The Journal of Chemical Physics

Size-independent neural networks based first-principles method for accurate prediction of heat of formation of fuels

Cite as: J. Chem. Phys. **148**, 241738 (2018); https://doi.org/10.1063/1.5024442
Submitted: 01 February 2018 . Accepted: 22 May 2018 . Published Online: 08 June 2018

GuanYa Yang, Jiang Wu, ShuGuang Chen, WeiJun Zhou, Jian Sun, and GuanHua Chen

Size-independent neural networks based first-principles method for accurate prediction of heat of formation of fuels

GuanYa Yang, Jiang Wu, ShuGuang Chen, WeiJun Zhou, Jian Sun, and GuanHua Chen[a)]
Department of Chemistry, The University of Hong Kong, Pokfulam Road, Hong Kong, China

(Received 1 February 2018; accepted 22 May 2018; published online 8 June 2018)

Neural network-based first-principles method for predicting heat of formation (HOF) was previously demonstrated to be able to achieve chemical accuracy in a broad spectrum of target molecules [L. H. Hu *et al.*, J. Chem. Phys. **119**, 11501 (2003)]. However, its accuracy deteriorates with the increase in molecular size. A closer inspection reveals a systematic correlation between the prediction error and the molecular size, which appears correctable by further statistical analysis, calling for a more sophisticated machine learning algorithm. Despite the apparent difference between simple and complex molecules, all the essential physical information is already present in a carefully selected set of small molecule representatives. A model that can capture the fundamental physics would be able to predict large and complex molecules from information extracted only from a small molecules database. To this end, a size-independent, multi-step multi-variable linear regression-neural network–B3LYP method is developed in this work, which successfully improves the overall prediction accuracy by training with smaller molecules only. And in particular, the calculation errors for larger molecules are drastically reduced to the same magnitudes as those of the smaller molecules. Specifically, the method is based on a 164-molecule database that consists of molecules made of hydrogen and carbon elements. 4 molecular descriptors were selected to encode molecule's characteristics, among which raw HOF calculated from B3LYP and the molecular size are also included. Upon the size-independent machine learning correction, the mean absolute deviation (MAD) of the B3LYP/6-311+G(3df,2p)-calculated HOF is reduced from 16.58 to 1.43 kcal/mol and from 17.33 to 1.69 kcal/mol for the training and testing sets (small molecules), respectively. Furthermore, the MAD of the testing set (large molecules) is reduced from 28.75 to 1.67 kcal/mol. *Published by AIP Publishing.* https://doi.org/10.1063/1.5024442

I. INTRODUCTION

Reaching chemical accuracy is still a great challenge for theoretical and computation chemists. In recent years, experimentalists rely heavily on accurate first-principles calculations to help find new materials and design new microscopic functional structures. Although there are well-developed *ab initio* methods, only the methods with proper treatment of the electron correlation can possibly reach chemical accuracy. These methods, however, are usually very expensive and their use is limited to molecules of limited numbers of atoms. However, the study of large molecules such as enzyme is more fascinating and mysterious. Even for the smallest known enzyme, 4-Oxalocrotonate tautomerase (consisting of 62 amino acids),[1] the computational complexity for a real quantum mechanical first-principles calculation is beyond currently available computational resources. Computational chemists almost always resort to drastic approximations due to the steep scaling of computational complexity with respect to system size. For example, the full configuration interaction (full-CI) method[2] scales factorially with respect to the number of electrons in the system.

The computational complexity can be greatly reduced by employing density-functional theory (DFT), which states that the ground-state electron density function rather than the many-electron wave function is enough to determine all electronic properties. The DFT method has been widely employed to calculate chemical properties of various molecules and has made great success. However, its accuracy is yet to be improved further. Particularly, computational errors of almost all *ab initio* first-principles calculations (includes DFT) are systematic, which can be corrected by the definition. One particular property of our interest is the heat of formation (HOF) of hydrocarbon. In DFT calculations where B3LYP or other widely used functionals are adopted, such error in the Heat of Formation (HOF)[3–6] and bond energy[7–10] is mainly attributable to the poorly estimated long-range nonbonded attractive effects.[11] We also observed a strong correlation between the raw DFT calculation error and the system size by indicating correlation coefficients among physical descriptors of molecules and in hydrocarbons.

In 2003, for the first time, we proposed a Neural Network (NN)-based DFT approach[12,13] which is able to reach chemical accuracy.[14] This type of algorithm[14–17] was adopted and improved by several groups around the world and was extended to predicting a variety of physical properties such as bond dissociation energy,[18] enthalpies of reaction,[19] potential energy surface,[20,21] DFT energies,[22] and electronic structures. For instance, the X1 method makes use of an extended database that includes open-shell molecules and radicals to

[a)]Author to whom correspondence should be addressed: ghc@everest.hku.hk

improve the accuracy of the calculated $\Delta_f H^{\ominus}_{298}$ (HOF).[23] Besides the pursuit of accuracy, the machine learning methods have also been applied to speed up the first-principles methods. In 2012, Rupp *et al.* employed machine learning method and designed coulomb matrix to bypass solving Schrödinger equation, which successfully reduced the calculation time of molecular atomization energies to several seconds based on nuclear charges and atomic positions only.[24] Later in 2014, we further improved the NN-based DFT method by enlarging the database and quantifying the uncertainty of each calculated value.[25] The state-of-the-art algorithm can achieve even better accuracy, for example, with more complicated descriptors. Zhou *et al.* (2016) reported errors of 0.7 and 0.6 kcal/mol for the G4 and X_{3D} methods, respectively, on the Chen/13 database.[26] These studies opened a new direction for first-principles quantum mechanical methods and an alternative approach to reach the chemical accuracy. On the other hand, instead of further bringing down the overall error, our current work will be focusing on demonstrating a crucial idea of capturing and insulating the size-dependency in the model, making size-invariant predictions possible (predicting large molecules with small molecules only, without extra errors).

As reported in Ref. 25, the mean absolute deviation (MAD) error of the NN-based DFT prediction is greatly reduced to 1.31 kcal/mol in the testing set, but the existing systematic error against size-related physical descriptors is disappointing. Therefore, in this manuscript, we aim to achieve two goals. The first is to improve the overall prediction accuracy for both large and small molecules by correcting the systematic error in the DFT calculation. The second goal is to develop an improved NN-based DFT methodology that is size-independent so that the physical properties such as HOF of large molecules can be accurately predicted using the data from small molecules. After performing apreliminary test on the performance of various machine learning algorithms in reducing size-dependent errors, we found that Multi-Variable Linear Regression (MVLR) is a simple and efficient method to reduce the initial calculation errors. To reduce the error further, we introduce a neural network to correct remaining non-linear errors. As comparisons, DFT and size-independent NN approaches are tested on a database composed of hydrocarbons. The two-step MVLR-NN method is expected to remove the trend of increasing errors with the increasing molecular sizes while minimizing the overall error at the same time.

II. METHODOLOGY

A. Construction of database

In our previous work, the Chen/13 database was well prepared to include accurate experimental data and ensure its wide coverage and high distribution density, which served as a good starting point for this work.[25] To focus on fuel molecules, the Chen/13 database is tailored to remove the radicals and keep only hydrocarbons (164 molecules). The purpose of restricting our dataset to hydrocarbon molecules was mainly to limit the dimension of the descriptor space and to avoid complications and extra uncertainty due to the scarcity of O/N/S/Cl-containing molecules. The new database of 164 molecules, which is termed the Chen/13-Fuel database, is employed in this study. Each molecule is characterized by a vector $(x_1, x_2, \ldots, x_4)$, in which x_1 to x_4 represent the 4 physical descriptors of the molecule. Then the Chen/13-Fuel database can be represented by a 4×164 matrix, with each column as a vector containing 4 descriptors of one of the 164 molecules. In our previous machine learning method,[25] a prediction model is made to minimize the overall error in the testing set, which is assumed to be generated by the same distribution with the training set. However, in this manuscript, we aim to predict large molecules with information extracted only from small molecules. To evaluate the predictive power and generalizability, 13 molecules containing over 11 carbon atoms are extracted as the testing set (large molecules). Same number of molecules containing less than 11 carbon atoms, randomly extracted from the dataset, are used as the testing set (small molecules). The remaining 138 molecules containing less than 11 carbon atoms are used as the training set (Fig. 1). Details of the data are listed in Table SIII of the supplementary material.

B. Computational methods

The geometries of all 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 (no imaginary frequencies). The raw B3LYP/6-311+G(d,p) ZPE is employed in calculating $\Delta_f H^{\ominus}_{298}$. The strategies to calculate $\Delta_f H^{\ominus}_{298}$ are adopted from the previous work.[25] The difference between the DFT-calculated HOF and the experimental HOF (ε_{DFT}) will enter our model as the target; this differs from the models in our previous publications, where experimental HOF serves as the target and DFT calculated HOF serves as one of the descriptors. Bond order of a molecule is defined as the sum of (I − 1), where I represents each bond's order whose

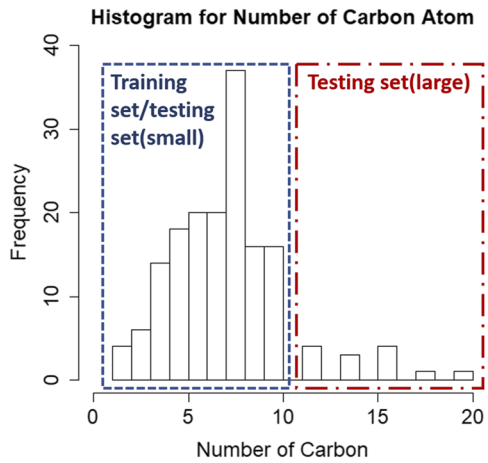

FIG. 1. Histogram of the number of carbon atoms in the Chen/13-Fuel dataset. The blue dashed line indicates molecules in the training set (138 small molecules) and the testing set (13 small molecules), which contain less than 11 carbon atoms. The red dot dashed line represents the testing set (13 small molecules) which contains more than 11 carbon atoms.

value is larger than 1. I is obtained from Natural bond orbital (NBO) analysis.[27] Spin multiplicity for a molecule is given by 2S + 1, where S is the total spin for the molecule. Gaussian 03 software package[28] is employed in the calculations.

III. SIZE-DEPENDENT ERROR CORRECTION APPROACH USING MULTI-VARIABLE LINEAR REGRESSION

A. Physical descriptors and the structure of multivariable linear regression

Descriptor subset selection is employed to automatically select the optimal descriptor number and combination among the 10 descriptors [Number of Hydrogen (N_H), Number of Carbon (N_C), Bond Order (B.O.), Spin Multiplicity (S.M.), Number of Valence electron (N_{Ve}), Molecular Weight (M.W.), Spin Multiplicity per Valence Electron (S.M.N_{Ve}), Number of Electrons per Atom (N_e%), Degree of Unsaturation (DU), and Degree of Unsaturation per Carbon Atom (DU%)]. 3 descriptors {B.O., S.M., and DU} are selected using best subset selection for MVLR.

1. Calculation of degree of unsaturation of molecule

The Degree of Unsaturation (DU) is defined as

$$DU = \frac{2N_C + 2 - N_H}{2}, \tag{1}$$

where N_C and N_H represent number of carbon and hydrogen atoms.

2. Best subset selection

Best subset selection is performed using an efficient algorithm called "leap and bounds."[29] This algorithm aims to find for each k ∈ {0, 1, 2, . . . p} the subset of size k, which gives the smallest residual sum or squares or other criterion. The results of best subset selection are shown in Fig. S1 and Tables SI and SII of the supplementary material, where RSS (residual sum of squares), Adjusted RSq (adjusted R-square value), CP (Mallow's Cp), and BIC (Bayesian information criterion) are used as the criteria for selecting descriptors. Results showed that when 3 descriptors {N_C, B.O., and DU} or 4 descriptors {B.O., DU%, S.M., and DU} are chosen, models gave best results (by performance as well as parsimony). Notice that the 3 descriptor set is not a subset of the 4 descriptor set. By this information alone, choosing either set for the MVLR would not make a significant difference; however, since the training set and testing set are selected based on the number of carbon atoms, choosing the 4 descriptor set instead of the 3 descriptor set {N_C, B.O., and DU} avoids the complexity of testing set contamination from this descriptor. Additionally, since the correlation between DU and DU% is too string and their information is to some extent redundant, we manually removed DU% (DU is more "size-dependent" than DU%) from the 4 descriptor set. Linear regression was recalculated with the 3 descriptors {B.O., S.M., and DU} alone, and their coefficients are shown in Table I.

3. Structure of multi-variable linear regression

The multi-variable linear regression is performed by "stats" package in R studio,[30] which employs the least-squares method to minimize the sum of squared residue errors over the training set (138 molecules) and determine the optimal coefficients. Equation (3) and determined coefficients (see Table I) are as follows. Instead of using the experimental HOF value as the target, in this study the MVLR model is employed to calibrate the difference (ε_{DFT}) between the DFT-calculated HOF (Raw DFT $\Delta_f H^{\ominus}_{298}$) and the experimental value (Expt $\Delta_f H^{\ominus}_{298}$) [see Eq. (2)]; the resulted MVLR prediction value is termed as D_{MVLR}, which is then used as one of the input layers to train the size-independent NN [see Eq. (3)], and the output of the size-independent NN-predicted difference is termed as $D_{Size-independent\ NN}$. Errors of MVLR (ε_{MVLR}) and size-independent ($\varepsilon_{Szie-independent\ NN}$) are calculated using Eqs. (4) and (5).

TABLE I. The coefficient and intercept of the MVLR model.

Intercept (β_0)	Spin multiplicity (β_1)	Degree of unsaturation (β_2)	Bond order (β_3)
−6.3138	−4.5654	2.3003	0.8887

$$\varepsilon_{DFT} = \text{Raw DFT } \Delta_f H^{\ominus}_{298} - \text{Expt } \Delta_f H^{\ominus}_{298}, \tag{2}$$

$$D_{MVLR} = \beta_0 + \beta_1\, Spin\ Multiplicity + \beta_2\, Degree\ of\ Unsaturation + \beta_3 Bond\ Order + \varepsilon, \tag{3}$$

$$\varepsilon_{MVLR} = \text{Raw DFT } \Delta_f H^{\ominus}_{298} - \text{Expt } \Delta_f H^{\ominus}_{298} - D_{MVLR}, \tag{4}$$

$$\varepsilon_{Szie-independent\ NN} = \text{Raw DFT } \Delta_f H^{\ominus}_{298} - \text{Expt } \Delta_f H^{\ominus}_{298} - D_{Size-independent\ NN}. \tag{5}$$

IV. CONSTRUCTION OF NEURAL NETWORK

A. Physical descriptors and the structure of neural network

As in a previous study,[25] a feed-forward NN model is adopted with a three-layer architecture, including the input layer, hidden layer, and output layer. According to the previous benchmark, 5 hidden neurons are used to avoid potential overfitting. Size-independent NN is constructed with 4 descriptors in its input layers [see Fig. 2(a)], including the output of MVLR (D_{MVLR}) [see Eq. (3)] and the percentages of carbon atoms in each molecule (N_C %). Following our work in 2014, one of the remaining two descriptors is chosen to be the bond order. Additionally, the degree of unsaturation per carbon atom (degree of unsaturation %) is used as the forth input descriptor. Since the size-related information is encoded in the output value of multi-variable linear regression itself as the first input neuron of the size-dependent NN, the rest of the input neurons are chosen to be size-invariant physical descriptors. As a comparison, size-dependent NN is also studied [see Fig. 2(b)]. Instead of using size-invariant descriptors, size-dependent NN is constructed with size-related descriptors, such as raw DFT-calculated HOF (DFT $\Delta_f H^{\ominus}_{298}$), number of hydrogen ($N_H$) and carbon atoms ($N_C$), degree of unsaturation, and bond order [Fig. 2(b)].

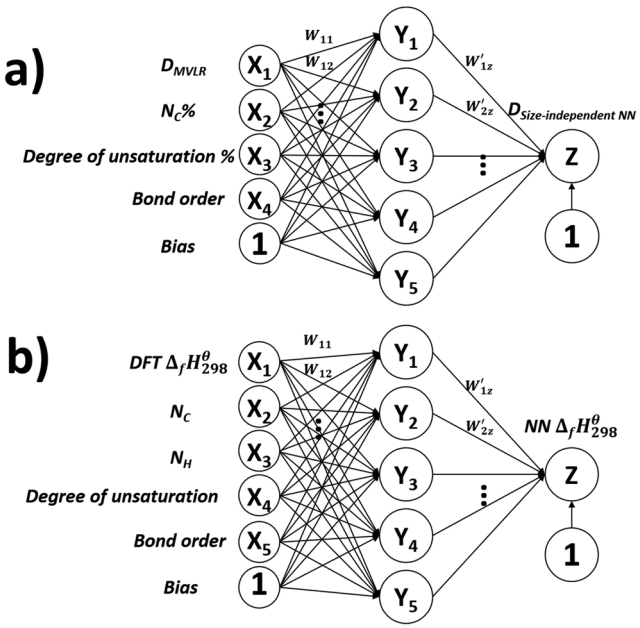

FIG. 2. Structure of the size-independent NN (a) and size-dependent NN (b). $\mathbf{W_{ih}}$ and $\mathbf{W'_{hz}}$ are the weight coefficients connecting the input ($X_{1\text{-}5}$) and hidden ($Y_{1\text{-}5}$) layers and the hidden and output layers (Z), respectively. N_α is the number of α-type atoms, and N_α% is the percentage of α-type atoms in molecules. Bias equals to 1. The output values of size-independent NN and size-dependent NN are termed as $D_{\text{Size-independent NN}}$ and NN $\Delta_f H^{\ominus}_{298}$, respectively.

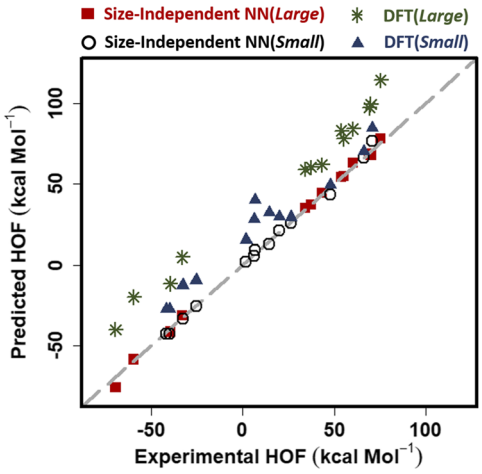

FIG. 3. Validation of HOF value in the overall testing set. Solid red squares and black circles represent large and small molecules of the testing set, respectively, which are generated by size-independent NN. Green eight-spoked asterisks and blue triangles represent large and small molecules calculated by DFT, respectively. The dashed line is the identity line.

B. Training neural network

The Bayesian regulation back-propagation method[31] is adopted to train the neural network, which can fully utilize the training dataset and avoid potential over-fitting. The "brnn" function of the neural network package[32] in R is employed to perform the training. Before training, all the inputs are scaled and mapped to [−1, +1] as

$$x_{i,j} \rightarrow \frac{2(x_{i,j} - x_i(min))}{x_i(max) - x_i(min)} - 1, \tag{6}$$

where i runs over 4 descriptors, j runs over the 164 molecules in the training set, and x_i (max) and x_i (min) are the minimum and maximum elements of the raw inputs of the i-th descriptor in the training set, respectively. The output value of the output neuron, Z, is related to the inputs of the j-th molecule, $x_j = (x_{1,j}, x_{2,j}, \ldots, x_{4,j})$, as

$$Z = \sum_{h=1,5} W_{y_h} Sig\left(\sum_{i=1,4} W_{ih} x_i + b_h\right) + b_z, \tag{7}$$

where the transfer function $Sig(v) = \frac{2}{1+e^{-2v}} - 1$; W_{ih} is the weight coefficient connecting the input neuron x_i and the hidden neuron y_h, b_h is the bias on hidden neuron y_h, W'_{hz} is the weight coefficient connecting the hidden neuron y_h and the output neuron Z, and b_z is the bias on output neuron Z (Fig. 2).

V. RESULTS AND DISCUSSION

A. Analysis of size-independent NN-based first-principles results in testing set

Figure 3 shows the fitting plot of the experimental HOF value against the predicted value of HOF by DFT and size-independent NN in the overall testing set (26 molecules). DFT-calculated large molecules (green eight-spoked asterisk) are located far from the identity line compared with the DFT-calculated small molecules (blue triangles). After performing size-independent NN, not only the overall error decreased but also the trend of size discrimination in prediction accuracy is no longer visible, indicating a successful removal of size-related residue errors from raw DFT prediction. After the first step, the neural network regression can fully focus on the size-invariant residue errors; as can be seen in the plot, the errors in both small molecules (black circle) and large molecules (solid red triangle) were further brought down by the NN step, in a uniform manner.

To further validate prediction models, Mean Absolute Deviations (MAD) are calculated for three HOF calculation approaches regarding the training set and testing set and shown in Table II. As expected, DFT achieves great errors on large molecules (MAD = 28.75) compared with small molecules (MAD = 17.33), which indicates that the existence of systematic errors limits the usage of DFT from further implementation on large systems. Moreover, size-independent NN showed its satisfactory performance of prediction in large and small molecule groups of the testing set as the error is reduced to less than 1.67 and 1.69 kcal/mol, respectively. Especially, the percentage of improvements by size-independent

TABLE II. Mean Absolute Deviations (MAD) for 164 molecules obtained from the data set in this study (all data are in kcal/mol).

MAD (kcal/mol)	DFT	MVLR	Size-dependent NN[a]	Size-independent NN[a]
Testing (large)[b]	28.75	2.73	2.70 (90.6%)	1.67 (94.2%)
Testing (small)[b]	17.33	2.00	1.70 (90.2%)	1.69 (90.2%)
Training	16.58	1.78	1.51 (89.3%)	1.43 (91.4%)

[a]Percentage of improvement by the NN correction is given in parentheses.
[b]testing set is divided into two groups based on the number of carbon atoms in each molecule, which are labeled as "large" ($N_C > 11$) (13 molecules) and "small" ($N_C < 11$) (13 molecules).

TABLE III. Comparison of the size-dependency of the raw DFT calculation, MVLR, and size-independent NN using R-square value analysis.

R^2	ε_{DFT}	ε_{MVLR}	$\varepsilon_{Size-independent\ NN}$
N_{at}	0.83	0.32	0.00044
N_e	0.87	0.13	0.019
N_C	0.78	0.06	0.037

NN in large molecules is 94.2. In comparison, size-dependent NN is also constructed and the performance of this one-step NN is shown in Table II. MAD of the testing set (small molecules) is reduced to 1.70 kcal/mol from 17.33 kcal/mol (DFT calculation), with a 90.2% improvement, which is the same as with size-independent NN. However, in the testing set (large molecule), 2.70 kcal/mol is still much more larger compared with the MAD value of the testing set (small molecules).

B. Error analysis for larger size of fuel molecules in testing set

To validate the performance and to quantitate linearity between errors of models and system size, R^2 values (see Table III) are generated between errors of DFT (ε_{DFT}), errors of MVLR (ε_{MVLR}), errors of size-independent NN ($\varepsilon_{Size-independent\ NN}$) [see Eqs. (2), (4), and (5)], and three size-related descriptors [total number of atoms (N_{at}), number of electrons (N_e), and number of carbon atoms (N_C)] in the testing set (13 large molecules and 13 small molecules). It shows that the total number of atoms, number of electrons, as well as the number of carbon atoms all give a high R-square value against the DFT error, which demonstrates that the error DFT calculation greatly depends on the size of the system. However, the R^2 value of the three descriptors decreases dramatically after applying MVLR and size-independent NN. The systematic error is well corrected by size-independent NN since there is nearly no correlation between the error of size-independent NN and size-related physical descriptors. Notice that the MVLR step alone does not have to eliminate all size-dependency in its residue (column 2) since, afterward, when performing the following neural network step, the size-dependent information still exists but only exists in such a way that can be eliminated through the correlation with the first descriptor (the MVLR output). In a sense, all relevant size-dependency was encoded by the MVLR output that serves as the only entry route for size-dependent information into the neural network step.

Figure 4 expends the prediction error of the raw DFT calculation versus that of our size-independent model along 3 size-related variables, number of atoms, number of electrons, and number of carbon atoms, respectively. In addition to dramatic improvements in the overall magnitude of error, our model successfully removes visible linear or non-linear trends from all 3 plots, where the red triangles (size-independent NN) distribute uniformly along the entire range of the data. In conclusion, our model significantly reduces the prediction error over DFT calculations and displays impressive size-independency.

Prediction results of MVLR and size-independent NN are compared by the analysis of polynomial regression, where the black dots and solid lines represent the error and fitting curve of MVLR results. The red triangles and dot dashed lines are generated from size-independent NN results (Fig. 5). All the fitting curves are plotted using the second-/third-order polynomial regression. Fitting curves of size-independent NN (red dot dashed curves) are much more flat compared with curves that fitted using MVLR prediction results (black solid lines). Conclusion can be made that our second step neural network successfully removes the non-linear error that still exists after implying the linear regression algorithm on DFT calculation.

Error is a complicated object. An overall error decomposes usually into different components depending on their sources or statistical properties. In the case of HOF, an unknown deterministic function dictates each molecule's target value. The only limitation lies in the fact that the sample

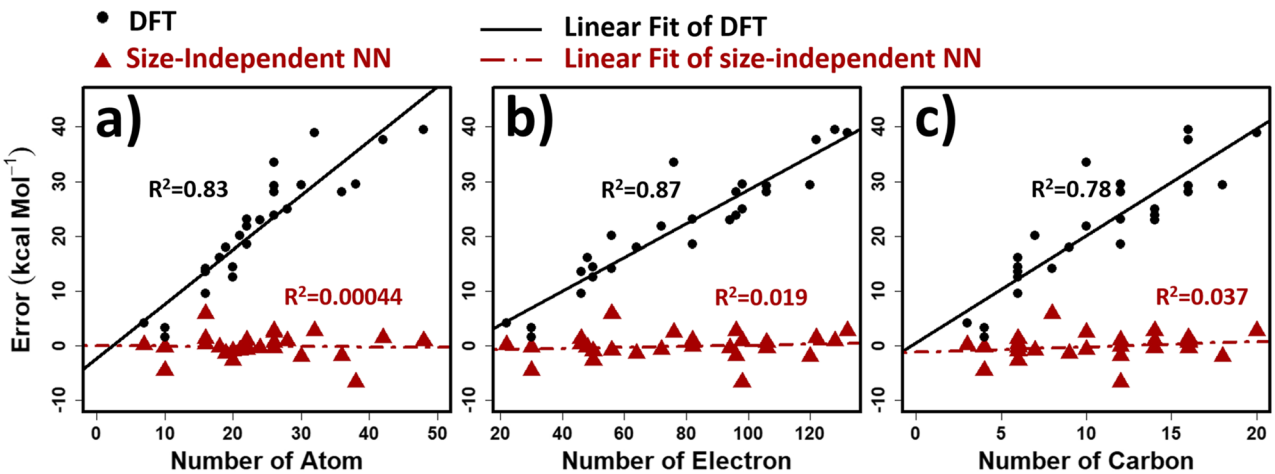

FIG. 4. Errors of DFT and size-independent NN versus (a) total number of atoms, (b) number of electrons, and (c) number of carbon atoms. R^2 values are calculated by linear regression for the errors of the two methods on each of the three descriptors. The six regression lines are plotted using coefficients and intercepts from six linear regressions. Solid black lines demonstrate the linear fit of the DFT calculation and dot dashed lines represent the linear fit of the size-independent NN calculation.

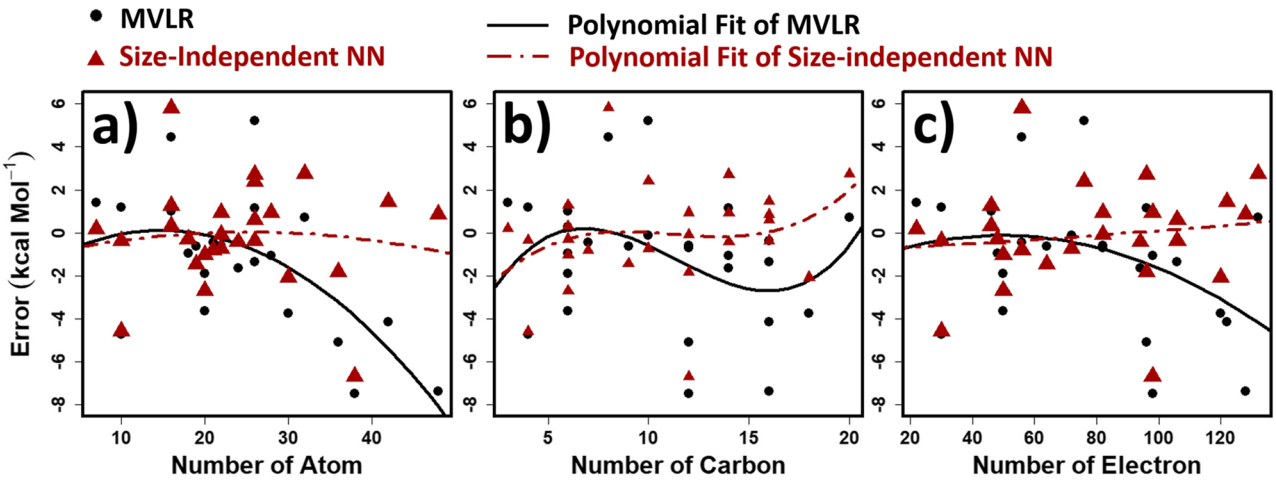

FIG. 5. Polynomial regression between N_{at} and N_C and N_e against errors of MVLR and size-independent NN. (a) Errors of MVLR and size-independent NN versus the number of carbon atoms; (b) errors of MVLR and size-independent NN versus the number of carbon atoms; (C) errors of MVLR and size-independent NN versus the number of electrons. Black solid lines represent the polynomial fit curve of MVLR results and red dot dashed lines are the fitting curve of size-independent NN calculation.

set is finite and came from a sampling process that introduces extra randomness. All error components naturally form a hierarchy of increasing complexity. Methods targeting simpler error components are easier to implement and usually generalize better out-of-sample while skipping many details. On the other hand, methods targeting the complex error components perform better overall while deteriorating quickly out of data range (e.g., when extrapolating). In our case, a one-step neural network Bayesian training is by itself complicated enough to capture the fine details, as demonstrated by our earlier studies.[14–17,25] However, it does not have the capability to automatically distinguish error components of different complexity and may regard all errors as complex, therefore losing some degree of generalizability. On the other hand, in our implementation, a MVLR extracts and reserves the simpler trend leaving only the "harder" components for neural network training. To prevent the neural network from reverting the previously captured trend in favor of removing complex error components, the Bayesian regularizer[31] automatically switches on to protect generalizability and prevent overfitting. Our results undoubtedly demonstrate the advantages of our multi-step regression algorithm, with its small-to-large extrapolation giving 1.67 kcal/mol in the testing set (large molecules), out-performing the 2.70 kcal/mol from the one-step method counterpart.

VI. CONCLUSION

In this paper, we report a size-independent NN-based first-principles method for HOF of fuel molecules (Chen/13-Fuel, 164 molecules in total). MAD of the testing set (large molecules) is reduced from 28.75 kcal/mol to 1.67 kcal/mol for the B3LYP/6-311+G(3df,2p) calculation. Our method yields the accurate HOFs for large molecules, while training is performed on the database of the small molecules. Results show that the MVLR-NN method is indeed size-independent, removing the trend of increasing errors versus the increasing molecular sizes.

SUPPLEMENTARY MATERIAL

See supplementary material for the details on the Chen/13-fuel database and results of the subset selection.

ACKNOWLEDGMENTS

We thank the Research Grant Council of HKSAR (Grant Nos. AoE/P-04/08, HKU700913P, and 17316016) and The University of Hong Kong (Seed Funding for Strategic Research Theme) for their support.

[1] L. H. Chen, G. L. Kenyon, F. Curtin, S. Harayama, M. E. Bembenek, G. Hajipour, and C. P. Whitman, J. Biol. Chem. **267**, 17716 (1992).
[2] J. B. Foresman, A. E. Frisch, and I. Gaussian, *Exploring Chemistry with Electronic Structure Methods* (Gaussian, Inc., 1996).
[3] B. K. Raghavachari, B. B. Stefanov, and L. A. Curtiss, Mol. Phys. **91**, 555 (1997).
[4] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, J. Chem. Phys. **106**, 1063 (1997).
[5] P. Winget and T. Clark, J. Comput. Chem. **25**, 725 (2004).
[6] M. Saeys, M. Reyniers, G. B. Marin, V. Van Speybroeck, and M. Waroquier, J. Phys. Chem. A **107**, 9147 (2003).
[7] Y. Feng, L. Liu, J. T. Wang, H. Huang, and Q. X. Guo, J. Chem. Inf. Comput. Sci. **43**, 2005 (2003).
[8] F. Yao, X. Y. Dong, Y. M. Wang, L. Liu, and Q. X. Guo, Chin. J. Chem. **23**, 474 (2005).
[9] E. J. Brändas and E. S. Kryachko, *Fundamental World of Quantum Chemistry: A Tribute to the Memory of Per-Olov Löwdin Volume III* (Springer Netherlands, 2004).
[10] E. I. Izgorodina, M. L. Coote, and L. Radom, J. Phys. Chem. A **109**, 7558 (2005).
[11] M. D. Wodrich, C. Corminboeuf, and P. von Ragué Schleyer, Org. Lett. **8**, 3631 (2006).
[12] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[13] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[14] L. H. Hu, X. J. Wang, L. H. Wong, and G. H. Chen, J. Chem. Phys. **119**, 11501 (2003).
[15] X. Zheng, L. H. Hu, X. J. Wang, and G. H. Chen, Chem. Phys. Lett. **390**, 186 (2004).
[16] X. J. Wang, L. H. Hu, L. H. Wong, and G. H. Chen, Mol. Simulat. **30**, 9 (2004).
[17] 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).
[18] J. Wu and X. Xu, J. Chem. Phys. **129**, 164103 (2008).

[19]M. D. Wodrich and C. Corminboeuf, J. Phys. Chem. A **113**, 3285 (2009).
[20]J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[21]J. Behler, R. Martoňák, D. Donadio, and M. Parrinello, Phys. Rev. Lett. **100**, 185501 (2008).
[22]R. M. Balabin and E. I. Lomakina, J. Chem. Phys. **131**, 074104 (2009).
[23]J. Wu and X. Xu, J. Chem. Phys. **127**, 214105 (2007).
[24]M. Rupp, A. Tkatchenko, K. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[25]J. Sun, J. Wu, T. Song, L. H. Hu, K. L. Shan, and G. H. Chen, J. Phys. Chem. A **118**, 9120 (2014).
[26]Y. Zhou, J. Wu, and X. Xu, J. Comput. Chem. **37**, 1175 (2016).
[27]A. E. Reed, L. A. Curtiss, and F. Weinhold, Chem. Rev. **88**, 899 (1988).
[28]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, and J. C. Burant, GAUSSIAN 03, Revision C. 02, Gaussian, Inc., Wallingford, CT, 2004.
[29]G. M. Furnival and R. W. Wilson, Technometrics **16**, 499 (1974).
[30]R. C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing (2017).
[31]D. J. C. MacKay, Neural Comput. **4**, 415 (1991).
[32]P. P. Rodriguez and D. Gianola, brnn: Bayesian Regularization for Feed-Forward Neural Networks, R package version 0.6 (2016).