| $CH_3O$ | 36.2 | −1.6 | −2.4 | −0.8 |
| $CH_3S$[f] | 43.1 | −1.1 | −1.8 | −0.1 |
| $CH_2S$ | 10.7 | 4.9 | 2.9 | −0.6 |
| $CH_2CN$ | 35.6 | 0.6 | −0.2 | 1.1 |
| $CH_2NC$ | 24.4 | 1.5 | 0.4 | 1.0 |
| CHCO | 54.2 | −1.8 | −2.1 | −0.6 |
| $CH_2CHO$ | 42.1 | 0.2 | −0.4 | 0.9 |
| $CH_3CO$ | 9.8 | −0.6 | −2.0 | −3.3 |
| $CH_3CH_2O$ | 39.5 | −0.2 | −0.7 | 0.3 |
| $CH_3CH_2S$ | 45.0 | −1.4 | −1.9 | −1.1 |
| LiH | 7.9 | 2.3 | −0.2 | −1.1 |
| HNO | 7.8 | 6.9 | 4.9 | 1.3 |
| $HO_2$ | 24.9 | −2.5 | −3.5 | −2.5 |

[a] All data are in the units of kcal·mol⁻¹. [b] Experimental data are taken from ref 6. [c] Differences between the raw calculated and experimental EAs. [d] Differences between the calculated and experimental EAs for DFT−MLR calculation. [e] Differences between the calculated and experimental EAs for DFT-NEURON calculation. [f] Molecules belong to the testing set.

increasing number of repeating units. The largest negative deviation is −0.50 eV (molecule 30, see Table 4). The largest positive deviation is 0.84 eV (molecule 36). Among molecules 25−30, the experimental absorption energies red-shift slightly as the number of repeating units increases. These oligomers do not have good conjugated structures, as the optimized geometries

Density-Functional Theory Calculation

*J. Phys. Chem. A, Vol. 108, No. 40, 2004* **8521**

**TABLE 4: The Structure, Experimental Absorption Energies and the Differences between the Calculated and Experimental Values of 60 Molecules (All Data Are in the Units of eV)**

| No. | Structures | Expt.[a] | Deviation[b] | Deviation[c] | Deviation[d] |
|---|---|---|---|---|---|
| 1 | | 5.10 | 0.83 | 0.32 | 0.21 |
| 2 | | 4.11 | -0.08 | -0.11 | -0.13 |
| 3 | | 3.50 | -0.23 | -0.07 | -0.04 |
| 4 | 1-6 (n=1-6) | 3.18 | -0.34 | -0.07 | -0.07 |
| 5 | | 2.98 | -0.41 | -0.08 | -0.12 |
| 6 | | 2.87 | -0.49 | -0.12 | -0.19 |
| 7 | | 5.96 | 0.80 | 0.08 | 0.02 |
| 8 | | 4.49 | 0.25 | 0.05 | 0.06 |
| 9* | 7-11 (n=1-3, 5 ,7) | 3.91 | 0.02 | 0.02 | 0.09 |
| 10 | | 3.38 | -0.17 | 0.00 | 0.12 |
| 11* | | 3.25 | -0.34 | -0.10 | 0.07 |
| 12 | | 5.93 | 0.58 | -0.07 | -0.14 |
| 13 | | 4.40 | 0.17 | 0.01 | 0.01 |
| 14 | 12-15 (n=1-4) | 3.78 | -0.07 | -0.02 | 0.03 |
| 15 | | 3.43 | -0.20 | -0.02 | 0.04 |
| 16* | 16 | 4.90 | 0.47 | 0.10 | 0.06 |
| 17* | | 3.76 | -0.20 | -0.12 | -0.11 |
| 18 | | 3.19 | -0.18 | 0.04 | 0.04 |
| 19 | 17-19 (n=0-2) | 2.96 | --0.33 | 0.02 | -0.08 |
| 20 | | 3.81 | -0.03 | 0.00 | -0.01 |
| 21 | | 3.23 | -0.11 | 0.08 | 0.09 |
| 22 | 20-23 (n=0-3) | 2.99 | -0.24 | 0.04 | 0.00 |
| 23 | | 2.83 | -0.32 | 0.02 | -0.07 |
| 24 | 24 | 5.58 | 0.46 | -0.08 | -0.10 |
| 25 | | 4.96 | 0.01 | -0.26 | -0.23 |
| 26 | | 4.58 | 0.08 | -0.12 | 0.02 |
| 27 | | 4.44 | -0.23 | -0.32 | -0.15 |
| 28 | 25-30 (n=2-6, 8) | 4.35 | -0.28 | -0.34 | -0.06 |
| 29 | | 4.34 | -0.42 | -0.45 | -0.08 |
| 30 | | 4.32 | -0.50 | -0.52 | 0.03 |
| 31 | | 4.82 | 0.50 | 0.15 | 0.09 |
| 32* | | 3.87 | 0.03 | 0.02 | 0.03 |
| 33 | 31-33 (n=1-3) | 3.10 | 0.05 | 0.22 | 0.18 |
| 34* | 34 | 4.38 | 0.31 | 0.11 | 0.02 |
| 35* | 35 | 3.83 | 0.28 | 0.23 | 0.19 |
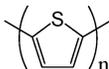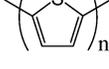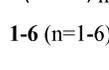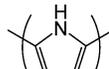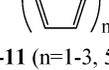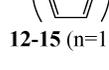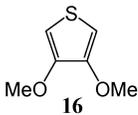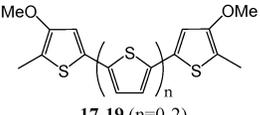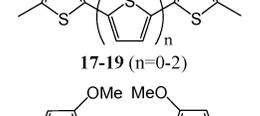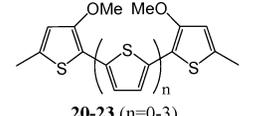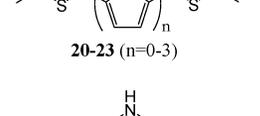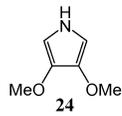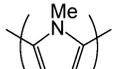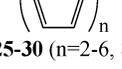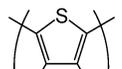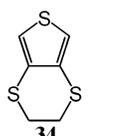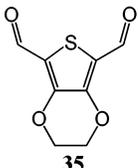
**TABLE 4 (Continued)**

| No. | Structures | Expt.[a] | Deviation[b] | Deviation[c] | Deviation[d] |
|---|---|---|---|---|---|
| 36 | | 5.58 | 0.84 | 0.21 | 0.23 |
| 37* | | 5.93 | 0.53 | -0.12 | -0.07 |
| 38 | | 5.90 | 0.55 | -0.10 | -0.03 |
| 39* | | 3.45 | -0.01 | 0.10 | 0.14 |
| 40 | | 3.60 | -0.05 | 0.03 | 0.10 |
| 41 | | 3.32 | -0.14 | 0.03 | 0.01 |
| 42 | | 3.43 | -0.14 | 0.01 | 0.00 |
| 43 | | 3.63 | -0.18 | -0.08 | 0.07 |
| 44* | | 3.01 | -0.21 | 0.05 | -0.01 |
| 45 | | 2.89 | -0.21 | 0.09 | 0.07 |
| 46 | | 2.73 | -0.28 | 0.07 | 0.03 |
| 47 | | 3.15 | -0.25 | 0.00 | 0.07 |
| 48 | | 2.98 | -0.20 | 0.07 | 0.00 |
| 49 | | 2.69 | -0.24 | 0.11 | 0.07 |
| 50 | | 4.11 | 0.02 | -0.06 | 0.01 |
| 51 | | 3.46 | -0.12 | 0.00 | 0.03 |
| 52 | | 3.21 | -0.33 | -0.10 | -0.13 |
| 53 | | 2.95 | -0.33 | -0.04 | -0.07 |
| 54 | | 4.07 | 0.05 | -0.03 | 0.07 |
| 55 | | 3.43 | -0.10 | 0.01 | 0.01 |
| 56 | | 3.09 | -0.24 | -0.02 | -0.01 |
| 57 | | 3.09 | -0.21 | 0.01 | 0.00 |
| 58 | | 2.71 | -0.11 | 0.17 | 0.14 |
| 59 | | 3.05 | -0.21 | 0.02 | -0.02 |
| 60 | | 2.88 | -0.34 | -0.04 | -0.05 |
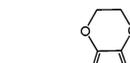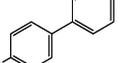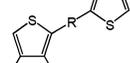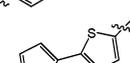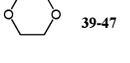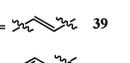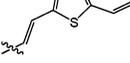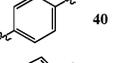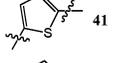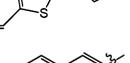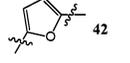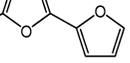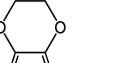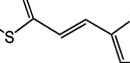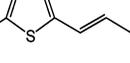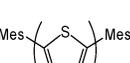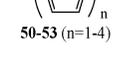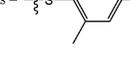
[a] Experimental data are taken from ref 5. [b] Differences between the raw calculated and experimental values. [c] Differences between calculated and experimental values for DFT−MLR calculation. [d] Differences between calculated and experimental values for DFT-NEURON calculation. * Molecules belong to the testing set.

Density-Functional Theory Calculation

*J. Phys. Chem. A, Vol. 108, No. 40, 2004* **8523**
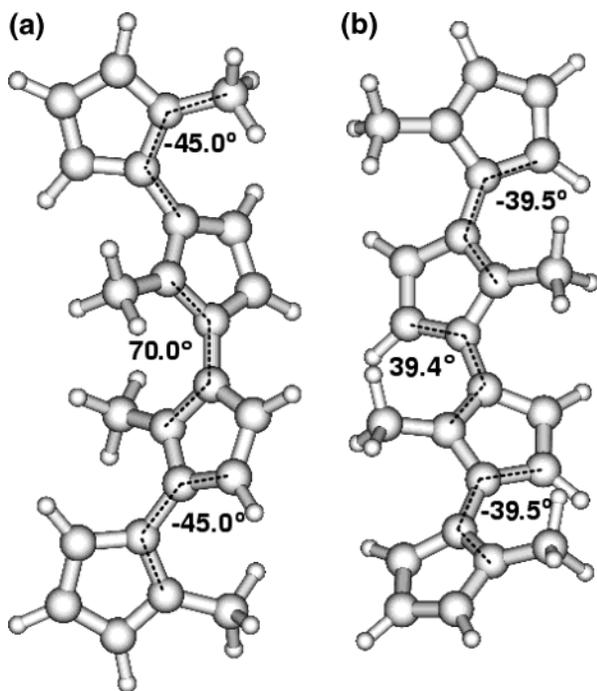


**Figure 8.** Two different possible structures of molecule **27**. (a) Dimeric structure. (b) Nondimeric structure.

have dimeric structures with their average dihedral angles lying between 45.0° and 70.0° (see Figure 8a). Their excited-state energies red-shift slightly compared to their nondimeric isomers (with average dihedral angles of 39.5°, see Figure 8b). We emphasize that the calculated properties of excited states depend on the reliability of geometry optimization.

The raw calculated absorption energy $\Delta$'s versus the experimental measured $\Delta_{ex}$'s are shown in Figure 9a. The vertical coordinate is the experimental $\Delta_{ex}$, and the horizontal coordinate is the calculated $\Delta$. The dashed line is where the vertical and horizontal coordinates are equal.

To employ the MLR or neural networks correction approach to improve the raw calculated $\Delta$, we need to identify the relevant physical descriptors. As discussed in section II, the raw calculated $\Delta$ is the primary descriptor. For the same set of oligomers, when the number of repeating units is small (for example, 1 or 2), the raw calculated absorption energies are higher than the experimental counterparts, and when the number of repeating units is large, the calculated absorption energies red-shift strongly compared to their experimental counterparts. In other words, the oligomer size correlates strongly with the deviation between the raw $\Delta$ and $\Delta_{ex}$. The number of electrons $N_e$ is thus taken as the second physical descriptor. The oscillator strength $O_s$ is a measure of absorption magnitude and is selected as the third and last descriptor. We divide the 60 molecules randomly into a training set (50 molecules) and a testing set (10 molecules), and the training set is further divided into five subsets of equal size for the cross-validation.

Similarly, two hidden neurons yield the overall best results for the neural networks correction approach. The differences between calculated and experimental absorption energies for the MLR and neural networks correction results are tabulated in Table 4. After the MLR and neural networks corrections, RMS deviations of the TDDFT/B3LYP calculations are reduced from 0.33 to 0.14 and 0.09 eV, respectively. For the neural networks correction approach, RMS deviations for the training and testing sets are 0.09 eV, while those of the MLR correction are 0.15 and 0.11 eV, respectively. Figures 9a–c are for the
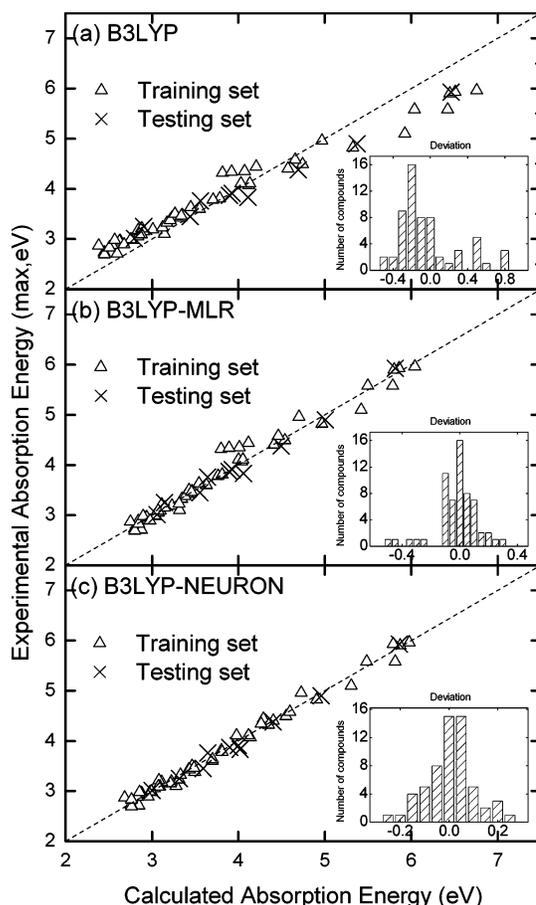


**Figure 9.** Calculated absorption energy versus experimental absorption energy $\Delta_{ex}$ for all 60 compounds. Parts a, b, and c are for raw B3LYP/RPA, and MLR- and neural-networks-corrected B3LYP/RPA results, respectively. Triangles ($\triangle$) are for the training set, and crosses ($\times$) are for the testing set. Insets are the histograms for the differences between the experimental and calculated absorption energies. All values are in units of eV.

raw, MLR-corrected and neural-networks-corrected absorption energies, respectively. Compared to the raw calculated results, the neural-networks-corrected values are much closer to the experimental values for both training and testing sets. More importantly, the systematic deviation in Figure 9a is eliminated. This can be further demonstrated by the error analysis performed for all 60 molecules. In Figure 9, the insets are the histograms for the deviations. Note that the error distribution after the neural networks correction is of an approximate Gaussian-type. For the raw TDDFT calculation, the deviations of smaller oligomers are positive, while those of large oligomers are negative with the magnitude increasing with increasing molecular size. After the neural networks correction, the calculated absorption energies of large and small oligomers have similar magnitudes of deviations.

## IV. Analysis

In Table 5, we list the values of synaptic weights $Wx_{ij}$ and $Wy_j$ for the four neural networks employed. For $\Delta G_f°$, we list the synaptic weights for approach B. To identify the significance of a particular physical descriptor $X_i$ in correcting the raw calculated results, we compute the partial derivative $(\partial Z/\partial X_i)$. In Table 6, we tabulate the resulting $\partial Z/\partial X_i$ for the four neural networks. The derivatives are computed at $X_i = 0.5$. For all cases, $\partial Z/\partial X_1$ has the maximum magnitude, and $X_1$ corresponds to the calculated properties of interest or the primary descriptor.

**TABLE 5: Optimized Values of Synaptic Weights $Wx_{ij}$ and $Wy_j$**

|            | $\Delta G_f^{\circ\ a}$ | IP     | EA     | $\Delta$ |
|------------|--------|--------|--------|--------|
| $Wx_{11}$  | −0.426 | −0.086 | 0.108  | −1.451 |
| $Wx_{12}$  | 0.698  | 0.650  | −0.708 | 0.617  |
| $Wx_{21}$  | 0.154  | −0.032 | −0.128 | −0.419 |
| $Wx_{22}$  | −0.277 | 0.001  | −0.076 | −0.089 |
| $Wx_{31}$  | 0.398  | 0.429  | 0.523  | 1.237  |
| $Wx_{32}$  | 0.001  | 0.145  | 0.359  | 0.243  |
| $Wx_{41}$  | −0.051 | −0.239 | 0.299  | −0.729 |
| $Wx_{42}$  | −0.405 | −0.062 | 0.182  | −0.396 |
| $Wx_{51}$  | 0.002  | 0.187  | 0.006  |        |
| $Wx_{52}$  | 0.393  | −0.248 | 0.105  |        |
| $Wy_1$     | −0.660 | −0.834 | 1.622  | −0.836 |
| $Wy_2$     | 1.601  | 1.781  | −1.724 | 1.333  |

$^a$ Approach B.

**TABLE 6: Derivatives of the Normalized Values of the Observed Properties with Respect to the Normalized Physical Descriptors**

| $i$ | $X_i$ | $\partial\Delta G_f^\circ/\partial X_i{}^a$ | $X_i$ | $\partial(\mathrm{IP})/\partial X_i$ | $X_i$ | $\partial(\mathrm{EA})/\partial X_i$ | $X_i$ | $\partial\Delta/\partial X_i$ |
|---|-------|--------|-------------|--------|-------------|--------|-------|
| 1 | $\Delta G_f^\circ$ | 1.444  | IP    | 1.315  | EA    | 1.153  | $\Delta$ | 1.767 |
| 2 | $N_t$ | −0.872 | $N_{ve}$ | 0.017  | $N_{ve}$ | 0.024  | $N_e$ | 0.232 |
| 3 | ZPE   | 0.798  | $g_s$ | −0.153 | $g_s$ | −0.052 | $O_s$ | 0.262 |
| 4 | $N_H$ | −0.155 | $E_g$ | −0.048 | $E_g$ | 0.060  |       |       |

$^a$ Approach B.

For IP, EA, or absorption energy, $|\partial Z/\partial X_1|$ is significantly larger than $|\partial Z/\partial X_2|$, $|\partial Z/\partial X_3|$, or $|\partial Z/\partial X_4|$, which indicates that the raw calculated IP, EA, and absorption energy are indeed the most important physical descriptors in determining the exact values of IP, EA, and absorption energy, respectively. For $\Delta G_f^\circ$, $|\partial Z/\partial X_2|$ and $|\partial Z/\partial X_3|$ are slightly larger than one-half of $|\partial Z/\partial X_1|$. However, $\partial Z/\partial X_3$ has the opposite sign of $\partial Z/\partial X_2$ and $\partial Z/\partial X_4$. $N_t$, $N_H$, and ZPE are highly correlated and are approximately proportional to each other. $|\partial Z/\partial X_2 + \partial Z/\partial X_3 + \partial Z/\partial X_4| \approx 0.23$, which is much less than $|\partial Z/\partial X_1| \approx 1.44$ for $\Delta G_f^\circ$. This verifies that the raw calculated $\Delta G_f^\circ$ is much more important than the other three descriptors. All of these confirm that the calculated value captures the essence of the property of interest and that the other physical descriptors provide finer tuning. Note that the derivatives with respect to $N_{ve}$ are very small for IP and EA. This indicates that $N_{ve}$ is not an important physical descriptor for improving the calculated IP or EA and may thus be neglected. Indeed, we employ the 4-2-1 neural networks for IP and EA by employing only the physical descriptors IP/EA, $g_s$, and $E_g$, and find that their resulting RMS deviations are 3.3 and 1.9 kcal·mol$^{-1}$, respectively, which are virtually the same as those of the 5-2-1 neural network.

For MLR correction, the partial correction coefficient measures the significance of each descriptor and is expressed as follows

$$V_j = \sqrt{1 - q/Q_j}$$

where $j = 1, 2, ..., m$

$$Q_j = \sum_{i=1}^{n}\left[Y_i - \left(C_0 + \sum_{\substack{k=1 \\ k \neq j}}^{m} C_k X_{ki}\right)\right]^2$$

$$q = \sum_{i=1}^{n}\left[Y_i - \left(C_0 + \sum_{k=1}^{m} C_k X_{ki}\right)\right]^2$$

The larger $V_j$ is, the more important the descriptor $X_j$ is. In Table 7, we tabulate the partial correction coefficients in the

**TABLE 7: The Partial Correction Coefficients in MLR for Different Properties with Respect to the Normalized Physical Descriptors**

| $I$ | $X_i$ | $V_i^a$ | $X_i$ | $V_i^b$ | $X_i$ | $V_i^c$ | $X_i$ | $V_i^d$ |
|---|-------|--------|-------|--------|-------|--------|-------|--------|
| 1 | $\Delta G_f^\circ$ | 0.998 | IP    | 1.000 | EA    | 0.998 | $\Delta$ | 0.999 |
| 2 | $N_t$ | 0.998  | $N_{ve}$ | 0.710 | $N_{ve}$ | 0.190 | $N_e$ | 0.331 |
| 3 | ZPE   | 0.998  | $g_s$ | 0.731 | $g_s$ | 0.368 | $O_s$ | 0.065 |
| 4 | $N_H$ | 0.986  | $E_g$ | 0.081 | $E_g$ | 0.225 |       |       |

$^a$ The partial correction coefficients for $\Delta G_f^\circ$ in approach B. $^b$ The partial correction coefficients for IP. $^c$ The partial correction coefficients for EA. $^d$ The partial correction coefficients for absorption energy.

MLR correction approach. $X_1$ always has the maximum partial correction coefficient for the MLR correction, which is consistent with that of the neural networks correction approach. For $\Delta G_f^\circ$, all four descriptors have very large partial correction coefficients, which illuminates the fact that all four descriptors are very important in correcting the raw $\Delta G_f^\circ$.

## V. Discussion and Conclusion

Compared to what we did in ref 8, we have greatly generalized the DFT-NEURON method and its applications. Besides the various ground-state properties, we have demonstrated that it can be used to improve the calculated properties of the excited states. The statistical correction approach developed here can be further generalized. For instance, it can be extended to construct the exchange-correlation functional approach for DFT and TDDFT.[22,23] DFT maps a many-electron problem onto an effective single-electron problem. The mapping is exact if the exact exchange-correlation functional is known. Unfortunately, only approximated exchange-correlation functionals exist. Neural networks or linear regression fit can be employed to construct accurate exchange-correlation functional values by discovering regularities among the available experimental data. Work along this direction is in progress.[23]

Although the MLR correction approach may improve the raw calculated results, the DFT-NEURON method usually yields better values. This is because the neural networks method is much more versatile. The MLR correction approach works well when the raw calculated results are very close to the exact values. Unfortunately, this is often not the case. The raw calculated values can deviate considerably from the experimental values. Correction beyond the linear regression fit is thus required. The neural networks method provides a better and more general solution. Another important feature of the statistical-correction-based first-principles methods is that they do not discriminate against large molecules, because the bias at the raw calculation level has been corrected at the training stage.

So far, one neural network or MLR fit is to be trained, tested, and determined for each property of interest. This is tedious and time-consuming. We are working to construct a general statistical correction model that improves the calculated energy directly. Because all physical properties are ultimately determined by energy, this general statistical correction model could be used to calculate accurately the properties other than energy itself. We thus need to train only one statistical correction model, instead of one model per physical property.

To summarize, we have developed the DFT-NEURON method into any statistical-correction-based first-principles method and expanded it to compute a variety of ground- and excited-state properties of small-to-medium-sized molecules or atoms. The accuracy of any statistical-correction-based first-principles method can be systematically improved as more or better experimental data are available. This combined first-principles calculation and statistical correction approach is

potentially a powerful tool in computational science, and it may open the possibility for first-principles methods to be employed routinely as predictive tools in materials research and development.

**Supporting Information Available:** Table for the experimental $\Delta G_f°$ (298 K) values and the differences between the calculated and experimental values for 180 compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

## References and Notes

(1) Cramer, C. J. *Essentials of computational chemistry: theories and models*; John Wiley: West Sussex, England, 2002.

(2) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.

(3) Foresman, J. B.; Frisch, A. *Exploring chemistry with electronic structure method*, 2nd ed.; Gaussian, Inc.: Pittsburgh, PA, 1996.

(4) Irikura, K. K.; Frurip, D. J. *Computational thermochemistry: prediction and estimation of molecular thermodynamics*; American Chemical Society: Washington, DC, 1998.

(5) Hutchison, G. R.; Ratner, M. A.; Marks, T. J. *J. Phys. Chem. A* **2002**, *106*, 10596.

(6) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J Chem. Phys.* **1998**, *109*, 7764.

(7) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *Chem. Phys. Lett.* **1997**, *270*, 419.

(8) Hu, L. H.; Wang, X. J.; Wong, L. H.; Chen, G. H. *J. Chem. Phys.* **2003**, *119*, 11501.

(9) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

(10) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Nature* **1986**, *323*, 533.

(11) Haykin, S. *Neural networks: a comprehensive foundation*; Prentice Hall: Upper Saddle River, NJ, 1999.

(12) Yao, X.; Zhang, X.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. *Comput. Chem.* **2001**, *25*, 475.

(13) Pompe, M.; Novič, M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 59.

(14) Dong, L.; Yan, A.; Chen, X.; Xu H.; Hu, Z. *Comput. Chem.* **2001**, *25*, 551.

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

(16) Lide, D. R. *CRC Handbook of Chemistry and Physics*, 82nd Edition; CRC Press: Boca Raton, FL, 2002.

(17) Dean, J. A. *Lange's Handbook of Chemistry*, 15th ed.; McGraw-Hill: New York, 1999.

(18) Chase, M. W.; Davies, C. A.; Downey, J. R.; Frurip, D. J.; McDonald, R. A.; Syverud, A. N. *J. Phys. Chem. Ref. Data* **1985**, *14*, Supplement 1.

(19) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*; Gaussian, Inc.: Pittsburgh, PA, 1998.

(20) Hus, C.-P.; Hirata, S.; Head-Gordon, M. *J. Phys. Chem. A* **2001**, *105*, 451.

(21) Hirata, S.; Lee, T. J.; Head-Gordon, M. *J. Chem. Phys.* **1999**, *111*, 8904.

(22) Tozer, D. J.; Ingamells, V. E.; Handy, N. C. *J. Chem. Phys.* **1996**, *105*, 9200.

(23) Zheng, X.; Hu, L. H.; Wang, X. J.; Chen, G. H. *Chem. Phys. Lett.* **2004**, *390*, 186.