COMMUNICATIONS

Combined first-principles calculation and neural-network correction approach for heat of formation

LiHong Hu, XiuJun Wang, LaiHo Wong, and GuanHua Chen
Department of Chemistry, The University of Hong Kong, Hong Kong, China

(Received 17 March 2003; accepted 14 October 2003)

Despite their success, the results of first-principles quantum mechanical calculations contain inherent numerical errors caused by various intrinsic approximations. We propose here a neural-network-based algorithm to greatly reduce these inherent errors. As a demonstration, this combined quantum mechanical calculation and neural-network correction approach is applied to the evaluation of standard heat of formation $\Delta_f H^{\ominus}$ for 180 small- to medium-sized organic molecules at 298 K. A dramatic reduction of numerical errors is clearly shown with systematic deviation being eliminated. For example, the root-mean-square deviation of the calculated $\Delta_f H^{\ominus}$ for the 180 molecules is reduced from 21.4 to 3.1 kcal mol^{-1} for B3LYP/6-311+G(d,p) and from 12.0 to 3.3 kcal mol^{-1} for B3LYP/6-311+G($3df,2p$) before and after the neural-network correction.
 [DOI: 10.1063/1.1630951]

One of the Holy Grails of computational science is to quantitatively predict properties of matter prior to experiments. Despite the fact that the first-principles quantum mechanical calculation[1,2] has become an indispensable research tool and experimentalists have been increasingly relying on computational results to interpret their experimental findings, the practically used numerical methods by far are often not accurate enough in particular for complex systems. This limitation is caused by the inherent approximations adopted in the first-principles methods. Because of computational cost, electron correlation has always been a difficult obstacle for first-principles calculations. Finite basis sets chosen in practical computations are not able to cover entire physical space and this inadequacy introduces also inherent computational errors. Effective core potential is frequently used to approximate the relativistic effects, resulting inevitably in errors for systems that contain heavy atoms. The accuracy of a density functional theory (DFT) calculation is mainly determined by the exchange-correlation (XC) functional being employed,[1] whose exact form is however unknown.

Nevertheless, the results of first-principles quantum mechanical calculation can capture the essence of physics. For instance, the calculated results, despite that their absolute values may agree poorly with measurements, are usually of the same tendency among different molecules as their experimental counterparts. The quantitative discrepancy between the calculated and experimental results depends predominantly on the property of primary interest and, to a less extent, also on other related properties of the material. There exists thus a sort of quantitative relation between the calculated and experimental results, as the aforementioned approximations to a large extent contribute to the systematic errors of specified first-principles methods. Can we develop general ways to eliminate the systematic computational errors and further to quantify the accuracies of numerical methods used? It has been proven an extremely difficult task to determine the calculation errors from the first-principles. Alternatives must be sought.

We propose here a neural-network-based algorithm to determine the quantitative relationship between the experimental data and the first-principles calculation results. The determined relation will subsequently be used to eliminate the systematic deviations of the calculated results and thus reduce the numerical uncertainties. Since its beginning in the late fifties, neural networks has been applied to various engineering problems, such as robotics, pattern recognition, speech, etc.[3,4] As the first application of neural networks to quantum mechanical calculations of molecules, we choose the standard heat of formation $\Delta_f H^{\ominus}$ at 298.15 K as the property of interest.

A total of 180 small- or medium-sized organic molecules, whose $\Delta_f H^{\ominus}$ values are well documented in Refs. 5–7, are selected to test our proposed approach. The three tabulated values of $\Delta_f H^{\ominus}$ in the three references differ less than 1.0 kcal mol^{-1} for any one of the 180 molecules. The uncertainties of all $\Delta_f H^{\ominus}$ values are less than 1.0 kcal mol^{-1} in Refs. 5–7. These selected 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 has 32 atoms. We divide these molecules randomly into the training set (150 molecules) and the testing set (30 molecules). The geometries of 180 molecules are optimized via B3LYP/6-311+G(d,p)[8] calculations, and the zero point energies (ZPEs) are calculated at the same level. The enthalpy of each molecule is calculated at both B3LYP/6-311+G(d,p) and B3LYP/6-311+G($3df,2p$).[8] B3LYP/6-311+G($3df,2p$) 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 $\Delta_f H^{\ominus}$ calculations. The strategies in Ref. 9 are adopted to calculate $\Delta_f H^{\ominus}$. The calculated $\Delta_f H^{\ominus}$s for B3LYP/6-311

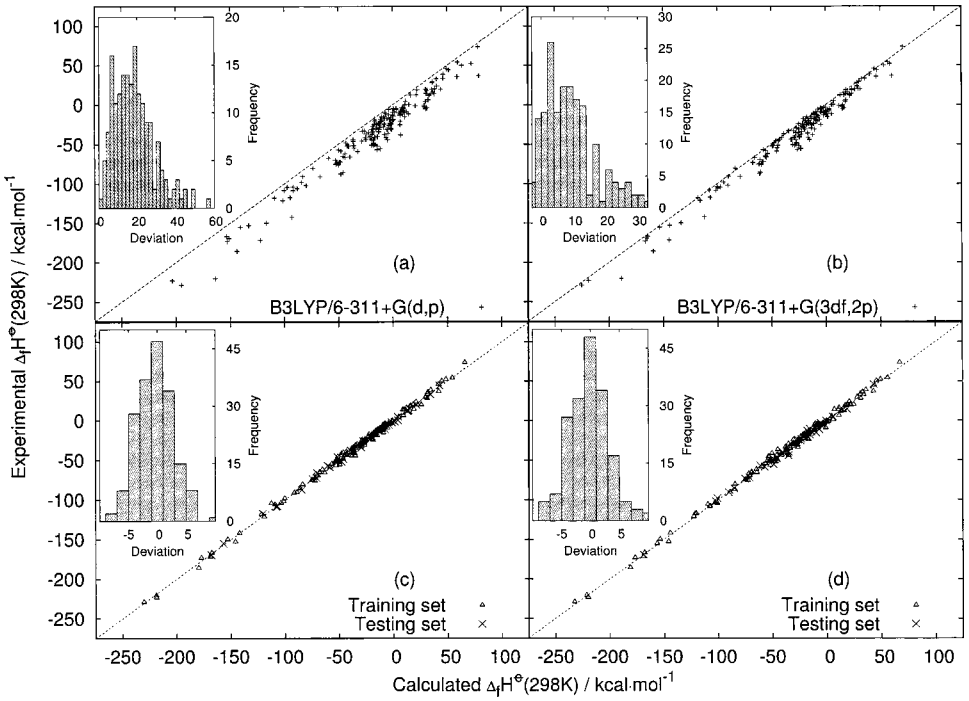

FIG. 1. Experimental $\Delta_f H^{\ominus}$ vs the calculated $\Delta_f H^{\ominus}$ for all 180 compounds. (a) and (b) are the comparisons of the experimental $\Delta_f H^{\ominus}$s to their raw B3LYP/6-311+G(d,p) and B3LYP/6-311+G$(3df,2p)$ results, respectively. (c) and (d) are the comparisons of the experimental $\Delta_f H^{\ominus}$s to the neural-network corrected B3LYP/6-311+G(d,p) and B3LYP/6-311+G$(3df,2p)$ $\Delta_f H^{\ominus}$s, respectively. In (c) and (d), the triangles are for the training set and the crosses for the testing set. The correlation coefficients of the linear fits are 0.998 and 0.994 in (c) and (d), respectively. Insets are the histograms for the differences between the experimental and calculated $\Delta_f H^{\ominus}$s (before and after the neural-network corrections). All values are in the units of kcal mol^{-1}.

+G(d,p) and B3LYP/6-311+G$(3df,2p)$ are compared to their experimental data in Figs. 1(a) and 1(b). The horizontal coordinates are the raw calculated data and the vertical coordinates are the experimental values. The dashed lines are where the vertical and horizontal coordinates are equal, i.e., where the B3LYP calculations and experiments would have the perfect match. The raw calculation values are mostly below the dashed line, i.e., most raw $\Delta_f H^{\ominus}$s are larger than the experimental data. In other words, there are systematic deviations for B3LYP $\Delta_f H^{\ominus}$. Compared to the experimental measurements, the root-mean-square (RMS) deviations for $\Delta_f H^{\ominus}$ are 21.4 and 12.0 kcal mol^{-1} for B3LYP/6-311+G(d,p) and B3LYP/6-311+G$(3df,2p)$ calculations, respectively. In Table I we compare the B3LYP and experimental $\Delta_f H^{\ominus}$s for 180 molecules with first seven small molecules reported in Ref. 9. Overall, B3LYP/6-311+G$(3df,2p)$ 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$(3dp,2p)$ calculations result in very small deviations from the experiments. For instance, the $\Delta_f H^{\ominus}$ deviations for CH_4 and CS_2 are only -0.3 and 0.4 kcal mol^{-1}, respectively. The results in Ref. 9 are slightly better than the B3LYP/6-311+G$(3df,2p)$ results for small molecules, and this is because the ZPEs in Ref. 9 are scaled by an empirical factor 0.98. For large molecules, both B3LYP/6-311+G(d,p) and B3LYP/6-311+G$(3df,2p)$ calculations yield quite large deviations from their experimental counterparts. To improve the comparison with the experiment, different empirical scaling factors have been employed for large molecules.

Our neural network adopts a three-layer architecture which has an input layer consisting of inputs from the physical descriptors and a bias, a hidden layer containing a number of hidden neurons, and an output layer that outputs the corrected values for $\Delta_f H^{\ominus}$ (see Fig. 2). The number of hidden neurons is to be determined. 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 $\Delta_f H^{\ominus}$ contains the essence of exact $\Delta_f H^{\ominus}$, and is thus an obvious choice of the primary descriptor for correcting $\Delta_f H^{\ominus}$. We observe that the size of a molecule affects the accuracies of calculations. We plot the difference between the calculated $\Delta_f H^{\ominus}$ and the measured $\Delta_f H^{\ominus}$ versus the number of atoms N_t in Fig. 3. Roughly speaking, the more atoms a molecule has, the larger the deviation. The deviation is approximately proportional to N_t, although the proportionality is by no means strict. This is consistent with the general observations of others in the field.[9] N_t of a molecule is thus chosen as the second descriptor for the molecule. ZPE is an important parameter in calculating $\Delta_f H^{\ominus}$. Its calculated value is often rescaled in evaluating $\Delta_f H^{\ominus}$,[9] and it is thus

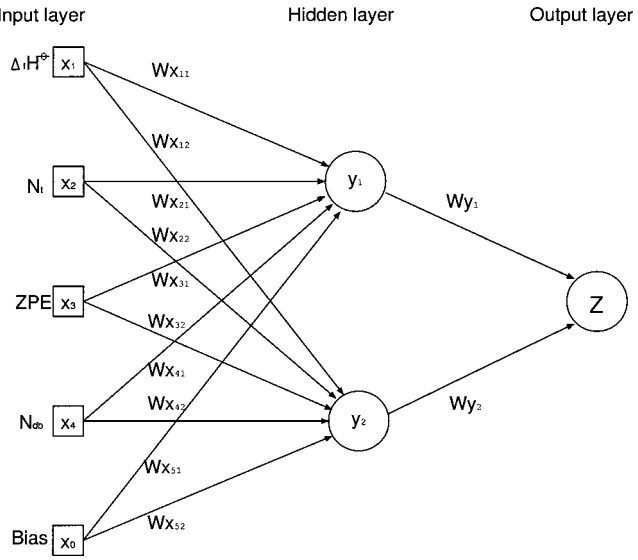

FIG. 2. Structure of our neural network.

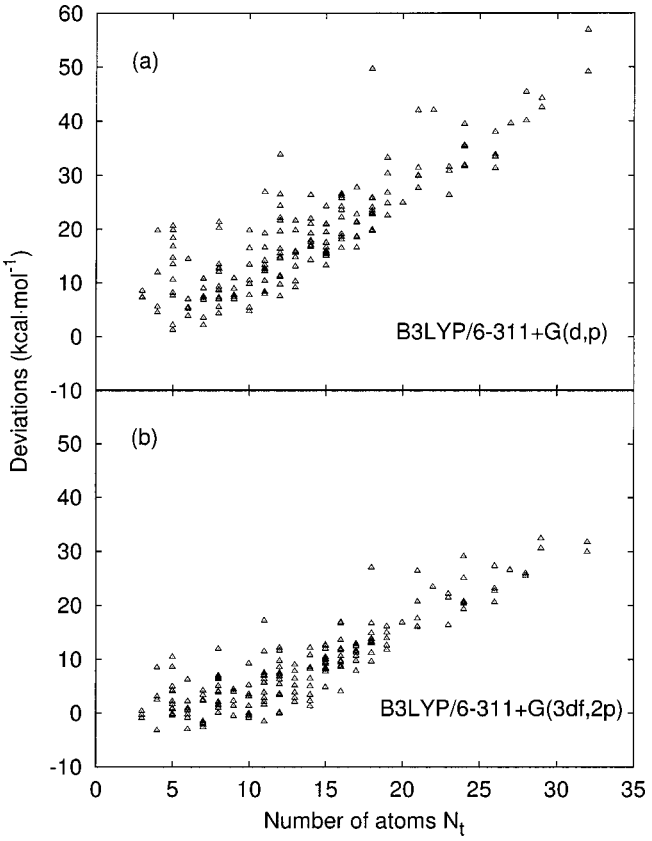

FIG. 3. Deviations (theory–expt.) vs the number of atoms (a) for B3LYP/6-311+G(d,p); (b) for B3LYP/6-311+G($3df$,$2p$).

taken as the third physical descriptor. Finally, the number of double bonds, N_{db}, is selected as the fourth and last descriptor to reflect the chemical structure of a molecule. To ensure the quality of our neural network, a cross-validation procedure is employed to determine our neural network including its structure and weights.[10] We randomly divide further 150 training molecules into five subsets of equal size. Four of them are used to train the neural network, and the fifth to validate its predictions. This procedure is repeated five times in rotation. The number of neurons in the hidden layer is varied from 1 to 10 to decide the optimal structure of our neural network. We find that the hidden layer containing two neurons yields the best overall results. Therefore, the 5-2-1 structure is adopted for our neural network as depicted in Fig. 2. The input values at the input layer, x_1, x_2, x_3, x_4, and x_5, are scaled $\Delta_f H^{\ominus}$, N_t, ZPE, N_{db}, and bias, respectively. The bias x_5 is set to 1. The weights $\{Wx_{ij}\}$s connect the input neurons $\{x_i\}$ and the hidden neurons y_1 and y_2, and $\{Wy_j\}$s connect the hidden neurons and the output Z which is the scaled $\Delta_f H^{\ominus}$ upon neural-network correction. The output Z is related to the input $\{x_i\}$ as

$$Z=\sum_{j=1,2} Wy_j \ \mathrm{Sig}\left(\sum_{i=1,5} Wx_{ij}\ x_i\right), \tag{1}$$

where $\mathrm{Sig}(v)=1/[1+\exp(-\alpha v)]$ and α is a parameter that controls the switch steepness of Sigmoidal function $\mathrm{Sig}(v)$. An error back-propagation learning procedure[4] is used to optimize the values of Wx_{ij} and Wy_j ($i=1,2,3,4,5$ and $j=1,2$). In Figs. 1(c) and 1(d) we plot the comparison between the neural-network corrected $\Delta_f H^{\ominus}$s and their experimental values (with the vertical coordinates for the experimental values and the horizontal coordinates for the calculated $\Delta_f H^{\ominus}$s). The triangles belong to the training set and the crosses belong to the testing set. 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 for $\Delta_f H^{\ominus}$ in Figs. 1(a) and 1(b) are eliminated, and the resulting numerical deviations are reduced substantially. This can be further demonstrated by the error analysis performed for the raw and neural-network corrected $\Delta_f H^{\ominus}$s of all 180 molecules. In the insets of Fig. 1, we plot the histograms for the deviations (from the experiments) of the raw B3LYP $\Delta_f H^{\ominus}$s and their neural-network corrected values. Obviously, the raw calculated $\Delta_f H^{\ominus}$s have large systematic deviations: 21.4 and 12.0 kcal mol^{-1} for $\Delta_f H^{\ominus}$s at B3LYP/6-311+G(d,p) and B3LYP/6-311+G($3df$,$2p$), respectively. On the contrary, the neural-network corrected $\Delta_f H^{\ominus}$s have virtually no systematic deviations. Moreover, the remaining numerical deviations are much smaller. Upon the neural-network corrections, the RMS deviations of $\Delta_f H^{\ominus}$s are reduced from 21.4 to 3.1 kcal mol^{-1} and 12.0 to 3.3 kcal mol^{-1} for B3LYP/6-311+G(d,p) and B3LYP/6-311+G($3df$,$2p$), respectively. Note that the error distributions after the neural-network correction are of approximate Gaussian distributions [see Figs. 1(c) and 1(d)]. Although the raw B3LYP/6-311+G(d,p) results have much larger deviations than those of B3LYP/6-311+G($3df$,$2p$), the neural-network corrected values of both calculations have deviations of the same magnitude. This implies that it may be sufficient to employ the smaller basis set 6-311+G(d,p) in our combined DFT calculation and neural-network correction (or DFT-NEURON) approach. The neural-network-based algorithm can correct easily the deficiency of a small basis set. Therefore, the DFT-NEURON approach can potentially be applied to much larger systems. In Table I we distinguish the molecules of the testing sets. The deviations of large molecules are of the same magnitude as those of small molecules. Unlike other quantum mechanical calculations that usually yield worse results for larger molecules than for small ones, the DFT-NEURON approach does not discriminate against the large molecules.

In Table II we list the values of $\{Wx_{ij}\}$ and $\{Wy_j\}$ of the two neural networks for correcting $\Delta_f H^{\ominus}$ of B3LYP/6-311+G(d,p) and B3LYP/6-311+G($3df$,$2p$) calculations. Analysis of our neural network reveals that the weights connecting the input for $\Delta_f H^{\ominus}$ have the dominant contribution in all cases. This confirms our fundamental assumption that the calculated $\Delta_f H^{\ominus}$ captures the essential value of exact $\Delta_f H^{\ominus}$. The bias contributes significantly to the correction of systematic deviations in the raw calculated data, and always has the second largest weights for all cases. The input for the second physical descriptor, N_t, has also large weights in all cases, in particular for 6-311+G(d,p). This is because the raw $\Delta_f H^{\ominus}$ deviations are roughly proportional to N_t as shown in Fig. 3, which confirms the importance of N_t as a significant descriptor of our neural network. ZPE has been often rescaled to account for the discrepancies between cal-

TABLE I. Experimental and calculated $\Delta_f H^{\ominus}$ (298 K) for 180 compounds (all data are in the units of kcal mol^{-1}).

Formula	Name	$\Delta_f H^{\ominus}$ (298 K)					
		Expt.[a]	DFT1[b]	NN1[c]	DFT2[d]	NN2[e]	DFT3[f]
CF_2O	carbonyl fluoride	−152.7	−132.9	−146.0	−144.2	−146.2	−143.6
CH_2Cl_2	dichloromethane	−22.8	−12.2	−19.2	−17.8	−17.9	−18.2
CH_2F_2[g]	difluoromethane	−108.2	−100.1	−107.2	−107.5	−107.5	−107.7
CH_4	methane	−17.9	−16.6	−16.7	−18.2	−16.8	−19.5
CH_4O[g]	methanol	−48.1	−42.9	−46.2	−47.2	−47.3	−48.1
CS_2	carbon disulfide	28.0	36.5	31.2	28.4	31.1	28.2
C_6H_6[g]	benzene	19.8	36.2	21.1	26.7	21.1	24.2
$CBrCl_3$	bromotrichloromethane	−9.3	11.3	−5.3	1.2	−1.1	–
$CBrF_3$[g]	bromotrifluoromethane	−155.1	−140.4	−156.9	−153.4	−156.7	–
$CClF_3$[g,h]	chlorotrifluoromethane	−169.2	−150.9	−168.0	−165.0	−169.1	–
$CClN$[g]	cyanogen chloride	33.0	40.3	34.3	32.7	33.8	–
CCl_2O	phosgene	−52.3	−40.4	−49.8	−49.2	−48.6	–
CF_4	carbon tetrafluoride	−223.0	−203.2	−218.6	−218.9	−220.1	–
$CHCl_3$	chloroform	−24.2	−7.4	−18.8	−15.6	−16.7	–
CHF_3	trifluoromethane	−166.7	−153.2	−167.6	−164.5	−168.2	–
CH_2O_2	formic acid	−90.5	−82.9	−89.2	−89.6	−89.2	–
CH_3Br	methyl bromide	−9.0	−6.8	−10.3	−9.2	−8.4	–
CH_3NO_2	nitromethane	−17.9	−11.0	−20.1	−20.5	−22.0	–
CH_3NO_2	methyl nitrite	−15.3	−7.9	−17.8	−17.4	−19.1	–
CH_4S[g,h]	methyl mercaptan	−5.5	1.4	−4.0	−3.3	−3.7	–
CH_5N	methylamine	−5.5	−3.4	−6.5	−7.3	−8.0	–
COS	carbonyl sulfide	−33.1	−25.8	−29.4	−34.0	−30.6	–
C_2H_2	acetylene	54.2	59.8	54.2	56.7	55.7	–
$C_2H_2Cl_2$[g]	1,1-dichloroethylene	0.6	15.0	4.4	6.8	5.8	–
$C_2H_2F_2$	1,1-difluoroethylene	−80.5	−73.5	−83.9	−83.5	−84.7	–
$C_2H_2O_4$	oxalic acid	−173.0	−152.7	−177.3	−166.5	−176.9	–
C_2H_3Br	vinyl bromide	18.7	22.6	15.6	18.5	18.1	–
C_2H_3ClO[h]	acetyl chloride	−58.3	−47.5	−58.3	−55.0	−57.0	–
$C_2H_3ClO_2$	chloroacetic acid	−104.3	−97.3	−112.6	−97.3	−101.2	–
$C_2H_3Cl_3$[g]	1,1,1-trichloroethane	−34.0	−12.7	−29.4	−22.0	−26.7	–
C_2H_3F	vinyl fluoride	−33.2	−27.7	−33.8	−34.0	−34.1	–
C_2H_4[g]	ethylene	12.5	16.3	13.5	13.1	13.7	–
$C_2H_4Br_2$	1,2-dibromoethane	−9.3	−0.0	−12.4	−4.3	−7.8	–
$C_2H_4Cl_2$	1,1-dichloroethane	−31.0	−17.6	−29.7	−24.3	−27.9	–
$C_2H_4Cl_2$	1,2-dichloroethane	−31.0	−18.2	−29.9	−24.6	−28.1	–
$C_2H_4F_2$	1,1-difluoroethane	−118.0	−109.3	−121.9	−117.9	−121.7	–
C_2H_4O	ethylene oxide	−12.6	−3.6	−9.8	−10.3	−11.7	–
$C_2H_4O_2$	acetic acid	−103.9	−91.9	−103.8	−100.1	−103.4	–
C_2H_4S	thiacyclopropane	19.7	30.3	22.4	23.9	21.8	–
C_2H_5Br	bromoethane	−15.3	−9.8	−18.0	−13.2	−15.9	–
C_2H_5Cl	ethyl chloride	−26.7	−18.1	−26.0	−22.6	−25.2	–
C_2H_5N	ethyleneimine	29.5	36.8	29.4	30.5	27.7	–
C_2H_5NO[g,h]	acetamide	−57.0	−49.6	−60.5	−57.5	−61.1	–
$C_2H_5NO_2$	nitroethane	−24.2	−14.4	−28.8	−25.1	−30.7	–
$C_2H_5NO_3$	ethyl nitrate	−36.8	−24.2	−42.3	−38.4	−45.1	–
C_2H_6	ethane	−20.2	−15.9	−20.2	−18.8	−20.5	–
C_2H_6O	dimethyl ether	−44.0	−36.3	−44.5	−42.6	−46.2	–
C_2H_6S	dimethyl sulfide	−9.0	1.9	−7.9	−4.6	−8.3	–
C_2H_7N	dimethylamine	−4.5	0.9	−7.2	−4.5	−8.6	–
C_2H_7N[h]	ethylamine	−11.0	−6.3	−14.2	−11.4	−15.5	–
$C_2H_8N_2$	ethylenediamine	−4.1	3.4	−8.2	−4.0	−10.5	–
C_2N_2	cyanogen	73.8	78.4	65.7	70.7	66.9	–
C_3H_3NO	oxazole	−3.7	8.9	−1.6	−2.1	−3.9	–
C_3H_4[g]	methylacetylene	44.3	51.5	42.5	46.8	43.7	–
C_3H_4[h]	propadiene	45.9	49.4	41.5	44.4	42.9	–
$C_3H_4O_3$	ethylene carbonate	−121.2	−101.4	−119.3	−115.9	−122.7	–
$C_3H_5Cl_3$	1,2,3-trichloropropane	−44.4	−17.5	−38.3	−27.2	−35.1	–
C_3H_6[g]	cyclopropane	12.7	21.6	14.1	16.7	13.4	–
C_3H_6[g]	propylene	4.9	11.9	4.4	7.2	4.6	–
$C_3H_6Br_2$	1,2-dibromopropane	−17.4	−5.2	−22.5	−10.5	−17.5	–
$C_3H_6Cl_2$[h]	1,2-dichloropropane	−39.6	−20.4	−37.2	−28.1	−35.1	–
C_3H_6O	acetone	−52.0	−41.6	−53.6	−48.5	−53.2	–
$C_3H_6O_2$[h]	methyl acetate	−98.0	−83.9	−100.3	−94.1	−100.8	–
$C_3H_6O_2$[g]	propionic acid	−108.4	−91.8	−108.3	−101.3	−108.1	–
C_3H_6S[h]	thiacyclobutane	14.6	31.1	18.8	23.9	18.6	–

TABLE I (*Continued*).

Formula	Name	$\Delta_f H^{\ominus}$ (298 K) Expt.[a]	DFT1[b]	NN1[c]	DFT2[d]	NN2[e]	DFT3[f]
C_3H_7Br	1-bromopropane	−21.0	−10.7	−23.7	−15.4	−21.5	–
C_3H_7Br[h]	2-bromopropane	−23.2	−12.8	−26.1	−17.5	−23.6	–
C_3H_7Cl	isopropyl chloride	−35.0	−21.7	−34.7	−27.5	−33.6	–
C_3H_7Cl	n-propyl chloride	−31.1	−18.9	−31.7	−24.8	−30.8	–
C_3H_7F	1-fluoropropane	−67.2	−58.9	−71.5	−65.6	−71.7	–
C_3H_7NO[g]	N,N-dimethylformamide	−45.8	−36.2	−51.5	−45.9	−52.8	–
$C_3H_7NO_2$	1-nitropropane	−29.8	−15.0	−33.7	−27.0	−35.3	–
$C_3H_7NO_2$	2-nitropropane	−33.5	−17.9	−36.8	−29.6	−38.1	–
$C_3H_7NO_3$	propyl nitrate	−41.6	−24.9	−47.7	−40.3	−50.5	–
$C_3H_7NO_3$	isopropyl nitrate	−45.6	−28.1	−51.3	−43.4	−53.7	–
C_3H_8[h]	propane	−24.8	−16.8	−25.9	−21.0	−26.2	–
C_3H_8O	methyl ethyl ether	−51.7	−40.6	−53.7	−48.1	−55.1	–
C_3H_8S	n-propyl mercaptan	−16.2	−1.6	−16.5	−8.7	−16.0	–
C_3H_8S	isopropyl mercaptan	−18.2	−2.7	−17.8	−9.7	−17.0	–
C_3H_8S	ethyl methyl sulfide	−14.2	0.7	−13.8	−7.0	−14.1	–
C_3H_9N	n-propylamine	−17.3	−7.0	−19.7	−13.4	−21.0	–
C_3H_9N	isopropylamine	−20.0	−9.7	−22.7	−16.2	−23.8	–
C_3H_9N	trimethylamine	−5.7	3.4	−9.8	−3.6	−11.2	–
$C_3H_{10}N_2$[h]	1,2-propanediamine	−12.8	0.4	−16.1	−8.0	−17.9	–
$C_4H_4N_2$	succinonitrile	50.1	63.5	44.5	53.2	45.3	–
C_4H_6	1,2-butadiene	38.8	46.5	34.4	40.1	35.7	–
C_4H_6O[h]	divinyl ether	−3.3	10.0	−4.5	−0.5	−5.6	–
C_4H_8[h]	1-butene	−0.0	11.3	−0.8	5.3	−0.6	–
C_4H_8O	isobutyraldehyde	−51.5	−35.6	−52.3	−43.7	−51.7	–
$C_4H_8O_2$	ethyl acetate	−105.9	−88.0	−109.4	−99.4	−109.8	–
C_4H_9Br	1-bromobutane	−25.6	−11.4	−29.3	−17.5	−27.0	–
C_4H_9Cl	tert-butyl chloride	−43.8	−24.6	−42.9	−31.7	−41.3	–
$C_4H_{10}O$	sec-butanol	−69.9	−52.4	−70.5	−60.3	−70.7	–
$C_4H_{10}O_2$[h]	1,4-butanediol	−102.0	−79.8	−101.3	−90.3	−102.1	–
$C_4H_{10}S$	isobutyl mercaptan	−23.2	−2.4	−22.3	−10.7	−21.4	–
$C_4H_{10}S$	methyl propyl sulfide	−19.5	−0.1	−19.4	−9.1	−19.6	–
$C_4H_{11}N$	tert-butylamine	−28.6	−12.1	−30.4	−19.8	−31.0	–
C_5H_5N	pyridine	33.5	46.2	31.2	35.6	30.6	–
C_5H_6S	2-methylthiophene	20.0	44.3	25.6	31.7	24.6	–
C_5H_8	trans-1,3-pentadiene	18.6	31.7	16.0	23.7	16.7	–
$C_5H_8O_2$[h]	acetylacetone	−90.8	−66.6	−91.4	−78.8	−90.2	–
C_5H_{10}[h]	cyclopentane	−18.5	0.9	−14.9	−5.8	−15.5	–
C_5H_{10}	2-methyl-1-butene	−8.7	7.9	−9.2	0.5	−8.9	–
C_5H_{10}	2-methyl-2-butene	−10.2	5.9	−11.6	−1.7	−11.2	–
C_5H_{10}	3-methyl-1-butene	−6.9	10.6	−6.5	3.3	−6.1	–
C_5H_{10}	1-pentene	−5.0	10.4	−6.5	3.1	−6.2	–
C_5H_{10}	cis-2-pentene	−6.7	9.0	−8.1	1.6	−7.7	–
C_5H_{10}[h]	trans-2-pentene	−7.6	7.5	−9.7	0.1	−9.3	–
$C_5H_{10}O$	2-pentanone	−61.8	−43.8	−65.5	−53.3	−64.8	–
$C_5H_{10}O$	valeraldehyde	−54.4	−35.3	−56.6	−44.7	−56.1	–
$C_5H_{10}O_2$[g]	valeric acid	−117.2	−94.4	−120.6	−106.5	−120.6	–
$C_5H_{10}S$[h]	thiacyclohexane	−15.1	11.4	−9.2	1.8	−9.9	–
$C_5H_{10}S$[g,h]	cyclopentanethiol	−11.5	14.7	−7.3	5.2	−6.7	–
$C_5H_{11}Br$	1-bromopentane	−30.9	−12.2	−35.0	−19.6	−32.5	–
$C_5H_{11}Cl$	1-chloropentane	−41.8	−20.5	−43.0	−29.0	−41.9	–
$C_5H_{11}N$	piperidine	−11.7	9.7	−9.6	0.8	−11.3	–
C_5H_{12}	isopentane	−36.9	−18.4	−37.4	−25.3	−37.4	–
C_5H_{12}	n-pentane	−35.0	−18.4	−37.2	−25.2	−37.3	–
$C_5H_{12}O$	2-methyl-1-butanol	−72.2	−49.2	−71.9	−58.4	−72.2	–
$C_5H_{12}O$[g,h]	3-methyl-1-butanol	−72.2	−48.2	−70.7	−57.4	−71.1	–
$C_5H_{12}O$	3-methyl-2-butanol	−75.1	−52.1	−75.1	−61.4	−75.2	–
$C_5H_{12}O$	2-pentanol	−75.0	−52.0	−74.7	−61.1	−74.9	–
$C_5H_{12}O$	3-pentanol	−75.7	−49.8	−72.6	−59.0	−72.8	–
$C_5H_{12}O$	ethyl propyl ether	−65.1	−45.4	−68.2	−55.4	−69.2	–
$C_5H_{12}O_4$	pentaerythritol	−185.6	−143.7	−179.6	−159.3	−181.2	–
$C_5H_{12}S$[g]	n-pentyl mercaptan	−25.9	−3.2	−27.8	−12.9	−27.0	–
$C_5H_{12}S$	butyl methyl sulfide	−24.4	−0.9	−25.1	−11.2	−25.2	–
C_6F_6	hexafluorobenzene	−228.6	−194.8	−230.3	−225.2	−232.6	–
$C_6H_4Cl_2$	m-dichlorobenzene	6.3	32.8	9.3	18.5	11.1	–
$C_6H_4F_2$[g,h]	p-difluorobenzene	−73.4	−51.3	−74.8	−67.1	−75.3	–

TABLE I (*Continued*).

Formula	Name	$\Delta_f H^{\ominus}$ (298 K) Expt.[a]	DFT1[b]	NN1[c]	DFT2[d]	NN2[e]	DFT3[f]
C_6H_5Cl	monochlorobenzene	12.4	34.0	14.8	22.1	15.7	–
C_6H_5F	fluorobenzene	−27.9	−8.3	−26.8	−21.0	−27.3	–
$C_6H_5NO_2$	nitrobenzene	16.2	37.1	12.1	19.6	11.0	–
$C_6H_6N_2O_2$	m-nitroaniline	14.0	38.2	8.8	18.0	6.9	–
C_6H_6O	phenol	−23.0	−1.5	−20.3	−14.0	−21.1	–
$C_6H_6O_2$	1,3-benzenediol	−65.7	−39.3	−63.1	−54.8	−64.5	–
C_6H_7N	2-methylpyridine	23.6	40.7	20.8	28.7	20.5	–
$C_6H_8N_2$	adiponitrile	35.7	59.3	31.7	46.4	32.6	–
C_6H_{10}[g]	1-methylcyclopentene	−1.3	20.9	1.4	12.3	1.6	–
C_6H_{10}	1,5-hexadiene	20.1	38.7	18.2	29.6	19.1	–
$C_6H_{10}O_3$	propionic anhydride	−149.7	−116.5	−153.0	−133.6	−153.9	–
$C_6H_{11}NO$[g]	e-caprolactam	−58.8	−32.1	−58.4	−45.0	−59.9	–
C_6H_{12}	trans-3-hexene	−13.0	6.9	−15.1	−1.8	−14.6	–
$C_6H_{12}O$	butyl vinyl ether	−43.7	−18.9	−44.9	−31.1	−45.9	–
$C_6H_{12}O$[g]	3-hexanone	−66.4	−43.8	−70.2	−54.6	−69.5	–
C_6H_{14}[g,h]	3-methylpentane	−41.0	−16.1	−39.8	−24.2	−39.7	–
$C_6H_{14}S$	methyl pentyl sulfide	−29.3	−1.7	−30.7	−13.3	−30.7	–
C_7H_5N	benzonitrile	52.3	72.1	48.4	58.7	49.2	–
C_7H_6O	benzaldehyde	−8.8	13.2	−8.6	−0.3	−8.1	–
C_7H_8[g]	toluene	11.9	32.9	12.7	22.0	13.1	–
C_7H_8O	o-cresol	−30.7	−5.0	−29.1	−18.8	−29.5	–
C_7H_9N	2,6-dimethylpyridine	14.0	35.2	10.2	21.9	10.3	–
C_7H_{14}	cis-1,2-dimethylcyclopentane	−31.0	−1.0	−26.9	−10.3	−27.0	–
$C_7H_{15}Br$	1-bromoheptane	−40.2	−13.8	−46.2	−23.8	−43.6	–
C_7H_{16}[h]	3,3-dimethylpentane	−48.2	−17.4	−46.2	−26.7	−45.7	–
C_7H_{16}	2,2,3-trimethylbutane	−48.9	−17.4	−46.6	−26.8	−45.9	–
$C_7H_{16}S$	n-heptyl mercaptan	−35.8	−4.1	−38.4	−16.5	−37.5	–
$C_8H_6O_4$	terephthalic acid	−171.6	−121.9	−169.0	−144.6	−169.6	–
C_8H_8O	acetophenone	−20.7	7.0	−19.9	−7.8	−19.3	–
C_8H_{10}[h]	o-xylene	4.5	30.3	5.1	17.9	5.6	–
$C_8H_{10}O$[g]	3,4-xylenol	−37.4	−7.1	−36.7	−22.4	−36.8	–
C_8H_{16}	cis-1,2-dimethylcyclohexane	−41.1	−5.7	−36.1	−16.1	−36.2	–
C_8H_{16}[h]	trans-1,4-dimethylcyclohexane	−44.1	−4.7	−35.0	−15.0	−35.1	–
C_8H_{16}[h]	2,4,4-trimethyl-2-pentene	−25.1	6.8	−25.4	−4.5	−24.3	–
C_8H_{18}	2,3-dimethylhexane	−51.1	−17.7	−51.3	−28.5	−50.9	–
C_8H_{18}	3-ethylhexane	−50.4	−16.6	−50.0	−27.3	−49.7	–
C_8H_{18}	4-methylheptane	−50.7	−19.4	−52.8	−30.1	−52.5	–
C_8H_{18}[g]	2,3,4-trimethylpentane	−52.0	−14.0	−47.6	−24.7	−47.1	–
$C_8H_{18}O$[g]	2-ethyl-1-hexanol	−87.3	−47.7	−84.4	−60.8	−84.6	–
$C_8H_{18}S_2$	dibutyl disulfide	−37.9	7.5	−36.2	−11.9	−38.2	–
C_9H_{12}	m-ethyltoluene	−0.5	29.3	−0.8	15.6	−0.1	–
C_9H_{12}[g]	1,2,3-trimethylbenzene	−2.3	29.1	−1.3	15.3	−0.5	–
$C_9H_{18}O$[h]	diisobutyl ketone	−85.5	−45.4	−86.3	−60.0	−85.3	–
C_9H_{20}[h]	3,3-diethylpentane	−55.4	−11.2	−49.4	−23.0	−48.8	–
C_9H_{20}	2,2,3,4-tetramethylpentane	−56.6	−14.1	−52.8	−26.1	−52.0	–
$C_{10}H_{14}$	sec-butylbenzene	−4.0	31.5	−3.1	16.7	−2.4	–
$C_{10}H_{14}$	isobutylbenzene	−4.9	30.4	−4.2	15.5	−3.5	–
$C_{10}H_{18}O_4$	sebacic acid	−220.3	−163.4	−218.8	−188.6	−221.4	–
$C_{10}H_{20}O_2$	n-decanoic acid	−142.0	−92.9	−142.2	−112.1	−144.6	–
$C_{12}H_{10}$	acenaphthene	37.0	79.0	41.6	60.5	42.7	–

[a]The experimental data were taken from Ref. 5.
[b]The calculated $\Delta_f H^{\ominus}$ by using B3LPY/6-311+G(d,p) geometries, zero point energies and enthalpies.
[c]The calculated $\Delta_f H^{\ominus}$ by B3LYP/6-311+G(d,p)-neural networks approach.
[d]The calculated $\Delta_f H^{\ominus}$ by using the 6-311+G(d,p) geometries and zero point energies, and recalculated enthalpies by 6-311+G($3df,2p$) basis.
[e]The calculated $\Delta_f H^{\ominus}$ by B3LYP/6-311+G($3df,2p$)-neural networks approach.
[f]The $\Delta_f H^{\ominus}$s were taken from Ref. 9, where the zero point energies were corrected by a scale factor 0.98.
[g]The molecule belongs to the testing set in NN1 calculation.
[h]The molecule belongs to the testing set in NN2 calculation.

culations and experiments,[9] 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 the calculations. In all cases the number of double bonds, N_{db}, has the smallest but non-negligible weights.

Our DFT-NEURON approach has a RMS deviation of ~ 3 kcal mol^{-1} for the 180 small- to medium-sized organic molecules. The physical descriptors adopted in our neural network, the raw calculated $\Delta_f H^{\ominus}$, the number of atoms N_t, the number of double bonds N_{db}, and the ZPE are quite

TABLE II. Weights of DFT-neural networks for $\Delta_f H^{\ominus}$.

	NN1[a]		NN2[b]	
Weights	y_1	y_2	y_1	y_2
Wx_{1j}	−2.48	0.79	0.85	−2.77
Wx_{2j}	0.25	−0.50	−0.18	0.09
Wx_{3j}	0.01	0.38	0.09	0.20
Wx_{4j}	0.26	0.01	0.01	0.20
Wx_{5j}	0.41	−0.49	−0.56	0.59
Wy_j	−0.21	1.41	1.43	−0.20

[a]NN1 refers B3LYP/6-311+G(d,p)-neural networks approach.
[b]NN2 refers B3LYP/6-311+G($3df,2p$)-neural networks approach.

general, and are not limited to special properties of organic molecules. The DFT-NEURON approach developed here is expected to yield a RMS deviation of $\sim 3\,\mathrm{kcal\,mol^{-1}}$ for $\Delta_f H^{\ominus}$s of any small- to medium-sized organic molecules. G2 (Ref. 9) and G3 (Ref. 11) methods result in more accurate $\Delta_f H^{\ominus}$ for small molecules. Like the DFT-NEURON approach, G2 and G3 methods have empirical fittings to the high order electron correlations and ZPEs, and are thus not entirely *ab initio*. Our approach is much more efficient and can be applied to much larger systems. To improve the accuracy for the DFT-NEURON approach, we need more and better experimental data, and possibly, more and better physical descriptors for the molecules. Besides $\Delta_f H^{\ominus}$, the DFT-NEURON approach can be generalized to calculate other properties such as Gibbs free energy, ionization energy, dissociation energy, absorption frequency, band gap, etc. The raw first-principles property of interest contains the essence of its exact value, and is thus always the primary descriptor. As the raw calculation error accumulates with increasing molecular size, the number of atoms N_t should thus be selected for other DFT-NEURON calculations. Additional physical descriptors should be chosen according to their relations to the property of interest and to the physical and chemical characteristics of the compounds. Others have used neural networks to determine the quantitative relationship between the experimental measured thermodynamic properties and the structure parameters of the molecules.[10] We distinguish our work from them by utilizing specifically the first-principles methods and with the objective to improve quantum mechanical results. Since the first-principles calculations capture readily the essences of the properties of interest, our approach is more reliable and covers much a wider range of molecules or compounds.

To summarize, we have developed a promising new approach to improve the results of first-principles quantum mechanical calculations and to calibrate their uncertainties. The accuracy of DFT-NEURON approach can be systematically improved as more and better experimental data are available. As the systematic deviations caused by small basis sets and less sophisticated methods adopted in the calculations can be easily corrected by neural networks, the requirements on first-principles calculations are modest. Our approach is thus highly efficient compared to much more sophisticated first-principles methods of similar accuracy, and more importantly, is expected to be applied to much larger systems. The combined first-principles calculation and neural-network correction approach developed in this work is potentially a powerful tool in computational physics and chemistry, and may open the possibility for first-principles methods to be employed practically as predictive tools in materials research and design.

We thank Professor YiJing Yan for extensive discussion on the subject and generous help in manuscript preparation. 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.

[1] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989), and references therein.
[2] H. F. Schaefer III, *Methods of Electronic Structure Theory* (Plenum, New York and London, 1977), and references therein.
[3] B. D. Ripley, *Pattern Recognition and Neural Networks* (Cambridge University Press, New York, 1996).
[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Nature (London) **323**, 533 (1986).
[5] C. L. Yaws, *Chemical Properties Handbook* (McGraw–Hill, New York, 1999).
[6] D. R. Lide, *CRC Handbook of Chemistry and Physics*, 3rd electronic ed. (CRC, Boca Raton, FL, 2000).
[7] J. B. Pedley, R. D. Naylor, and S. P. Kirby, *Thermochemical Data of Organic Compounds*, 2nd ed. (Chapman and Hall, New York, 1986).
[8] M. J. Frisch, G. W. Trucks, H. B. Schlegel *et al.*, GAUSSIAN 98, Revision A.11.3, Gaussian, Inc., Pittsburgh, PA, 2002.
[9] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, J. Chem. Phys. **106**, 1063 (1997).
[10] X. Yao, X. Zhang, R. Zhang, M. Liu, Z. Hu, and B. Fan, Comput. Chem. **25**, 475 (2001).
[11] L. A. Curtiss, K. Raghavachari, P. C. Redfern, V. Rassolov, and J. A. Pople, J. Chem. Phys. **109**, 7764 (1998).