frontiers in Chemistry

ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fchem.2020.00162

Toward the Prediction of Multi-Spin State Charges of a Heme Model by Random Forest Regression

Wei Zhao†, Qing Li†, Xian-Hui Huang, Li-Hua Bie and Jun Gao**

Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China

The random forest regression (RFR) model was introduced to predict the multiple spin state charges of a heme model, which is important for the molecular dynamic simulation of the spin crossover phenomenon. In this work, a multiple spin state structure data set with 39,368 structures of the simplified heme–oxygen binding model was built from the non-adiabatic dynamic simulation trajectories. The ESP charges of each atom were calculated and used as the real-valued response. The conformational adapted charge model (CAC) of three spin states was constructed by an RFR model using symmetry functions. The results show that our RFR model can effectively predict the on the fly atomic charges with the varying conformations as well as the atomic charge of different spin states in the same conformation, thus achieving the balance of accuracy and efficiency. The average mean absolute error of the predicted charges of each spin state is <0.02 e. The comparison studies on descriptors showed a maximum 0.06 e improvement in prediction of the charge of Fe^{2+} by using 11 manually selected structural parameters. We hope that this model can not only provide variable parameters for developing the force field of the multi-spin state but also facilitate automation, thus enabling large-scale simulations of atomistic systems.

Keywords: spin crossover, heme model, force field, machine learning, ESP charge

Edited by:
Yong Wang,
Ningbo University, China

Reviewed by:
Tong Zhu,
East China Normal University, China
Chaoyuan Zhu,
National Chiao Tung University, Taiwan

***Correspondence:**
Li-Hua Bie
biebie@mail.hzau.edu.cn
Jun Gao
gaojun@mail.hzau.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Theoretical and Computational Chemistry, a section of the journal Frontiers in Chemistry

Received: 27 January 2020
Accepted: 24 February 2020
Published: 31 March 2020

Citation:
Zhao W, Li Q, Huang X-H, Bie L-H and Gao J (2020) Toward the Prediction of Multi-Spin State Charges of a Heme Model by Random Forest Regression. Front. Chem. 8:162. doi: 10.3389/fchem.2020.00162

1. INTRODUCTION

Coordinated compounds of transition metal ions can exhibit a switching phenomenon under certain conditions related to changes in temperature, pressure, light, or magnetic field; the central metal ion changes the spin states (the so-called high-spin, HS, and low-spin, LS, configurations), which is the spin transition (ST) or spin crossover (SCO) (Bousseksou et al., 2011; Gutlich et al., 2013). Since Cambi et al. first reported the thermally induced change of spin states in 1931, (Cambi and Szegö, 1931) many more SCO complexes have been synthesized thereafter and have been applied to various domains, including molecular switches, memory elements (Jureschi et al., 2014; Shao et al., 2015), temperature sensors (Gütlich and Goodwin, 2004; Doukov et al., 2011), nanomaterials (Nagl et al., 2008; Hauser, 2013), and so on (Bousseksou et al., 2011; Cong et al., 2018; Yuan et al., 2018; Meyer et al., 2019).

In the switching phenomenon, the change of spin state is accompanied by a switch of electron configurations of the central ions, which often leads to marked changes in the physical and chemical properties of the entire complex (Gütlich and Goodwin, 2004; Habenicht and Prezhdo, 2012; Gutlich et al., 2013). Meanwhile, the reorganization of electrons among atoms and the formation of molecules are complex and multifaceted processes, and their full description is only possible within the boundaries of quantum mechanics (QM) (Bristow et al., 2014; Sanvito, 2019). Density functional theory (DFT) is the most common choice for routine ground-state calculations; however, the number of valence electrons scaled cubically, increasing the computational costs significantly (Engler et al., 2019). It will therefore not be suitable, especially when one needs to sample extended size and time scales.

Molecular dynamics (MD) simulation can handle system sizes of typically 10^7 atoms and above, and this has been used for decades to explore chemical and biochemical problems at an atomic level (Liu et al., 2017; Riniker, 2018). The classical MD predominantly uses simplified atomistic models called force fields (FFs) to describe the exact ground-state potential energy surface (PES) of a system. The bonded parameters are represented in terms of equilibrium bond distances, bond and dihedral angles, force constants, and rotation barriers; the non-bonded interactions are typically described by atom-centered point charges and Lennard-Jones potential (Ivanov et al., 2015) while disregarding the explicit treatment of electronic polarizability (De et al., 2018; Sahoo and Nair, 2018; Heid et al., 2019). It is not capable of capturing a restricted but essential number of chemical features, including spin crossover, wherein the molecular system is required to "hop" from one PES of the initial spin state onto another of the product state.

In order to better understand the effect of molecular properties on their electronic ground or excited states, the potential parameter set needs to be extended by a multi-spin state in which at least two issues should be taken into account. Firstly, the geometric configuration at energy minima of the excited state is different from that of the ground state in most cases. This issue can be fixed by adjust the parameters in bonding terms. For example, Meyer's Group has modified force constants for bond stretching and bending terms according to DFT calculation for atomistic molecular dynamics simulations of the HS and LS states of the Fe^{2+} containing model (Meyer et al., 2019). Secondly, it is well-known that the charge distribution in the excited state is different from the ground state, and it will change with molecular structures; it is important for the force field to provide the charges of two spin states. In this regard, an increasing number of schemes have been proposed in addition to the polarized force field, such as the SSAPs method (Xu et al., 2018).

In recent years, many efforts have been directed to the efficient improvement of force fields. In particular, machine learning combined with molecular simulation has been verified by many groups to be effective to develop force field including inferring charges based on a set of reference molecules (Botu et al., 2016; Chen et al., 2018; Inokuchi et al., 2018; Engler et al., 2019; Hu et al., 2019; Roman et al., 2019; Sanvito, 2019; Unke and Meuwly, 2019; Ye et al., 2019). Among these, the random forest regression (RFR) method has been proven to be feasible for the prediction of atomic charge without expending much effort on parameter tuning or descriptor selection. As a classification and regression tool, the Random Forest algorithm was first introduced by Breiman (2001), inspired by the earlier work of Amit and Geman (1997). It uses bootstrap samples of the training data and random feature selection in tree induction. Each tree in the ensemble produces an output according to the molecular descriptors or properties, and outputs from all trees are aggregated to produce the final prediction by average (Breiman, 2001; Cutler et al., 2011). This procedure can reduce overfitting and offer some unique features, including built-in performance assessment and measures of variable importance (Svetnik, 2003; Klusowski, 2018), which make it suitable for quantitative structure-activity relationship (QSAR) tasks (Svetnik, 2003; D Richard et al., 2007; Statnikov et al., 2008; Genuer et al., 2010). For instance, Rai and Bakken (2013) combined random forest regression with ESP charges from high-level QM calculations to predict the partial atomic charge of H, C, N, O, F, S, and Cl. Building on their work, Bleiziffer et al. (2018) further presented a conformational robust charge extraction scheme DDEC to predict partial charges and achieved accuracy beyond a HF/6-31G* setup. Our group developed a conformational adaptive charges (CAC) model based on atom type symmetry function (ATSF), which was, in turn, based on the RFR method (Wang and Gao, 2020). These machine learning approaches in tandem with quantum mechanics have many merits in developing flexible and adaptive force fields, including low cost, accuracy, and versatility. Yet, they are mainly used to predict charges on the single potential energy surface of the equilibrium configuration of the molecule. The performances of these method on multi-spin state charges remains unreported.

In our previous work (Liu et al., 2017; Du et al., 2018), the spin-forbidden dioxygen binding dynamics in a simplified heme model were investigated by the non-adiabatic trajectory surface-hopping dynamics, and this involved the coupled singlet, triplet, and quintuplet states. The results revealed that there existed dominant long-lived, kinetically meta-stable states during the dynamics trajectories, and each meta-stable pattern showed a distinct partial charge population. Based on this geometric dependence of the partial charge population on the excited state, we proposed to extend the conformation adapted charge (CAC) model and RFR method to the multi-spin state charges of the heme model. The fixed-point charge in the traditional force field can be modified according to the conformation on the fly, and thus the key to the multi-spin state is transformed into the change of charge in the multi-spin state. We hope that this model can not only provide variable parameters for constructing the force field of the multi-spin state but also facilitate automation, thus enabling large-scale simulations of atomistic systems.

2. MATERIALS AND METHODS

In this work, we targeted the simplified heme model (see **Figure 1**), introduced a random forest regression (RFR) algorithm using Behler-Parrinello symmetry functions as descriptors (Behler et al., 2007; Hagai et al., 2010; Behler, 2011a), performed model training by fitting ESP charges of different spin states, and achieved high-quality predictions. The key steps of the workflow are shown in **Figure 2**. The samples sufficient molecular conformations were obtained from *ab initio* dynamic trajectories of previous work (Du et al., 2018), which covered a wide range of conformations related to the spin crossover. Different descriptors were then extracted,

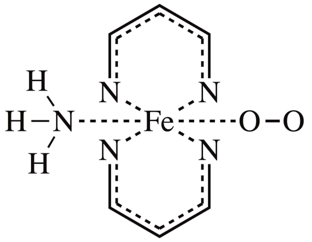

FIGURE 1 | The molecular model of this work. The simplified heme model $Fe^{2+}(C_3N_2)_2NH_3$ complex with O_2 binding was adopted.

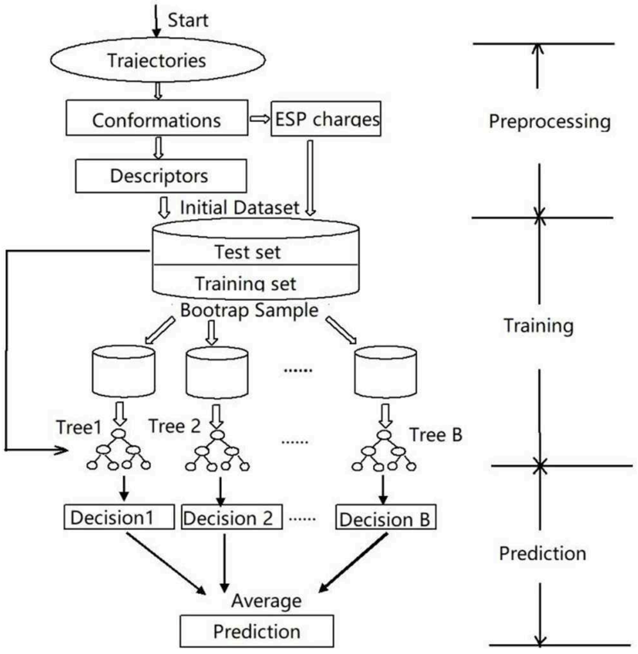

FIGURE 2 | The work flow of construction data set and prediction process of the RFR model.

and the ESP charges of three spin states of each atom in each conformation were calculated using the density function theory method, and together these constitute the initial dataset. After this preprocessing was completed, half of the data were selected randomly as the training set to build the RFR model, and the remaining half of the data were used to test the model's ability to reproduce the atomic partial charge under different spin states and thereby to analyze and assess the performance of the model.

2.1. Data Set Preparing

A total of 33 stable trajectories of open-shell singlet state were selected from a non-adiabatic trajectory surface-hopping dynamics simulation from our previous work. The B3LYP/6-31G* level of the method (Reiher et al., 2001; Salomon et al., 2002) was used to calculate the ESP atomic charge of each structure in the singlet, triplet, and quintuplet state. We finally achieved 39,368 converged structures owing to the convergence of the calculation. The data preparation was time consuming. By and large, it took 2 weeks to complete all the calculations of the 39,368 structures for each spin state with four computer nodes; each node had dual Intel 2683v3 CPUs. All the electronic structure calculations were implemented with a Gaussian 16 package (Frisch et al., 2016), and the detail charge distribution of each atom in the different spin states were analyzed and shown in the section 3.

2.2. Random Forest Regression Model Training

The raw dataset was preprocessed firstly to extract appropriate features, such as the descriptors of structures and input of model. Specifically, each RFR model was constructed separately under certain spin states for each atom according to the flow shown in **Figure 2**. Since there were 14 atoms in the simplified heme model, 14 independent RFR models were constructed by training for each spin state. There were 42 models in total.

Let $D = \{(x_1, y_1), \cdots\cdots, (x_N, y_N)\}$ denote the training data, with $N = 39368/2$, $x_i = \left(x_{i,1}, \cdots\cdots x_{i,p}\right)^T$representing the information relative to atom i in each structure described with p features, and y_i denoting the ESP charge. During the training process, for each decision tree in the forest, a bootstrap sample D_j from the training data of N molecules was drawn first. Starting with all observations $(x_1, y_1)\cdots\cdots(x_N, y_N)$, of D_j at each node, m predictors were selected at random from the p predictors (m<p), and the node was split into two descendant nodes using the best split among the remaining predictors. This process was repeated until no further splits ere possible to grow a tree, and the steps were repeated again until all the trees were grown.

Although Random Forests can obtain good results using the default parameters in most cases, appropriate parameters can further improve the accuracy for particular situations. There is only one parameter to which random forests is somewhat sensitive—m. This denotes the number of randomly selected predictor variables at each node. The default value of m is often set by $p/3$. In the RFR model, combined with symmetry functions, different values of m were tested, and, finally, $m = 5$ was determined by comparing the Pearson correlation coefficient (r) between the predicted charges and the ESP charges of Fe^{2+}. Another parameter, B, which represents the number of trees in the forest, can be chosen to be as large as desired; Breiman (2001) showed the generalization error for random forests converges almost surely to a limit as B increases. Here, B was set as 200.

When the training is completed, the prediction charge of a given atom i in a new geometry structure will be given by the average prediction of all individual trees. Thus, the predicted charge is assigned as Equation (1):

$$\bar{q_i} = \frac{\sum_{j=1}^{B} T_j(x_i)}{B} \tag{1}$$

The standard deviation of the predicted charge for atom i by the tree T is defined as Equation 2:

$$\sigma_i = \sqrt{\frac{\sum_j^B \left(q_i\left(T_j\right) - \bar{q}_i\right)^2}{B}} \tag{2}$$

where $q_i\left(T_j\right)$ is the partial charge predicted by tree T_j. The RFR algorithm was implemented using the scikit-learn module in Python.

2.3. Descriptor Selection

To encode the physical features and the mandatory symmetries of the problem, many descriptors have been introduced (Imbalzano et al., 2018). For example, Huan et al. (2017) utilized a d-dimensional vector $V_{i,\alpha}$, representing the atomic environment of atom i viewed along the Cartesian α direction (Huan et al., 2017). Heid et al. (2019) used the type of each atom and its connectivity as the input for the neural network. Schutt et al. (2017) introduced a vector of nuclear charges and a matrix of atomic distances to describe the molecular structures. In addition, molecules can be represented as Coulomb matrices (Rupp et al., 2012; Lilienfeld, 2015), scattering transforms (Hansen et al., 2015), bags of bonds (Bartók et al., 2010; Bartók et al., 2013), and so on. Among these various descriptors, atomic-based symmetric function, which was first proposed by Behler et al. (2007), has been widely used in machine learning (Behler et al., 2007; Behler, 2011a,b). Here, we adopted this method to describe the molecular structure.

Atom-based symmetric functions describe the chemical environment of atom i in terms of radial and angular terms. Therefore, each atom's Cartesian coordinate $R_i = \left(x_i, y_i, z_i\right)$ needs to be converted into the so-called symmetric function form of Equation (3):

$$\begin{aligned} R_i &= \{G_i^{angular}, G_i^{radial}\} \\ &= \Big\{G_{i,\,E_1}^{radial}, \cdots,\ G_{i,E_n}^{radial}, G_{i,E_1,E_1}^{angular}, \cdots, \\ &\quad G_{i,E_1,E_n}^{angular}, G_{i,E_2,E_2}^{angular} \cdots, G_{i,E_2,E_n}^{angular}, \cdots, G_{i,E_n,E_n}^{angular}\Big\} \end{aligned} \tag{3}$$

where $G_{i,\,E_1}^{radial}$ represents the total contribution of the distance between all the surrounding atoms, and atom i, and $G_{i,E_i,E_j}^{angular}$ represents the angular relationship between any two surrounding atoms and itself. All atoms are distinguished according to their element E_i, and the set of symmetric functions of two atoms belonging to the same element are thus the same.

In this study, Equation (4) was used to describe the distance component of each atom, where R_{ij} represents the distance between atom i and j. The cutoff function $f_c\left(R_{ij}\right)$ was introduced in Equation 5 because the atoms in the molecular dynamic simulation may enter or leave the cutoff distance, which can lead to the number of neighbor atoms to be variable. Here, R_c was thus set to 99 Å to include all the atoms, and R_s and η were both set to 1.0.

$$G_{i,J}^{radial} = \sum_{j \neq i}^{j\ in\ J} e^{-\eta\left(R_{ij}-R_s\right)^2} f_c\left(R_{ij}\right) \tag{4}$$

$$f_c\left(R_{ij}\right) = \begin{cases} 0.5 \times \left[cos\left(\frac{\pi R_{ij}}{R_c}\right)+1\right] & for\ R_{ij} \leq R_c \\ 0 & for\ R_{ij} \geq R_c \end{cases} \tag{5}$$

Equation (6) is the angular component, which defines the angular distribution centered on each reference atom; here, $\lambda = 1.0$, $\zeta = 1.0$.

$$\begin{aligned} G_{i,j,k}^{angular} &= 2^{1-\zeta} \sum_{j,k\neq i}^{j\in J\&k\in K} \left(1+\lambda coa\theta_{ijk}\right)^{\zeta} \\ &\quad \times\ e^{-\eta\left(R_{ij}^2+R_{ik}^2+R_{jk}^2\right)} f_c\left(R_{ij}\right) f_c\left(R_{ik}\right) f_c\left(R_{jk}\right) \end{aligned} \tag{6}$$

Therefore, through coordinate transformation, the symmetric functions for each atom can be obtained and combined with the ESP charge to finally form the training set as the input of model.

Meanwhile, in order to compare the effect of descriptor selection on prediction performance, 11 structural parameters were manually selected and used as descriptors to train the model. Specifically, the 11 parameters included eight distance values (Fe-N1, Fe-N2, Fe-N3, Fe-N4, Fe-N11, Fe-O12, Fe-O13, and O12-O13), one angle value (Fe-O12-O13), and two dihedral angles (N2-Fe-N1-C10 and N1-Fe-N2-C5). According to our chemical perception, these 11 parameters reflect the features of molecule structure, so they can well-describe different conformations.

3. RESULTS AND DISCUSSION

3.1. Charge Distribution of Multi-Spin State in the Initial Data Set

It can be seen in **Figure 3** that most variations range from 0.5 to 0.7e; the fluctuation of Fe^{2+} was the most significant, as it was close to 2e. The variation of O12 was larger than that of O13. It can also be found that there was a slight tendency for the mean value of N to decrease and the mean value of C to increase. For Fe and the coordinating O12 and O13, the difference among the mean values under different spin states was relatively more significant. Specifically, the atomic charge of Fe^{2+} in the singlet state was distributed around 1.2 and 1.5e in quintuplet. Further analysis of the charge distribution of different spin states showed that the triplet charge of Fe^{2+} in most structures was greater than the singlet charge ($\Delta 31 > 0$, see **Figure 4**), with the difference being at the highest probability concentrated at 0.1e, while, for the quintuplet and triplet spin state, the difference reached 0.2e. The results confirmed that different spin states in the same structure had distinct charge distributions.

Additionally, it should be noted that the atomic charge of each atom fluctuated within a certain range, among which Fe^{2+} fluctuated the most. Just taking the singlet state as an example, the variation ranged from 0.3 to 2.2e, which implied that the charge distribution of a certain atom in a specific spin state was conformation dependent.

3.2. Charge Prediction of RFR Model With Symmetric Functions

In order to better distinguish between different molecular structures, the atom-based symmetry functions were used to convert atomic coordinates into a series of function values, which embed the atoms in their neighborhood depending on

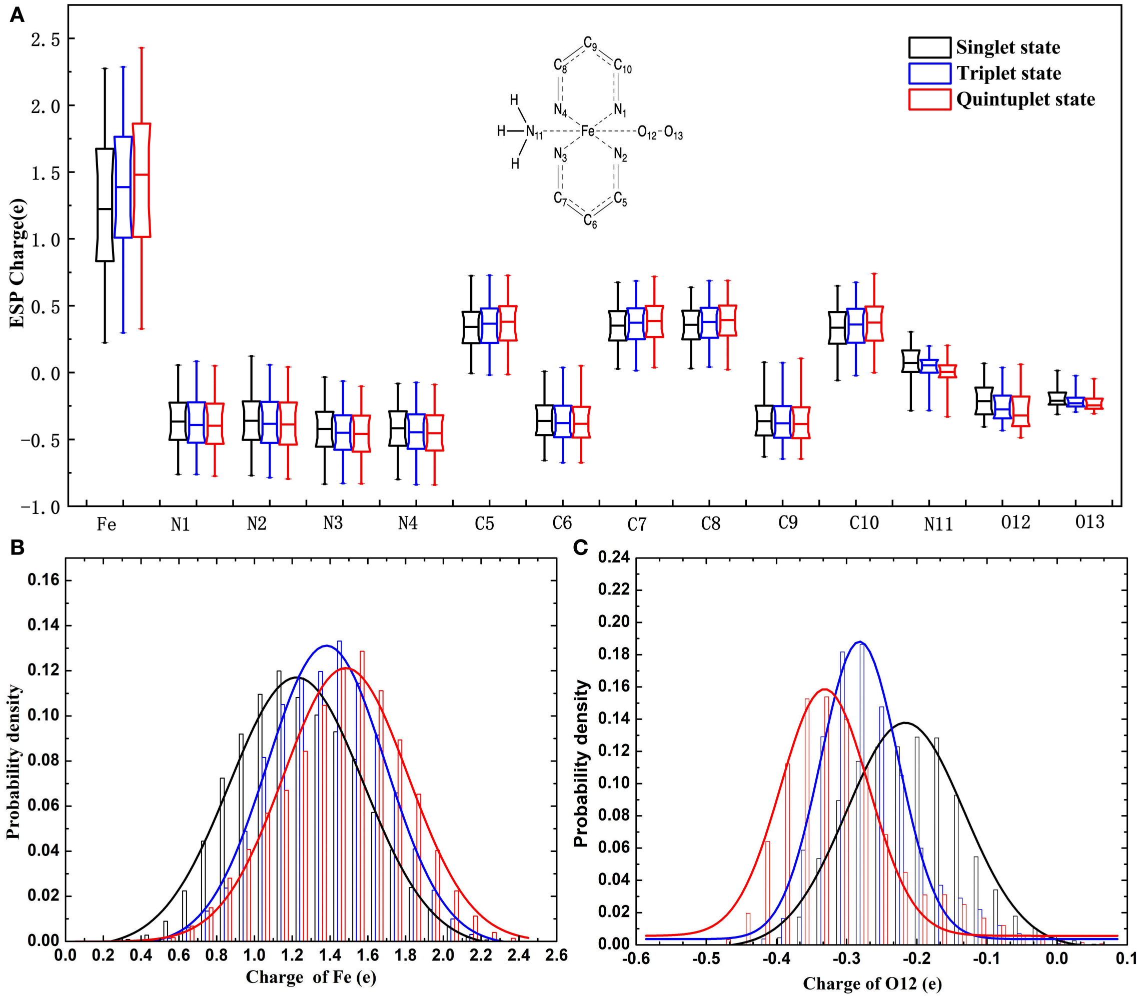

FIGURE 3 | Charge distribution of multi-spin state for 14 atoms of the model in the data set. **(A)** Is boxplot representation of charge distribution of 14 atoms in the heme system in different spin states. **(B)** Is probability density of charge distribution of Fe^{2+}. **(C)** Is the probability density of the charge distribution of the O12 atom.

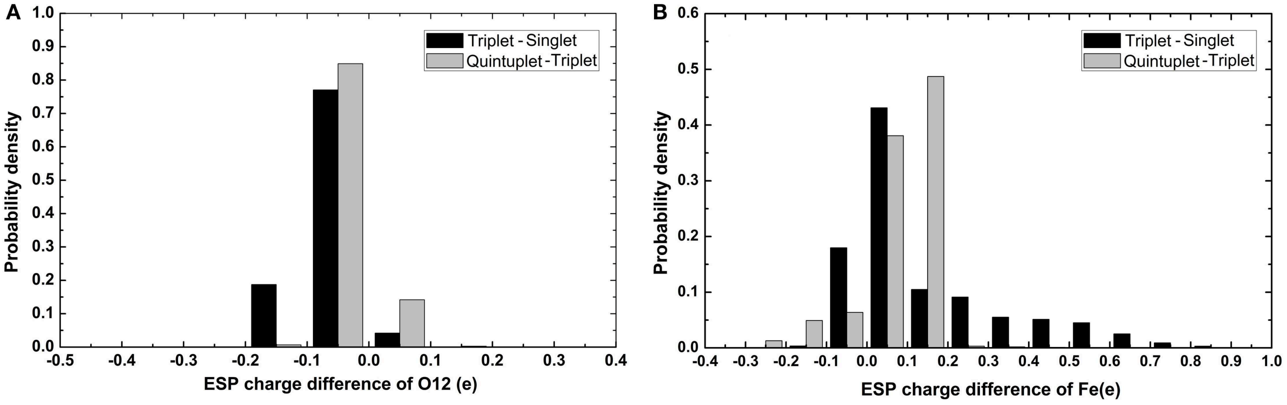

FIGURE 4 | Histogram of charge differences between two spin states of selected atoms. **(A)** Is O12 atom and **(B)** is Fe^{2+}. The black color code is the charge difference of singlet and triplet and gray color code is charge difference of quintuplet and triplet state.

the element type (Schutt et al., 2017). It is an efficient way to consider the chemical environments that the invariances, such as translation, rotation, and permutation, can be guaranteed to be exploited by. By doing so, the RFR model combined with symmetry functions and ESP charge was constructed.

As mentioned above, although complex parameter tuning is not required in the RFR model, it is sensitive to the number of descriptors. To this end, we tested and compared the predicted charge of Fe^{2+} at different values of m (i.e., the number of features selected from p descriptors at random; here $p = 19$) and then calculated the correlation between the predictions and the ESP charges. The results are shown in **Table 1**. It can be seen from **Table 1** that, when $m = 5$, the correlation between the predicted value and the fitted value is the largest (0.9784), which indicated that prediction gave the best performance. The parameter m was consequently set to five in the subsequent analysis.

To assess the prediction performance of the charge models, the mean absolute error (MAE) was calculated for each atom in the three spin states, and the standard deviation of the error was given as well (**Table 2**). According to **Table 2**, the MAEs of the predicted charges in three spin states are all within 0.015e for all the spin states. There was no obvious difference between two states. For each state, most of the MAEs of the atoms were within 0.02e as well, except for Fe^{2+}, which reached a maximum of 0.047e. Moreover, the Pearson correlation coefficient (r) between the predicted charges and the ESP charges of the RFR model in all three states was above 0.96. These data demonstrated that the model had high prediction accuracy, especially for N1 and N2. At the same time, the MAE and error standard deviation were close in the three states, indicating that our RFR model had good stability.

TABLE 1 | Tests on the number of features selected in the RFR model.

m	Pearson correlation coefficient
5	0.9784
0.2	0.9750
log2	0.9771
$\sqrt{p}$	0.9771
19	0.8729

When using the symmetric function method, each molecular structure is described by 19 features (p = 19), and m is the max features randomly selected from it to fit a tree.

For clarity, we further selected three atoms—Fe^{2+}, N11, and O12—to plot their charge distributions for comparison (**Figure 5**). As shown in **Figure 5**, the predicted charges of the RFR model are basically gathered around the straight line $y = x$; they were very close to the high-precision charges calculated by DFT, indicating that our model achieved satisfactory accuracy.

By carefully comparing the distribution of each atom in different spin states, it can be found that the predicted values of Fe^{2+} have a good aggregation and few scattered points. However, the predictions are larger when the corresponding ESP charges are <1.5e and smaller when they are above 1.5e. The aggregation centers in the three states were different but essentially distributed in 0.5–2.25e, which is consistent with the analysis in **Figure 5**. For N11 atom, the predictions were more concentrated in the singlet and triplet, as there existed a few scattered points in the quintuplet. For O12 atoms, the correlation coefficients in all three spin states exceeded 0.98, and although there were some scattered points, the distribution was uniform. Finally, by comparing the distribution of different atoms in the same spin state, it can be found that the predictions in the triplet state were more concentrated overall. In summary, it was demonstrated that our model can predict the atomic charge of most structures well.

TABLE 2 | The performance of prediction using RFR model with symmetric functions for three spin states.

Atoms	Predicted values (e)			MAE			Error std.			Pearson coefficient		
	Singlet	Triplet	Quintuplet	Singlet	Triplet	Quintuplet	Singlet	Triplet	Quintuplet	Singlet	Triplet	Quintuplet
Fe^{2+}	1.390	1.382	1.459	0.048	0.046	0.047	0.051	0.049	0.050	0.978	0.980	0.982
N1	−0.390	−0.382	−0.390	0.013	0.014	0.013	0.014	0.014	0.014	0.991	0.991	0.991
N2	−0.387	−0.378	−0.385	0.014	0.014	0.014	0.014	0.014	0.014	0.991	0.990	0.991
N3	−0.457	−0.449	−0.458	0.013	0.013	0.012	0.014	0.014	0.014	0.988	0.988	0.989
N4	−0.451	−0.443	−0.452	0.014	0.014	0.014	0.015	0.015	0.015	0.986	0.986	0.986
C5	0.350	0.357	0.374	0.015	0.015	0.015	0.016	0.016	0.016	0.985	0.984	0.984
C6	−0.369	−0.371	−0.377	0.018	0.018	0.018	0.019	0.019	0.019	0.973	0.972	0.972
C7	0.363	0.368	0.384	0.016	0.016	0.016	0.017	0.018	0.017	0.978	0.977	0.977
C8	0.371	0.375	0.390	0.015	0.016	0.015	0.016	0.017	0.016	0.978	0.977	0.978
C9	−0.372	−0.373	−0.378	0.018	0.018	0.019	0.019	0.019	0.020	0.971	0.970	0.967
C10	0.348	0.354	0.370	0.016	0.016	0.015	0.017	0.017	0.016	0.983	0.983	0.984
N11	0.052	0.051	0.007	0.004	0.005	0.006	0.005	0.005	0.008	0.988	0.985	0.971
O12	−0.217	−0.266	−0.306	0.005	0.007	0.008	0.006	0.009	0.012	0.990	0.987	0.987
O13	−0.229	−0.224	−0.238	0.002	0.003	0.004	0.004	0.006	0.006	0.987	0.970	0.978
Mean				0.015	0.015	0.015	0.016	0.017	0.017	0.983	0.981	0.981

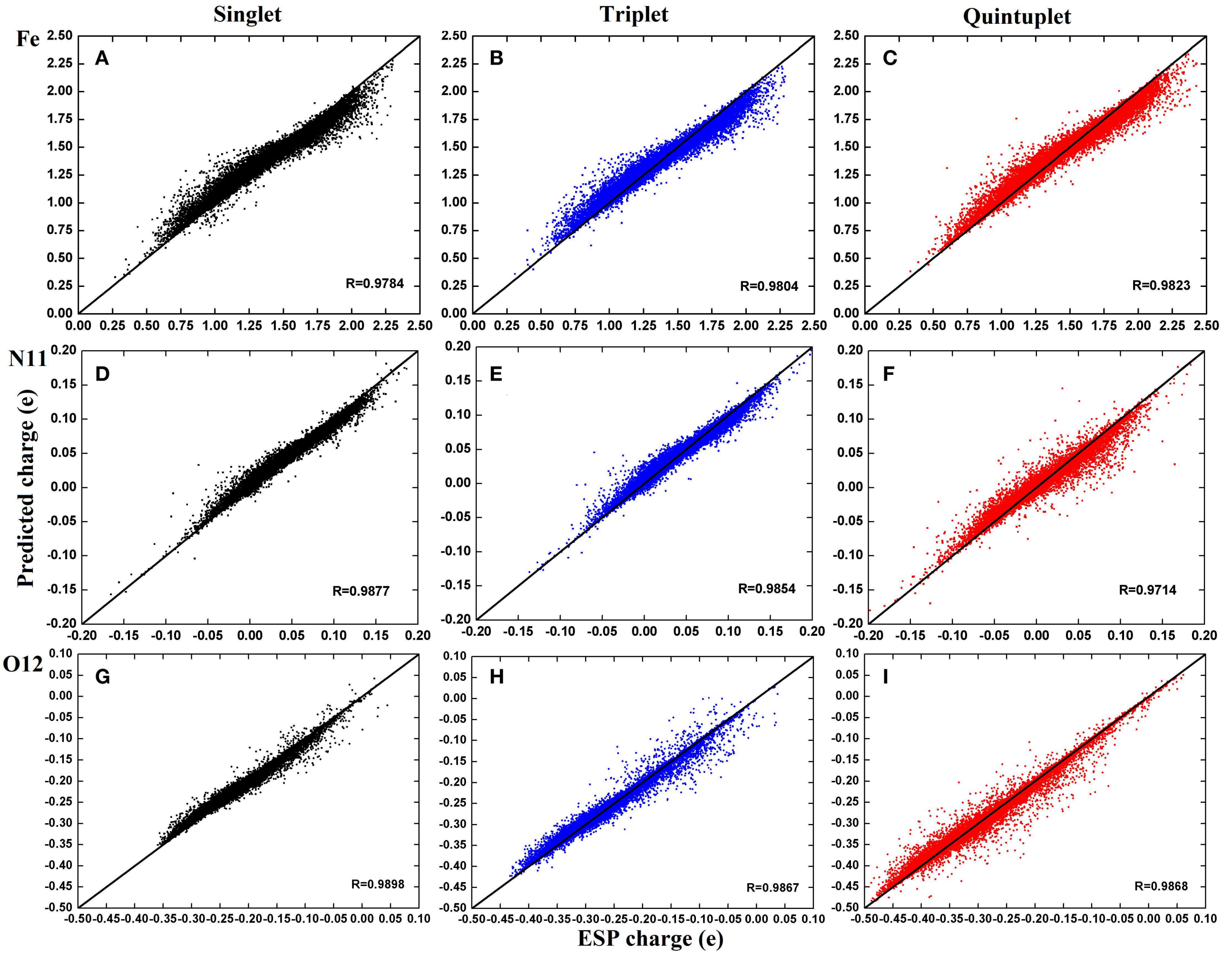

FIGURE 5 | Distribution of predicted charge of atoms Fe^{2+}, N11, and O12 in three spin states. The illustration use **(A–C)** for Fe^{2+}; **(D–F)** for N11; **(G–I)**. The color codes are Black for singlet state, Blue for triplet state, and Red for quintuplet state.

3.3. Charge Prediction of RFR Model With Manually Selected Structural Parameters

To compare the performance of the RFR models with different descriptors, we manually screened 11 parameters to describe the molecular structure, including eight bond lengths (Fe-N1,Fe-N2, Fe-N3, Fe-N4, Fe-N11, Fe-O12, Fe-O13, and O12-O13), one bond angle (Fe-O12-O13), and two dihedrals angles (N2-Fe-N1-C10 and N1-Fe-N2-C5). The same process and method were used for RFR model training and prediction. A comparison of the prediction performance of selected atoms (Fe^{2+}, N11, and O12) is shown in **Figure 6**, and, for comparison, the root mean square error RMSE of the prediction of each atom in different spin states was further calculated and is shown in **Figure 7**.

It can be seen clearly from **Figure 6** that both models have good prediction performances, and the same model has a similar RMSE of predictions for different spin states. When 11 structural parameters were used as descriptors, however, the prediction values were more concentrated, and the model prediction performance was better than with in the case of symmetry functions. The average RMSE and the RMSE of each atom were reduced. Among these, the RMSE of Fe^{2+} reduced from 0.07 to 0.0035e, which is a maximum 0.06e improvement. We think that this is partially due to the use of a dihedral angle as the descriptor, which is a four-body term and is not included in the symmetry function.

In conclusion, choosing different descriptors will affect the prediction performance of the RFR model; the 11 manually selected parameters can better describe the molecular structure and thus achieved better results. At the same time, however, it should be noted that the difference between the two cases is not significant. As shown in **Figure 7**, the RMSE of Fe^{2+} is relatively larger, but its fluctuations are still below 0.04e, and the variations of RMSE for other atoms are all below 0.02e. This indicates that, even if there is no empirical experience involved, the RFR model with symmetry functions can achieve satisfactory predictions, and the advantage is that it can be automatized.

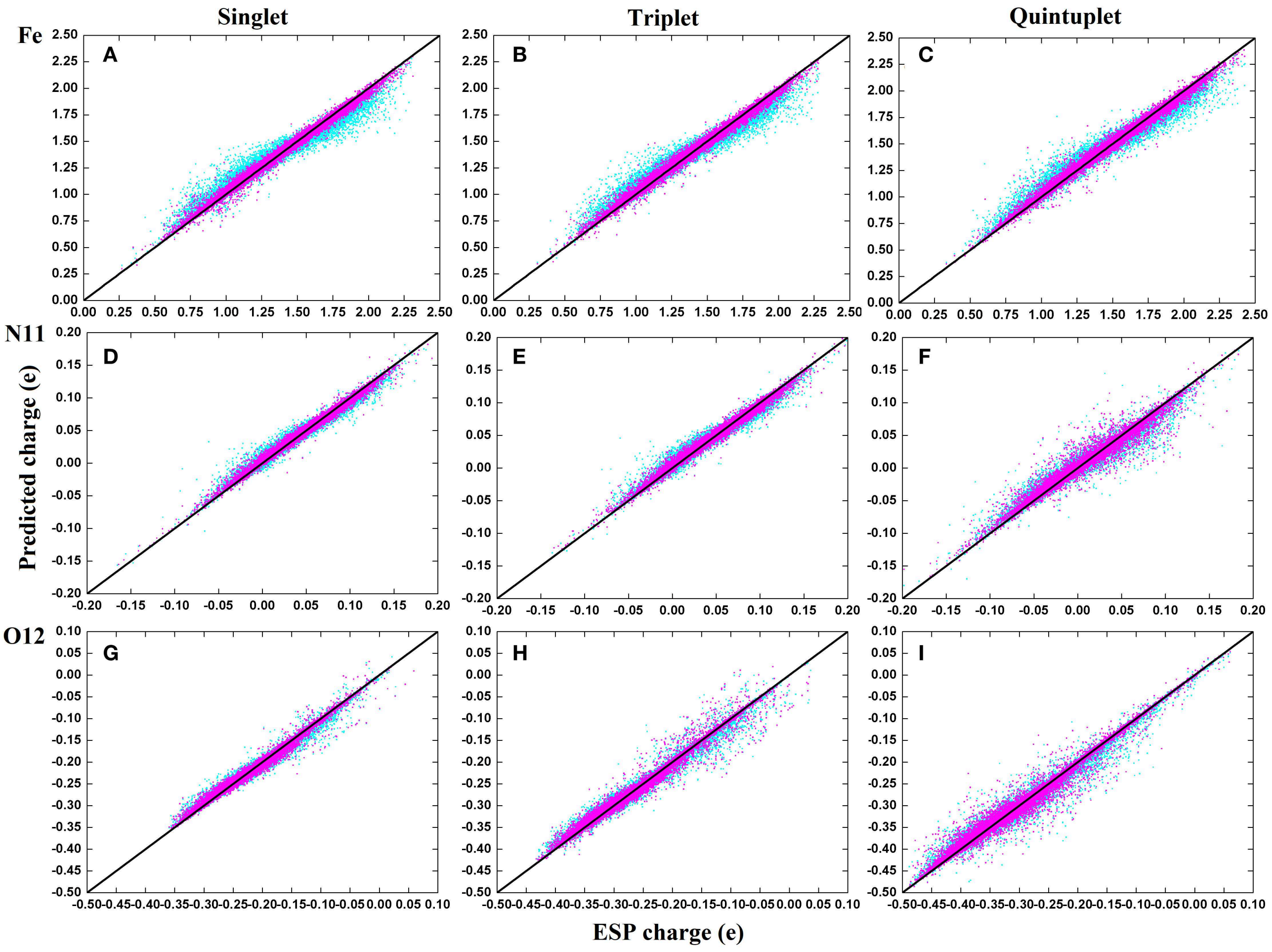

FIGURE 6 | Comparison on the performance of two descriptors of RFR predictions in three spins states. The illustration use **(A–C)** for Fe^{2+}; **(D–F)** for N11; **(G–I)**. Cyan corresponds to the RFR with a symmetric function, and magenta represents the RFR with 11 structural parameters.

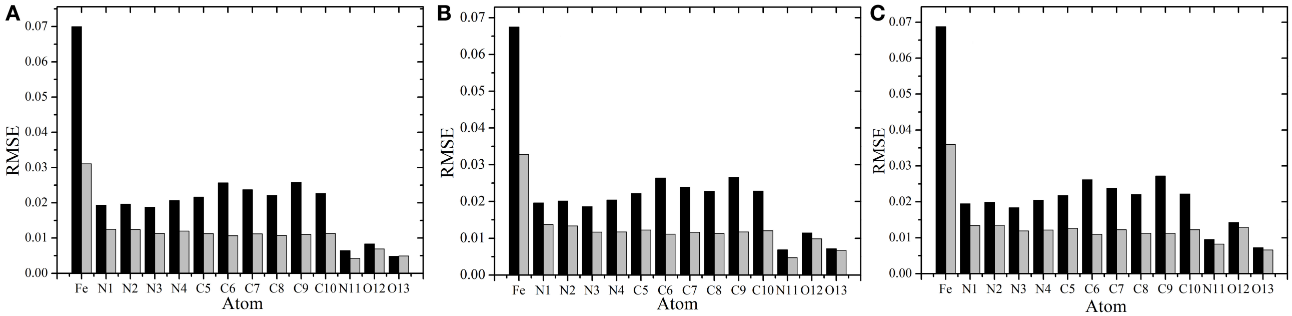

FIGURE 7 | Comparison of the root mean square error(RMSE) using different RFR models. **(A–C)** Represent the RMSE of each atom in singlet, triplet, and quintuplet state, respectively. The black bar is the RFR model using symmetric functions, and the gray bar is the RFR model with 11 manual selected structural parameters.

4. CONCLUSIONS

This study aimed at exploring the spin crossover phenomenon in the model heme system according to the characteristics of atomic charge distribution in different spin states with conformation. The random forest method was introduced to construct a prediction model of multi-spin variable charge, which can provide a separate prediction for a single atom.

In this model, symmetry functions were used as descriptors to describe the atomic chemical environment. The model was trained in conjunction with the ESP charges to predict the atomic charge in different spin states. Meanwhile, in order to compare the prediction performance, 11 artificially selected structural parameters were also used as the input of RFR model. The results showed that, when the 11 selected parameters were adopted, the prediction was more accurate, but it was not suitable for automation considering the involvement of human experience. In contrast, the RFR model using symmetry functions can achieve a good trade-off between calculation accuracy and efficiency, realize automatic processing, and provide separate prediction for a single atom. It should be noted that, in this method, the transformation of coordinates is a time-consuming pre-processing process, but it avoids the problem of inconsistent calculation of energy or force in Cartesian coordinates. When the number of descriptors is large enough, the random forest algorithm is very effective. This study is only a preliminary exploration of the heme force field, and there are still many deficiencies. In future work, we will further improve the calculation method of the multi-spin state variable charge force field parameters.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

JG: conceptualization. QL, WZ, X-HH, and L-HB: methodology. WZ and L-HB: validation. WZ, L-HB, and JG: writing–original draft, writing–review, and editing. JG: project administration and funding acquisition.

FUNDING

This work was supported by the National Key R&D Program of China (Grant No. 2017YFB0203405), the National Natural Science Foundation of China (No. 21873034), and the Fundamental Research Funds for the Central Universities (Project 2662018JC027).

REFERENCES

Amit, Y., and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Comput*. 9, 1545–1588. doi: 10.1162/neco.1997.9.7.1545

Bartók, A. P., Kondor, R., and Csányi, G. (2013). Publisher's note: on representing chemical environments [phys. Rev. B 87, 184115 (2013)]. *Phys. Rev. B* 87:219902. doi: 10.1103/PhysRevB.87.219902

Bartók, A. P., Payne, M. C., Kondor, R., and Csányi, G. (2010). Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett*. 104:136403. doi: 10.1103/PhysRevLett.104.136403

Behler, J. (2011a). Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys*. 134:074106. doi: 10.1063/1.3553717

Behler, J. (2011b). Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys*. 13, 17930–17955. doi: 10.1039/c1cp21668f

Behler, J., Lorenz, S., and Reuter, K. (2007). Representing molecule-surface interactions with symmetry-adapted neural networks. *J. Chem. Phys*. 127:014705. doi: 10.1063/1.2746232

Bleiziffer, P., Schaller, K., and Riniker, S. (2018). Machine learning of partial charges derived from high-quality quantum-mechanical calculations. *J. Chem. Inf. Model*. 58, 579–590. doi: 10.1021/acs.jcim.7b00663

Botu, V., Batra, R., Chapman, J., and Ramprasad, R. (2016). Machine learning force fields: construction, validation, and outlook. *J. Phys. Chem. C* 121, 511–522. doi: 10.1021/acs.jpcc.6b10908

Bousseksou, A., Molnar, G., Salmon, L., and Nicolazzi, W. (2011). Molecular spin crossover phenomenon: recent achievements and prospects. *Chem. Soc. Rev*. 40, 3313–3335. doi: 10.1039/c1cs15042a

Breiman, L. (2001). Random forests. *Mach. Learn*. 45, 5–32. doi: 10.1023/A:1010933404324

Bristow, J. K., Tiana, D., and Walsh, A. (2014). Transferable force field for metal-organic frameworks from first-principles: BTW-FF. *J. Chem. Theory Comput*. 10, 4644–4652. doi: 10.1021/ct500515h

Cambi, L., and Szegö, L. (1931). Über die magnetische susceptibilität der komplexen verbindungen. *Ber. Deutsch. Chem. Gesellsch*. 64, 2591–2598. doi: 10.1002/cber.19310641002

Chen, W.-K., Liu, X.-Y., Fang, W.-H., Dral, P. O., and Cui, G. (2018). Deep learning for nonadiabatic excited-state dynamics. *J. Phys. Chem. Lett*. 9, 6702–6708. doi: 10.1021/acs.jpclett.8b03026

Cong, Y., Li, Y., Jin, K., Zhong, S., Zhang, J. Z. H., Li, H., et al. (2018). Exploring the reasons for decrease in binding affinity of hiv-2 against hiv-1 protease complex using interaction entropy under polarized force field. *Front. Chem*. 6:380. doi: 10.3389/fchem.2018.00380

Cutler, A., Cutler, D., and Stevens, J. (2011). Random forests. *Mach. Learn*. 45, 157–176. doi: 10.1007/978-1-4419-9326-7_5

Cutler, D. R., Edwards, T. C. Jr., Beard, K. H., Culter, A., Hess, K. T., Gibson, J., et al. (2007). Random forests for classification in ecology. *Ecology* 88, 2783–2792. doi: 10.1890/07-0539.1

De, S., Chamoreau, L. M., El Said, H., Li, Y. L., Flambard, A., Boillot, M. L., et al. (2018). Thermally-induced spin crossover and liesst effect in the neutral [FeII(Me$_{\mathrm{bik}}$)$_2$(NCX)$_2$] complexes: variable-temperature structural, magnetic, and optical studies (X = S, Se; Me$_{\mathrm{bik}}$ = bis(1-methylimidazol-2-yl)ketone). *Front. Chem*. 6:15. doi: 10.3389/fchem.2018.00326

Doukov, T., Li, H., Sharma, A., Martell, J. D., Soltis, S. M., Silverman, R. B., et al. (2011). Temperature-dependent spin crossover in neuronal nitric oxide synthase bound with the heme-coordinating thioether inhibitors. *J. Am. Chem. Soc*. 133, 8326–8334. doi: 10.1021/ja201466v

Du, L., Liu, F., Li, Y., Yang, Z., Zhang, Q., Zhu, C., et al. (2018). Dioxygen activation by iron complexes: the catalytic role of intersystem crossing dynamics for a heme-related model. *J. Phys. Chem. C* 122, 2821–2831. doi: 10.1021/acs.jpcc.7b11462

Engler, M. S., Caron, B., Veen, L., Geerke, D. P., Mark, A. E., and Klau, G. W. (2019). Automated partial atomic charge assignment for drug-like molecules: a fast knapsack approach. *Algorithms Mol. Biol*. 14, 1–10. doi: 10.1186/s13015-019-0138-7

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al. (2016). *Gaussian 16 Revision A.03*. Wallingford, CT: Gaussian Inc.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recogn. Lett*. 31, 2225–2236. doi: 10.1016/j.patrec.2010.03.014

Gutlich, P., Gaspar, A. B., and Garcia, Y. (2013). Spin state switching in iron coordination compounds. *Beilstein J. Org. Chem*. 9, 342–391. doi: 10.3762/bjoc.9.39

Gütlich, P., and Goodwin, H. A. (2004). *Spin Crossover in Transition Metal Compounds III*. Berlin; Heidelberg: Springer Berlin Heidelberg.

Habenicht, B. F., and Prezhdo, O. V. (2012). *Ab initio* time-domain study of the triplet state in a semiconducting carbon nanotube: intersystem crossing, phosphorescence time, and line width. *J. Am. Chem. Soc*. 134, 15648–15651. doi: 10.1021/ja305685v

Hagai, E., Rustam, Z. K., Thomas, D. K., Jorg, B., and Michele, P. (2010). *Ab initio* quality neural-network potential for sodium. *Phys. Rev. B* 81:184107. doi: 10.1103/PhysRevB.81.184107

Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., von Lilienfeld, O. A., Müller, K.-R., et al. (2015). Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett*. 6, 2326–2331. doi: 10.1021/acs.jpclett.5b00831

Hauser, A. (2013). Spin-crossover materials. properties and applications. *Angew. Chem. Int. Ed*. 52:10419. doi: 10.1002/anie.201306160

Heid, E., Fleck, M., Chatterjee, P., Schroder, C., and MacKerell, A. D., J. (2019). Toward prediction of electrostatic parameters for force fields that explicitly treat electronic polarization. *J. Chem. Theory Comput*. 15, 2460–2469. doi: 10.1021/acs.jctc.8b01289

Hu, W., Ye, S., Zhang, Y., Li, T., Zhang, G., Luo, Y., et al. (2019). Machine learning protocol for surface-enhanced raman spectroscopy. *J. Phys. Chem. Lett*. 10, 6026–6031. doi: 10.1021/acs.jpclett.9b02517

Huan, T. D., Batra, R., Chapman, J., Krishnan, S., Chen, L., and Ramprasad, R. (2017). A universal strategy for the creation of machine learning-based atomistic force fields. *NPJ Comput. Mater*. 3:37. doi: 10.1038/s41524-017-0042-y

Imbalzano, G., Anelli, A., and Giofré, D. (2018). Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys*. 148:241730. doi: 10.1063/1.5024611

Inokuchi, T., Li, N., Morohoshi, K., and Arai, N. (2018). Multiscale prediction of functional self-assembled materials using machine learning: high-performance surfactant molecules. *Nanoscale* 10, 16013–16021. doi: 10.1039/C8NR03332C

Ivanov, M. V., Talipov, M. R., and Timerghazin, Q. K. (2015). Genetic algorithm optimization of point charges in force field development: challenges and insights. *J. Phys. Chem. A* 119, 1422–1434. doi: 10.1021/acs.jpca.5b00218

Jureschi, C. M., Rusu, I., Codjovi, E., Linares, J., Garcia, Y., and Rotaru, A. (2014). Thermo- and piezochromic properties of [fe(hyptrz)]a2·h2o spin crossover 1d coordination polymer: towards spin crossover based temperature and pressure sensors. *Phys. B Phys. Condensed Matter* 449, 47–51. doi: 10.1016/j.physb.2014.04.081

Klusowski, J. M. (2018). Sharp analysis of a simple model for random forests. *arXiv. [Preprint]*. arXiv:1805.02587.

Lilienfeld, R. R. A. V. (2015). Machine learning, quantum mechanics, and chemical compound space. *Phys. Chem. Chem. Phys*. 15, 501–509. doi: 10.1002/9781119356059.ch5

Liu, F., Du, L., Zhang, D., and Gao, J. (2017). Direct learning hidden excited state interaction patterns from *ab initio* dynamics and its implication as alternative molecular mechanism models. *Sci. Rep*. 7:8737. doi: 10.1038/s41598-017-09347-2

Meyer, R., Mücksch, C., Wolny, J. A., Schünemann, V., and Urbassek, H. M. (2019). Atomistic simulations of spin-switch dynamics in multinuclear chain-like triazole spin-crossover molecules. *Chem. Phys. Lett*. 733:136666. doi: 10.1016/j.cplett.2019.136666

Nagl, J., Gerald Auböck, G., Hauser, A. W., Allard, O., Callegari, C., and Ernst, W. E. (2008). High-spin alkali trimers on helium nanodroplets: spectral separation and analysis. *J. Chem. Phys*. 128:154320. doi: 10.1063/1.2906120

Rai, B. K., and Bakken, G. A. (2013). Fast and accurate generation of *ab initio* quality atomic charges using nonparametric statistical regression. *J. Comput. Chem*. 34, 1661–1671. doi: 10.1002/jcc.23308

Reiher, M., Oliver, S., and Bernd Artur, H. (2001). Reparameterization of hybrid functionals based on energy differences of states of different multiplicity. *Theor. Chem. Acc*. 107, 48–55. doi: 10.1007/s00214-001-0300-3

Riniker, S. (2018). Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: an overview. *J. Chem. Inform. Model*. 58, 565–578. doi: 10.1021/acs.jcim.8b00042

Roman, Z., Justin, S. S., and Leszczynski, J., Isayev, O. (2019). Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv*. 5:eaav6490. doi: 10.1126/sciadv.aav6490

Rupp, M., Tkatchenko, A., Müller, K.-R., and von Lilienfeld, O. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett*. 108, 58301–58300. doi: 10.1103/PhysRevLett.108.058301

Sahoo, S. K., and Nair, N. N. (2018). Interfacing the core-shell or the drude polarizable force field with car-parrinello molecular dynamics for qm/mm simulations. *Front. Chem*. 6:275. doi: 10.3389/fchem.2018.00275

Salomon, O., Reiher, M., and Hess, B. A. (2002). Assertion and validation of the performance of the b3lyp* functional for the first transition metal row and the g2 test set. *J. Chem. Phys*. 117:4729. doi: 10.1063/1.1493179

Sanvito, A. L. S. (2019). A unified picture of the covalent bond within quantum-accurate force fields: from organic molecules to metallic complexes' reactivity. *Sci. Adv*. 5:eaaw2210. doi: 10.1126/sciadv.aaw2210

Schutt, K. T., Arbabzadah, F., Chmiela, S., Muller, K. R., and Tkatchenko, A. (2017). Quantum-chemical insights from deep tensor neural networks. *Nat. Commun*. 8:13890. doi: 10.1038/ncomms13890

Shao, X. D., Zhang, X., Shi, C., Yao, Y.-F., and Zhang, W. (2015). Switching dielectric constant near room temperature in a molecular crystal. *Adv. Sci*. 2:1500029. doi: 10.1002/advs.201500029

Statnikov, A., Wang, L., and Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics* 9:319. doi: 10.1186/1471-2105-9-319

Svetnik, V. (2003). Random forest a classification and regression tool for compound classification and qsar modeling. *J. Chem. Inf. Comput. Sci*. 2003, 1947–1958. doi: 10.1021/ci034160g

Unke, O. T., and Meuwly, M. (2019). Physnet: a neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput*. 15, 3678–3693. doi: 10.1021/acs.jctc.9b00181

Wang, X., and Gao, J. (2020). Atomic partial charge predictions for furanoses by random forest regression with atom type symmetry function. *RSC Adv*. 10, 666–673. doi: 10.1039/C9RA09337K

Xu, T., Wang, W., and Yin, S. (2018). Electrostatic polarization energies of charge carriers in organic molecular crystals: a comparative study with explicit state-specific atomic polarizability based amoeba force field and implicit solvent method. *J. Chem. Theory Comput*. 14, 3728–3739. doi: 10.1021/acs.jctc.8b00132

Ye, S., Hu, W., Li, X., Zhang, J., Zhong, K., Zhang, G., et al. (2019). A neural network protocol for electronic excitations of n-methylacetamide. *Proc. Natl. Acad. Sci. U.S.A*. 116, 11612–11617. doi: 10.1073/pnas.1821044116

Yuan, S., Feng, L., Wang, K., Pang, J., Bosch, M., Lollar, C., et al. (2018). Stable metal-organic frameworks: design, synthesis, and applications. *Adv. Mater*. 30:e1704303. doi: 10.1002/adma.201704303

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.