JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 1849−1857

pubs.acs.org/jcim

Efficient Corrections for DFT Noncovalent Interactions Based on Ensemble Learning Models

Wenze Li,[†] Wei Miao,[†] Jingxia Cui,[‡] Chao Fang,[†] Shunting Su,[†] Hongzhi Li,[†] LiHong Hu,*[,†] Yinghua Lu,*[,†,‡] and GuanHua Chen[§]

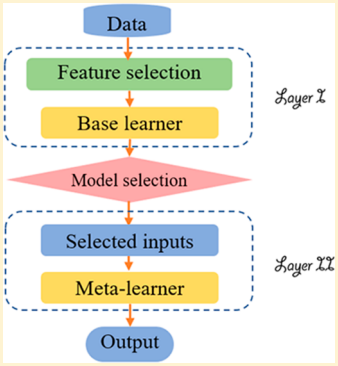
[†]School of Information Science and Technology, Northeast Normal University, Changchun, 130117, China

[‡]Institute of Functional Material Chemistry, Faculty of Chemistry, Northeast Normal University, Changchun, 130024, China

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

Supporting Information

ABSTRACT: Machine learning has exhibited powerful capabilities in many areas. However, machine learning models are mostly database dependent, requiring a new model if the database changes. Therefore, a universal model is highly desired to accommodate the widest variety of databases. Fortunately, this universality may be achieved by ensemble learning, which can integrate multiple learners to meet the demands of diversified databases. Therefore, we propose a general procedure for learning ensemble establishment based on noncovalent interactions (NCIs) databases. Additionally, accurate NCI computation is quite demanding for first-principles methods, for which a competent machine learning model can be an efficient solution to obtain high NCI accuracy with minimal computational resources. In regard to these aspects, multiple schemes of ensemble learning models (Bagging, Boosting, and Stacking frameworks), are explored in this study. The models are based on various low levels of density functional theory (DFT) calculations for the benchmark databases S66, S22, and X40. All NCIs computed by the DFT calculations can be improved to high-level accuracy (root-mean-square error RMSE = 0.22 kcal/mol in contrast to CCSD(T)/CBS benchmark) by established ensemble learning models. Compared with single machine learning models, ensemble models show better accuracy (RMSE of the best model is further lowered by ∼25%), robustness and goodness-of-fit according to evaluation parameters suggested by the OECD. Among ensemble learning models, heterogeneous Stacking ensemble models show the most valuable application potential. The standardized procedure of constructing learning ensembles has been well utilized on several NCI data sets, and this procedure may also be applicable for other chemical databases.



1. INTRODUCTION

Noncovalent interactions (NCIs) are the dominant driving force in many significant systems, such as biological molecules,[1] supermolecules,[2,3] and nanomaterials.[4,5] For large molecular systems, they are ubiquitous and represent a very important binding mode. Their striking manifestations include the properties of liquids,[6] functional materials,[4,5] molecular recognition,[7] or the structure and function of biomacromolecules,[8] to name just a few. The direct experimental measurement of NCIs has been extremely difficult, especially for large molecules, so computational methods have become important tools for investigating NCIs.[9,10]

Currently, CCSD(T)/CBS is considered the gold standard of NCI calculations, so its results can provide benchmark data that can be used for investigating the nature of NCIs.[11,12] However, it is a quite demanding calculation, and even unrealistic for molecules with tens of atoms.[13,14] Recently, great progress has been made: Neese et al. have developed a linear scaling DLPNO−CCSD(T), which provides CCSD calculations for large molecules with high efficiency and accuracy,[15−17] and machine learning potentials have also been effectively used for improving computational methods.[18−21] In spite of these advances, the need for better efficiency is not yet satisfied. Density functional theory (DFT) methods are the most often used quantum chemical methods because of their low cost and satisfactory performance,[22] and DFT is one of the most promising tools for large molecules. However, DFT methods with NCI computations are still one of the major application obstacles.[13,14,22−24] Lately, progress in machine learning has begun to soar again, which opens up new opportunities for improving DFT NCI calculations.

Recently, artificial intelligence/machine learning has been widely adopted in various research fields and industries, and its spread and influence are still growing. There are many algorithms for regression modeling. According to the complexity of learning frameworks, they can be divided into three categories: weak, ensemble and deep learning. Although single machine learning is simple and weak, it is the basis of complex machine learning methods. Conventional techniques, such as back-propagation, activation function, and neural network, are

Special Issue: Women in Computational Chemistry

Received: November 30, 2018
Published: March 26, 2019

DOI: 10.1021/acs.jcim.8b00878
J. Chem. Inf. Model. 2019, 59, 1849−1857

frequently used in deep learning.[25] In machine learning methods, linear and nonlinear algorithms are generally adopted. Linear regression analysis methods include ordinary least-squares regression or multiple linear regression (MLR), principal component regression, partial least-squares (PLS) regression, ridge regression, least absolute shrinkage and selection operator elastic net, and linear support vector regression (SVR). On the other hand, popular nonlinear methods, such as artificial neural networks, deep learning, nonlinear SVR, and random forests (RF), are often used to solve difficult problems.[26−28]

In practice, machine learning adoptions depend on the scale and attributes of databases. At present, deep learning is a hotspot, but it is not applicable to all databases. Small databases can only be modeled by single and ensemble learning methods, since deep learning needs to be fed by big data. However, single machine learning is usually unstable and its generalization ability is limited because of its data sensitivity. While ensemble learning is a machine learning paradigm and a collection of a finite number of single machine learning algorithms trained for the same task,[29] it aims to obtain better prediction performance than that from any of its component weak learning (i.e., base learner) algorithms. Plenty of applications have shown that ensemble learning generally outperforms a single learning model. In addition, it has been found very useful in diverse domains, such as computer-aided medical diagnosis, computer vision, network security, and so on.[30,31] Currently, even deep learning is applied under ensemble frameworks for better performances.[32]

Bagging, boosting, and stacking are three basic schemes of learning ensembles. In principle, when constructing a capable ensemble learning method, two important things should be considered carefully: the diversity of base learners and a suitable combination rule to achieve a final output.[33] The common source of diversities can be determined from different features or training subsets, or perhaps various settings of base learners. For early bagging and boosting methods, the diversity is acquired by the resample strategy. The classifiers included in bagging are trained with various data subsets, which are randomly sampled from the original data set. Generally, a majority voting scheme is applied as the combining method to make a collective decision in bagging. Boosting mostly uses a weighted resample strategy, where the weights of all instances are initialized equally, whereas if an instance is misclassified, its weight will be increased. Thus, it is more likely to select the misclassified instances into the next training subset. A stacking ensemble is based on at least two levels. In the first level, multiple base learners are trained. The diversity can be introduced by using the same or different base learners with various feature spaces or training subsets. In the second layer, a combination learning algorithm (i.e., meta-learner) is applied to generate the final prediction. The meta-learner takes the outputs of first-level base learners as inputs to get the real label instead of the vote in bagging and boosting.

Our correction model combines machine learning and quantum chemical calculations. The quantum chemical methods are carried out to obtain molecular descriptors, which are taken as the inputs of machine learning for performing modeling. In our models, the calculated value of the property/target by a low-level theory (e.g., DFT calculated NCI) is also taken as a molecular descriptor. Because quantum chemical descriptors, especially the calculated target value, capture the major physical essence, a machine learning model is quite likely to establish a solid relationship between the inputs and targets. In this way, the gap between the calculated value and the target value is easy to reduce.[26] Moreover, running a machine learning model usually only takes hours at most, so in our study the time spent on obtaining high accuracy results is similar to that of the low-level calculation. Therefore, low-level quantum chemical calculations can be remarkably and efficiently improved to high accuracy. In our previous study, general regression neural networks (GRNN) were applied for NCI databases, in which different DFT NCI calculations were corrected to high level computational methods.[34] The root-mean-square errors (RMSEs) of the DFT calculated NCIs were improved by at least 70%. However, the stability and generalization of correction models need to be further improved. Addressing these problems, herein, a variety of ensemble learning models for NCI calculation corrections are developed, and general ensemble modeling strategies are explored. The established learning ensembles further improve the DFT NCI calculations. Among ensemble models, a typical structure, two-layer heterogeneous stacking ensembles, shows high performance on tested databases. To simplify ensemble model applications, a general procedure for learning ensemble establishment has been proposed. The two-layer stacking model is able to be standardized. In the first layer, the number and variety of base learners are two key factors. The number can be determined by meta-learner training through forward selection. The variety acquired by different feature selection methods significantly contributes to the accuracy of model predictions. In the second layer, meta-learners are trained by inputs determined in the first layer. Therefore, to achieve an exceptional learning ensemble, this standardized procedure may include several basic steps: multiple feature selection, hyperparameter determination and base/meta learner training. The details of each step are discussed in section 4.

The rest of this article is organized as follows. Section 2 gives a brief introduction to the theoretical methods of all the algorithms used in NCI corrections. Section 3 describes three major ensemble frameworks, Bagging, Boosting and Stacking. In section 4, the performance of built ensemble models for NCI corrections are compared, and it is discussed how to obtain the key factors—the number and the variety of base learners—in Stacking learning ensembles. And finally, a general procedure for Stacking ensemble establishment is summarized. Concluding remarks are given in section 5.

2. MATERIALS AND METHODS

2.1. Data Sets. The database was generated from publicly available NCI benchmark databases S22, S66, and X40 contributed by the Hobza group.[14,35] The detailed description of the database is in our previous study.[34] The descriptor list is in the support information Table S1. In this study, six commonly used DFT methods (M062X, ωB97XD, PBE, PBE-D3, B3LYP, B3LYP-D3) are utilized to compute the database in vacuum. Both Pople and Dunning types of basis sets (6-31G*, 6-31+G*, cc-pVDZ, and aug-cc-pVDZ)[36,37] are adopted in DFT calculations, which created nine data sets as shown in Table 1, where the data sets are abbreviated D1−D9 for convenience. The B3LYP, B3LYP-D3, PBE, and PBE-D3 calculations for data sets were performed by the ORCA3.0 quantum chemical program, and the rest of the DFT calculations were carried out using the Gaussian 09 program package.[38,39] For some data sets in our database, NCI obtained

DOI: 10.1021/acs.jcim.8b00878
J. Chem. Inf. Model. 2019, 59, 1849−1857

Table 1. DFT Methods for Noncovalent Interactions Data Sets

data sets	DFT method	basis sets
D1	M062X	6-31G*
D2	M062X	6-31+G*
D3	M062X	cc-pVDZ
D4	M062X	aug-cc-pVDZ
D5	ωB97XD	6-31G*
D6	PBE	6-31G*
D7	PBE-D3	6-31G*
D8	B3LYP	6-31G*
D9	B3LYP-D3	6-31G*

by ORCA3.0 is a little better than that in Gaussian 09 due to different calculation settings, such as HF exchange ratio, so those data sets generated from ORCA3.0 results were kept for machine learning modeling.

The data sets were divided into training and test sets by using set partitioning based on joint $x-y$ distances (SPXY) algorithm,[40] where the numbers of data in training and test sets were set to 91 and 30, respectively. In addition, all the data were normalized to [0,1] using the Minmax standardization method so that the values of all features and the response variable were of the same order of magnitude.

2.2. Feature Selection. Feature selection is a simple way to create diversified base learners to build an ensemble, where the diversity of base learners is from different feature subspaces.[41] In our study, because the scale of databases is quite small, feature selection is regularly adopted to reduce dimensionality to prevent the overfitting problem. So herein, feature selection is not only for dimension reduction but also for generating varieties of base learners. Moreover, multiple feature selection algorithms might be able to improve knowledge discovery, since important features can be discovered by different algorithms. However, it is clear that an ensemble of feature subset models affects the economy of representation.[42]

Traditional feature selection algorithms for generic data can be grouped into four categories: similarity based, information theoretical based, sparse learning-based and statistical-based methods.[42] One method has been chosen in each category as a representative feature selection method in our experiments, which are ReliefF, joint mutual information (JMI), least absolute shrinkage and selection operator (Lasso), and correlation-based feature selection (CFS). Because all of these four feature selection methods are filters, a wrapper method, forward selection (FS),[43] was added for enhancing the variety of selections.

2.3. Regression Methods. *2.3.1. Random Forest (RF).* The RF algorithm was proposed by Breiman, and it is an assembled recursive partitioning method.[44] In an RF regression model, when a new input sample enters, each decision tree in the forest is utilized to determine one prediction for the sample, and then the average of predictions is taken as an output. It has been widely employed for QSAR.[28,31] In our experiments, the default values of *ntree*(500) and *mtry*(p/2) are used, where p is the number of variables used in the model.

2.3.2. Gradient Boosting Decision Tree (GBDT). GBDT, one of the gradient Boosting algorithms, was developed by Jerome Friedman.[45] It is a combination of gradient boosting and decision tree and has been widely used in machine learning techniques for regression and classification problems.

Boosting relies on weighted majority voting of models in the ensemble, where new base learners are trained with respect to the error of the former ensemble model in each iteration. Thus, later base learners may generate a lower error rate than previous base learners.[46] The gradient-descent based on boosting algorithms and the corresponding models are called gradient boosting machines.[45] Classification and regression tree (CART) is a typical regression model. Mathematically, the model can be viewed as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{1}$$

where K is the number of trees, f is a function in the functional space F that is the set of all possible regression trees, and $\hat{y}_i$ is the predicted value.

2.3.3. Support Vector Machine (SVM). SVM was proposed by Vapnik in 1995 and is popularly used for pattern classification and nonlinear regression.[47] In particular, it can handle nonlinear and high-dimensional problems for small databases, because samples can be mapped to high-dimensional space where they are converted into linear problems. However, this may result in very complex calculations that used to be solved by using the kernel function to directly transfer features from low to high dimensions. This can also overcome the local minima problem. Therefore, SVM has become one of the most popular classification and regression methods due to its outstanding performance.[27] In the paper, the fitting function and the RBF kernel function are $f(x) = \sum \alpha_i y_i K(x,x_i) + \beta_0 (0 \le \alpha_i \le C)$ and $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, respectively, where the parameters γ and C are obtained from a grid search.

2.3.4. Multiple Linear Regression (MLR). MLR is a simple linear statistical method. The relation between response variable y and independent variable x is

$$y = \beta_0 - x\beta \tag{2}$$

where β_0 and β are obtained by minimizing the least-squares in eq 3.

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - x_i\beta)^2 \tag{3}$$

2.4. Model Assessment. Validation parameters recommended by OECD are used to evaluate the performance of ensemble models.[48] They are root mean squared error (RMSE) in eq 4, mean absolute errors (MAE, eq 5), predictive squared correlation coefficient (q^2, eq 6), and q^2 in K-fold cross-validation (q_{cv}^2). All RMSE and MAE values are in kcal/mol.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i| \tag{5}$$

$$q^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} \tag{6}$$

where Y_i, $\hat{Y}_i$, $\bar{Y}$, and n are the response, the predicted value, the averaged values (the training set), and the number of all

DOI: 10.1021/acs.jcim.8b00878
J. Chem. Inf. Model. 2019, 59, 1849−1857

samples, respectively. In the cross-validation, the q^2 values of 10 subset models are averaged to q_{cv}^2.

The sum of ranking differences (SRD) is often used to compare models, methods, analytical techniques, and panel members. To further compare the performance of DFT NCI correction models, SRDs are computed.[49] The closer the SRD value is to zero, the better the model.

3. ENSEMBLE LEARNING ON NCI CORRECTION

The frameworks of ensemble learning are constructed with either sequential or parallel base learners, which can be same (homogeneous) or multiple (heterogeneous) types in bagging and stacking ensembles, whereas boosting can only be homogeneous because of its algorithm. Through combination rules, an ensemble can grow to a strong learning machine by taking majority results of weak machine learning methods.

3.1. Bagging. Bagging, a bootstrap aggregating, is one of the earliest ensemble algorithms proposed by Breiman.[50] A subset randomly generated from a training data set by bootstrapping is used to train base learners. Generally, the individual outputs of base learners vote or are averaged for the final output of ensemble models. Bagging has shown fair performances in the variance reduction effect.[29] RF is a representative of the state-of-art ensemble methods, which is an extension of Bagging. Therefore, RF is chosen as the Bagging technique.

3.2. Boosting. Boosting works by training a set of learners sequentially, where later learners focus more on the mistakes of earlier learners. To achieve the goal, each training sample is assigned a weight. The whole training process tends to concentrate on "hard" instances through updating weights accordingly. The boosting algorithm derives its ultimate result by combining many base learners into a single final output through weights, which tends to reduce the bias.[51] GBDT is chosen as a representative of boosting technique in experiments.

3.3. Stacking. Stacking[29] is a general two-level framework that uses a learning algorithm as a specific combination method. Stacking can be built using the same or different learning algorithms in the first level, which thus forms homogeneous or heterogeneous stacking ensembles, respectively. A combiner algorithm (meta-learner) is trained to make the ultimate predictions using the predictions generated by the base learners in the previous layer.

In this study, the designed framework, a two-layer stacking structure, is presented in Figure 1. The established homogeneous and heterogeneous stacking ensembles are abbreviated as homo-SE and hete-SE, respectively. In the first layer, five feature subsets are generated by five feature selection methods. Then, three base learners (either the same in homo-SE or different in hete-SE) are trained by taking five feature subsets as inputs, respectively. In homo-SE, base learners are SVM (i.e., $Model_{1-n}$ all use the SVM method) and in hete-SE, base learners use three algorithms, where $Model_{1-5}$, $Model_{6-10}$, and $Model_{11-15}$ adopt SVM, RF, and MLR, respectively. When training homo-SE with each selected feature subset, a 10-fold cross-validation is applied, where the top three SVM models proceed to a model pool for further forward selection. Therefore, in total 15 models are preserved in the first layer. In hete-SE, three types of machine learning methods are trained based on feature subsets selected by five feature selection methods, respectively, thus generating 15 models as well in the first layer. In stacking ensembles, the output of the first layer is taken as the input of the second layer. To get optimal inputs for the meta-learner, forward selection (taking results of models one by one according to RMSEs that were ranked in the descending order) is used to select the best models from the model pool generated in the first layer. Finally, a meta-learner based on selected models is constructed in the second layer.

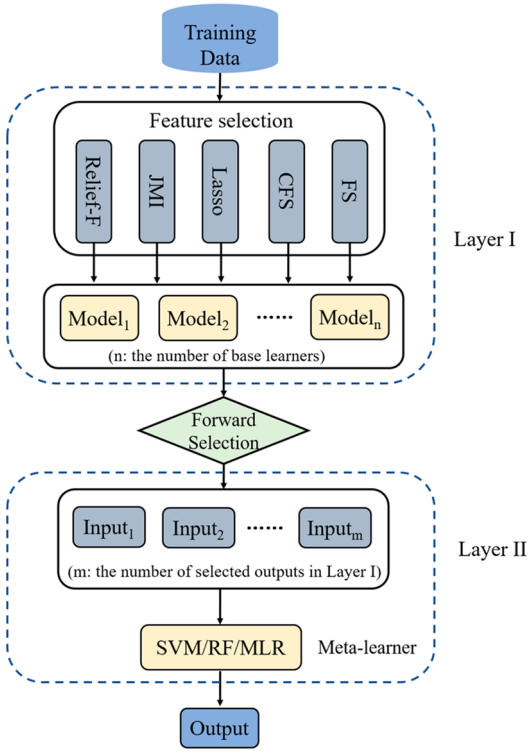


Figure 1. Stacking ensemble framework

4. RESULTS AND DISCUSSION

4.1. DFT Calculations. Four basis sets are used with DFT calculations on NCI computations. The results are shown in Table 2. In M062X calculations, Pople basis sets (6-31G* and 6-31+G*) only cost approximately 1/60 of the time spent by corresponding Dunning basis sets (cc-pVDZ and aug-cc-pVDZ), but the accuracy is similar for the corresponding basis sets (D1:1.66 vs D3:2.01, D2:1.07 vs D4:0.94 kcal/mol). Comparing D1 with D2 and D3 with D4, DFT calculations are significantly improved through adding diffuse functions on both types of basis sets in vacuum. However, it is unexpected that D7 and D9 are worse than D6 and D8, respectively. The reason may be that the magnitude of dispersion correction is a comparably large value, which makes the results with or without the dispersion correction differ greatly.[24,52]

4.2. Ensemble Learning Models. *4.2.1. Number of Base Learners.* The number of base learners is an important parameter when building an ensemble model. To a great extent, it depends on the algorithm in the meta-learners. Herein, SVM is used as the meta-learner, so the base learner number is optimized based on SVM. In the model pool, the preserved 15 models generated in the first layer are sorted according to RMSEs of validation data sets, and then the top m base learners are selected to constitute the input of the meta-learners.

DOI: 10.1021/acs.jcim.8b00878
J. Chem. Inf. Model. 2019, 59, 1849−1857

Table 2. Evaluation Parameters of DFT and Different Correction Models[a]

	RMSE/MAE					
	q_{cv}^2/q^2					
data sets	DFT	GRNN	RF	GBDT	homo-SE	hete-SE
D1	1.66/1.34	0.50/0.35	0.42/0.30	0.30/0.20	0.34/0.22	0.31/0.20
	−/0.83	0.95/0.98	0.96/0.98	0.93/0.99	0.99/0.99	0.99/0.99
D2	1.07/**0.80**	**0.44**/0.36	0.35/0.24	0.27/0.21	0.34/0.19	0.25/0.16
	−/0.93	0.97/0.98	0.93/0.99	0.99/0.99	0.99/0.99	0.99/0.99
D3	2.01/1.54	0.61/0.39	0.59/0.36	0.50/0.30	0.52/0.27	0.50/0.30
	−/0.74	0.89/0.96	0.93/0.95	0.96/0.99	0.97/0.98	0.98/0.97
D4	**0.94/0.80**	0.47/0.34	**0.34/0.22**	**0.24/0.18**	**0.24/0.14**	**0.22/0.15**
	−/0.95	0.96/0.98	0.96/0.99	0.98/0.99	0.99/0.99	0.99/0.99
D5	1.77/1.55	0.47/0.35	0.41/0.28	0.35/0.23	0.32/0.21	0.29/0.20
	−/0.82	0.96/0.98	0.97/0.98	0.96/0.99	0.99/0.99	0.99/0.99
D6	1.96/1.52	0.47/**0.25**	0.59/0.42	0.39/0.26	0.37/0.26	0.37/0.25
	−/0.80	0.93/0.98	0.93/0.96	0.91/0.99	0.99/0.98	0.99/0.98
D7	2.64/2.19	0.46/0.32	0.50/0.32	0.41/0.27	0.37/0.24	0.31/0.21
	−/0.62	0.93/0.98	0.94/0.97	0.94/0.99	0.99/0.98	0.99/0.99
D8	1.96/1.57	0.48/0.34	0.54/0.38	0.36/0.26	0.36/0.23	0.34/0.24
	−/0.79	0.93/0.98	0.89/0.97	0.90/0.99	0.99/0.98	0.99/0.99
D9	2.34/2.05	0.48/0.32	0.43/0.28	0.27/0.19	0.37/0.25	0.36/0.26
	−/0.68	0.98/0.98	0.96/0.98	0.97/0.99	0.99/0.99	0.99/0.98

[a]RMSE and MAE in kcal/mol, q_{cv}^2 and q^2.

The results on the number of base learners on D4 (the RMSE is smallest among all data sets) are shown in Figure 2,

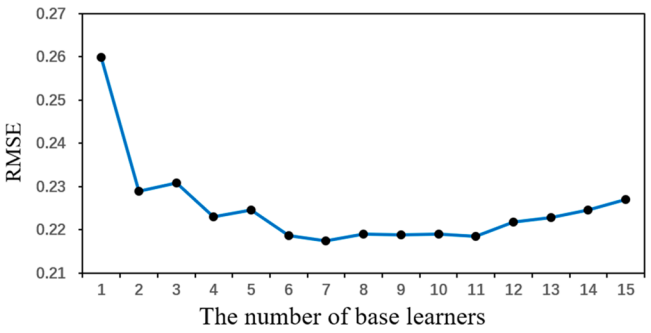


Figure 2. Hete-SE results based on different numbers of base learners on D4 data set.

where the relation between the number of base learners and RMSE of ensemble models can be observed. When the number of base learners is smaller than 7, RMSE goes down quickly. After that RMSEs become small and stable between 7 and 11; however, RMSE starts slowly going up starting at 12, which suggests that 7−11 is an optimal range of the number of base learners. The same calculation is also applied to other data sets (as shown in Figures S1−S8). For other data sets, the optimal range is probably different.

Since feature selection methods adopt different algorithms, the selected features are more or less diversified. According to the results of five feature selection methods, several features show more statistical significance than others, such as the arrangement of two NCI parts in dimers, the ratio of SP3 carbon atoms in molecules, the energy of HOMO, dipole moment, and nuclear potential.

4.2.2. Learning Ensemble Construction. Various ensemble models are constructed for DFT NCI data sets. All the proposed models were built in MATLAB 2014.[53] Three strategies of ensemble learning are compared for NCI correction. The single GRNN correction model[34] is also built with all data sets for comparison with ensemble models. The evaluation parameters of all models are listed in Table 2.

In Table 2, from the data set point of view, the best model is based on the best DFT calculation D4 (M062X/aug-cc-PVDZ); however, the difference between D2 and D4 is quite small and D2 (M062X/6-31+G*) costs a lot less than D4. In the perspective of models, NCI DFT calculations can be better corrected by ensemble models than the single GRNN model except for RF models based on D6−8 data sets. The RMSEs of ensemble models were reduced from 0.94 to 2.64 to 0.22−0.59 and MAEs decreased from 0.80 to 2.19 to 0.14−0.42. The RMSE of D7 data set, which has the largest RMSE (2.64) of DFT in all data sets, was reduced to 0.50 maximum (RF model) and 0.31 minimum (hete-SE model), that is, the RMSE was reduced by 81%−88%. The largest correction was achieved by combinations of hete-SE with DFT in D7 and GBDT with DFT in D9, which reduced RMSE by 88%. The least error result (RMSE = 0.22) was obtained by hete-SE correction in D4, which was based on the best DFT results. The ensemble model prediction based on D2 (RMSE S66 = 0.18 and S22 = 0.18) is much better than the single GRNN model (RMSE S66 = 0.43 and S22 = 0.28), and similar to MP2.5 (RMSE S66 = 0.16 and S22 = 0.22).[9] It can be seen that the performance of most hete-SE is the best, which suggests the importance of the base learner diversity, that is, the difference between homo-SE and hete-SE. In addition, boosting GBDT models are better than bagging RF, which proves that the combination rule is one of the key points of ensemble modeling. Because of the randomness of input features in the root tree of GBDT and RF, the results of base learners are not stable in their ensemble construction; therefore, q_{cv}^2 of GBDT and RF are smaller than other ensembles in Table 2. Additionally, the predictive power of models q^2 of all the ensemble models is larger than GRNN in all data sets, which indicates good predictability of ensemble learning.

DOI: 10.1021/acs.jcim.8b00878
J. Chem. Inf. Model. 2019, 59, 1849−1857

The difference of ensemble and single models in hete-SE method of D4 (the RMSE is the smallest in all data sets) is displayed in Figure 3. It shows that performances of ensemble models are superior to all single models, even though the 13th single model is very close to ensemble models. Apparently, RMSEs of single models fluctuate much of the time when changing feature subsets or base learners, whereas ensemble models can greatly avoid the instability of single models.

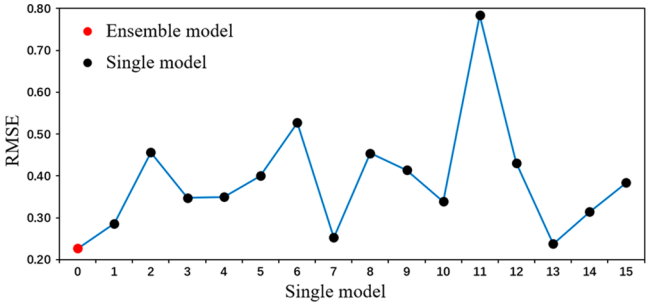


Figure 3. RMSE of the ensemble model vs single models in hete-SE of D4 data set. Horizontal and verticals are single models and RMSE, respectively.

The violin plots of errors between benchmark and various calculation results for all data sets are shown in Figure 4. It shows that all correction models give better results than DFT calculations and that the ensemble correction models are obviously superior to simple machine learning algorithms, especially on D4 and D5 data sets, underscoring the exceptional performance of the ensemble models. In Figure 4, hete-SE results are clearly better than any other method in D1, D2, D5, and D7 data sets, and similar to other methods in other data sets. From the overall evaluation, hete-SE is the best among all models.

It is further verified by the ranking of models by SRD values listed in Table 3, where the ranks of hete-SE and GRNN are mostly the first and last, respectively. SRD comparisons also demonstrate that ensemble models can more properly correct the DFT NCIs than a simple machine learning algorithm.

4.2.3. Base Learner and Meta-Learner in Stacking Ensemble. In the homo-SE scheme, homo-SVM chooses SVM as both base learner and meta-learner. For comparison, homo-SE with RF and MLR (homo-RF, homo-MLR) are also built, where the types of the base learner and meta-learner are same, that is, either RF or MLR, respectively.

In the hete-SE scheme, SVM, RF, and MLR are acting as the meta-learner, respectively, while the base learners include all three types of base learners, where each feature subset is modeled by all three methods (SVM/RF/MLR). The results of different algorithms used in homo-SE and hete-SE are shown in Table 4.

In Table 4, the results of nonlinear models (homo-SVM and homo-RF) are better than homo-MLR in homo-SE, but the results of three hete-SE models are similar. It is noted that under the same meta-learner conditions, homo-SEs are based on the same corresponding type of base learners, thus the

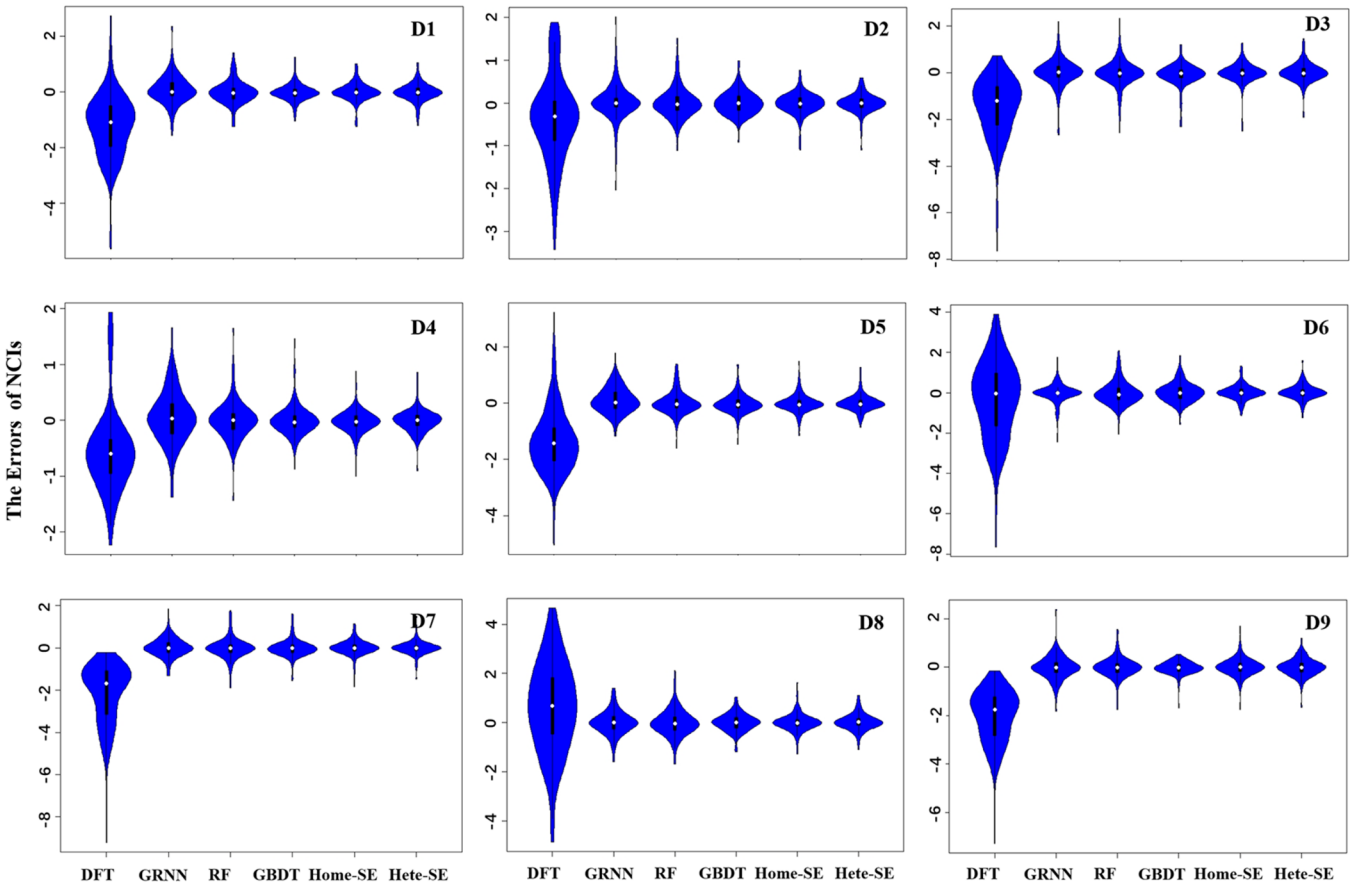


Figure 4. Violin plots of prediction errors for all data sets. The errors of data sets are calculated with $NCI_{Benchmark}-NCI_{DFT}$ and $NCI_{Benchmark}-NCI_{Model}$.

DOI: 10.1021/acs.jcim.8b00878
J. Chem. Inf. Model. 2019, 59, 1849−1857

Table 3. SRD Values and Ranks of GRNN, RF, GBDT, Homo-SE, and Hete-SE Correction Models of NCI Data Sets

	SRD					SRD rank				
data sets	GRNN	RF	GBDT	homo-SE	hete-SE	GRNN	RF	GBDT	homo-SE	hete-SE
D1	9.03	5.22	4.19	5.14	5.09	5	4	1	3	2
D2	6.75	4.71	5.07	4.74	4.15	5	2	4	3	1
D3	8.32	6.92	6.27	6.02	6.51	5	4	2	1	3
D4	8.04	4.00	3.52	3.29	3.25	5	4	3	2	1
D5	8.81	5.02	4.24	5.72	4.97	5	3	1	4	2
D6	8.94	7.91	7.90	6.50	6.08	5	4	3	2	1
D7	8.33	4.88	4.77	5.32	4.57	5	3	2	4	1
D8	9.71	6.98	6.27	6.25	5.68	5	4	3	2	1
D9	8.63	4.98	3.92	5.40	5.54	5	2	1	3	4

Table 4. RMSE of Different Meta-learners Used in Homo-SE and Hete-SE

	homo-SE			hete-SE		
data set	SVM	RF	MLR	SVM	RF	MLR
D1	0.34	0.33	0.84	0.31	0.28	0.31
D2	0.34	0.25	0.71	0.25	0.22	0.27
D3	0.52	0.45	0.85	0.50	0.43	0.49
D4	0.24	0.30	0.34	0.23	0.20	0.20
D5	0.32	0.30	0.71	0.32	0.30	0.32
D6	0.37	0.39	0.65	0.37	0.30	0.35
D7	0.37	0.31	0.60	0.32	0.29	0.30
D8	0.36	0.33	0.81	0.33	0.32	0.35
D9	0.37	0.41	0.59	0.35	0.31	0.35

variety of base learners in each homo-SE is smaller than that in hete-SEs, which are based on different types of base learners. This suggests that the variety of base learners is more important than that of a meta-learner in the stacking technique. The results for all data sets by hete-SE are listed in Tables S2 and S3.

4.3. General Procedure for Hete-SE. In summary, to establish a learning ensemble, a general procedure can be simply adopted as follows: (1) Multiple feature selections/feature subspaces are used to obtain high varieties of base learners. (2) Determination of the hyperparameters in which the type and the number of base learners on the basis of meta-learners. (3) Setting a threshold for selecting good base learners as inputs of meta-learners to achieve the high accuracy prediction.

5. CONCLUSION

NCI is an intractable problem for both theoretical and experimental studies. Machine learning is probably an easy way to solve this problem. Ensemble learning is a paradigm of machine learning and a great strategy for both simple machine learning and deep learning algorithms, which avoids the instability and improves the performance of single models. In this study, ensemble learning has been established for NCI DFT calculation corrections; moreover, a general framework for hete-SE in chemical databases is proposed. In comparison with single models, the best hete-SE improves the accuracy further by 25% except for the robustness and generalization of the correction models. The standardized model building procedure may be adaptive for other types of chemical databases.

According to our experiments, we offer the following suggestions for obtaining high accuracy NCIs. (1) The ensemble models, especially hete-SE, are superior to a single machine learning algorithm in NCI corrections. An economic computation on high-accuracy NCIs can use hete-SE based on M062X/6-31+G* calculations. (2) Building a hete-SE can be performed with the following procedure: select multiple feature sets, model base learners, optimize the number of inputs based on meta-learners, and finally, obtain outputs with meta-learners. (3) The variety of base learners is more important than that of meta-learners.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00878.

Hete-SE results based on different numbers of base learners on different data sets besides D4, full list of the molecular descriptors, and Hete-SE calculated NCIs on all data sets (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*Tel.: +86-431-8453-6230. Fax: +86-431-8453-6331. E-mail: lhhu@nenu.edu.cn.

*E-mail: luyh@nenu.edu.cn.

ORCID

LiHong Hu: 0000-0003-3792-2917

GuanHua Chen: 0000-0001-5015-0902

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from National Natural Science Foundation of China (21473025) and the Science and Technology Development Planning of Jilin Province (20160204044GX).

■ REFERENCES

(1) Geng, L. J.; Yu, X. D.; Li, Y. J.; Wang, Y. Q.; Wu, Y. Q.; Ren, J. J.; Xue, F. F.; Yi, T. Instant hydrogel formation of terpyridine-based complexes triggered by DNA via non-covalent interaction. *Nanoscale* **2019**, *11*, 4044−4052.

(2) Otero-De-La-Roza, A.; Johnson, E. R. Predicting Energetics of Supramolecular Systems Using the XDM Dispersion Model. *J. Chem. Theory Comput.* **2015**, *11*, 4033−4040.

(3) Ksenofontov, A. A.; Bichan, N. G.; Khodov, I. A.; Antina, E. V.; Berezin, M. B.; Vyugin, A. I. Novel non-covalent supramolecular systems based on zinc(II) bis (dipyrromethenate)s with fullerenes. *J. Mol. Liq.* **2018**, *269*, 327−334.

DOI: 10.1021/acs.jcim.8b00878
J. Chem. Inf. Model. 2019, 59, 1849−1857

(4) Nell, K. M.; Fontenot, S. A.; Carter, T. G.; Warner, M. G.; Warner, C. L.; Addleman, R. S.; Johnson, D. W. Non-covalent functionalization of high-surface area nanomaterials: a new class of sorbent materials. *Environ. Sci.: Nano* **2016**, *3*, 138−145.
(5) Cooper, V. R.; Lam, C. N.; Wang, Y. Y.; Sumpter, B. G. Noncovalent Interactions in Nanotechnology. In *Non-Covalent Interactions in Quantum Chemistry and Physics: Theory and Applications*, 1st ed.; Otero-De-La-Roza, A., DiLabio, G. A., Eds.; Elsevier: Cambridge, MA, 2017; Chapter 14, pp 417−451.
(6) Roohi, H.; Ghauri, K.; Salehi, R. Non-covalent green functionalization of boron nitride nanotubes with tunable aryl alkyl ionic liquids: A quantum chemical approach. *J. Mol. Liq.* **2017**, *243*, 22−40.
(7) Huang, G. B.; Liu, W. E.; Valkonen, A.; Yao, H.; Rissanen, K.; Jiang, W. Selective recognition of aromatic hydrocarbons by endo-functionalized molecular tubes via C/N-H···Π interactions. *Chin. Chem. Lett.* **2018**, *29*, 91−94.
(8) Ryde, U.; Söderhjelm, P. Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods. *Chem. Rev.* **2016**, *116*, 5520−5566.
(9) Caldeweyher, E.; Bannwarth, C.; Grimme, S. Extension of the D3 dispersion coefficient model. *J. Chem. Phys.* **2017**, *147*, 034112.
(10) Prasad, V. K.; Otero-de-la-Roza, A.; DiLabio, G. A. Atom-Centered Potentials with Dispersion-Corrected Minimal-Basis-Set Hartree−Fock: An Efficient and Accurate Computational Approach for Large Molecular Systems. *J. Chem. Theory Comput.* **2018**, *14*, 726−738.
(11) Sedlak, R.; Riley, K. E.; Řezáč, J.; Pitoňák, M.; Hobza, P. MP2.5 and MP2.X: approaching CCSD(T) quality description of non-covalent interaction at the cost of a single CCSD iteration. *ChemPhysChem* **2013**, *14*, 698−707.
(12) Řezáč, J.; Hobza, P. Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the "Gold Standard," CCSD(T) at the Complete Basis Set Limit? *J. Chem. Theory Comput.* **2013**, *9*, 2151−2155.
(13) Hobza, P.; Müller-Dethlefs, K. In *Non-covalent Interactions: Theory and Experiment*; Hirst, J., Jordan, K., Lim, C., Theil, W., Eds.; RSC Theoretical and Computational Chemistry Series; The Royal Society of Chemistry: Cambridge, UK, 2009.
(14) Hobza, P. Calculations on Noncovalent Interactions and Databases of Benchmark Interaction Energies. *Acc. Chem. Res.* **2012**, *45*, 663−672.
(15) Riplinger, C.; Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **2013**, *138*, 034106.
(16) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. Sparse maps—A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *J. Chem. Phys.* **2016**, *144*, 024109.
(17) Neese, F.; Hansen, A.; Liakos, D. G. Efficient and accurate approximations to the local coupled cluster singles doubles method using a truncated pair natural orbital basis. *J. Chem. Phys.* **2009**, *131*, 064103.
(18) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(19) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: A data set of 20M off-equilibrium DFT calculations for organic molecules. *Sci. Data* **2017**, *4*, 170193.
(20) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(21) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.
(22) Pastorczak, E.; Corminboeuf, C. Perspective: Found in translation: Quantum chemical tools for grasping non-covalent interactions. *J. Chem. Phys.* **2017**, *146*, 120901.
(23) Hoja, J.; Tkatchenko, A. First-principles stability ranking of molecular crystal polymorphs with the DFT+MBD approach. *Faraday Discuss.* **2018**, *211*, 253−274.
(24) Kruse, H.; Goerigk, L.; Grimme, S. Why the Standard B3LYP/6-31G* Model Chemistry Should Not Be Used in DFT Calculations of Molecular Thermochemistry: Understanding and Correcting the Problem. *J. Org. Chem.* **2012**, *77*, 10824−10834.
(25) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, 2017.
(26) Hu, L. H.; Wang, X. J.; Wong, L. H.; Chen, G. H. Combined first-principles calculation and neural-network correction approach for heat of formation. *J. Chem. Phys.* **2003**, *119*, 11501−11507.
(27) Li, H. Z.; Zhong, Z. Y.; Li, L.; Gao, R.; Cui, J. X.; Gao, T.; Hu, L. H.; Lu, Y. H.; Su, Z. M.; Li, H. A cascaded QSAR model for efficient prediction of overall power conversion efficiency of all-organic dye-sensitized solar cells. *J. Comput. Chem.* **2015**, *36*, 1036−1046.
(28) Li, H. Z.; Li, W. Z.; Pan, X. F.; Huang, J. Q.; Gao, T.; Hu, L. H.; Li, H.; Lu, Y. H. Correlation and redundancy on machine learning performance for chemical databases. *J. Chemom.* **2018**, *32*, No. e3023.
(29) Zhou, Z. H. In *Ensemble Methods: Foundations and Algorithms*; Herbrich, R., Graepel, T., Eds.; Machine Learning & Pattern Recognition Series; Taylor & Francis Group, LLC: Boca Raton, FL, 2012.
(30) Xia, J.; Hsieh, J. H.; Hu, H.; Wu, S.; Wang, X. S. The Development of Target-Specific Pose Filter Ensembles to Boost Ligand Enrichment for Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2017**, *57*, 1414−1425.
(31) Li, H.; Cui, Y. Y.; Liu, Y. L.; Li, W. Z.; Shi, Y.; Fang, C.; Li, H. Z.; Gao, T.; Hu, L. H.; Lu, Y. H. Ensemble Learning for Overall Power Conversion Efficiency of the All-Organic Dye-Sensitized Solar Cells. *IEEE Access* **2018**, *6*, 34118−34126.
(32) Xu, Y. Y.; Wang, L.; et al. Ensemble One-Dimensional Convolution Neural Networks for Skeleton-Based Action Recognition. *IEEE Signal Proc. Lett.* **2018**, *25*, 1044−1048.
(33) Zhou, Z. H. *Machine Learning* (in Chinese); Tsinghua University: Beijing, 2016.
(34) Gao, T.; Li, H. Z.; Li, W. Z.; Li, L.; Fang, C.; Li, H.; Hu, L. H.; Lu, Y. H.; Su, Z. M. A machine learning correction for DFT non-covalent interactions based on the S22, S66 and X40 benchmark databases. *J. Cheminf.* **2016**, *8*, 24.
(35) Řezáč, J.; Riley, K. E.; Hobza, P. Benchmark calculations of noncovalent interactions of halogenated molecules. *J. Chem. Theory Comput.* **2012**, *8*, 4285−4292.
(36) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971**, *54*, 724−728.
(37) Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(38) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision D.01; Gaussian, Inc.: Wallingford, CT, 2009.
(39) Neese, F. The ORCA program system. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73−78.

DOI: 10.1021/acs.jcim.8b00878
J. Chem. Inf. Model. 2019, 59, 1849−1857

(40) Galvao, R. K. H.; Araujo, M. C. U.; Jose, G. E.; Pontes, M. J. C.; Silva, E. C.; Saldanha, T. C. B. A method for calibration and validation subset partitioning. *Talanta* **2005**, *67*, 736−740.
(41) Brown, G.; Wyatt, J.; Harris, R.; Yao, X. Diversity creation methods: a survey and categorization. *Inform. Fusion* **2005**, *6*, 5−20.
(42) Li, J. D.; Cheng, K. W.; Wang, S. H.; Morstatter, F.; Trevino, R. P.; Tang, J. L.; Liu, H. Feature Selection: A Data Perspective. *Acm Comput. Surv.* **2017**, DOI: 10.1145/3136625.
(43) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Benchmarking Variable Selection in QSAR. *Mol. Inf.* **2012**, *31*, 173−179.
(44) Breiman, L. Random forests. *Mach. learn.* **2001**, *45*, 5−32.
(45) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189−1232.
(46) Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobotics* **2013**, *7*, 21.
(47) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273−297.
(48) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694−701.
(49) Héberger, K.; Kollár-Hunek, K. Sum of ranking differences for method discrimination and its validation: Comparison of ranks with random numbers. *J. Chemom.* **2011**, *25*, 151−158.
(50) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123−140.
(51) Schapire, R. E.; Freund, Y. *Boosting: Foundations and Algorithms*; MIT Press: Cambridge, Massachusetts, 2012.
(52) Hostaš, J.; Řezáč, J. Accurate DFT-D3 Calculations in a Small Basis Set. *J. Chem. Theory Comput.* **2017**, *13*, 3575−3585.
(53) *Matlab R2014b Neural Network Toolbox*, version R2014b; The Mathworks: Natick, MA, 2014.

DOI: 10.1021/acs.jcim.8b00878
J. Chem. Inf. Model. 2019, 59, 1849−1857