The X1 method for accurate and efficient prediction of heats of formation

Jianming Wu and Xin Xu[a)]

State Key Laboratory of Physical Chemistry of Solid Surfaces and Center for Theoretical Chemistry, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China

(Received 31 July 2007; accepted 25 September 2007; published online 5 December 2007)

We propose the X1 method which combines the density functional theory method with a neural network (NN) correction for an accurate yet efficient prediction of heats of formation. It calculates the final energy by using B3LYP/6-311+G($3df$,$2p$) at the B3LYP/6-311+G(d,p) optimized geometry to obtain the B3LYP standard heats of formation at 298 K with the unscaled zero-point energy and thermal corrections at the latter basis set. The NN parameters cover 15 elements of H, Li, Be, B, C, N, O, F, Na, Mg, Al, Si, P, S, and Cl. The performance of X1 is close to the Gn theories, giving a mean absolute deviation of 1.43 kcal/mol for the G3/99 set of 223 molecules up to 10 nonhydrogen atoms and 1.48 kcal/mol for the X1/07 set of 393 molecules up to 32 nonhydrogen atoms. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2800018]

I. INTRODUCTION

The accurate prediction of heats of formation (ΔH_f^θ) is one of the central topics in computational chemistry. The Gn(n=1–3) family of model chemistries represents one of the most successful methods up to date.[1–6] The G2/97 set contains 148 molecules of less than 6 heavy atoms, whose experimental ΔH_f^θ are accurately known. Calibrated with the G2/97 set (148 molecules up to 6 heavy atoms), the G2 method leads to a mean absolute deviation (MAD) of 1.57 kcal/mol.[4] The G3/99 set included 75 additional heats of formation of larger molecules (up to 10 heavy atoms).[6] The G3 method improves over G2, giving a MAD of 1.05 kcal/mol for the G3/99 set (223). Nevertheless, these methods are based on coupled-cluster-type treatments [QCISD(T)]. Hence, they are very computationally resource demanding and computationally time consuming.

Density functional theory (DFT) offers a promising alternative. In fact, Becke's three parameter hybrid method[7] (B3LYP) has been widely recognized as a cost-effective method and has been successfully applied to many chemically interesting systems.[8] However, there are increasing evidences, showing that B3LYP degrades rapidly with the increase of the molecular size.[9] Thus, Curtiss *et al.*[6] showed that B3LYP provides a MAD of 3.1 kcal/mol for the G2/97 set, whereas the MAD is more than doubled to 8.2 kcal/mol for the G3-3 set (75) of larger molecules.

DFT, in principle, can take into full account of all complex many-body effects at a computational cost characteristic of mean field approximations. However, the exact exchange-correlation functional is unknown; approximate functionals have to be used in practice. To make things worse, there lacks a systematic way to improve an approximate functional for greater accuracy. Some correction schemes on top of DFT have been proposed.[10–18] Noteworthily, Hu *et al.*[14] proposed a neural network (NN) scheme to correct the errors of the B3LYP method to calculate heats of formation of neutral organic molecules. Their results are promising. Calibrated with their own compilation of 180 organic molecules for the B3LYP/6-311+G($3df$,$2p$) method, the root-mean-square deviations of the calculated ΔH_f^θ were reduced from 12.0 to 3.3 kcal/mol upon the NN correction. Their training set, however, contained no open-shell molecules, no radicals, and no inorganic molecules, which limited the applicability of their NN.

In the present work, we set up a new NN (the X1 method) based on the well established G3/99 set for heats of formation, plus 170 additional molecules (the X1-1 set) with more diverse chemical environment up to 32 heavy atoms (n-$C_{32}H_{66}$). Our NN reduces the MAD from 5.58 to 1.43 kcal/mol for the G3/99 set and that from 13.60 to 1.53 kcal/mol for the X1-1 set.

II. DATA SETS OF HEATS OF FORMATION

The G3/99 set includes 223 neutral molecules, where the experimental heats of formation were accurately known. It covers a wide range of different chemical structures and bonding situations, containing not only the closed-shell organic molecules but also radicals, open-shell molecules, inorganic compounds, etc. The largest molecules are up to 10 heavy atoms ($C_{10}H_8$, naphthalene, and azulene).

We have compiled the X1/07 data set (393), which added X1-1 [see Table I (Refs. 19–24)], a new set of 170 neutral molecules up to 32 heavy atoms (n-$C_{32}H_{66}$), to the G3/99 set. We included more radicals (e.g., NO_3 and SF_5) and inorganic compounds (e.g., N_3 and C_3O_2), and we added more complicated structures such as larger polyhalogenated species [e.g., $C(NF_2)_4$ and C_6F_{10}], highly branched alkanes [e.g., $(CH_3)_2CHCH(CH_3)C(CH_3)_3$], and strained organic systems [e.g., housane (C_5H_8)]. We have extended the size of the polynuclear aromatic hydrocarbons, including molecules such as pyrene ($C_{16}H_{10}$), chrysene ($C_{18}H_{12}$), perylene ($C_{20}H_{12}$), etc. We have also included some more phosphine

[a)] Author to whom correspondence should be addressed. Electronic mails: xinxu@xmu.edu.cn and xinxu@wag.caltech.edu

0021-9606/2007/127(21)/214105/8/$23.00

TABLE I. The X1-1 set: Experimental data for heats of formation at 298 K.

No.	Molecule	Expt.	Reference	No.	Molecule	Expt.	Reference
1	1*H*-tetrazole	78.06	19	86	*n*-undecane	−64.56	20
2	5-aminotetrazole	77.39	19	87	Acenaphthene	37.30	21
3	Methyl nitrate	−29.16	19	88	Biphenyl	43.40	21
4	Thiourea	5.47	19	89	Diphenyl sulfide	55.30	21
5	Hydrazinecarbothioamide	30.60	19	90	$C_6H_5{-}S{-}S{-}C_6H_5$	58.20	21
6	Methyl chlorosilane	−50.20	21	91	Diethyl phthalate	−164.40	20
7	Carbonothioic dihydrazide	62.20	19	92	1-dodecene	−39.50	20
8	$C(NF_2)_4$	0.20	21	93	*n*-dodecanoic acid	−153.30	20
9	Tetranitromethane	19.70	21	94	Acenaphthylene	62.10	21
10	Carbonyl chloride	−52.30	20	95	Carbazole	50.00	21
11	$CH_2{=}CF_2$	−80.50	21	96	Methyl dodecanoate	−146.20	20
12	1,2,2-trichloroethane	−35.40	19	97	Acridine	65.50	21
13	1,1,1-trifluoroethane	−178.00	20	98	Phenanthrene	49.50	21
14	Oxamide	−92.52	19	99	(*Z*)-stilbene	60.30	21
15	Ethanedithioamide	19.80	19	100	1,2-diphenylethane	34.20	20
16	1*H*-tetrazole, 5-methyl-	67.09	19	101	Fluoranthene	69.10	21
17	1*H*-tetrazole, 1-methyl-	77.17	19	102	Dibutyl phthalate	−179.30	20
18	Thiolacetic acid	−41.80	22	103	*n*-decylbenzene	−32.80	20
19	Ethylnitate	−36.80	22	104	*n*-hexadecanoic acid	−176.00	20
20	Ethanethioamide	2.30	23	105	*n*-hexadecane	−89.20	20
21	Urea, methyl-	−55.80	19	106	1-hexadecanol	−125.50	20
22	Dimethyl sulfite	−115.50	19	107	Benz[a]anthracene	70.00	21
23	Dimethyl sulfate	−164.10	20	108	*p*-terphenyl	66.20	20
24	2,3-dithiabutane	−5.80	19	109	Triphenylphosphine	76.50	21
25	Chlorodimethylsilane	−67.34	21	110	*n*-octadecane	−99.00	22
26	Ethylenediamine	−4.07	19	111	Perylene	75.40	21
27	Dimethylsilane	−22.60	20	112	*n*-$C_{32}H_{66}$	−166.50	22
28	1,2-propanediamine	−12.80	19	113	Aluminum dimer	116.40	23
29	Trimethylsiliconthydroxide	−120.00	20	114	Aluminum trichloride dimer	−310.00	21
30	Malononitrile	63.65	19	115	Aluminum chloride	−12.30	21
31	1,3,5-triazine	53.98	19	116	B_2Cl_4	−117.00	21
32	Oxazole	−3.70	19	117	B_2F_4	−342.00	21
33	1,3-dithiolan-2-one	−30.10	21	118	OBBO	−109.00	21
34	1,3-dithiolan-2-thione	22.40	19	119	$B_3O_3Cl_3$	−390.00	21
35	1,3,5-trioxane	−111.32	19	120	$B_3O_3F_3$	−565.00	21
36	Thiacyclobutane	14.60	19	121	Beryllium dimer	152.30	19
37	Sarcosine	−87.77	19	122	$O(BeF)_2$	−288.00	21
38	Urea, ethyl-	−61.53	19	123	$BeCl_2$	−86.10	21
39	Dimethoxymethane	−83.20	19	124	BeO	32.60	19
40	Methylethyl sulfone	−97.60	19	125	MgF_2	−174.00	21
41	1-propanthiol	−16.40	19	126	Sulfur tetrafluoride	−182.00	21
42	Methylethylthioether	−14.40	19	127	Sulfur pentafluride	−217.12	23
43	Propane-1,3-dithiol	−7.10	19	128	Aluminumtrifluoride dimer	−629.50	23
44	Trimethyl phosphite	−169.00	21	129	Aluminum fluoride	−63.50	19
45	Carbon suboxide	−22.40	19	130	Aluminum chloride fluoride	−104.20	24
46	Sec-butanol	−70.00	19	131	Aluminum dichloridefluoride	−189.00	21
47	1,4-butanediol	−102.00	19	132	MgF($^2\Sigma^+$)	−56.60	21
48	Diethyl peroxide	−46.10	23	133	NaF	−69.42	19
49	Diethyl sulfoxide	−49.10	19	134	PF (triplet)	−12.50	19
50	*Trans*-2-butenedinitrile	81.30	19	135	(*Z*)-diazine	50.90	23
51	Maleic anhydride	−95.20	20	136	Sulfuric acid	−175.13	19
52	Diketene	−45.47	19	137	Hydrogen disulfide	3.70	19
53	4-methylthiazole	26.70	19	138	Hydrogen tetrasulfide	10.60	21
54	Ethylacetylene	39.50	19	139	Hydrogen pentasulfide	13.80	21
55	2,3-butadione	−78.10	19	140	BH_3	25.50	19
56	*s*-ethylthioacetate	−54.50	19	141	Si_2H_4	65.70	21
57	2-butynedinitrile	126.50	21	142	Si_2H_5	53.30	21
58	1-chloropentane	−41.90	19	143	$P(SiH_3)H_2$	1.80	21
59	$Si(CH_3)_3OC_2H_5$	−119.00	21	144	B_2H_6	8.74	24
60	$(CH_3)_3SiN(CH_3)_2$	−59.30	21	145	Trisilane	28.90	22

TABLE I. (*Continued.*)

No.	Molecule	Expt.	Reference	No.	Molecule	Expt.	Reference
61	Bicyclo[2.1.0]-pentane	37.30	19	146	B_5H_9	17.50	21
62	Perfluorocyclohexene	−461.90	23	147	LiOH	−54.73	19
63	1-methylcyclopentene	−0.86	19	148	Hydrazoic acid	70.30	19
64	1,5-hexadiene	20.10	20	149	Nitrous acid, *trans*	−18.80	21
65	Triethyl phosphate	−194.40	21	150	Nitric acid	−32.10	21
66	$Si(CH_3)_2(OC_2H_5)_2$	−186.00	21	151	Fluorosulfonic acid	−180.00	19
67	Triethylene tetramine	0.80	20	152	NaOH	−45.65	19
68	Hexamethyldisiloxane	−185.80	20	153	$(LiCl)_2$	−143.10	21
69	1,3-dichlorobenzene	6.10	19	154	$(LiCl)_3$	−233.10	24
70	Nitrobenzene	16.20	20	155	LiNa	43.40	21
71	3,3-dimethylpentane	−48.20	20	156	Magnesium dimer	68.75	19
72	*n*-heptyl mercaptan	−35.80	22	157	$MgCl_2$	−93.80	21
73	Benzothiazole	48.80	21	158	Azide	99.00	23
74	2-chlorobenzaldehyde	−15.10	21	159	Na_2Cl_2	−135.30	19
75	Benzaldehyde	−8.80	20	160	Nitrate radical	17.00	19
76	Benzoic acid	−69.30	20	161	Nitrosylfluoride	−15.70	21
77	2,4,4-trimethyl-2-pentene	−25.10	22	162	Aluminumdioxide	−20.60	23
78	2,3,4-trimethylpentane	−52.00	22	163	Aluminumoxide dimer	−94.30	23
79	Dibutyl sulfide	−40.05	22	164	Sulfuryl fluoride	−181.30	23
80	Tetraethylsilane	−71.00	21	165	Dialuminum oxide	−34.70	21
81	Tetraethylenepentamine	3.01	20	166	Phosphorus oxyfluoride	−299.80	23
82	Styrene	35.10	19	167	Na_2O	−8.60	21
83	2,6-dimethyl-4-heptanone	−85.50	20	168	Phosphorus oxide	−6.66	19
84	2,2,3,4-tetramethylpentane	−56.60	20	169	Disulfur monoxide	−13.39	19
85	*n*-decyl mercaptan	−50.50	20	170	Sulfur octamer (cyclic)	24.00	19

and silane compounds which are less abundant in the G3/99 set. The X1/07 data set is found to be more balanced for training the NN parameters.

We have optimized the equilibrium geometry of each molecule at the level of B3LYP/6-311+G(d,p). This level of theory has been shown to give reliable geometric prediction.[14,25] Analytical harmonic frequency was calculated at the same level to give the zero-point energy (ZPE) and thermocorrections and to ensure that each geometry corresponded to a true local minimum. Enthalpies at 298 K for molecules were obtained by single point energy calculations at the level of B3LYP/6-311+G($3df$,$2p$), corrected by the unscaled ZPE and thermocorrection of heats H_{0-298} of 6-311+G(d,p). The standard heats of formation ΔH_f^θ at 298 K were calculated in the same manner as Curtiss *et al.*[4,6] by first subtracting the calculated atomization energies from the known experimental heats of formation of the isolated atoms and then adding the thermocorrections. Our method differs from that of Curtiss *et al.*[4,6] in that they have adopted the MP2(full)/6-31G(d) geometries and the scaled ZPEs at the HF/6-31G(d) level when they made the assessment of B3LYP. All calculations were performed by using the GAUSSIAN 03 suite of programs.[26]

III. THE NEURAL NETWORK APPROACH

In their NN scheme, Hu *et al.*[14] chose $\Delta H_f^{\mathrm{B3LYP}}$ (the B3LYP calculated heats of formation), N_a (the total number of atoms in a molecule), ZPE (the calculated zero-point vibrational energy), and N_{db} (the number of double bonds). We agree that $\Delta H_f^{\mathrm{B3LYP}}$, N_a, and ZPE are the reasonable physical descriptors for enthalpy calculations (or corrections), but we discard N_{db} as it may bias against other bonding types such as single bonds and triple bonds. It is widely recognized that the theoretical error for ΔH_f^θ per atom tends to accumulate in large systems; This error depends on the specific atom type.[15,27,28]The other well-known observation is that as the number of the electrons increases, higher-order electron correlation effects are difficult to account for increase.[29] So we decide to include the total number of electrons in the system as a descriptor, N_e, and the number of each constituent elements (e.g., N_{H}, N_{C}, N_{N}, N_{O}, N_{F}, N_{Si}, N_{P}, N_{S}, and N_{Cl}) as the other descriptors. This final choice echoes many successful schemes based on atom additivity.[15,28]

We notice that there are other ways to choose the descriptors.[16–18] Very recently, Friesner *et al.* presented a localized orbital correction (LOC) model,[18] which significantly improves the accuracy of DFT methods for the prediction of ΔH_f^θ of neutral molecules. Their analysis on the DFT residual errors are very illuminating. Based on a classical valence bond picture, the LOC model made the atomic corrections as a function of hybridization states, and the bond corrections plus the bond environmental and radical environmental corrections. They showed that the corrections were so powerful that the B3LYP-LOC method led to MAD of only 0.8 kcal/mol for the G3/99 set. In the present work, the bond type corrections are not included for simplicity. We plan to include them in our future work in order to achieve an accurate description of isomerization energy.

Our NN adopts a three-layer architecture [see Fig. 1 (Refs. 14, 28, and 30)], which has an input layer consisting of inputs from the physical descriptors, a hidden layer con-

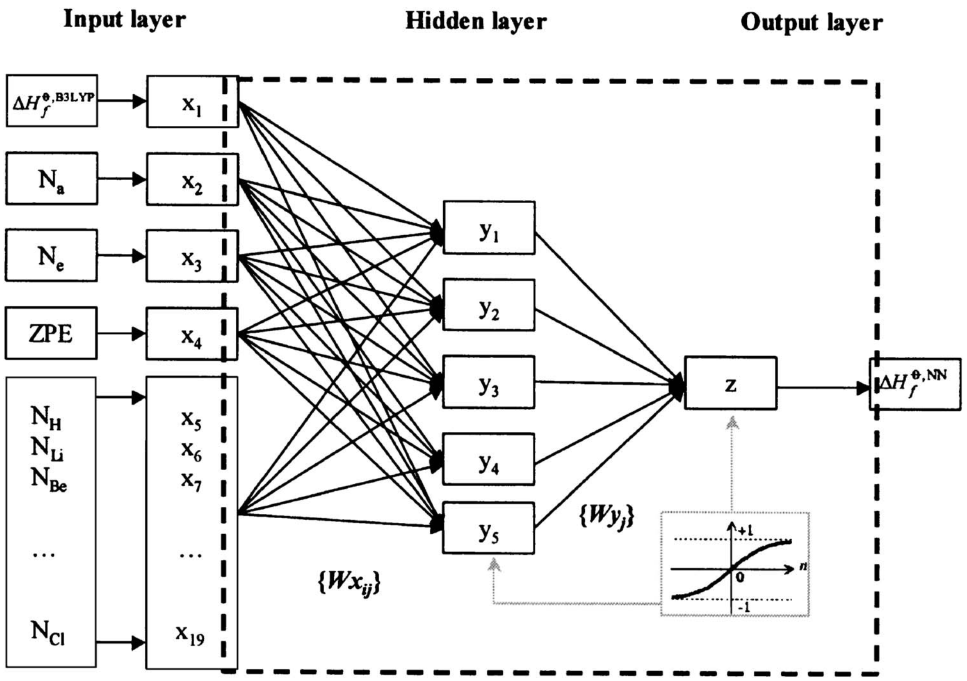

FIG. 1. Topological structure for neural network.

taining a number of hidden neurons, and an output layer that outputs the corrected values for ΔH_f^θ. The number of hidden neurons is to be determined. We find that the hidden layer containing five neurons yields the best overall results. $\{Wx_{ij}\}$ and $\{Wy_j\}$ are sets of the connection weights, where $\{Wx_{ij}\}$ connects the input neurons and the hidden neurons, and $\{Wy_j\}$ connects the hidden neurons and the output neuron. The connection weights $\{Wx_{ij}\}$ and $\{Wy_j\}$ are optimized by using the general back propagation (BP) algorithm against the training set.[30] To prevent optimization from trapping into a local minimum, we combined BP with a genetic algorithm, which is used to find good initial connection weights.[30] The final weights are summarized in Table II.

Upon applications, the input data will be scaled into x_i in the range from −0.9 to 0.9 by using Eq. (1),

$$X = \frac{\text{Input value}}{\text{Model maximum}} \times 0.9. \tag{1}$$

For $\Delta H_f^{\mathrm{B3LYP}}$, ZPE, N_a, N_e, N_{H}, N_{C}, and N_{F}, the model maxima are 1000.0, 2.0, 100, 1000, 100, 100, and 20. We set 10 for the maximum numbers of the other elements. Application beyond the model maxima should be avoided. The training set does not contain any charged species, to which an application of the present method is not recommended.

The hyperbolic tangent sigmoid transfer function is employed in this work, which has the form as

$$f(a) = \frac{2}{1 + e^{-2a}} - 1. \tag{2}$$

The value of a hidden neuron (y_j) and the NN output (z) can be obtained by the following functions with the connection weights $\{Wx_{ij}\}$ and $\{Wy_j\}$ listed in Table I:

$$y_j = f\left(\sum_{i=1}^{19} Wx_{ij} \times x_i + \text{Bias } X_j\right), \tag{3}$$

$$z = f\left(\sum_{j=1}^{5} Wy_j \times y_j + \text{Bias } Y\right). \tag{4}$$

The output z can be recovered to give $\Delta H_f^{\theta,\mathrm{NN}}$ by using Eq. (5),

$$\Delta H_f^\theta = \frac{z}{0.0009}. \tag{5}$$

Details on how to apply the NN correction may be found on the Web site.[31]

IV. RESULTS AND DISCUSSION

Accurate prediction of ΔH_f^θ demands a high level calculation of the correlation energy. This is a nontrivial problem as the correlation energy only occupies a tiny fraction of the total energy. At this point, it should be emphasized that the Gn theory is a composite method. The G2 theory aims to reproduce effectively the quadratic configuration interaction QCISD(T,FC)/6-311+G($3df$,$2p$) energies through a series of calculations at a lower level. The G3 energy is effective at the QCISD(T, Full)/G3large level if the different additivity approximations work well. The G3 large basis set is similar to 6-311+G($3d2f$,$2df$,$2p$), which uses $3d2f$ polarization on the second row, $2df$ polarization on the first row, and $2p$ on hydrogen. Even when the extrapolation in the one-particle and many-particle spaces is done to these levels, there are still MADs as high as 15.22 kcal/mol for ΔH_f^θ in the G3/99 set for the G3 theory and 11.36 kcal/mol in the G2/97 set for the G2 theory (see Table III). A higher level correction (HLC) procedure was designed to compensate the remaining deficiencies of the method, which was parametrized against some experimental data.[1–6] Specifically, HLC of G3 counted the numbers of valence electrons with α and β spins for atomic and molecular systems separately and all four parameters were fitted against the experimental energies of the

TABLE II. Final weights and biases.

$(W_{x_{ij}})$	y_1	y_2	y_3	y_4	y_5
x_1 ($\Delta H_f^{\mathrm{B3LYP}}$)	−1.260 80	1.198 69	−0.086 57	0.826 17	0.881 19
x_2 (N_a)	−2.458 86	1.942 32	−0.874 87	0.173 19	1.796 08
x_3 (N_e)	−1.050 17	1.283 23	−0.342 49	−0.176 54	1.114 54
x_4 (ZPE)	−0.103 60	2.011 83	0.334 67	−1.263 84	0.642 16
x_5 (N_{H})	−2.343 21	1.568 38	0.153 28	0.327 57	0.177 04
x_6 (N_{Li})	0.187 42	0.367 22	0.190 59	−0.266 87	0.430 58
x_7 (N_{Be})	1.030 58	0.600 51	−0.005 20	−0.565 85	−0.887 65
x_8 (N_{B})	−0.590 06	0.632 16	−0.042 02	−0.049 41	0.174 34
x_9 (N_{C})	−0.705 88	0.027 72	−1.032 87	−0.299 35	1.555 82
x_{10} (N_{N})	0.086 62	−0.199 50	−0.282 37	−0.045 57	0.424 11
x_{11} (N_{O})	−0.396 30	0.582 47	1.314 14	0.290 25	−1.278 13
x_{12} (N_{F})	−1.357 23	1.560 90	1.586 57	0.335 24	−1.395 52
x_{13} (N_{Na})	1.296 70	0.358 38	1.066 40	−0.425 61	−1.614 81
x_{14} (N_{Mg})	−1.080 01	−0.260 65	0.203 20	0.461 69	0.799 95
x_{15} (N_{Al})	0.654 65	0.311 34	0.682 35	−0.170 77	−1.470 72
x_{16} (N_{Si})	0.798 69	−0.189 29	2.367 10	0.351 37	−2.161 69
x_{17} (N_{P})	−0.489 78	0.724 82	0.189 32	−0.013 74	−0.617 37
x_{18} (N_{S})	−1.149 97	1.219 86	−1.144 89	−0.336 16	0.470 43
x_{19} (N_{Cl})	−0.636 87	0.615 87	−0.372 42	−0.143 66	0.327 50
Bias*X*	−0.609 73	1.237 97	0.521 79	0.141 30	1.763 60
(W_{y_j})	z				
y_1	0.774 73				
y_2	2.061 69				
y_3	−0.291 43				
y_4	1.154 92				
y_5	1.021 85				
Bias*Y*	−2.311 86				

complete G2/97 set including 85 ionization potentials, 58 electron affinities, 8 proton affinities, as well as 148 heats of formation.[6] After HLC, the final G3 accuracy on average for heats of formation of 223 molecules is just 1.05 kcal/mol, while that for G2 of the 148 molecules is 1.57 kcal/mol. This demonstrated that the remaining errors for *ab initio* methods are very systematic and HLC for electron pairs is highly effective. In a manner exactly analogous to those used for the G2 and G3 theories, Curtiss *et al.*[6] have tried to derive HLC for B3LYP. Unfortunately, it turned out that the G3 HLC-like procedure was less effective for B3LYP, reducing MAD for the G3/99 set only from 4.27 to 3.31 kcal/mol. Here, we show that in B3LYP combined with a NN correction, MADs of 1.36 and 1.43 kcal/mol are achieved for the prediction of heats of formation in the G2/97 set and the G3/99 set, respectively (see Table III). The NN approach involves nonlinear functions to correct the calculated data, which partially explains its success over the linear procedure of G3 HLC-like correction.

Figure 2 shows the histogram for various methods against the G3/99 set. About 63% (141/223) of the G3 deviations fall within the range from −1 to 1 kcal/mol. This is substantially better than the G2 theory for which 39% (87/223) of the deviations fall in this range. Impressively, there are 48% (106/223) of the X1 deviations clustering

TABLE III. Summary of mean absolute deviations (MADs) (kcal/mol) for different methods against different data sets. For each entry, the maximum MAD is given in parentheses.

	G3	G3-HLC[a]	G2	G2-HLC[a]	B3LYP	X1
G2/97	0.93	10.84	1.57	11.36	3.40	1.36
(148,6)[b]	(4.90)	(23.97)	(8.20)	(30.05)	(20.27)	(6.49)
G3/99	1.05	15.22	1.88	16.58	5.58	1.43
{223,10}[b]	(7.10)	(42.78)	(9.39)	(48.42)	(22.22)	(6.49)
X1-1	⋯	⋯	⋯	⋯	13.60	1.53
{170,32}[b]	⋯	⋯	⋯	⋯	(85.08)	(7.80)
X1/07	⋯	⋯	⋯	⋯	9.05	1.48
{393,32}[b]	⋯	⋯	⋯	⋯	(85.08)	(7.80)

[a] G*n*-HLC indicates that the higher level correction is removed from the G*n* method.

[b] Followed with the set name are the number of molecules and the maximum number of nonhydrogen atoms.

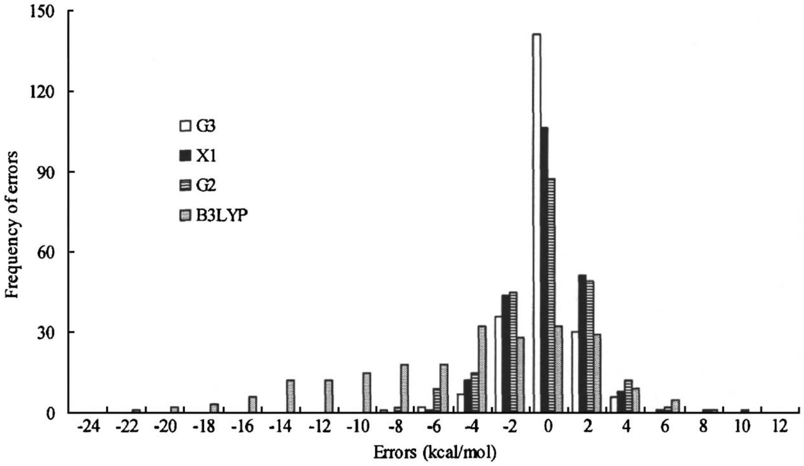

FIG. 2. Histogram of G3, G2, B3LYP, and X1 for the heats of formation of the G3 set. Each vertical bar represents deviations in 2 kcal/mol range. Errors are defined as (Expt.-Theor.).

within the “−1−1” interval. This is a significant improvement over the B3LYP method, where only 14% (32/223) of the errors fall within this interval. The maximum negative deviations [−7.10(G3), −7.18(G2), and −5.56(X1) kcal/mol] occur at PF_5, SiF_4, and CF_3CN, respectively, whereas the maximum positive deviations [4.90 (G3), 9.39 (G2), and 6.49(X1) kcal/mol] occur at C_2F_6, C_2F_4, and C_2F_4, respectively. The situation with X1 is much improved when compared to that with B3LYP, for which the error distribution is between −17.4 [*n*-octane (C_8H_{18})] and 8.0(BeH) kcal/mol. It is important to note that for the G*n* theory without HLC, there is a clear underbinding tendency for the *ab initio* methods; the error intervals are between −42.8 [*n*-octane (C_8H_{18})] and 1.4(Na_2) kcal/mol for G3-HLC and −48.4 [*n*-octane (C_8H_{18})] and −0.5(Na_2) kcal/mol for G2-HLC.

Figure 3 shows the histogram of B3LYP before and after the NN correction for the X-1 set. More than 45% (77/170) of the X1 deviations fall within the range from −1 to 1 kcal/mol. Significantly, there are 86% (147/170) of the X1 deviations clustering within the “−2−2” interval. This is a substantial improvement over the original B3LYP method without NN correction, where only 5% (9/170) of the errors fall within this interval and the −2−2 interval only covers 15% (26/170). Disturbingly, B3LYP gives 28% (48/170) of the molecules whose errors for the prediction of heats of formation are higher than 20.0 kcal/mol. The maximum negative deviations [−85.08 (B3LYP) and −7.80(X1) kcal/mol] occur at *n*-$C_{32}H_{66}$ and 2,2,3,4-tetramethylpentane (C_9H_{20}), respectively, whereas the maximum positive deviations [11.56 (B3LYP) and 5.84(X1) kcal/mol] occur at HN_3 and C_3O_2, respectively. There is an increasing underbinding tendency, i.e., increased frequency of negative errors, for the B3LYP method as the size of the molecule is increased, whereas X1 leads to a balanced error distribution.

Table IV summarizes the occurrence of improvement versus degradation after NN correction for the G3/99 set and the X1-1 set. For the G3/99 set, nearly 82% (182/223) is improved, while only 18% (41/223) is degraded upon NN correction. Among the degraded systems, 56% (23/41) is within the “0.0–1.0” interval, which is, thus, within the experimental uncertainty. Only 8 out of 41 are notably degraded by 2.0−4.0 kcal/mol [$CH(CH_3)_2$(2.07), OCS (2.56), CO_2(2.75), ClF_3(2.79), C_2F_4(2.89), $ClCHCH_2$(2.99), Si_2(3.16), and CF_3CN (3.56 kcal/mol)]. These would be the difficult molecules where the present NN corrections are less effective. Among the improved systems, 26% (46/182) is within the “5.0–10.0” interval and 33 out of 182 are improved by more than 10.0 kcal/mol. Improvement is most significant for some hypervalent molecules such as SO_2Cl_2(15.2), PF_5(16.2), PCl_5(16.3), and SF_6 (20.3 kcal/mol). The X-1 set contains molecules of larger size (e.g., *n*-$C_{32}H_{66}$, l-hexadecanol, and *n*-dodecanoic acid)

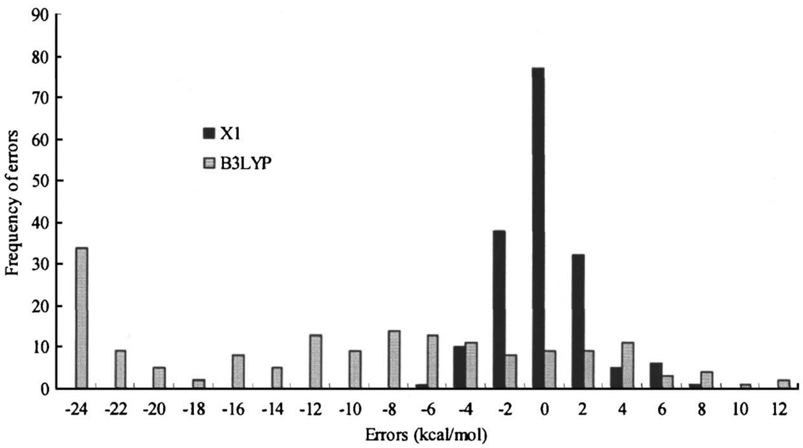

FIG. 3. Histogram of B3LYP and X1 for the heats of formation of the X-1 set. Each vertical bar represents deviation in 2 kcal/mol range. For B3LYP, hits with MAD larger than 24 kcal/mol are grouped together. Errors are defined as (Expt.-Theor.).

TABLE IV. Number of occurrence for improvement or degradation after NN correction on top of the B3LYP/6-311+G($3df,2p$) results.

	The G3/99 set (223)	The X-1 set (170)
$I(>10.0)$[a]	33	77
$I(5.0-10.0)$	46	35
$I(2.0-5.0)$	44	26
$I(1.0-2.0)$	24	11
$I(0.0-1.0)$	35	9
$D(0.0-1.0)$	23	9
$D(1.0-2.0)$	10	2
$D(2.0-4.0)$[b]	8	1

[a] For the G3/99 set, there are 33 molecules for which the NN correction reduces the error by more than 10.0 kcal/mol.

[b] For the X-1 set, there is only one molecule for which the NN correction degrades the B3LYP result by 2.0–4.0 kcal/mol.

or with more demanding chemical environment [e.g., S_8, Al_2F_6, H_2SO_4, $(CH_3)_3SiN(CH_3)_2$, and benz(a)anthracene]. In these cases, B3LYP leads to notorious errors, whereas the X1 method shows its power and efficiency. For the X-1 set, nearly 93% (158/170) is improved. Among the improved systems, 22% (35/158) is within the 5.0–10.0 interval and 77 out of 158 are improved by more than 10.0 kcal/mol. Some impressive examples are B3LYP gives errors of −29.44(Al_2Cl_6), −31.36(Al_2F_6), −16.55(C_6F_{10}), −29.22 [benz(a)anthracene ($C_{18}H_{12}$)], −28.01 [pyrene ($C_{16}H_{10}$)], −21.88[$(CH_3)_3Si-N(CH_3)_2$], and −36.52 kcal/mol [triphenylphosphine ($C_{18}H_{15}P$)], while X1 reduces the errors to 0.47, −0.21, −0.32, 1.66, −0.87, 0.31, and −0.15 kcal/mol, respectively. Only 7% (12/170) is degraded upon NN correction. In fact, 9 out of 12 for the degraded systems are within the experimental uncertainty, i.e., within the 0.0–1.0 interval. There is only one system [i.e., sarcosine (CH_3NHCH_2COOH)] for which B3LYP gives error of 0.67 kcal/mol, while NN increases the error to 4.12 kcal/mol. When we calculated this molecule with the G3 theory, we obtained $\Delta H_f^{\theta}=-92.8$ kcal/mol. This number should be compared with values obtained with other methods, i.e., −91.9 (X1), −88.4 (B3LYP), and −87.8 (experimental). The agreement between G3 and X1 makes us to challenge the reliability of the experimental data. Our NN corrections are very significant and yet very effective with no additional cost upon application. Hence, the X1 method can be used as a valuable complement to the experiment for thermochemistry of large molecules.

V. CONCLUSION

In this paper, we present a composite method for the calculation of the standard heats of formation at 298 K for neutral molecular species in gas phase, containing H and first row (Li–F) and second row (Na–Cl) elements. We combine the DFT results at the level of B3LYP/6-311+G($3df,2p$)//B3LYP/6-311+G(d,p) with a NN correction. This B3LYP-NN method significantly eliminates the notorious size dependent errors of B3LYP and reduces its MAD from 5.58 to 1.43 kcal/mol for the G3/99 set, which is comparable to the accuracy of the Gn theories [1.88 (G2) and 1.05(G3) kcal/mol]. This method also gives good performance for a newly compiled data set (X1-1). This set contains additional 170 neutral molecules, many of which have larger sizes or more diverse chemical structures. While the original MAD for B3LYP is as high as 13.60 kcal/mol, it reduces to 1.53 kcal/mol after NN correction.

We name the B3LYP-NN method the X1 method, as it is our first version, developed by the Xiamen group. The X1 method offers an express way to calculate accurate heats of formation of larger molecules, inaccessible by accurate wavefunction-based methods such as G2 or G3. The X1 method greatly extends the reliability and applicability of the B3LYP method. Works on improving the DFT description of charged species, noncovalent bondings, isomerization energy, energy gaps, and reaction barriers are in progress.

ACKNOWLEDGMENTS

This work is supported by NSFC (Nos. 20525311, 20533030, and 10774126) and the Ministry of Science and Technology (Nos. 2007CB815206 and 2004CB719902).

[1] J. A. Pople, M. Head-Gordon, D. J. Fox, K. Raghavachari, and L. A. Curtiss, J. Chem. Phys. **90**, 5622 (1989).
[2] L. A. Curtiss, C. Jones, G. W. Trucks, K. Raghavachari, and J. A. Pople, J. Chem. Phys. **93**, 2537 (1990).
[3] L. A. Curtiss, K. Raghavachari, G. W. Trucks, and J. A. Pople, J. Chem. Phys. **94**, 7221 (1991).
[4] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, J. Chem. Phys. **106**, 1063 (1997).
[5] L. A. Curtiss, K. Raghavachari, P. C. Redfern, V. Rassolov, and J. A. Pople, J. Chem. Phys. **109**, 7764 (1998).
[6] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, J. Chem. Phys. **112**, 7374 (2000).
[7] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993); C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988); A. D. Becke, Phys. Rev. A **38**, 3098 (1988).
[8] W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory* (Wiley-VCH Verlag GmbH, Weinheim, 2001).
[9] P. R. Schreiner, A. A. Fokin, R. A. Pascal, Jr., and A. De Meijere, Org. Lett. **8**, 3635 (2006); S. Grimme, C. Diedrich, and M. Korth, Angew. Chem., Int. Ed. **45**, 625 (2006); C. E. Check and T. M. Gilbert, J. Org. Chem. **70**, 9828 (2005).
[10] J. Cioslowski, G. Liu, and P. Piskorz, J. Phys. Chem. A **102**, 9890 (1998).
[11] H. Li, L. L. Shi, M. Zhang, Z. M. Su, X. J. Wang, L. H. Hu, and G. H. Chen, J. Chem. Phys. **126**, 144101 (2007).
[12] J. M. Seminario, M. G. Maffei, L. A. Agapito, and P. F. Salazar, J. Phys. Chem. A **110**, 1060 (2006).
[13] N. L. Haworth and G. B. Bacskay, J. Chem. Phys. **117**, 11175 (2002).
[14] L. Hu, X. Wang, L. Wong, and G. Chen, J. Chem. Phys. **119**, 11501 (2003).
[15] P. Winget and T. Clark, J. Comput. Chem. **25**, 725 (2004).
[16] X.-M. Duan, G.-L. Song, Z.-H. Li, X.-J. Wang, G.-H. Chen, and K.-N. Fan, J. Chem. Phys. **121**, 7086 (2004).
[17] D. A. Long and J. B. Anderson, Chem. Phys. Lett. **402**, 524 (2005).
[18] R. A. Friesner, E. H. Knoll, and Y. Cao, J. Chem. Phys. **125**, 124107 (2006).
[19] NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101, edited by R. D. Johnson III, 2005 (http://srdata.nist.gov/cccbdb).
[20] J. A. Dean, *Lange's Handbook of Chemistry*, 15th ed. (McGraw-Hill, New York, 1999).
[21] J. Cioslowski, M. Schimeczek, G. Liu, and V. Stoyanov, J. Chem. Phys. **113**, 9377 (2000).
[22] D. R. Lide, *CRC Handbook of Chemistry and Physics*, 82nd ed. (CRC, Boca Raton, 2001).
[23] H. Y. Afeefy, J. F. Liebman, and S. E. Stein, Neutral Thermochemical Data, in NIST Chemistry WebBook, NIST Standard Reference Database Number 69, edited by P. J. Linstrom and W. G. Mallard, National Institute of Standards and Technology, Gaithersburg, MD, 20899, 2005

(http://webbook.nist.gov).

[24] B. J. McBride, M. J. Zehe, and S. Gordon, NASA Glenn Coefficients for Calculating Thermodynamic Properties of Individual Species (http://gltrs.grc.nasa.gov/GLTRS).

[25] M. P. Andersson and P. Uvdal, J. Phys. Chem. A **109**, 2937 (2005).

[26] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, R. F. K. Toyota, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, GAUSSIAN 03, Revision D.01, Gaussian, Inc., Wallingford, CT, 2004.

[27] P. C. Redfern, P. Zapol, L. A. Curtiss, and K. Raghavachari, J. Phys. Chem. A **104**, 5850 (2000).

[28] S. Kondo, A. Takahashi, K. Tokuhashi, A. Sekiya, Y. Yamada, and K. Saito, J. Fluorine Chem. **117**, 47 (2002).

[29] L. A. Curtiss, P. C. Redfern, and K. Raghavachari, J. Chem. Phys. **123**, 124107 (2005).

[30] X. Yao, Proc. IEEE **87**, 1423 (1999).

[31] See http://www.pcoss.org/users/xinxu