THE JOURNAL OF
PHYSICAL CHEMISTRY A

Article

pubs.acs.org/JPCA

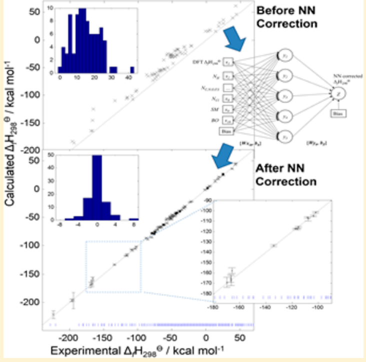
Alternative Approach to Chemical Accuracy: A Neural Networks-Based First-Principles Method for Heat of Formation of Molecules Made of H, C, N, O, F, S, and Cl

Jian Sun,*[†] Jiang Wu,[†] Tao Song,[†] LiHong Hu,[‡] KaiLu Shan,[†] and GuanHua Chen*[†]

[†]Department of Chemistry, The University of Hong Kong, Hong Kong, China

[‡]School of Computer Science and Information Technology, Northeast Normal University, Changchun, Jilin, China

S *Supporting Information*



ABSTRACT: The neural network correction approach that was previously proposed to achieve the chemical accuracy for first-principles methods is further developed by a combination of the Kennard−Stone sampling and Bootstrapping methods. As a result, the accuracy of the calculated heat of formation is improved further, and moreover, the error bar of each calculated result can be determined. An enlarged database (Chen/13), which contains a total of 539 molecules made of the common elements H, C, N, O, F, S, and Cl, is constructed and is divided into the training (449 molecules) and testing (90 molecules) data sets with the Kennard−Stone sampling method. Upon the neural network correction, the mean absolute deviation (MAD) of the B3LYP/6-311+G(3df,2p) calculated heat of formation is reduced from 10.92 to 1.47 kcal mol^{-1} and 14.95 to 1.31 kcal mol^{-1} for the training and testing data sets, respectively. Furthermore, the Bootstrapping method, a broadly used statistical method, is employed to assess the accuracy of each neural-network prediction by determining its error bar. The average error bar for the testing data set is 1.05 kcal mol^{-1}, therefore achieving the chemical accuracy. When a testing molecule falls into the regions of the "Chemical Space" where the distribution density of the training molecules is high, its predicted error bar is comparatively small, and thus, the predicted value is accurate as it should be. As a challenge, the resulting neural-network is employed to discern the discrepancy among the existing experimental data.

■ INTRODUCTION

First-principles quantum mechanical methods have become indispensable research tools in chemistry, condensed matter physics, materials science, molecular biology, electronics, and many other fields.[1−4] Experimentalists rely increasingly on these methods to interpret their findings. And recently, there have been increasing interests in applying first-principles methods to design or discover novel material,[5] which is also referred as "Computational Materials Design/Discovery".[6] Despite their successes, first-principles methods are not accurate enough yet to reach the chemical accuracy, i.e., ∼1 kcal mol^{-1} in energy, the thermal energy at the room temperature. As such, the ultimate objective of the field, to predict the properties prior to the measurement, is yet to be realized. To reach the chemical accuracy, more accurate theories and numerical methodologies have been developed from first-principles. For instance, on the basis of a coupled-cluster method (QCISD(T)),[7] the G$_n$ (n = 1−4) methods are quite accurate; however, they are limited to small systems as the computational resources are very demanding.[8−12] For density-functional theory (DFT), efforts have been focused on constructing better exchange−correlation functionals.[13,14] Various exchange−correlation functionals have been developed, and yet, the goal of achieving the chemical accuracy is still elusive.

In 2003, an alternative approach was proposed for first-principles methods to reach the chemical accuracy. Chen and co-workers introduced a neural network-based method to improve the accuracy of the B3LYP calculated heat of formation ($\Delta_f H_{298}^{\theta}$) for organic molecules.[15] The key idea is that all the errors of first-principles calculations are systematic and can thus be corrected in principle. A neural network (NN) is introduced to establish the quantitative relationship between the characteristic properties (or descriptors) and the experimental $\Delta_f H_{298}^{\theta}$ of a molecule. A total of 180 experimental results were employed to train and test the NN. As a result, the accuracy of the calculated $\Delta_f H_{298}^{\theta}$ is improved almost by 1 order of magnitude. For instance, for the selected 180 organic molecules, the root-mean-square deviation (RMS) of the calculated $\Delta_f H_{298}^{\theta}$ was successfully reduced from the original 21.4 to 3.1 kcal mol^{-1} for B3LYP/6-311+G(d,p) calculation.

Following the work in 2003, several research groups around the world employed the NN and other machine learning or artificial intelligence methods to calibrate and improve various

Special Issue: International Conference on Theoretical and High Performance Computational Chemistry Symposium

Received: February 28, 2014
Revised: June 30, 2014
Published: June 30, 2014

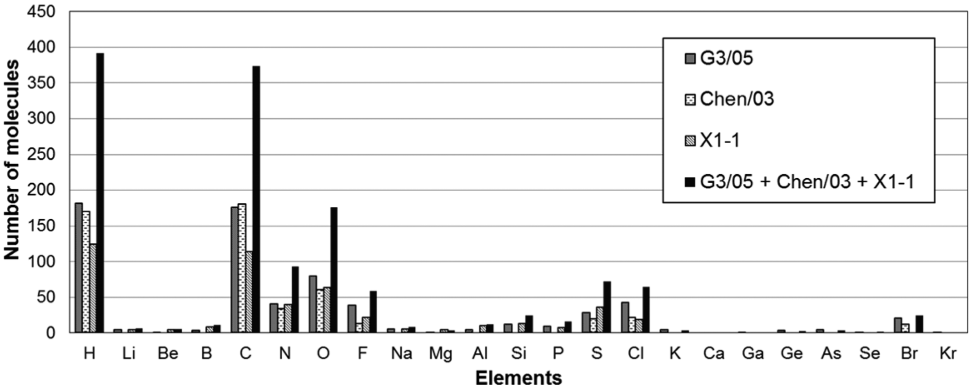


Figure 1. Number of molecules containing the specified element in the G3/05, Chen/03, and X1-1 databases.

first-principles calculated properties, such as bond dissociation energy,[16] enthalpies of reaction,[17] potential energy surface,[18,19] DFT energies,[20] and electronic structure.[21] Xu et al. continued to improve the accuracy of the calculated $\Delta_f H_{298}{}^{\theta}$ and developed X1 method.[22] In their study, an enlarged database (X1/07) containing 393 molecules that are made of 15 different elements was constructed to cover closed-shell organic molecules, radicals, and inorganic compounds. Combining the NN-based correction method and DFT calculation reduces the mean absolute deviation (MAD) of X1 results to a value (1.43 kcal mol^{-1}) close to those of the G_2 (1.88 kcal mol^{-1}) and G_3 (1.05 kcal mol^{-1}) methods.[22,23] Besides accuracy, machine learning methods are also applied to speed up the first-principles methods. In 2012, Rupp and co-workers employed machine learning method and their designed coulomb matrix to bypass solving Schrödinger equation and successfully reduced the calculation time of molecular atomization energies to several seconds.[24]

These works opened up a new direction for the field of first-principles quantum mechanical methods, and an alternative approach to reach the "Heaven of Chemical Accuracy".[25] However, to firmly establish the NN-based first-principles method as a competitive approach in the race for the "Chemical Accuracy", more issues need to be addressed. As the NN-based approach is a statistical method, it is essential to quantify its applicable range and to ensure the reliability of its prediction. For instance, if a testing molecule is quite different from those contained in the training data set, the predicted result can be wrong. Therefore, when a molecule of interest is given, it is important to determine whether the molecule falls into the application range covered by the NN-based method, and if so, the NN-calibrated result is expected to be reliable. This calls for quantitatively defining the applicable range of the NN-based first-principles method in the "Chemical Space". Another related issue is the error bars of the predicted values. Often when an experimental value is given, its error is provided to quantify its accuracy. The situation is quite different in the field of first-principles calculation: only the calculated value is reported, and no error bar is quantified to assess its accuracy, despite the common knowledge that the reported value is only an approximation. It is thus highly desirable to quantify the accuracy of each first-principles calculated result by reporting its error bar. As it is statistics in nature, the NN-based first-principles method may thus be extended to quantify the uncertainty of each prediction.

In this study, a database containing 539 molecules is constructed from several well-established databases, G3/05,[26] Chen/03[15] and X1-1,[22] and three published handbooks[27−29] with well-tabulated experimental heats of formation $\Delta_f H_{298}{}^{\theta}$. The uncertainty of experimental $\Delta_f H_{298}{}^{\theta}$ for each molecule in the database is less than 1.0 kcal mol^{-1}. Only the molecules made of H, C, N, O, F, Cl, and S elements are included as their distribution density in the "Chemical Space" is high enough so that the NN so constructed is expected to yield reliable $\Delta_f H_{298}{}^{\theta}$ for the similar molecules made of these elements. Kennard−Stone (KS) sampling method[30] is employed to select evenly the training and testing molecules among the new database. Finally, the Bootstrapping method,[31] a well-known resampling method, is employed to assess the accuracy of the NN-based first-principles method by determining the error bar of each prediction.

■ DATABASE AND COMPUTATIONAL METHODS

Database of Heat of Formation ($\Delta_f H_{298}{}^{\theta}$). A database with accurate experimental values, high distribution density and wide coverage is a prerequisite to guarantee the reliability and accuracy of the NN-based first-principles method.[12,15] However, this was not emphasized in early studies.[15−22] For instance, in Figure 1 we plot the number of the molecules versus the specific element that they contain in the G3/05, Chen/03, and X1-1 databases. The molecules containing Li, Be, B, Na, Mg, Al, Si, P, and third row elements are scarce, and thus, the NNs that cover these elements are not expected to predict accurately the $\Delta_f H_{298}{}^{\theta}$ of the molecules containing these elements. In light of the scarcity of reliable experimental $\Delta_f H_{298}{}^{\theta}$ for compounds containing these elements,[32] it is decided that only molecules made of H, C, N, O, F, Cl, and S elements are included in new database for this study. This ensures sufficient data density to construct a reliable NN that is capable of calibrating the discrepancy between the calculated and experimental $\Delta_f H_{298}{}^{\theta}$s.

In this study, the database is compiled from four sources, including G3/05,[26] Chen/03,[15] X1-1,[22] and additional experimental $\Delta_f H_{298}{}^{\theta}$'s from the three handbooks[27−29] used earlier.[15] First, we combine two well-established databases, G3/05[26] and Chen/03[15] databases, in which the standard deviation of experimental $\Delta_f H_{298}{}^{\theta}$ for each molecule is less than 1.0 kcal mol^{-1}. To increase the diversity of the chemical species of the database, the X1-1 set[22] is also included, for it contains more radicals, inorganic compounds, larger molecules, and versatile structures than those in G3/05[26] and Chen/03[15] databases. During the compilation, we eliminate the duplicated molecules, keep only the molecules made of H, C, N, O, F, Cl or S elements, and exclude the molecules with experimental uncertainties larger than 1.0 kcal mol^{-1}, which results in 353 molecules. Furthermore, we add additional 186 organic molecules, whose $\Delta_f H_{298}{}^{\theta}$ values are well documented in three different handbooks[27−29] used in our previous work[15] with consistent values and small standard deviation (less than 1.0 kcal mol^{-1}). The whole database of 539 molecules, which is termed the Chen/13 database, covers a wide range of different chemical structures and bonding situations, containing not only the hydrocarbons (167 molecules) and substituted hydrocarbons (290 molecules) but also non-hydrogens (44 molecules), inorganic hydrides (11 molecules), and radicals (27 molecules) (see Table S1 of the Supporting Information for the details).

Computational Methods. The geometries of all 539 molecules are optimized via B3LYP/6-311+G(d,p) calculations, and the zero point energies (ZPEs) are calculated at the same level. The single point energy of each molecule is calculated at the B3LYP/6-311+G(3df,2p) level. Vibrational frequencies of all optimized structures are determined to confirm that they are at local minima (having all real frequencies). The raw B3LYP/6-311+G(d,p) ZPE is employed in calculating $\Delta_f H_{298}{}^{\theta}$. The strategies to calculate $\Delta_f H_{298}{}^{\theta}$ are adopted from previous work.[33] The bond order is calculated via the natural bond orbital (NBO) analysis.[34] Gaussian 03 software package[35] is employed in the calculations.

■ NEURAL-NETWORK AND BOOTSTRAPPING

Physical Descriptors of Molecules and NN-Chemical Space. As with previous studies,[15,22,36−39] our NN adopts a three-layer architecture, which has an input layer consisting of 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 $\Delta_f H_{298}{}^{\theta}$ (Figure 2). We select 10 physical descriptors to describe each molecule, and they are the raw DFT $\Delta_f H_{298}{}^{\theta}$, the number of H atoms in the molecule (N_H), the number of *C* atoms in the molecule (N_C), the number of *N* atoms in the molecule (N_N), the number of O atoms in the molecule (N_O), the number of *F* atoms in the molecule (N_F), the number of *S* atoms in the molecule (N_S), the number of Cl atoms in the molecule (N_{Cl}), spin multiplicity (SM), and bond order (BO). The descriptors are selected on the basis of our previous work in 2003.[15] In this work, we substitute the total number of atoms of the molecule (N_t) with numbers of each constituent element (e.g., N_H, N_C, N_N, N_O, N_F, N_S, and N_{Cl}), and this has been shown to substantially improve the performance of the NN-based first-principles method.[22] Furthermore, ZPE is excluded as a descriptor in this study. Number of double bonds (N_{db}) is substituted by BO, which is defined as the sum of (I − 1), where I represents the bond order whose value is larger than 1. The bond orders I are obtained from NBO analysis.[34] SM for a molecule is given by 2S + 1, where S is the total spin for the molecule. This descriptor is designed to improve the performance of the NN on the radicals.

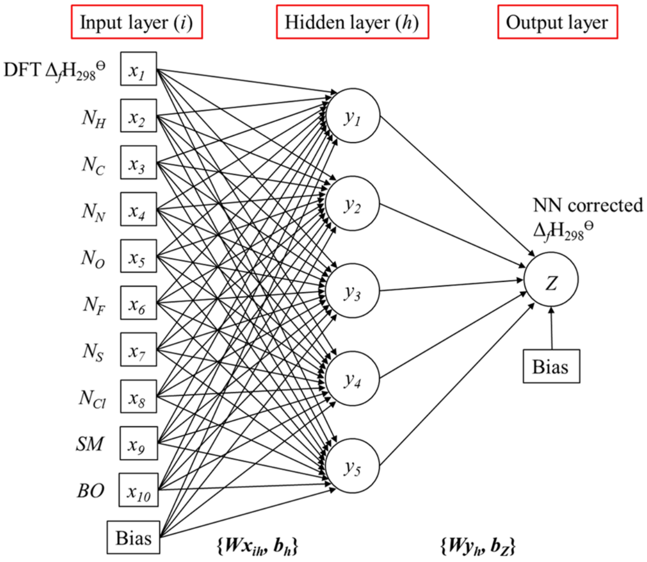


Figure 2. Structure of the neural network. Wx_{ih}/b_h and Wy_h/b_z are the weight coefficients connecting the input and hidden layers, and the hidden and output layers, respectively. The Bias equals 1.

As shown in Figure 2, each molecule is characterized by an array (x_1, x_2, ..., x_{10}), in which x_1 to x_{10} represent the 10 physical descriptors of the molecule. The 10-dimensional space spanned by the 10 descriptors (x_1, x_2, ..., x_{10}) is defined as the NN-Chemical Space. The Chen/13 database can be represented by a 10 × 539 matrix shown in eq 1, with each column as a vector containing 10 descriptors of one of the 539 molecules,

$$X(\text{Chen}/13) = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,539} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,539} \\ \vdots & \vdots & \ddots & \vdots \\ x_{10,1} & x_{10,2} & \cdots & x_{10,539} \end{bmatrix} \tag{1}$$

Standardization of NN inputs is required, just as in the previous works.[22,36,40] In this study, we adopt the *mapstd* function in the Matlab package[41] to standardize the inputs; the mean and standard deviation of each row in eq 1 are mapped to 0 and 1, respectively, by scaling each element $x_{i,j}$ in eq 1 as follows,

$$x_{i,j} \rightarrow \frac{(x_{i,j} - \bar{x}_i)}{\sigma_i} \tag{2}$$

where i runs over 10 descriptors, j runs over 539 molecules in the Chen/13 database, $\bar{x}_i$, and σ_i is the mean and standard deviation of the ith row of eq 1, respectively. The 539 standardized data distribute in the 10-dimensional NN-Chemical Space. To visualize the distribution, we map the 10-dimensional NN-Chemical Space onto a 2-dimensional space characterized by a polar coordinate, as shown in Figure 3. The radius for a data point j is defined as the Euclidean norm of its standardized coordinates ($x_{1,j}$, $x_{2,j}$, ..., $x_{10,j}$) and the angle is the raw DFT $\Delta_f H_{298}{}^{\theta}$ mapped linearly onto [0, 2π] where 0 and 2π correspond to the minimum (i.e., −313.86 kcal mol^{-1} for C_2F_6) and maximum (140.98 kcal mol^{-1} for $CH^{\bullet}$) raw DFT $\Delta_f H_{298}{}^{\theta}$'s, respectively. Figure 3 shows the resulting distribution of the 539 data points in the 2-dimensional space. It is observed that the 539 data points of Chen/13 are not distributed uniformly. Molecules are denser in the angular range [π, 3π/2], corresponding to the raw DFT $\Delta_f H_{298}{}^{\theta}$ [−86.44, +27.27] kcal mol^{-1}.

Because data points are distributed unevenly in the NN-Chemical Space, it is important to divide carefully the training and testing data sets. In this study, the training set with 449 molecules is selected from the Chen/13 database sequentially and evenly by KS sampling method. KS sampling method is designed to pick the data points evenly in the space of

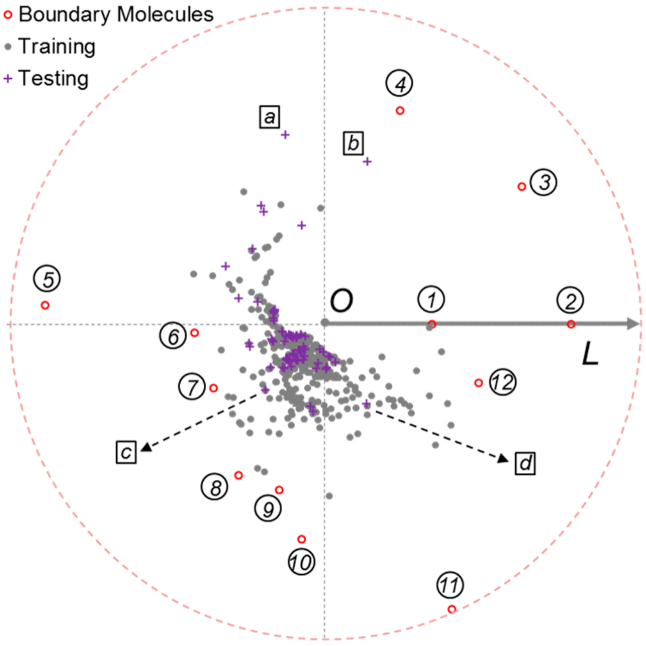


Figure 3. Chen/13 database (539 molecules) in a 2-dimensional space presented by a polar coordinate, in which the radius for a data point j is the Euclidean norm of its standardized coordinates ($x_{1,j}$, $x_{2,j}$, ..., $x_{10,j}$) and the angle is the raw DFT $\Delta_f H_{298}{}^{\theta}$ mapped linearly onto $[0, 2\pi]$. The red open circles represent the boundary molecules. The gray dots and purple crosses represent the training molecules (beside the boundary molecules) and the testing molecules, respectively. (The points 1−10 in the circles and a−d in the rectangles are illustrated in Tables 1 and 8, respectively.)

interests.[30] The remaining 90 molecules in the Chen/13 database, labeled by an asterisk (*) in Table S1 (Supporting Information), are used as the testing set to validate the performance of the NN.

In our KS sampling, a few molecules are selected as the starting molecules, which in this study are the boundary molecules. A boundary molecule is defined as a molecule with one of its descriptors at either the maximum or minimum value for the corresponding descriptor, as shown in Table 1. The boundary molecules are labeled as the red circles in Figure 3. As a result, these 12 boundary molecules are served as the starting molecules for our KS sampling to determine the training data set. To select further the training molecules, we need to define the distance ($D_{\nu,\mu}$) between any two molecules (ν and μ) in the 10-dimensional NN-Chemical Space as

$$D_{\nu,\mu} = \sqrt{\sum_{m=1}^{10} (x_{m,\nu} - x_{m,\mu})^2} \tag{3}$$

where $x_{m,\nu}$ and $x_{m,\mu}$ are the values of the mth descriptors for molecules ν and μ, respectively.

To pick out the 13th molecule for the training set, the distances between the unselected 527 molecules in the Chen/13 database and the 12 starting molecules are calculated according to eq 3; i.e., for each unselected molecule, there are 12 distances calculated, among which the shortest one is kept. This results in 527 distances to represent 527 unselected molecules. The KS sampling selects one molecule with the longest distance as the 13th molecule. This ensures that the newly selected molecule is the one farthest from those already selected molecules. Now, there are 13 selected molecules and 526 unselected molecules. Performing the same procedure on the 526 unselected molecules, the 14th training molecule is so selected. Repeating this procedure, 437 molecules are selected. Together with the 12 starting molecules, the training set of 449 molecules is formed. We define the space spanned by the 539 molecules as the occupied subspace (OSS) in the 10-dimensional NN-Chemical Space. KS sampling selects the 449 training molecules and as well as the 90 testing molecules evenly in the OSS.

In Figure 3, one can visualize the 12 boundary molecules (defining the boundaries for the OSS). For any molecule falling into the OSS, its $\Delta_f H_{298}{}^{\theta}$ can be predicted accurately once the corresponding NN is constructed properly. For the molecules outside the OSS, the NN model cannot predict accurately the values of their $\Delta_f H_{298}{}^{\theta}$. Alternatively, the 449 molecules in the training data set may be selected randomly; the results of the KS and random sampling methods are compared later in the Results and Discussion.

Neural-Network Structure. We adopt an improved version back-propagation algorithm to train the NNs: Bayesian regulation back-propagation (*trainbr*).[42] Training is performed by employing Matlab.[41] Practically, a three-layer feed-forward neural network with certain number of hidden neurons is first initialized randomly and trained with the Bayesian regulation back-propagation (Matlab's *trainbr*).[41,42] The number of neurons in the hidden layer is varied from 1 to 10 to determine the optimal structure of our neural network. We find that the network containing 5 hidden neurons yields overall the best results (see the first section of Results and Discussion for details). Therefore, the 11−5−1 structure is adopted for our neural network as depicted in Figure 2. Before training, we adopt the *mapminmax* function in the Matlab package[41] to normalize the inputs of the training set, which shows a better performance than the *mapstd* function used in KS sampling; the minimum and maximum of 449 raw inputs of each descriptor in the training set are mapped to −1 and +1, respectively, by scaling each input $x_{i,j}$ as follows,

$$x_{i,j} \rightarrow \frac{2(x_{i,j} - x_i(\min))}{(x_i(\max) - x_i(\min))} - 1 \tag{4}$$

Table 1. Boundary Molecules Shown in Figure 3

	molecular formula	name	remark
1	$CH^{\bullet}$	CH radical	maximum $\Delta_f H_{298}{}^{\theta}$ (142.50 kcal mol^{-1})
2	C_2F_6	perfluoroethane	minimum $\Delta_f H_{298}{}^{\theta}$ (−321.30 kcal mol^{-1}); maximum N_F (6)
3	SF_6	sulfur hexafluoride	maximum N_F (6)
4	C_6F_6	hexafluorobenzene	maximum N_F (6)
5	$C_5H_8N_4O_{12}$	pentaerythritol tetranitrate	maximum N_N (4); maximum N_O (12)
6	$C_{16}H_{34}O$	1-hexadecanol	maximum N_H (34)
7	$C_{16}H_{34}$	hexadecane	maximum N_H (34)
8	CCl_4	carbon tetrachloride	maximum N_{Cl} (4)
9	C_2Cl_4	perchloroethene	maximum N_{Cl} (4)
10	CN_4O_8	tetranitromethane	maximum N_N (4)
11	S_8	sulfur octamer (cyclic)	maximum N_S (8)
12	$C_{20}H_{12}$	perylene	maximum N_C (20); maximum BO (7.92)

where i runs over 10 descriptors, j runs over 449 molecules in training set, and $x_i(\mathrm{min})$ and $x_i(\mathrm{max})$ are the minimum and maximum of the raw inputs of the ith descriptor, respectively. The output value of the output neuron, Z, is related to the inputs of the jth molecule, $x_j = (x_{1,j}, x_{2,j}, ..., x_{10,j})$, as

$$Z = \sum_{h=1,5} W_{y_h}\mathrm{Sig}\left(\sum_{h=1,10} W_{x_{ih}}x_{ij} + b_h\right) + b_z \qquad (5)$$

where the transfer function $\mathrm{Sig}(\nu) = [2/(1 + e^{-2\nu}) - 1]$; Wx_{ih} is the weight coefficient connecting the input neuron x_i and the hidden neuron y_h, b_h is the weight coefficient between the Bias and the neuron y_h, and Wy_h is the weight coefficient connecting the hidden neuron y_h and the output neuron Z. The final weights for the NN optimized to correct the raw B3LYP/6-311+G(3df,2p) results are summarized in Table 2. As mentioned earlier, Table 1 shows the boundary molecules of the OSS, and these molecules are depicted in Figure 3 as well. By definition, for the molecules within the boundary of the OSS, their $\Delta_f H_{298}{}^{\theta}$ can be calculated accurately, and for the molecules beyond the boundary, their $\Delta_f H_{298}{}^{\theta}$ may not be predicted accurately. It is thus important to quantify the boundary of the OSS and thus the OSS. Unfortunately, this is not possible at this moment. As stricter criteria are employed to construct the Chen/13 database, our OSS is smaller than those of previous studies.[22] As a positive consequence, the prediction of the resulting NN is expected to be more reliable.

Table 2. Final Weights and Biases for the NN Model in Our NNB Method Optimized To Correct the Raw B3LYP/6-311+G(3df,2p) Results

$W_{x_{ih}}$	y_1	y_2	y_3	y_4	y_5
DFT $\Delta_f H_{298}{}^{o}$ (x_1)	−0.4728	0.7283	−0.2655	0.4112	−0.2791
N_H (x_2)	0.8819	0.6254	0.182	0.061	−0.2824
N_C (x_3)	−0.6088	−0.2419	0.7553	−0.1927	0.1495
N_N (x_4)	−0.3918	−0.0465	−0.4284	−0.4006	0.0197
N_O (x_5)	−0.4434	−0.5073	0.3835	1.1422	0.6131
N_F (x_6)	−0.1965	−0.1683	−0.2515	0.4026	0.2715
N_S (x_7)	0.5701	−0.3275	0.744	−0.1042	−0.3156
N_{Cl} (x_8)	0.2404	0.0275	0.3355	0.1978	−0.0267
SM (x_9)	−0.1895	−0.0113	−0.5858	−0.3143	−0.0225
BO (x_{10})	0.4806	0.0554	−0.3556	0.1975	−0.1047
BiasX (b_h)	0.7526	0.6688	−0.1389	0.8465	0.5464
W_{y_h}	−0.6256	0.7136	−0.2280	0.5570	−1.3347
BiasY (b_Z)	−0.1075				

Bootstrapping Approach. Bootstrapping is a well-known resampling method applied in a broad spectrum of statistical problems.[31] It is used to estimate the uncertainty of a prediction by selecting a collection of new sample sets (named the Bootstrap samples) and constructing their respective new statistical models (i.e., the NNs in our case). The error bar of a prediction can be determined from the distribution of the predicted values out of these new statistical models. First, 300 sets of Bootstrap samples,[43] B_1 to B_{300}, with each one containing 449 training molecules (some of which can be the same), are randomly drawn with replacement from the training set. To construct a Bootstrap sample set, we randomly select one molecule out from the training set. The selected molecule is marked down and put back into the training set. This process is repeated for 449 times so that 449 molecules are selected to form a Bootstrap sample set, in which all the molecules are selected independently and a molecule may be selected more than once. We construct each Bootstrap sample set in the same manner. Then, 300 NNs are constructed for all 300 Bootstrap samples, respectively, as with the aforementioned method. Each of the 300 Bootstrap NNs is employed to correct the raw DFT calculated $\Delta_f H_{298}{}^{\theta}$s. The resulting 300 values are employed to evaluate the standard deviation for the corresponding predicted $\Delta_f H_{298}{}^{\theta}$. We term the combination of Neural Network and Bootstrapping method as the NNB method. Error bars predicted by the NNB method are shown for molecules belong to testing set in Table S1 (Supporting Information).

■ RESULTS AND DISCUSSION

Comparison of Different Sampling Methods. For a given database, it is important to choose proper molecules for the training and testing sets, which are used to construct/optimize and test the NN, respectively. Therefore, it is important to select the training molecules that distribute evenly and representatively in the database. In previous studies, the training molecules were selected randomly.[15,36−38,40,44,45] The uncontrolled random selection may omit key representative molecules in the database. In this study, we turn to the well-established sampling method, KS sampling method,[30] to determine the training and testing data sets. The detailed procedure of our KS sampling is described above. As a result, the training and testing sets distribute relatively evenly in the OSS of the 10-dimensional Chemical Space (Figure 3).

Figure 4 displays the performance of random and KS sampling methods. The mean absolute deviations (MAD) are plotted against the number of the hidden neurons. It is noted that the MAD of the training set selected by the KS sampling method is slightly higher than that by random sampling method. On the contrary, for the MAD of the testing set, which reflects the predictability of the NNB method, the random sampling has much higher values (0.6−1.2 kcal mol^{-1}) than those of the KS sampling method. For the random sampling method, the MAD of the testing set is also much higher than that of the training set; and for the KS sampling method, the

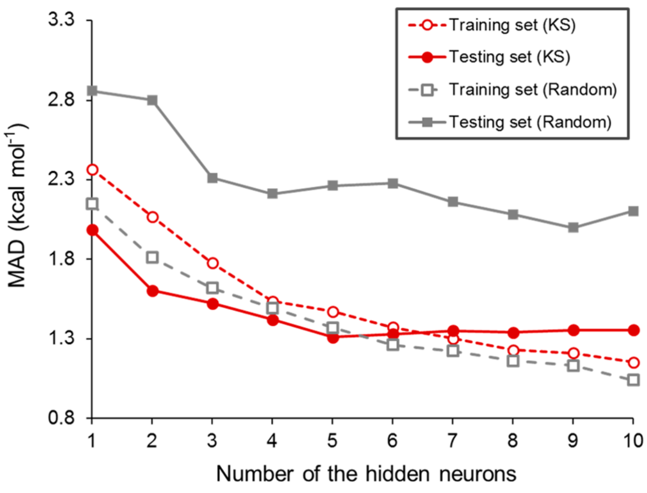


Figure 4. Comparison of two different sampling methods. The dark red open and filled circles represent the MADs of the training and testing sets by KS sampling, respectively. The gray open and filled rectangles represent the MADs of the training and testing sets by random sampling, respectively.

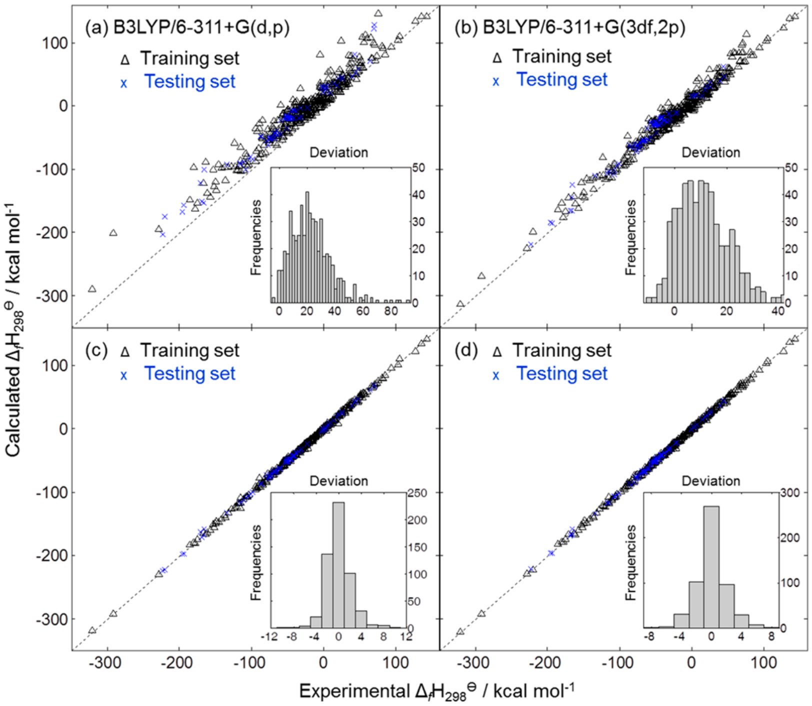


Figure 5. Experimental $\Delta_fH_{298}^\theta$ versus the calculated $\Delta_fH_{298}^\theta$ for all 539 molecules. (a) and (b) are the comparisons of the experimental $\Delta_fH_{298}^\theta$'s to their raw B3LYP/6-311+G(d,p) and B3LYP/6-311+G(3df,2p) results, respectively. (c) and (d) are the comparisons of the experimental $\Delta_fH_{298}^\theta$'s to the NN corrected B3LYP/6-311+G(d,p) and B3LYP/6-311+G(3df,2p) $\Delta_fH_{298}^\theta$'s, respectively. In (a)−(d), the triangles are for the training set and the crosses for the testing set. The correlation coefficients of the linear fit are 0.9679, 0.9857, 0.9993, and 0.9994 in (a)−(d), respectively. Insets are the histograms for the differences between the experimental and calculated $\Delta_fH_{298}^\theta$'s (calc − expt).

Table 3. Mean Absolute Deviations (MAD) and Root-Mean-Square (RMS) for 539 Molecules Obtained from Data Set in This Study (All Data in kcal mol^{-1})

	DFT1[a]		DFT2[b]		NN1[c]		NN2[d]	
	MAD	RMS	MAD	RMS	MAD[e]	RMS[e]	MAD[e]	RMS[e]
all	22.57	26.75	11.59	14.58	1.59 (93%)	2.21 (92%)	1.45 (88%)	1.99 (86%)
training	21.74	26.34	10.92	14.09	1.61 (93%)	2.23 (92%)	1.47 (87%)	2.01 (86%)
testing	26.69	28.69	14.95	16.81	1.46 (95%)	2.11 (93%)	1.31 (91%)	1.86 (89%)

[a] Raw DFT results from B3LYP/6-311+G(d,p) calculation. [b] Raw DFT results from B3LYP/6-311+G(3df,2p) calculation whereas the geometry is optimized at B3LYP/6-311+G(d,p) level. [c] NN corrected results on the results of DFT1 calculation. [d] NN corrected results on the results of DFT2 calculation. [e] Percentage of improvement by the NN correction is given in parentheses.

MADs of the testing set are about 0.2 kcal mol^{-1} lower than those of the training set when the number of hidden neurons is 5 or less. All these show that KS sampling is much more effective than random sampling. The MAD of training set decreases as the number of hidden neurons increases, whereas the MAD of the testing set remains unchanged when there are five hidden neurons or more. This indicates that 5 hidden neurons are sufficient, which is adopted for the subsequent calculations.

NN-Based Correction for Different Basis Sets. In Figure 5a,b, the calculated $\Delta_fH_{298}^\theta$'s under B3LYP/6-311+G(d,p) and B3LYP/6-311+(3df,2p) levels (raw calculated data) are compared to their experimental data. The horizontal coordinates are the experimental values and the vertical coordinates are the raw calculated data. The dashed line is where the vertical and horizontal coordinates are equal. Thus, if a certain point sits on the dashed line, its B3LYP calculation and experiment would have the perfect match. The raw calculated $\Delta_fH_{298}^\theta$ values in both DFT levels are mostly above the dashed line, showing that most of the raw $\Delta_fH_{298}^\theta$'s are larger than the experimental data. Compared to the experimental values, the MADs for $\Delta_fH_{298}^\theta$'s are 22.57 and 11.59 kcal mol^{-1} for B3LYP/6-311+G(d,p) and B3LYP/6-311+G(3df, 2p) calculations, respectively. In Table 3, the MADs and the root-mean-square (RMS) values for the training and test sets are listed.

In Figures 5c,d, the NN corrected $\Delta_fH_{298}^\theta$'s are compared to their experimental values. The black triangles are for the training set and the blue crosses for the testing set. The blue crosses uniformly distribute along the dashed line, indicating that the KS sampling method[30] selects evenly the testing set in the OSS. It follows that this testing set is a good sampling set to evaluate the performance of the NNB method. Compared to the raw DFT calculated results, the NN corrected values are much closer to the experimental values for both training and testing sets. As shown in Table 3, upon the NN corrections, the MAD of $\Delta_fH_{298}^\theta$'s for the whole data set are reduced from 22.57 to 1.59 kcal mol^{-1} (NN1, 93% improvement) and 11.59 to 1.45 kcal mol^{-1} (NN2, 88% improvement) for B3LYP/6-311+G(d,p) and B3LYP/6-311+G(3df, 2p), respectively. The RMS deviations are 2.21 and 1.99 kcal mol^{-1} for NN1 and NN2, respectively, which are substantially reduced compared to those of our previous work on 180 organic molecules, 3.1 and 3.3 kcal mol^{-1}, respectively.[15] For the deviations of the training and testing sets (Table 3), one can observe that the testing set has better improvement than the training set in both NN1 and NN2, which validates further KS sampling.

The performance of the NNB method can be further demonstrated by the error analysis employed for the raw and NN corrected $\Delta_f H_{298}^{\theta}$'s of all 539 molecules. In the insets of Figure 5, the histograms are plotted for the deviations from the experiments of the raw B3LYP $\Delta_f H_{298}^{\theta}$'s and their NN corrected values. Obviously, the error distributions after the NN correction are of approximate Gaussian distributions, as shown by the insets in Figure 5c,d. About 73% (394/539) and 74% (401/539) of the deviations of the NN corrected values fall within the range from −2 to +2 kcal mol^{-1} for NN1 and NN2, respectively. This is a significant improvement over the raw B3LYP methods, where only 3% (18/539) and 14% (75/539) of the errors fall within the same interval for DFT1 and DFT2, respectively.

Furthermore, from Table 3, we observe that NN2 (correcting the raw B3LYP/6-311+G(3df,2p) results) has better results than NN1 (correcting the raw B3LYP/6-311+G(d,p) results). Compared to the DFT1 method, DFT2 method only needs to take an extra step to calculate the single-point energy under the B3LYP/6-311+G(3df,2p) level, which is cheap under contemporary computational power. The subsequent results presented are of NN2, if not explicitly stated.

Analysis of NNB-Based First-Principles Results. As shown in Table 4, the 539 molecules are divided into five chemical species: non-hydrogens, hydrocarbons, substituted hydrocarbons, inorganic hydrides, and radicals. The MADs for raw B3LYP range from 2.17 to 17.44 kcal mol^{-1}, whereas the MADs after NN correction range from 1.07 to 1.95 kcal mol^{-1}. The MAD improvements against the raw B3LYP by NN correction can be as high as 92% (for hydrocarbons). The maximum and minimum MADs for the raw DFT results occur at hydrocarbons (17.44 kcal mol^{-1}) and inorganic hydrides (2.17 kcal mol^{-1}), respectively, which is most likely originated from the size differences of the molecules contained in these two species. The numbers of atoms in hydrocarbons range from 3 to 50, which are much larger than those of inorganic hydrides (ranging from 2 to 6). Moreover, hydrocarbons have 98% (163/167) molecules containing more than 6 atoms. This is consistent with previous observations that B3LYP performance deteriorates with increasing size.[15,33,46,47] Interestingly, the maximum and minimum improvements of the NN correction are for hydrocarbons (92%) and inorganic hydrides (12%), respectively. The performance is classified according to specific elements contained in molecules (Table 4). The MADs after NN correction for N, O, F, S, and Cl containing molecules range from 1.31 to 1.67 kcal mol^{-1}. The subset of S-containing molecules has the largest MAD for raw B3LYP (12.37 kcal mol^{-1}), but it has the largest improvement (89%).

As mentioned previously, the testing set serves to test the validity of the NNB method. In Table 5, we summarize the performance of the NNB method on the testing set (90 molecules) for different species. Not surprisingly, both the MADs and RMSs of the testing set have more improvement than the whole data set. For the MAD, the improvement ranges from 75% to 93%, whereas for the RMS, the range is 73% to 91%. Impressively, the MAD is less than 1.0 kcal mol^{-1} for the N-containing molecules (0.52 kcal mol^{-1}), S-containing molecules (0.97 kcal mol^{-1}), and Cl-containing molecules (0.97 kcal mol^{-1}). From Table 5, it is concluded that our NNB method can indeed predict accurately the $\Delta_f H_{298}^{\theta}$ for the molecules made of H, C, N, O, F, S, and Cl.

Table 4. Mean Absolute Deviations (MAD) and Root-Mean-Square (RMS) of Different Species for 539 Molecules of Chen/13[a] (All Data in kcal mol^{-1})

	mean absolute deviations		root-mean-square	
molecule types[b]	Raw DFT	nnb[c]	raw DFT	NNB[c]
all (539)	11.59	1.45 (88%)	14.58	1.99 (86%)
Divided by Chemical Species				
non-hydrogens (44)	6.33	1.95 (69%)	9.06	2.59 (71%)
hydrocarbons (167)	17.44	1.36 (92%)	19.71	1.93 (90%)
substituted hydrocarbons (290)	10.23	1.43 (86%)	12.57	1.96 (84%)
inorganic hydrides (11)	2.17	1.92 (12%)	2.60	2.29 (12%)
radicals (27)	2.48	1.07 (57%)	2.89	1.35 (53%)
Classified by Specific Element				
N-containing molecules (95)	5.03	1.54 (69%)	6.79	1.93 (72%)
O-containing molecules (197)	9.90	1.61 (84%)	12.52	2.24 (82%)
F-containing molecules (31)	4.97	1.67 (66%)	6.96	2.17 (69%)
S-containing molecules (71)	12.37	1.34 (89%)	14.60	1.77 (88%)
Cl-containing molecules (44)	7.69	1.31 (83%)	9.37	1.73 (82%)

[a]Calculated $\Delta_f H_{298}^{\theta}$'s are obtained from raw B3LYP/6-311+G(3df,2p) and the NN correction on this level. [b]Number of molecules contained in the group is given in parentheses. [c]Percentage of improvement by the NN correction is given in parentheses.

Table 5. Mean Absolute Deviations (MAD) and Root-Mean-Square (RMS) of Different Species for 90 Molecules of the Testing Set[a] (All Data in kcal mol^{-1})

	mean absolute deviations		root-mean-square	
molecule types[b]	raw DFT	NNB[c]	raw DFT	NNB[c]
all (90)	14.95	1.31 (91%)	16.81	1.86 (89%)
Divided by Chemical Species				
non-hydrogens (3)	5.71	1.45 (75%)	6.04	1.60 (73%)
hydrocarbons (38)	19.40	1.44 (93%)	20.52	1.87 (91%)
substituted hydrocarbons (49)	12.06	1.20 (90%)	13.78	1.87 (86%)
Classified by Molecules Containing Specific Element				
N-containing molecules (2)	3.28	0.52 (84%)	3.77	0.63 (83%)
O-containing molecules (39)	13.26	1.31 (90%)	14.85	2.05 (86%)
F-containing molecules (7)	4.57	1.00 (78%)	5.24	1.21 (77%)
S-containing molecules (3)	11.49	0.97 (92%)	11.50	1.20 (90%)
Cl-containing molecules (4)	8.01	0.97 (88%)	8.51	1.22 (82%)

[a]Calculated $\Delta_f H_{298}^{\theta}$'s are obtained from raw B3LYP/6-311+G(3df,2p) and the NN correction on this level. [b]Number of molecules contained in the group. [c]Percentage of improvement by the NN correction is given in parentheses.

Comparison with X1s and G3 Methods. Xu et al. developed the NN-based X1s method in 2010.[22,39] Being optimized for molecules with no more than 8 oxygen atoms, X1s is not applicable to pentaerythritol tetranitrate ($C_5H_8N_4O_{12}$) that contains 12 oxygen atoms. The experimental, raw DFT, and X1s corrected $\Delta_f H_{298}{}^{\theta}$'s for pentaerythritol tetranitrate are −92.45, −91.47, and −41.26 kcal mol^{-1}, respectively. The raw DFT yields a good result. In contrast, the X1s value is quite different from the experimental value, which implies that pentaerythritol tetranitrate is outside the OSS of X1s. On the other hand, the OSS of NNB model developed here covers pentaerythritol tetranitrate, and the corrected $\Delta_f H_{298}{}^{\theta}$ is −92.57 kcal mol^{-1}, which differs only 0.12 kcal mol^{-1} from the experimental value.

For a fair comparison, we exclude the pentaerythritol tetranitrate and compare the performance of the raw DFT, NNB, and X1s methods on the remaining database (538 molecules). Table 6 summarizes the comparison. The MADs for the raw DFT, NNB, and X1s methods are 11.61, 1.45, and 1.90 kcal mol^{-1}, respectively. Both NNB and X1s methods substantially improve over the raw DFT results, whereas the MAD and RMS of NNB corrected values are ~0.50 kcal mol^{-1} less than those of X1s. The maximum deviations, defined as the absolute of (theoretical value − experimental value), for the raw DFT, NNB, and X1s are 41.87 (1-hexadecanol, $C_{16}H_{28}O_2$), 8.40 (O=C=C=C=O), and 13.41 (C_6ClF_5) kcal mol^{-1}, respectively. NNB method has the smallest span of deviations among the three methods.

Table 6. Statistics for Raw DFT, NNB, and X1s Results on the Chen/13 Database[a]

	raw DFT[b]	NNB	X1s
maximum deviation	41.87 ($C_{16}H_{34}O$)[c]	8.40 (C_3O_2)[c]	13.41 (C_6ClF_5)[c]
MAD	11.61	1.45	1.90
RMS	14.60	1.99	2.65

[a] To make a fair comparison, the molecule $C_5H_8N_4O_{12}$, which is outside of the OSS of X1s, is excluded. The maximum deviations are defined as the absolute of (theoretical value − experimental value). Energies are in kcal mol^{-1}. [b] Raw DFT is obtained from B3LYP/6-311+G(3df,2p) level calculation. [c] $C_{16}H_{34}O$ (1-Hexadecanol); C_3O_2 (propa-1,2-diene-1,3-dione); C_6ClF_5 (1-chloropentafluorobenzene).

Table 7 compares the performances of the raw B3LYP, NNB, X1s, and G3 methods on the G3/05 portion (200 molecules) of the Chen/13 database. In general, the results of NNB, X1s, and G3 are significantly better than the raw B3LYP results. The MAD and RMS of NNB are 0.23 and 0.68 kcal mol^{-1} less than those of X1s, whereas only 0.32 and 0.22 kcal mol^{-1} more than those of G3 method, respectively. Considering G3 method is much more expensive than NNB, it is encouraging that NNB method achieves the comparable accuracy with much less computational resources.

Determination of Error Bar for Each Prediction. Often we report only the calculated value of the property of interests such as energy, geometry, or vibration frequency once the quantum mechanics calculation is carried out on a molecule, cluster, or crystal, and no assessment is usually made to quantify its accuracy. It is desirable to determine the error bar of a quantum mechanical calculation result. In this study, we attempt to rectify this. The Bootstrapping approach, a widely used statistical method[31] as described earlier, is implemented to evaluate the error bar of each NNB prediction. A total of 300 Bootstrap NNs are constructed. The error bars of the predicted $\Delta_f H_{298}{}^{\theta}$ are determined for all 90 testing molecules. The predicted values and their error bars are compared with the experimental values and shown in Figure 6.

As shown in Figure 6a, the NNB predicted values are plotted against the experimental values of the 90 testing molecules, together with the corresponding error bars. In the bottom of Figure 6, each short vertical line represents a training molecule with the corresponding experimental $\Delta_f H_{298}{}^{\theta}$ below. One can observe that the training molecules are dense in the $\Delta_f H_{298}{}^{\theta}$ range [−90.0, +60.0] kcal mol^{-1}, but sparse in the rest values, in particular, in the range [−340.0, −140.0] kcal mol^{-1}. Note that the error bar is inversely proportional to the density of the bars in the bottom, and this implies that the denser the training molecule distribution is in the OSS, the more accurate the predicted value is in the same region of the OSS. Although this is expected, it is good to know that our NNB indeed captures the systematic deviation hidden in the training data set. The more similarity a testing molecule has with the training molecules, the more accurate its predicted $\Delta_f H_{298}{}^{\theta}$ is. In general, the predicted error bar is small when a testing molecule falls into the region in the OSS where the training molecules are dense (i.e., $\Delta_f H_{298}{}^{\theta} \in [-90.0, +60.0]$ kcal mol^{-1}), whereas the error bar is large when the testing molecule belongs to the region where the training molecules is sparse (i.e., $\Delta_f H_{298}{}^{\theta} \in [-340.0, -140.0]$ kcal mol^{-1}). The experimental error bars of the 90 testing molecules are all available. We count the number of molecules in the testing set that has the overlap between [expt − experimental error bar, expt + experimental error bar] and [NN prediction − bootstrapping error bar, NN prediction + bootstrapping error bar], which is 51% (46/90). The average error bar for the testing set is 1.05 kcal mol^{-1}, indicating that our NNB method is capable to correct effectively the raw DFT $\Delta_f H_{298}{}^{\theta}$s.

Figure 6b shows the enlarged range of $\Delta_f H_{298}{}^{\theta}$ from −180.0 to −90.0 kcal mol^{-1}. We chose to show this range because it overlaps the dense and sparse regions of the training data sets. Correspondingly, the errors are small if the molecules are in the dense region, and large in the sparse region. An abnormal point

Table 7. Statistic Data for Raw DFT, NNB, X1s, and G3 Methods on the G3/05 Part of the Chen/13 Database (200 Molecules)[a]

	raw DFT[b]	NNB	X1s	G3
maximum deviation	22.22 (SF_6)[c]	5.70 (C_2H_2OS)[c]	13.41 (C_6ClF_5)[c]	9.20 ($ClFO_3$)[c]
MAD	5.45	1.35	1.58	1.02
RMS	7.18	1.79	2.47	1.55

[a] The maximum deviations are defined as the absolute of (theoretical value − experimental value). Energies are in kcal mol^{-1}. [b] Raw DFT is obtained from B3LYP/6-311+G(3df,2p) level calculation. [c] SF_6 (sulfur hexafluoride); C_2H_2OS (thiooxirane); C_6ClF_5 (1-chloro-pentafluorobenzene); $ClFO_3$ (perchloryl fluoride).

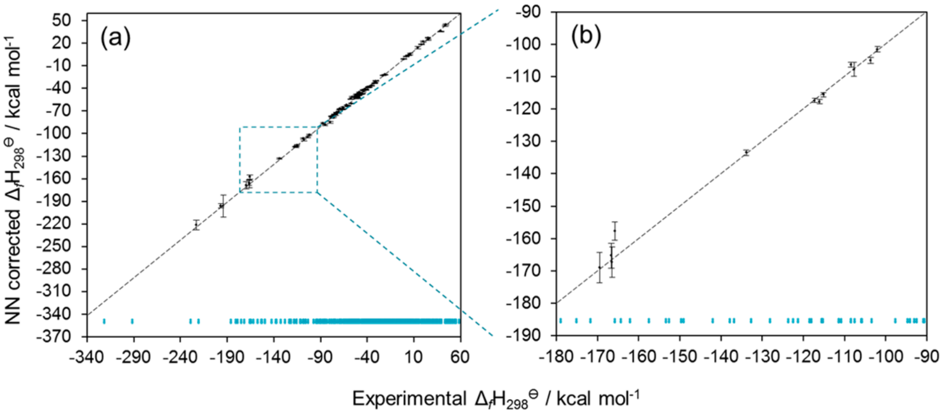


Figure 6. NN corrected $\Delta_f H_{298}^{\theta}$ versus experimental $\Delta_f H_{298}^{\theta}$ for 90 molecules in the testing set. Each vertical line in the bottom stands for a training molecule whose experimental $\Delta_f H_{298}^{\theta}$ equals the value below. The error bars are obtained with the Bootstrapping approach. All values are in the units of kcal mol^{-1}. (b) shows the enlarged range of $\Delta_f H_{298}^{\theta}$ from −180.0 to −90.0 kcal mol^{-1} in (a).

is noted at the NN corrected $\Delta_f H_{298}^{\theta}$ = 107.77 kcal mol^{-1}. It has a large error bar (2.17 kcal mol^{-1}) than those in the vicinity (0.62−0.87 kcal mol^{-1}). The reason may be that the NN-Chemical Space is defined not only by the raw DFT $\Delta_f H_{298}^{\theta}$'s but also by many other coordinates, for example, the number of H, C, N, O, F, S, Cl elements in the molecules, the spin multiplicity of molecules, and the bond orders. Therefore, we need a systematic way to measure the distribution of the training molecules in the 10-dimensional NN-Chemical Space and to quantify the density of each point in the OSS. We define the density of a testing point i, ρ_i, in the OSS as the sum of the reciprocals of Euclidean distance ($D_{i,m}$) between point i and every training molecule (m) in the NN-Chemical Space, i.e.,

$$\rho_i = \sum_{m=1}^{449} \frac{1}{D_{i,m}} \tag{6}$$

where i runs over 90 testing molecules, m runs over 449 training molecules, and $D_{i,m}$ is obtained from eq 3 with a cutoff at 1 (i.e., $D_{i,m}$ is set to 1 if its value is smaller than 1). Because the distance is very small between two isomers featuring almost the same descriptors, the density for a certain testing molecule may be high even if there is only its isomer nearby. Such a case can be eliminated by setting a cutoff to ensure that the more there are training molecules in the vicinity, the larger density ρ_i is. It follows that NNB method yields the accurate $\Delta_f H_{298}^{\theta}$ and thus the small error bar if the molecule i has a large ρ_i.

The validity of the Bootstrapping approach to calculate the error bar of the prediction of $\Delta_f H_{298}^{\theta}$ is confirmed in Figure 7, where the predicted error bar is plotted against the density ρ_i for all 90 testing molecules. Figure 7 shows that the error bar of a prediction is indeed inversely proportional to its density ρ_i. The largest error bar, 14.31 kcal mol^{-1}, shown in Figure 7, serves as a good example to demonstrate how the uncertainty of a prediction correlates to its density. The testing molecule C_6ClF_5 (1-chloropentafluorobenzene) has the largest error bar and corresponds point a in Figure 3. Its experimental $\Delta_f H_{298}^{\theta}$ is −194.10 kcal mol^{-1}. It is interesting to notice that in Figure 6a there are only few lines around $\Delta_f H_{298}^{\theta}$ = −194.10 kcal mol^{-1} and there are a few points nearby in Figure 3. Furthermore, this molecule contains 5 fluorine atoms, whereas the maximum number of fluorine atoms that any training molecule has is 6. We speculate that if more data points could be added into the training set, which have the experimental $\Delta_f H_{298}^{\theta}$ values close to −194.10 kcal mol^{-1} or have more molecules with about 5 fluorine atoms, better prediction can be expected for C_6ClF_5.

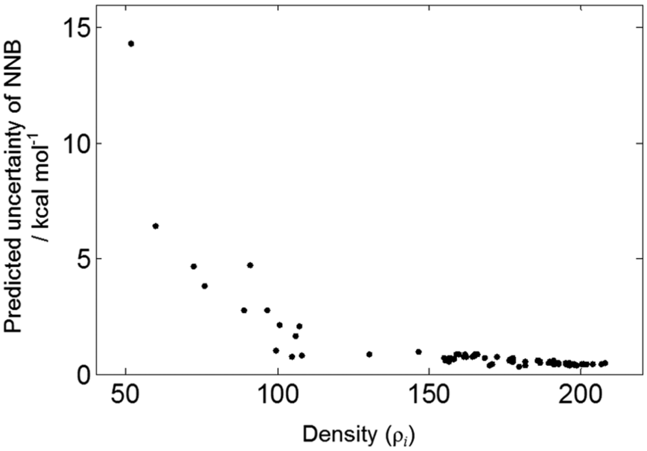


Figure 7. Error bars of the testing molecules versus the density ρ_i obtained by eq 6.

In Figure 3, we label four representative points a, b, c, and d of the testing set. The former two points (a and b) have very few training molecules nearby, whereas the latter two points (c and d) are surrounded by many more training molecules. The errors of the four molecules are listed in Table 8. Clearly there is a positive correlation between the high density in the OSS and the accuracy of a prediction. The error bars for a and b are quite large because the distribution of the training molecules in their vicinities are quite empty, whereas the error bars for c and d are small as the densities of the training molecules in their vicinities are high (Figure 3).

Table 8. Errors (kcal mol^{-1}) for Four Chosen Molecules in the Testing Set Corresponding to a−d in Figure 3

	molecules	expt	dev (DFT − expt)	dev (NN − expt)	error bar of NN
a	C_6ClF_5	−194.10	8.47	−2.25	14.31
b	CF_4	−223.04	4.16	1.53	6.42
c	$C_2H_4Cl_2$ (1,1-dichloroethane)	−31.00	6.67	−0.54	0.75
d	$C_{12}H_{10}$ (biphenyl)	43.40	18.59	−0.02	0.83

Table 9. Predicted $\Delta_f H_{298}^{\theta}$'s by NNB Method with Error Bars of 13 Molecules with Conflicting Experimental Values in Refs 27−29

molecules	name	Ref1[27]	Ref2[28]	Ref3[29]	NNB	raw DFT
$C_3H_4O_2$	2-oxetanone	−67.61	−67.61	−71.00	−67.90 ± 0.83	−62.80
$C_3H_8O_2$	1,2-propylene glycol	−102.72	−100.69	−100.74	−101.22 ± 0.93	−93.45
$C_3H_8O_2$	1,3-propylene glycol	−97.51	−93.71	−93.71	−95.29 ± 0.90	−87.48
C_6H_{12}	3,3-dimethyl-1-butene	−14.41	−14.46	−10.31	−11.52 ± 0.38	0.64
C_6H_{12}	2,3-dimethyl-2-butene	−16.28	−16.30	−14.15	−15.33 ± 0.35	−3.16
$C_6H_{12}O$	cyclohexanol	−68.40	−68.38	−70.40	−68.73 ± 0.72	−50.65
$C_6H_{14}O_2$	1,6-hexanediol	−110.23	−110.23	−112.29	−112.88 ± 0.94	−94.53
C_7H_9N	benzylamine	22.56	20.98	14.87	21.19 ± 0.66	30.29
$C_9H_{20}O$	1-nonanol	−89.99	−89.94	−92.47	−93.82 ± 0.98	−67.19
$C_{11}H_{10}$	2-methylnaphthalene	25.50	25.50	27.75	27.52 ± 0.49	45.80
$C_{12}H_{11}N$	diphenylamine	52.41	52.41	48.28	50.59 ± 1.75	67.52
$C_{12}H_{18}$	hexamethylbenzene	−18.50	−20.75	−25.26	−17.06 ± 1.14	12.89
$C_{13}H_{12}$	diphenylmethane	33.22	33.22	36.81	39.67 ± 0.91	62.11

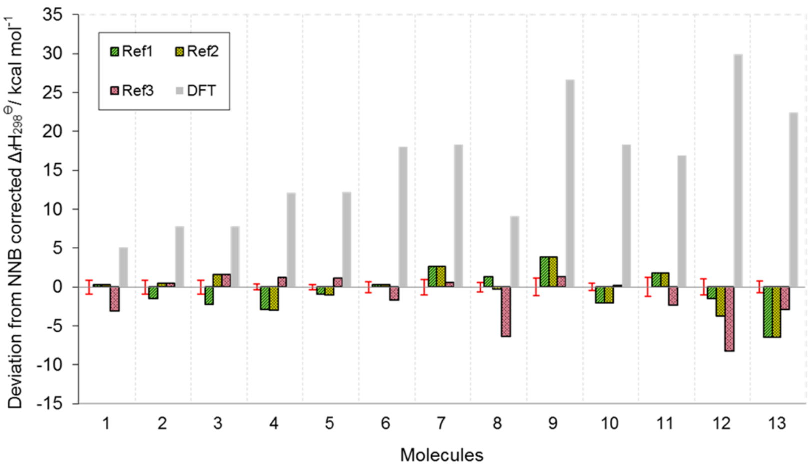


Figure 8. Deviations from the NNB corrected $\Delta_f H_{298}^{\theta}$ for the 13 molecules with conflicting experimental values. The error bars are obtained from the NNB method in this study. Ref1, Ref2, Ref3, and DFT correspond to the reported values for $\Delta_f H_{298}^{\theta}$ in refs 27−29, and the raw DFT result, respectively.

A Challenge to Experimentalists. Once established, the NNB method can be applied to predict the $\Delta_f H_{298}^{\theta}$ of the new molecules. It can also be applied to discern the conflicting experimental data reported in the literature. Table 9 lists a set of molecules whose reported experimental $\Delta_f H_{298}^{\theta}$s are not consistent from the three refs 27−29, and the differences are more than 2.0 kcal mol^{-1}. For example, the reported values of $\Delta_f H_{298}^{\theta}$ for benzylamine (C_7H_9N) are 22.56, 20.98, and 14.87 kcal mol^{-1} in refs 27−29, respectively. The maximum difference among them is as large as 7.69 kcal mol^{-1} (between refs 27 and 29). Our NNB method predicts that the heat formation of this molecule is 21.19 kcal mol^{-1} and the corresponding error bar is 0.66 kcal mol^{-1}. It follows that the experimental value from refs 27 and 28 are likely more reliable than that of ref 29, in particular, that of ref 28 is within the NNB predicted range. Furthermore, compared to the value predicted by the raw DFT (30.29 kcal mol^{-1}), which deviates much from all three experimental values, our NNB predicted value may be much more reliable. Figure 8 shows the deviations from the NNB predicted $\Delta_f H_{298}^{\theta}$ for the reported data from the three references and the raw DFT values for the 13 molecules listed in Table 9. We note that all raw DFT results deviate much from any reported values in the three references. Our NNB method successfully yields much closer values to the reported experimental data that are themselves not as accurate. We challenge the experimentalists to carry out again the measurement of the heat of formation for the 13 molecules listed in Table 9 and to test our NNB prediction!

CONCLUSION

In this study, we improve further the NN-based first-principles method for heat of formation ($\Delta_f H_{298}^{\theta}$) by constructing the enlarged experimental database (Chen/13, 539 molecules in total), employing Kennard−Stone sampling to select evenly the training molecules in the OSS of the NN-Chemical Space, utilizing Bayesian regression to train the NN, and applying the Bootstrapping method to determine the error bar of each NN prediction. Upon the NN correction, the MAD for the entire Chen/13 database decreases from 11.59 to 1.45 kcal mol^{-1} for the raw B3LYP/6-311+G(3df,2p) method, which shows that the NN-based calibration method can indeed be employed to reduce drastically the errors of first-principles calculation results. As the Chen/13 database contains only the molecules composed of the common elements H, C, N, O, F, S, and Cl, the NN model developed in this work is not applicable for molecules containing other elements. To extend the NNB to the molecules containing other elements, Chen/13 needs to be expanded further. An important development in this study is the determination of the error bar for each prediction. For the first time, the accuracy of individual first-principles calculation result is quantitatively assessed. With the error bar determined, the reliability of each calculation can thus be quantified. The average error bar for the testing set is 1.05 kcal mol^{-1}, and the magnitude of the error bar is inversely proportional to the density of the molecule in the OSS. It follows that the NNB method so developed can yield reliably, with the chemical accuracy, the $\Delta_f H_{298}{}^{\theta}$ for the molecules in the OSS. The NNB method developed in this work can surely be generalized to improve the accuracy of first-principles calculations on other properties such as solubility, melting temperature, and other thermodynamic properties. Work along such direction is ongoing.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Full data set used for training and testing the neural network and the final weights and biases of the resulted neural network model for B3LYP/6-311+G(d,p) results. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*Jian Sun: e-mail, sunjian@yangtze.hku.hk.
*GuanHua Chen: tel, +852-28592164; e-mail, ghc@everest.hku.hk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The financial support from the Hong Kong Research Grant Council (T23-713/11) is acknowledged.

■ REFERENCES

(1) Schaefer, H. F. *Methods of Electronic Structure Theory*; Plenum Press: New York, 1977.

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

(3) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*; Wiley: West Sussex, England, 2002.

(4) Cuevas, J.C.; Scheer, E. *Molecular Electronics: An Introduction to Theory and Experiment*; World Scientific: Singapore, 2010.

(5) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12*, 191−201.

(6) Saito, T. *Computational Materials Design*; Springer: Berlin, 1999.

(7) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. Quadratic Configuration Interaction. A General Technique for Determining Electron Correlation Energies. *J. Chem. Phys.* **1987**, *87*, 5968−5975.

(8) Curtiss, L. A.; Jones, C.; Trucks, G. W.; Raghavachari, K.; Pople, J. A. Gaussian-1 Theory of Molecular-Energies for 2nd-Row Compounds. *J. Chem. Phys.* **1990**, *93*, 2537−2545.

(9) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. Gaussian-2 Theory for Molecular Energies of First- and Second-Row Compounds. *J. Chem. Phys.* **1991**, *94*, 7221−7230.

(10) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. Gaussian-3 (G3) Theory for Molecules Containing First and Second-Row Atoms. *J. Chem. Phys.* **1998**, *109*, 7764−7776.

(11) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory. *J. Chem. Phys.* **2007**, *126*, 084108-1−084108-12.

(12) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gn Theory. *WIREs Comput. Mol. Sci.* **2011**, *1*, 810−825.

(13) Scuseria, G. E.; Staroverov, V. N. Chapter 24 - Progress in the Development of Exchange-Correlation Functionals. In *Theory and Applications of Computational Chemistry*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; pp 669−724.

(14) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215−241.

(15) Hu, L.; Wang, X.; Wong, L.; Chen, G. Combined First-Principles Calculation and Neural-Network Correction Approach for Heat of Formation. *J. Chem. Phys.* **2003**, *119*, 11501−11507.

(16) Wu, J.; Xu, X. Improving the B3LYP Bond Energies by Using the X1Method. *J. Chem. Phys.* **2008**, *129*, 164103-1−164103-11.

(17) Wodrich, M. D.; Corminboeuf, C. Reaction Enthalpies Using the Neural-Network-Based X1 Approach: The Important Choice of Input Descriptors. *J. Phys. Chem. A* **2009**, *113*, 3285−3290.

(18) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401-1−146401-4.

(19) Behler, J.; Martoňák, R.; Donadio, D.; Parrinello, M. Metadynamics Simulations of the High-Pressure Phases of Silicon Employing a High-Dimensional Neural Network Potential. *Phys. Rev. Lett.* **2008**, *100*, 185501.

(20) Balabin, R. M.; Lomakina, E. I. Neural Network Approach to Quantum-Chemistry Data: Accurate Prediction of Density Functional Theory Energies. *J. Chem. Phys.* **2009**, *131*, 074104-1−074104-8.

(21) Long, D. A.; Anderson, J. B. Bond-Based Corrections to Semi-Empirical and Ab Initio Electronic Structure Calculations. *Chem. Phys. Lett.* **2005**, *402*, 524−528.

(22) Wu, J.; Xu, X. The X1Method for Accurate and Efficient Prediction of Heats of Formation. *J. Chem. Phys.* **2007**, *127*, 214105-1−214105-8.

(23) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-3 and Density Functional Theories for a Larger Experimental Test Set. *J. Chem. Phys.* **2000**, *112*, 7374−7383.

(24) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301-1−058301-5.

(25) Perdew, J. P.; Ruzsinszky, A.; Tao, J. M.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. Prescription for the Design and Selection of Density Functional Approximations: More Constraint Satisfaction with Fewer Fits. *J. Chem. Phys.* **2005**, *123*, 062201-1−062201-9.

(26) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Assessment of Gaussian-3 and Density-Functional Theories on the G3/05 Test Set of Experimental Energies. *J. Chem. Phys.* **2005**, *123*, 124107-1−124107-12.

(27) Lide, D. R. *CRC handbook of chemistry and physics on CD-ROM*; CRC Press: Boca Raton, FL, 2009.

(28) Pedley, J. B.; R. D. N, and Kirby, S. P. *Thermochemical Data of Organic Compounds*; Chapman and Hall: New York, 1986.

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

(30) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137−148.

(31) Efron, B. The 1977 Rietz Lecture - Bootstrap Methods: Another Look At The Jackknife. *Ann. Stat.* **1979**, *7*, 1−26.

(32) Cioslowski, J.; Schimeczek, M.; Liu, G.; Stoyanov, V. A Set of Standard Enthalpies of Formation for Benchmarking, Calibration, and Parametrization of Electronic Structure Methods. *J. Chem. Phys.* **2000**, *113*, 9377−9389.

(33) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and Density Functional Theories for the Computation of Enthalpies of Formation. *J. Chem. Phys.* **1997**, *106*, 1063−1079.

(34) Reed, A. E.; Curtiss, L. A.; Weinhold, F. Intermolecular Interactions from a Natural Bond Orbital, Donor-Acceptor Viewpoint. *Chem. Rev.* **1988**, *88*, 899−926.

(35) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.;

Kudin, K. N.; Burant, J. C.; et al. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004.

(36) Zheng, X.; Hu, L.; Wang, X.; Chen, G. A Generalized Exchange-Correlation Functional: The Neural-Networks Approach. *Chem. Phys. Lett.* **2004**, *390*, 186−192.

(37) Duan, X. M.; Li, Z. H.; Song, G. L.; Wang, W. N.; Chen, G. H.; Fan, K. N. Neural Network Correction for Heats of Formation with a Larger Experimental Training Set and New Descriptors. *Chem. Phys. Lett.* **2005**, *410*, 125−130.

(38) Li, H.; Shi, L.; Zhang, M.; Su, Z.; Wang, X.; Hu, L.; Chen, G. Improving the Accuracy of Density-Functional Theory Calculation: The Genetic Algorithm and Neural Network Approach. *J. Chem. Phys.* **2007**, *126*, 144101-1−144101-8.

(39) Wu, J.; Ying Zhang, I.; Xu, X. The X1s Method for Accurate Bond Dissociation Energies. *ChemPhysChem* **2010**, *11*, 2561−2567.

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

(41) *MATLAB and Statistics Toolbox* Release 2012b; The MathWoks, Inc.: Natick, MA, United States.

(42) MacKay, D. J. C. Bayesian Interpolation. *Neural Comput.* **1992**, *4*, 415−447.

(43) Pattengale, N. D.; Alipour, M.; Bininda-Emonds, O. R. P.; Moret, B. M. E.; Stamatakis, A. How Many Bootstrap Replicates Are Necessary? *J. Comput. Biol.* **2010**, *17*, 337−354.

(44) Duan, X.-M.; Song, G.-L.; Li, Z.-H.; Wang, X.-J.; Chen, G.-H.; Fan, K.-N. Accurate Prediction of Heat of Formation by Combining Hartree–Fock/density Functional Theory Calculation with Linear Regression Correction Approach. *J. Chem. Phys.* **2004**, *121*, 7086−7095.

(45) Duan, X.-M.; Li, Z.-H.; Hu, H.-R.; Song, G.-L.; Wang, W.-N.; Chen, G.-H.; Fan, K.-N. Linear Regression Correction to First Principle Theoretical Calculations – Improved Descriptors and Enlarged Training Set. *Chem. Phys. Lett.* **2005**, *409*, 315−321.

(46) Schreiner, P. R.; Fokin, A. A.; Pascal, R. A.; de Meijere, A. Many Density Functional Theory Approaches Fail To Give Reliable Large Hydrocarbon Isomer Energy Differences. *Org. Lett.* **2006**, *8*, 3635−3638.

(47) Check, C. E.; Gilbert, T. M. Progressive Systematic Underestimation of Reaction Energies by the B3LYP Model as the Number of C−C Bonds Increases: Why Organic Chemists Should Use Multiple DFT Models for Calculations Involving Polycarbon Hydrocarbons. *J. Org. Chem.* **2005**, *70*, 9828−9834.