Molecular Simulation

ISSN: 0892-7022 (Print) 1029-0435 (Online) Journal homepage: http://www.tandfonline.com/loi/gmos20

A Combined First-principles Calculation and Neural Networks Correction Approach for Evaluating Gibbs Energy of Formation

XiuJung Wang , LiHong Hu , LaiHo Wong & GuanHua Chen

To cite this article: XiuJung Wang , LiHong Hu , LaiHo Wong & GuanHua Chen (2004) A Combined First-principles Calculation and Neural Networks Correction Approach for Evaluating Gibbs Energy of Formation, Molecular Simulation, 30:1, 9-15, DOI: 10.1080/08927020310001631098

To link to this article: https://doi.org/10.1080/08927020310001631098

Published online: 15 Aug 2006.

Molecular Simulation, Vol. 30 (1), 15 January 2004, pp. 9–15

A Combined First-principles Calculation and Neural Networks Correction Approach for Evaluating Gibbs Energy of Formation

XIUJUNG WANG, LIHONG HU, LAIHO WONG and GUANHUA CHEN*

Department of Chemistry, The University of Hong Kong, Pokfulam Road, Hong Kong, People's Republic of China

(Received September 2003; In final form October 2003)

Despite of their successes, the results of first-principles quantum mechanical calculations contain inherent numerical errors that are caused by inadequate treatment of electron correlation, incompleteness of basis sets, relativistic effects or approximated exchange-correlation functionals. In this work, we develop a combined density-functional theory and neural-network correction (DFT-NEURON) approach to reduce drastically these errors, and apply the resulting approach to determine the standard Gibbs energy of formation ΔG^0 at 298 K for small- and medium-sized organic molecules. The root mean square deviation of the calculated ΔG^0 for 180 molecules is reduced from 22.3 $kcal \cdot mol^{-1}$ to 3.0 $kcal \cdot mol^{-1}$ for B3LYP/6-311 + G(*d*,*p*). We examine further the selection of physical descriptors for the neural network.

Keywords: DFT; Neural network; Gibbs energy of formation; First-principles quantum mechanical methods

First-principles quantum mechanical methods have become indispensable research tools in chemistry, condensed matter physics, materials science and molecular biology [1,2]. Experimentalists increasingly rely 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 experimental measurements, in particular, for large systems. This is caused by the inherent approximations adopted in the first-principles methods. Because of the computational costs, electron correlation has always been a difficult obstacle for first-principles calculations. For instance, highly accurate full configuration interaction calculations have been limited to very small molecules [3]. Basis sets cannot cover entire physical space, and this introduces inherent computational errors [4]. In practice, limited by the computational resources we often adopt inadequate basis sets for large molecules. Effective core potential is frequently used to approximate the relativistic effects, which leads to inevitably the approximated calculation results for heavy element containing systems. Accuracy of a density-functional theory (DFT) calculation is mainly determined by the exchange-correlation (XC) functional that is employed [2]. No exact XC functional is known. All DFT calculations employ the approximated XC functionals, which lead to further calculation errors. Besides these inherent approximations, quantum mechanical calculations are often limited by finite computational resources and thus adopt less accurate methods. All these contribute to the discrepancies between the calculated and experimental results. An admirable objective 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 greatly the remaining numerical uncertainties. In addition, we need to quantify the accuracies of the numerical methods. Determining the calculation errors solely from the first principles has proven to be an extremely difficult task, and alternatives must be sought.

Despite the various approximations that the first-principles quantum mechanical calculations adopt, the calculated results capture the essence of the physics. For instance, although their absolute values

*Corresponding author. Fax: +852-2857-1586. E-mail: ghc@yangtze.hku.hk

ISSN 0892-7022 print/ISSN 1029-0435 online © 2004 Taylor & Francis Ltd
DOI: 10.1080/08927020310001631098

may not agree with the experimental data, the calculated results of different molecules often have the same relative tendency as the corresponding experimental results. To predict a physical property of a material, it may thus be sufficient to correct its raw result from the first-principles calculation. The discrepancies between the calculated and experimental results depend on the physical properties of the material. These physical properties include predominantly the property of interest and, to a lesser degree, other related properties of the materials which can again be evaluated via first-principles calculations. In other words, a quantitative relationship exists between the experimental results and calculated properties. Although it is exceedingly difficult to be determined from the first-principles, the quantitative relationship can be obtained empirically. We propose here a neural-network [5–8] based approach, DFT-NEURON, to determine this quantitative relationship which can be subsequently used to reduce drastically the numerical errors of first-principles calculations. As a demonstration, the standard Gibbs energies of formation ΔG^0s at 298.15 K of 180 small- or medium-sized organic molecules are evaluated via B3LYP calculations and subsequently corrected by the DFT-NEURON approach. Different physical descriptors are adopted to examine the validity of the DFT-NEURON approach.

The 180 small- or medium-sized organic molecules listed in Table I, are selected for the DFT-NEURON approach that combines first-principles calculation and neural network based correction. These molecules can be found in all three Refs. [9–11]. These molecules contain elements such as H, C, N, O, F, Si, S, Cl and Br. The heaviest molecule contains 14 heavy atoms, and the largest contains 32 atoms. We divide these molecules randomly into the training set containing 150 molecules and the testing set containing 30 molecules. The geometries of 180 molecules are optimized via B3LYP/6-311 + G(*d*,*p*) [12] calculations and zero-point energies (ZPE) are calculated subsequently at the same level. The Gibbs energy of formation of each molecule is calculated at both B3LYP/6-311 + G(*d*,*p*) and B3LYP/6-311 + G(3*df*,2*p*) levels. To calculate ΔG^0 of a molecule, we calculate first its ΔH^0 and its entropy S^0. For instance, for a molecule A_xB_y, its ΔG^0 at 298 K can be evaluated as:

$$\begin{aligned}\Delta G^0(\mathrm{A}_x\mathrm{B}_y, 298\,K) &= \Delta H^0(\mathrm{A}_x\mathrm{B}_y, 298\,K) - 298.15 \\ &\times \Big(S^0\big(\mathrm{A}_x\mathrm{B}_y, 298\,K\big)_{\mathrm{st}} - \Big(xS^0\big(\mathrm{A}, 298\,K\big)_{\mathrm{st}} \\ &+ yS^0\big(\mathrm{B}, 298\,K\big)_{\mathrm{st}}\Big)\Big) \qquad (1)\end{aligned}$$

where ΔH^0 (A_xB_y, 298 K) is the standard enthalpy of formation for A_xB_y at 298 K. S^0 (A_xB_y, 298 K) is the entropy of A_xB_y at 298 K, S^0 (A, 298 K)$_{\mathrm{st}}$ and S^0 (B, 298 K)$_{\mathrm{st}}$ are the standard entropies of species A and B relative to the elements in their reference states, respectively. S^0 (A, 298 K)$_{\mathrm{st}}$ and S^0 (B, 298 K)$_{\mathrm{st}}$ are taken from Ref. [13]. The strategies in Ref. [14] are adopted to calculate ΔH^0. B3LYP/6-311 + G(3*df*,2*p*) employs a larger basis set than B3LYP/6-311 + G(*d*,*p*). The unscaled B3LYP/6-311 + G(*d*,*p*) ZPE is employed in the ΔG^0 calculation. The ΔG^0 values for both B3LYP/6-311 + G(*d*,*p*) and B3LYP/6-311 + G(3*df*,2*p*) are compared to their experimental data in Fig. 1a,b, respectively. The horizontal coordinates are the experimental ΔG^0, and the vertical 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 above the dashed line, i.e. most calculated ΔG^0 are larger than the experimental data. Compared to the experimental measurements, the root mean square (RMS) deviations are 22.3 and 12.9 kcal · mol^{-1} for the raw B3LYP/6-311 + G(*d*,*p*) and B3LYP/6-311 + G(3*df*,2*p*) results, respectively. In Table II, we list the experimented and calculated ΔG^0 for 16 of 180 molecules. Overall, B3LYP/6-311 + G(3*df*,2*p*) calculations yield better agreements with the experiments than B3LYP/6-311 + G(*d*,*p*). In particular, for small molecules with few heavy elements B3LYP/6-311 + G(3*df*,2*p*) calculations result in very small deviations from the experiments. For instance, the deviations for CH_2F_2 and CS_2 are only 0.3 and 0.5 kcal · mol^{-1}, respectively. For large molecules, both B3LYP/6-311 + G(*d*,*p*) and B3LYP/6-311 + G(3*df*,2*p*) calculations yield quite large deviations from their experimental data.

Next we construct our neural network [15]. There is an input layer consisted of input from physical descriptors and a bias, a hidden layer, and an output layer [16]. The most important issue is to select the proper physical descriptors of our molecules, which are to be used as the input for our neural network. The calculated Gibbs energy of formation ΔG^0 contains the essential value of the experimental ΔG^0, and is thus an obvious choice for the primary descriptor. We observe that the size of a molecule affects the accuracy of the calculated ΔG^0. The more atoms a molecule has the worse calculated ΔG^0 is. This is consistent with the general observations in the field [14]. The total number of atoms N_t in a molecule is thus chosen as the second descriptor for the molecule. ZPE is an important component of ΔG^0. Its calculated value is often scaled in evaluating ΔH^0 [14,17], which has the dominant contribution for ΔG^0. It is thus taken as the third physical descriptor. Hydrogen atom is much lighter than other organic elements, and has the most quantum characteristics among all the elements. Its contributions to energy

TABLE I 180 molecules employed in the calculation

Formula	*Name*	*Formula*	*Name*	*Formula*	*Name*
$CBrF_3$	Bromotrifluoromethane	$CClF_3$	Chlorotrifluoromethane	CClN	Cyanogens chloride
CCl_2O	Phosgene	CF_2O	Carbonyl fluoride	CF_4	Carbon tetrafluoride
$CHCl_3$	Chloroform	CH_2Cl_2	Dichloromethane	CH_2F_2	Difluoromethane
CH_2O_2	Formic acid	CH_3Br	Methyl bromide	CH_3NO_2	Nitromethane
CH_3NO_2	Methyl nitrite	CH_4	Methane	CH_4O	Methanol
CH_4S	Methyl mercaptan	CH_5N	Methylamine	COS	Carbonyl sulfide
CS_2	Carbon disulfide	C_2H_2	Acetylene	$C_2H_2Cl_2$	1,1-Dichloroethylene
$C_2H_2O_4$	Oxalic acid	C_2H_3Br	Vinyl bromide	C_2H_3ClO	Acetyl chloride
C_2H_3F	Vinyl fluoride	C_2H_4	Ethylene	$C_2H_4Br_2$	1,2-Dibromoethane
$C_2H_4Cl_2$	1,1-Dichloroethane	C_2H_4O	Ethylene oxide	$C_2H_4O_2$	Acetic acid
C_2H_4S	Thiacyclopropane	C_2H_5Br	Bromoethane	C_2H_5N	Ethyleneimine
C_2H_5NO	Acetamide	$C_2H_5NO_2$	Nitroethane	C_2H_6	Ethane
C_2H_6O	Dimethyl ether	C_2H_6S	Dimethyl sulfide	C_2H_7N	Ethylamine
$C_2H_8N_2$	Ethylenediamine	C_2N_2	Cyanogen	C_3H_3NO	Oxazole
C_3H_4	Methylacetylene	C_3H_4	Propadiene	$C_3H_4O_3$	Ethylene carbonate
$C_3H_5Cl_3$	1,2,3-trichloropropane	C_3H_6	Cyclopropane	$C_3H_6Cl_2$	1,2-Dichloropropane
C_3H_6O	Acetone	$C_3H_6O_2$	Methyl acetate	$C_3H_6O_2$	Propionic acid
C_3H_6S	Thiacyclobutane	C_3H_7Br	1-Bromopropane	C_3H_7Br	2-Bromopropane
C_3H_7F	1-Fluoropropane	C_3H_7NO	*N*,*N*-dimethylformamide	$C_3H_7NO_2$	2-Nitropropane
$C_3H_7NO_3$	Propyl nitrate	$C_3H_7NO_3$	Isopropyl nitrate	C_3H_8	Propane
C_3H_8O	Methyl ethyl ether	C_3H_8S	*n*-Propyl mercaptan	C_3H_8S	Isopropyl mercaptan
C_3H_8S	Ethyl methyl sulfide	C_3H_9N	*n*-Propylamine	C_3H_9N	Isopropylamine
C_3H_9N	Trimethylamine	$C_3H_{10}N_2$	1,2-Propanediamine	$C_4H_4N_2$	Succinonitrile
C_4H_6	1,2-Butadiene	C_4H_8	1-Butene	C_4H_8O	Isobutyraldehyde
$C_4H_8O_2$	Ethyl acetate	C_4H_9Cl	tert-Butyl chloride	$C_4H_{10}O$	sec-Butanol
$C_4H_{10}O_2$	1,4-Butanediol	$C_4H_{10}S$	Isobutyl mercaptan	$C_4H_{10}S$	Methyl propyl sulfide
$C_4H_{11}N$	tert-Butylamine	C_5H_5N	Pyridine	C_5H_6S	2-Methylthiophene
$C_5H_8O_2$	Acetylacetone	C_5H_{10}	Cyclopentane	C_5H_{10}	2-Methyl-1-butene
C_5H_{10}	2-Methyl-2-butene	C_5H_{10}	3-Methyl-1-butene	C_5H_{10}	*trans*-2-Pentene
$C_5H_{10}O$	2-Pentanone	$C_5H_{10}O$	Valeraldehyde	$C_5H_{10}O_2$	Valeric acid
$C_5H_{10}S$	Thiacyclohexane	$C_5H_{10}S$	cyclopentanethiol	$C_5H_{11}Br$	1-Bromopentane
$C_5H_{11}Cl$	1-Chloropentane	$C_5H_{11}N$	Piperidine	C_5H_{12}	Isopentane
C_5H_{12}	*n*-Pentane	$C_5H_{12}O$	3-Methyl-1-butanol	$C_5H_{12}O$	3-Methyl-2-butanol
$C_5H_{12}O$	2-Pentanol	$C_5H_{12}O$	3-Pentanol	$C_5H_{12}O$	Ethyl propyl ether
$C_5H_{12}O_4$	Pentaerythritol	$C_5H_{12}S$	*n*-Pentyl mercaptan	C_6F_6	Hexafluorobenzene
$C_6H_4C_{l2}$	m-Dichlorobenzene	$C_6H_4F_2$	*p*-Difluorobenzene	C_6H_5Cl	Monochlorobenzene
C_6H_5F	Fluorobenzene	$C_6H_5NO_2$	Nitrobenzene	C_6H_6	Benzene
$C_6H_6N_2O_2$	m-Nitroaniline	C_6H_6O	Phenol	$C_6H_6O_2$	1,3-Benzenediol
C_6H_7N	2-Methylpyridine	$C_6H_{10}O_3$	Propionic anhydride	$C_6H_{11}NO$	e-Caprolactam
C_6H_{12}	*trans*-3-Hexene	$C_6H_{12}O$	Butyl vinyl ether	$C_6H_{12}O$	3-Hexanone
C_6H_{14}	3-Methylpentane	$C_6H_{14}S$	Methyl pentyl sulfide	C_7H_5N	Benzonitrile
C_7H_6O	Benzaldehyde	C_7H_8	Toluene	C_7H_8O	o-Cresol
C_7H_9N	2,6-Dimethylpyridine	C_7H_{14}	*cis*-1,2-Dimethylcyclopentane	$C_7H_{15}Br$	1-Bromoheptane
C_7H_{16}	3,3-Dimethylpentane	C_7H_{16}	2,2,3-Trimethylbutane	$C_7H_{16}S$	*n*-Heptyl mercaptan
$C_8H_6O_4$	Terephthalic acid	C_8H_8O	Acetophenone	C_8H_{16}	*cis*-1,2-Dimethylcyclohexane
C_8H_{16}	*trans*-1,4-Dimethylcyclohexane	C_8H_{18}	2,3-Dimethylhexane	C_8H_{18}	3-Ethylhexane
C_8H_{18}	4-Methylheptane	$C_8H_{18}S_2$	Dibutyl disulfide	C_9H_{12}	m-Ethyltoluene
C_9H_{12}	1,2,3-Trimethylbenzene	$C_9H_{18}O$	Diisobutyl ketone	C_9H_{20}	3,3-Diethylpentane
C_9H_{20}	2,2,3,4-Tetramethylpentane	$C_{10}H_{14}$	sec-Butylbenzene	$C_{10}H_{14}$	Isobutylbenzene
$C_{10}H_{18}O_4$	Sebacic acid	$C_{10}H_{20}O_2$	*n*-Decanoic acid	$C_{12}H_{10}$	Acenaphthene
$CBrC_{l3}$	Bromotrichloromethane	CHF_3	Trifluoromethane	$C_2H_2F_2$	1,1-Difluoroethylene
$C_2H_3ClO_2$	Chloroacetic acid	$C_2H_3Cl_3$	1,1,1-Trichloroethane	$C_2H_4Cl_2$	1,2-Dichloroethane
$C_2H_4F_2$	1,1-Difluoroethane	C_2H_5Cl	Ethyl chloride	$C_2H_5NO_3$	Ethyl nitrate
C_2H_7N	Dimethylamine	C_3H_6	Propylene	$C_3H_6Br_2$	1,2-Dibromopropane
C_3H_7Cl	Isopropyl chloride	C_3H_7Cl	*n*-Propyl chloride	$C_3H_7NO_2$	1-Nitropropane
C_4H_6O	Divinyl ether	C_4H_9Br	1-Bromobutane	C_5H_8	*trans*-1,3-Pentadiene
C_5H_{10}	1-Pentene	C_5H_{10}	*cis*-2-Pentene	$C_5H_{12}O$	2-Methyl-1-butanol
$C_5H_{12}S$	Butyl methyl sulfide	$C_6H_8N_2$	Adiponitrile	C_6H_{10}	1-Methylcyclopentene
C_6H_{10}	1,5-Hexadiene	C_8H_{10}	o-Xylene	$C_8H_{10}O$	3,4-Xylenol
C_8H_{16}	2,4,4-Trimethyl-2-pentene	C_8H_{18}	2,3,4-Trimethylpentane	$C_8H_{18}O$	2-Ethyl-1-hexanol

The upper panel is for the training set, and the lower panel is for the testing set.

and entropy are different from heavy elements, and need to be distinguished from other types of atoms. We thus set the number of hydrogen atoms of a molecule Nh as the fourth descriptor.

To ensure the quality of our Neural Network, a cross-validation procedure is adopted to determine our neural network [7,18–20]. We divide further randomly 150 training molecules into five subsets of equal size. Four of them are used to train the neural network, the fifth to validate its predictions. This procedure is repeated 5 times in rotation. The number of neurons in the hidden layer is

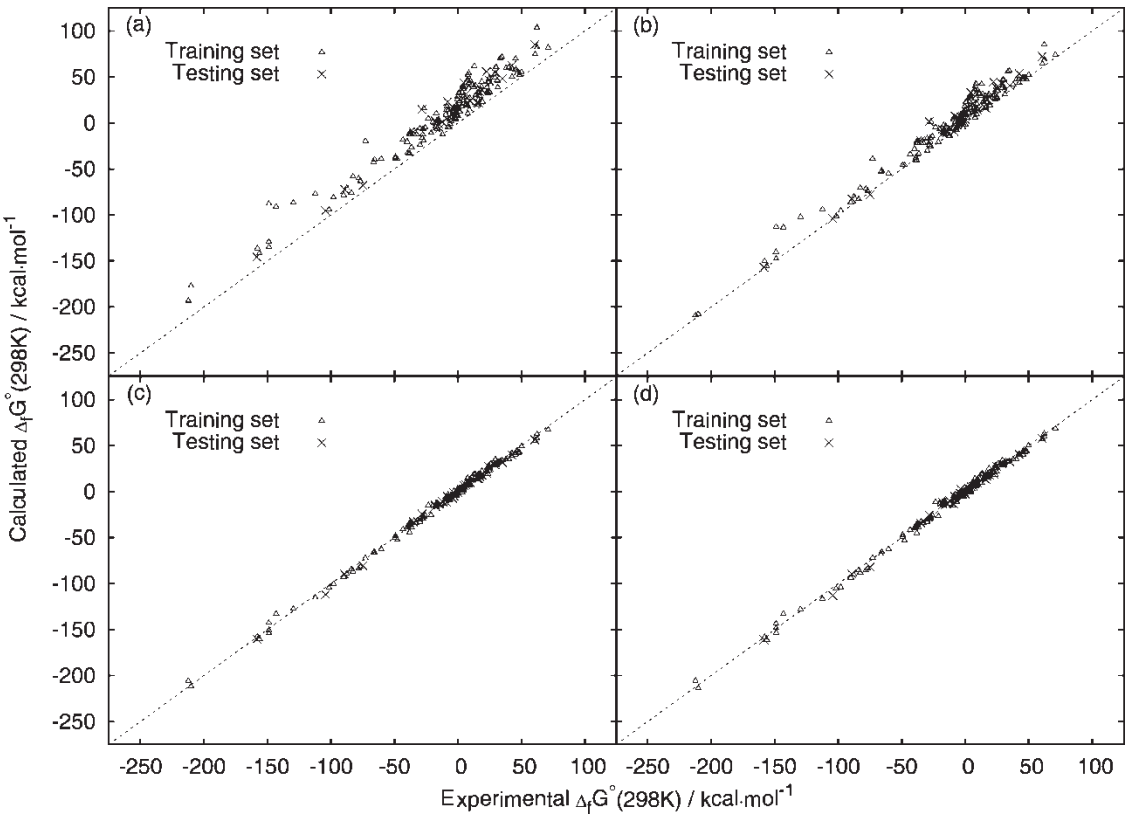

FIGURE 1 Calculated ΔG^0 versus experimental ΔG^0 for all 180 compounds. (a) and (b) are for raw B3LYP/6-311 + G(*d*,*p*) and B3LYP/6-311 + G(3*df*,2*p*) results, respectively. (c) and (d) are for neural-network corrected B3LYP/6-311 + G(*d*,*p*) and B3LYP/6-311 + G(3*df*,2*p*) ΔG^0s, respectively. Triangles are for the training set and crosses for the testing set. The dashed lines are where the calculations and experiments have perfect fits. The correlation coefficient of the linear fits are 0.996, 0.995 in (c) and (d), respectively. All values are in the units of kcal · mol^{-1}.

varied from 2 to 10 and the number of descriptors in the input layer is switched from 2 to 4 in order to determine the optimal structure of our neural network. We find that an input layer containing four descriptors (ΔG^0, N_t, ZPE and Nh) and a hidden layer containing two neurons yield the best overall results. The 5-2-1 structure is thus adopted for our neural network (See Fig. 2). It is found that

TABLE II Experimental and calculated Gibbs energies of formation for sixteen selected compounds (all data are in the units of kcal · mol^{-1})

			Deviation (Theory − Experiment)			
Formula	*Name*	*Experiment* ΔG^0 *(298.15 K)**	*DFT1*†	*DFT1-NN*‡	*DFT2*¶	*DFT2-NN*§
$CBrF_3$	Bromotrifluoromethane	− 148.82 ± 0.7	14.0	− 5.2	1.0	− 4.9
$CClF_3$	Chlorotrifluoromethane	− 156.30 ± 0.8	14.4	− 4.2	0.3	− 4.8
CF_2O	Carbonyl fluoride	− 148.99 ± 0.4	19.8	6.4	8.5	5.4
CF_4	Carbon tetrafluoride	− 212.34 ± 0.3	19.0	6.5	3.3	6.6
$CHCl_3$	Chloroform	− 16.38 ± 0.3	16.1	4.4	8.0	6.4
CHF_3	Trifluoromethane	− 158.48 ± 0.8	12.8	− 1.8	1.5	− 2.7
CH_2Cl_2	Dichloromethane	− 16.46 ± 0.3	10.2	3.0	4.6	4.3
CH_2F_2	Difluoromethane	− 101.66 ± 0.4	7.8	− 2.8	0.3	− 4.0
CH_4	Methane	− 12.15 ± 0.1	− 0.2	− 0.7	− 1.9	− 1.4
CS_2	Carbon disulfide	15.99 ± 0.2	8.6	2.2	0.5	1.5
C_5H_{12}	*n*-Pentane	− 2.00 ± 0.4	17.4	− 1.6	10.6	− 1.9
$C_5H_{12}O$	3-Methyl-2-butanol	− 37.37 ± 0.7	25.1	1.3	15.8	0.6
$C_5H_{12}O$	2-Pentanol	− 38.07 ± 0.6	25.7	2.3	16.5	1.4
C_6H_{14}	3-Methylpentane	− 0.51 ± 0.4	26.2	1.5	18.1	1.1
C_8H_{10}	o-Xylene	29.18 ± 0.6	25.8	− 0.2	13.4	0.6
C_9H_{12}	1,2,3-Trimethylbenzene	29.77 ± 0.6	31.1	− 0.9	17.3	− 0.7

*The experimental values are taken from Ref. [9] and the uncertainties are estimated from Refs. [13,21,22]. †The deviations of calculated ΔG^0 by using B3LPY/6-311 + G(*d*,*p*) geometries, zero point energies. ‡The deviations of calculated ΔG^0 by B3LYP/6-311 + G(*d*,*p*)—Neural Networks approach. ¶The deviations of calculated ΔG^0 by using the 6-311 + G(*d*,*p*) geometries, zero point energies and recalculated total energies by 6-311 + G(3df,2p) basis. §The deviations of calculated ΔG^0 by B3LYP/6-311 + G(3*df*,2*p*)—Neural Networks approach.

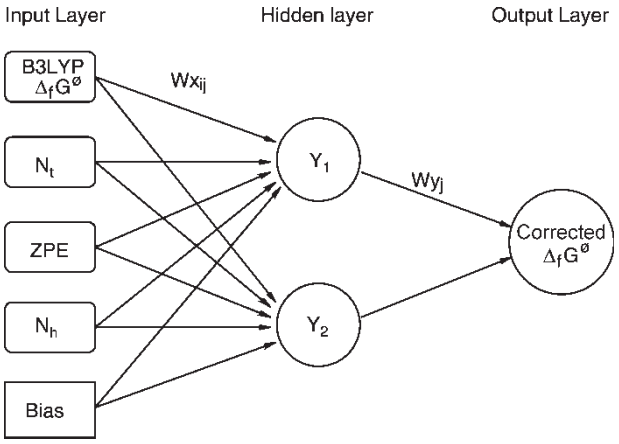

FIGURE 2 The structure of our neural network. The input for the bias is always 1.

our neural network can be tuned to fit the trained data and the corresponding experimental data very well for the training set. Moreover, for the testing set our neural network predicts the values that agree very well with experiments. The quantitative relationship between the calculated and the experimental data is thus established. In Fig. 1c,d, the triangles and crosses belong to the training and testing set, respectively. Compared to the raw calculated results, the neural-network corrected values are much closer to the experimental values for both training and testing sets. More importantly, the systematic deviations in Fig. 1a,b are eliminated, and the remaining numerical errors are reduced substantially. This can be further demonstrated by the error analysis performed for B3LYP/6-311 + G(*d*,*p*) ΔG^0 of all 180 molecules. In Fig. 3, we plot the histograms for the deviations (from the experiments) of the raw B3LYP/6-311 + G(*d*,*p*) ΔG^0s and their neural-network corrected values. Figure 3a,c shows the raw ΔG^0s and neural-network corrected ΔG^0s of training set, respectively. Figure 3b,d shows the raw ΔG^0s and neural-network corrected ΔG^0s of testing set, respectively. For the training set, the RMS errors before and after the neural-network correction are 22.6 and 3.1 $\text{kcal} \cdot \text{mol}^{-1}$, respectively; while for the testing set they are 20.7 and 2.8 $\text{kcal} \cdot \text{mol}^{-1}$, respectively. Overall, the raw calculated ΔG^0s have a large systematic deviations of 22.3 $\text{kcal} \cdot \text{mol}^{-1}$ while the neural-network corrected ΔG^0s have virtually no systematic deviation at all. Moreover, the remaining numerical deviations are much smaller for the neural-network corrected ΔG^0s, i.e. the RMS deviations is reduced from 22.3 to 3.0 $\text{kcal} \cdot \text{mol}^{-1}$ upon neural-network correction. Note that the error distribution after the neural-network correction is of an approximate Gaussian type (see Fig. 3c,d). We have performed the same error analysis for B3LYP/6-311 + G(3*df*,2*p*) ΔG^0s and have reached a similar conclusion. The RMS deviation for B3LYP/6-311 + G(3*df*,2*p*) ΔG^0s is reduced from 13.2 to 3.4 $\text{kcal} \cdot \text{mol}^{-1}$ upon neural-network correction. In Table II, we list the neural-network corrected ΔG^0s. The deviations of large molecules are of the same magnitude as those of small molecules. Our neural-network correction

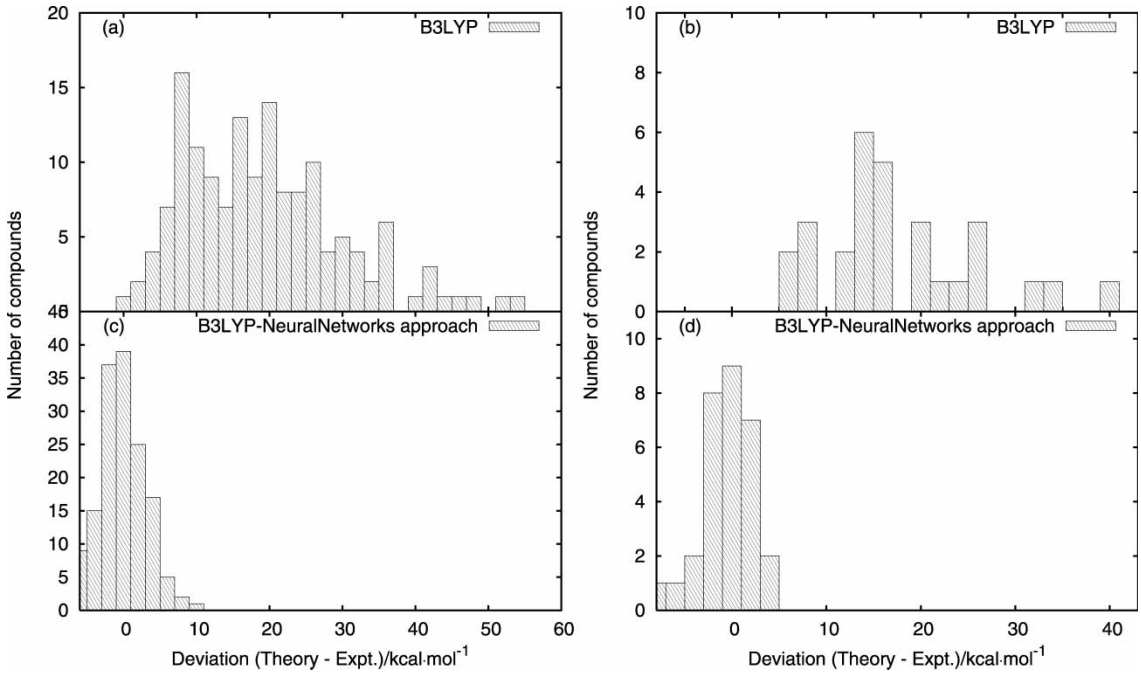

FIGURE 3 Histograms for the deviations of B3LYP/6-311 + G(*d*,*p*) calculated ΔG^0 for all 180 compounds. (a) and (b) are for the raw calculated ΔG^0 of training and testing set, respectively, and (c) and (d) are for the neural network corrected ΔG^0 of training and testing set, respectively. ΔG^0 is in the units of $\text{kcal} \cdot \text{mol}^{-1}$.

TABLE III The DFT-neural network RMS with different descriptors (all data are in the units of $kcal \cdot mol^{-1}$)

	B3LYP/6-311+G(d,p)*-NN*			*B3LYP/6-311+G*(3df,2p)*-NN*		
Method	*Training set*	*Testing set*	*Overall*	*Training set*	*Testing set*	*Overall*
A	4.3	5.0	4.4	3.7	4.3	3.8
B	3.1	2.7	3.1	3.7	4.0	3.7
C	3.1	2.8	3.1	3.5	3.6	3.5
D	3.1	2.8	3.0	3.4	3.4	3.4

Descriptors used in: *Method A*, ΔG^0 and N_t; *Method B*, ΔG^0, N_t and ZPE; *Method C*, ΔG^0, N_t, ZPE and N_{db}; *Method D*, ΔG^0, N_t, ZPE and N_H.

strategy does not discriminate against large molecules, unlike most calculations, which yield worse results for large molecules than for small ones. Although the raw B3LYP/6-311 + G(*d*,*p*) results have much larger deviations than those of B3LYP/6-311 + G(3*df*,2*p*), the neural-network corrected values of both calculations have deviations of the same magnitude. This implies that it is sufficient to employ the smaller basis set B3LYP/6-311 + G(*d*,*p*) in our DFT-NEURON approach. The neural-network step can easily correct the deficiency of a small basis set. Therefore, our approach can potentially be applied to much larger systems.

Analysis of our neural network reveals that the weights connecting the input for ΔG^0 have the dominant contribution in all cases. This confirms our fundamental assumption that the calculated ΔG^0 captures the essential values of exact ΔG^0. The input for the second physical descriptor, N_t, has quite large weights in all cases. In particular, when the smaller basis set 6-311 + G(*d*,*p*) is adopted in the B3LYP calculations, N_t has the second largest weights. It is found that the raw ΔG^0 deviations are roughly proportional to N_t, which confirms the importance of N_t as a significant descriptor of our neural network. We use the raw calculated ΔG^0 and N_t as the only two descriptors for the input layer, and find that the RMS deviations of the training and testing sets are corrected to 4.3, 5.0 $kcal \cdot mol^{-1}$ for B3LYP/6-311 + G(*d*,*p*), and 3.7, 4.3 $kcal \cdot mol^{-1}$ for B3LYP/6-311 + G(3*df*,2*p*) (See Table III), respectively. We thus conclude that raw ΔG^0 and N_t are the two main physical properties that determine the experimental values of ΔG^0. This is the main reason why our neural network approach is successful here. The bias contributes to the correction of systematic deviations in the raw calculated data, and has thus significant weights. When the larger basis set 6-311 + G(3df,2p) is used, the bias has the second largest weights for all cases. ZPE has been often scaled to account for the discrepancies of ΔH^0 between calculations and experiments [14,17], and it is thus expected to have large weights. This is indeed the case, especially when the smaller basis set 6-311 + G(*d*,*p*) is adopted in calculations. In all cases the number of hydrogen atoms, N_H, has the smallest but non-negligible weights. The number of double bonds, N_{db}, is the measurement of chemical structure of a molecule, and may be used as another physical descriptor. When N_H are replaced by N_{db}, the similar results are obtained as shown in Table III. This similarity can be rationalized. Both descriptors, i.e. the number of hydrogen N_H and the number of double bond N_{db}, are reflecting the structural information of individual molecules in different aspects.

The neural-network based approach adopted in this work reduces substantially the raw first-principles calculation errors, and can be used to quantify the uncertainty of the resulting ΔG^0. Our DFT-NEURON approach has a RMS error of $\sim 3\,kcal \cdot mol^{-1}$ for the 180 small- to medium-sized organic molecules. This is larger than the experimental error bars of our 180 molecules [9–11,13]. G2 method [14] results more accurate energies for small molecules. However, our approach is more efficient and can be applied to larger systems. The physical descriptors adopted in our neural network, the raw ΔG^0, N_t, ZPE and N_H, are quite general, and are not limited to the special motifs of our molecules. Our DFT-NEURON approach is therefore expected to yield the RMS deviations of $\sim 3\,kcal \cdot mol^{-1}$ for any small to medium sized organic molecules. To improve the accuracy for our DFT-NEURON approach, we may need more and better physical descriptors for the molecules, and possibly, more and better experimental data. The larger the experimental database is, the more accurately the DFT-NEURON approach predicts. In principle, the accuracy of our approach is limited only by the precision and size of the experimental database.

Besides the Gibbs energy of formation, our DFT-NEURON approach can be generalized to calculate other properties such as heat of formation, ionization energy, dissociation energy, absorption frequency and etc. The success of our approach is determined by the selection of the physical descriptors. The raw first-principles calculation property of interest contains its essential value, and is thus always the primary descriptor. The raw calculation error accumulates as the molecular size increases. The number of atom N_t should thus be selected for any Neural Networks correction procedure.

Additional physical descriptors should be chosen according to their relations to the property of interest and the physical and chemical structures of the compounds. Others have used Neural Networks to determine the quantitative structure relationship between the experimental measured thermodynamic properties and the structure parameters of the molecules [18]. We distinguish our work from others by utilizing specifically the first-principles methods. Since the first-principles calculations captures more readily the essences of the property of interest, our approach is more reliable and covers a much wider range of molecules or compounds.

To summarize, we have applied the DFT-NEURON approach to calculate the first-principles Gibbs energy of formation for small- and medium-sized organic molecules. The accuracy of the combined first-principle calculations and Neural Networks correction approach can be systematically increased, as more and better experimental data are available. Since the requirements on first-principles methods are modest, our approach is very efficient compared to the sophisticated first-principles methods of similar accuracy. More importantly, it can be easily extended to very large molecular systems.

Acknowledgements

Support from the Hong Kong Research Grant Council (RGC) and the Committee for Research and Conference Grants (CRCG) of the University of Hong Kong is gratefully acknowledged.

References

[1] Cramer, C.J. (2002) *Essentials of Computational Chemistry: Theories and Models* (John Wiley, West Sussex, England).
[2] Parr, R.G. and Yang, W. (1989) *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York).
[3] Foresman, J.B. and Frisch, A. (1996) *Exploring Chemistry with Electronic Structure Method*, 2nd Ed. (Gaussian Inc., Pittsburgh, PA).
[4] Irikura, K.K. and Frurip, D.J. (1998) *Computational Thermochemistry: Prediction and Estimation of Molecular Thermodynamics* (American Chemical Society, Washington, D.C.).
[5] Harvery, R.L. (1994) *Neural Network Principles* (Prentice-Hall, Englewood Cliffs, NJ).
[6] Ripley, B.D. (1996) *Pattern Recognition and Neural Networks* (Cambridge, New York).
[7] Haykin, S. (1999) *Neural Networks: a Comprehensive Foundation* (Prentice Hall, Upper Saddle River).
[8] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) "Learning representations by back-propagating errors", *Nature* **323**, 533.
[9] Yaws, C.L. (1999) *Chemical Properties Handbook* (McGraw-Hill, New York).
[10] Lide, D.R. (2000) *CRC Handbook of Chemistry and Physics*, 3rd Electronic ed. (CRC Press, Boca Raton, FL).
[11] Dean, J.A. (1999) *Lange's Handbook of Chemistry*, 15th Ed. (McGraw-Hill, New York).
[12] Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Zakrzewski, V.G., Montgomery, J.A., 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., Salvador, P., Dannenberg, J.J., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Cioslowski, J., Ortiz, J.V., Baboul, A.G., 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., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Andres, J.L., Gonzalez, C., Head-Gordon, M., Replogle, E.S. and Pople, J.A. (2002) *Gaussian* 98, Revision A.11.3, Gaussian, Inc., Pittsburgh, PA.
[13] Chase, M.W., Davies, C.A., Downey, J.R., Frurip, D.J., McDonald, R.A. and Syverud, A.N. (1985), *J. Phys. Chem. Ref. Data* **14**(Suppl. No. 1).
[14] Curtiss, L.A., Raghavachari, K., Redfern, P.C. and Pople, J.A. (1997) "Assessment of gaussian-2 and density functional theories for the computation of enthalpies of formation", *J. Chem. Phys.* **106**, 1063.
[15] Ivanciuc, O., Rabine, J.-P., Cabrol-Bass, D., Panaye, A. and Doucet, J.P. (1997) "^{13}C NMR chemical shift prediction of the sp^3 carbon atoms in the α position relative to the double bond in acyclic alkenes", *J. Chem. Inf. Comput. Sci.* **37**, 587.
[16] Nohair, M. and Zakarya, D. (2002) "Autocorrelation method adapted to generate new atomic environments: application for the prediction of 13-C chemical shifts of alkanes", *J. Chem. Inf. Comput. Sci.* **42**, 586.
[17] Curtiss, L.A., Raghavachari, K., Redfern, P.C. and Pople, J.A. (1997) "Investigation of the use of B3LYP zero-point energies and geometries in the calculation of enthalpies of formation", *Chem. Phys. Lett.* **270**, 419.
[18] Yao, X., Zhang, X., Zhang, R., Liu, M., Hu, Z. and Fan, B. (2001) "Prediction of enthalpy of alkanes by the use of radial basis function neural networks", *Comput. Chem.* **25**, 475.
[19] Pompe, M. and Novič, M. (1999) "Prediction of gas-chromatographic retention indices using topological descriptors", *J. Chem. Inf. Comput. Sci.* **39**, 59.
[20] Dong, L., Yan, A., Chen, X., Xu, H. and Hu, Z. (2001) "Research and prediction of coordination reactions between CPA-mA and some metal ions using artificial neural networks", *Comput. Chem.* **25**, 551.
[21] Afeefy, H. Y., Liebman, J. F. and Stein, S.E. (2003) "Neutral Thermochemical Data", In: Linstrom, P.J. and Mallard, W.G., eds, *NIST Chemistry WebBook* NIST Standard Reference Database Number 69, (National Institute of Standards and Technology, Gaithersburg MD, 20899), (http://webbook.nist.gov).
[22] Taylor, J.R. (1982) *An Introduction to Error analysis: The Study of Uncertainties in Physical Measurements* (University Science Books, Mill Valley, CA).