ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

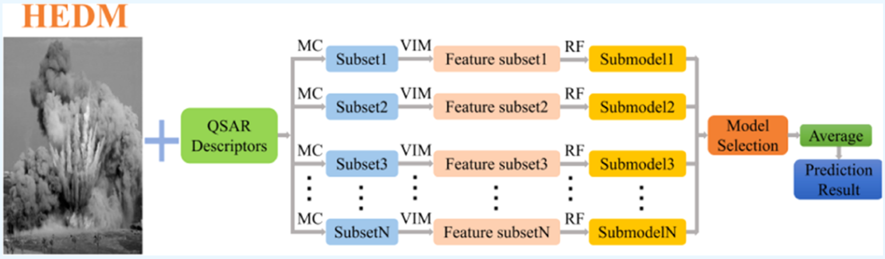
Novel Random Forest Ensemble Modeling Strategy Combined with Quantitative Structure−Property Relationship for Density Prediction of Energetic Materials

Maogang Li, Weipeng Lai, Ruirui Li, Jiajun Zhou, Yingzhe Liu, Tao Yu, Tianlong Zhang,* Hongsheng Tang, and Hua Li*

Cite This: *ACS Omega* 2023, 8, 2752−2759

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



ABSTRACT: With the further development of the concept of green chemistry, the new generation of energetic materials tends to exhibit detonation properties such as higher insensitivity, higher density, and higher energy. Therefore, the precise molecular design and green and efficient synthesis of energetic materials will be one of the serious challenges. For the purpose of accurate prediction of detonation performance of energetic materials, an ensemble modeling strategy based on the combination of Monte Carlo (MC) and variable importance measurement (VIM) improved random forest (RF) and quantitative structure−property relationship (QSPR) is proposed, which was successfully used for density prediction of energetic materials. First, the structure of 162 energetic compounds was optimized by Gaussian software, and the molecular descriptor data were calculated by CODESSA software based on the optimized molecular structure. Then, the MCVIMRF_Med ensemble model was constructed on the basis of the above molecular descriptor data and the corresponding energetic compound density index. The joint *X*−*Y* distance algorithm (SPXY) is used to partition the data set. And then, MC is used to further divide the calibration set data into multiple subsets for the construction of the ensemble model. The subset size and the number of iterations of the MCVIMRF_Med ensemble model were optimized through MC cross validation. The final output strategy of the ensemble model is optimized based on the optimized parameters, and an output optimization method based on median screening is proposed and successfully applied for the prediction performance optimization of the MCVIMRF_Med ensemble model. To further investigate the performance of the MCVIMRF_Med ensemble model, the performance of it was compared with partial least squares, RF, VIMRF, and MCVIMRF calibration models. It shows that the MCVIMRF_Med ensemble model can achieve a better prediction result for the density of energetic materials, with R^2_{CV} of 0.9596, RMSECV of 0.0437 g/cm^3, R^2_{P} of 0.9768, RMSEP of 0.0578 g/cm^3, and relative analysis deviation of prediction set of 3.951. Therefore, the MCVIMRF_Med ensemble modeling strategy combined with QSPR is an effective approach for the density prediction of energetic materials. This work is expected to provide new research ideas and technical support for accurate prediction of detonation performance of energetic materials.

1. INTRODUCTION

High energy density materials (HEDM),[1,2] also known as energetic materials, is a general term for a class of nitrogen (fluorine) compounds with higher insensitivity, higher density, and higher energy. Energetic materials can be divided into propellant, explosive, pyrotechnic agent, initiating explosive, and so forth. The upgrading of HEDM can remarkably improve the strike performance of weaponry and the effectiveness of space propulsion system and has become a significant indicator of the national core military level and military technology commanding heights.[3] In recent years, with the continuous promotion of the concept of green chemistry, the concepts of resource utilization optimization, ecological environment friendliness, and sustainable economic and social development have also been paid more attention in

Received: November 20, 2022
Accepted: December 23, 2022
Published: January 4, 2023

the research field of energetic materials, and thus, the requirements for detonation performance of new HEDM will be further improved.[4,5] Therefore, how to achieve precise molecular design, synthesis, assembly, and detection is one of the core issues in the development of new generation energetic materials.

Traditional molecular design methods usually require synthetic chemists to have a deeper understanding of the research materials and can be completed through a large number of structural design, synthesis route optimization, synthesis process condition optimization, product separation method research, product detection and analysis, and other processes, which is a very time-consuming and labor-intensive work. In addition, energetic compounds usually have high energy, so there will be greater experimental risk in their synthesis. Therefore, a more reasonable method for molecular design of energetic materials is urgently needed. In recent years, artificial intelligence has been used in the chemical field for related research, such as chemical literature retrieval and learning, intelligent laboratory robots, chemical process optimization, and so forth. Mark Waller's team successfully planned a new chemical synthesis route using a deep neural network and Monte Carlo (MC) tree algorithm.[6] Burger et al. designed and developed a more intelligent robot chemist to achieve automation of researchers rather than instrument automation.[7] Quantitative structure−activity/property relationship (QSAR)/(QSPR)[8,9] simulates the physical, chemical, biological, and other characteristics of organic molecules with the help of physical and chemical properties, structural parameters and other indicators, combined with mathematical, statistical, and other methods, which has been utilized extensively in the development and design of new drugs, environmental factor analysis, biological molecule action analysis, chemical agent toxicological characteristics research and other fields. Burello and Worth reviewed the application of QSAR in the research field of nanomaterials in recent years.[10] Ambure et al.[11] developed an ensemble software for QSAR modeling. Pontiki and Hadjipavlou-Litina[12] conducted relevant research on lipoxygenase inhibitors by QSAR. There are few reports on QSAR in the molecular design of energetic materials. In the previous research of our group, a method for quantitative analysis of explosive heat of energetic materials was reported based on QSPR and RF algorithm, and its prediction set determination coefficient (R^2_P) and average relative error (MREP) were 0.8801 and 10.52%, respectively.[13] This study preliminarily confirmed the feasibility of QSAR combined with machine learning method in predicting detonation performance of energetic compounds.

Based on theoretical calculation and molecular simulation, QSAR can calculate various descriptors of molecules to be studied, including topology, composition, electrostatics, geometry, quantum chemistry, WHIM, 3D Weiner, and other descriptor information, which will generate a lot of data. Therefore, QSAR data processing and model construction are one of the focuses of its research.[14] Common QSAR modeling methods include multiple linear regression (MLR),[15,16] partial least squares (PLS),[17,18] artificial neural network (ANN),[19,20] and support vector machine (SVM).[21,22] MLR and PLS are usually used to solve linear regression problems, so the performance of QSAR data modeling is generally mediocre. ANN and SVM often show good prediction performance in QSAR modeling, but due to the limitations of model parameters and structure, they often show low modeling efficiency in high-dimensional data processing. In recent years, the ensemble model strategy has attracted more and more attention in the construction of multiple regression models. Wang et al.[23] proposed a dual ensemble strategy of MC−LASSO−PLS for near-infrared spectral data modeling. Liu et al.[24] proposed a multi-dimensional ensemble method based on deep learning for short-term runoff prediction. It can be seen from previous studies that the ensemble model usually performs better than a single model in small sample data. The most popular resampling technologies include bagging, subagging, boosting, and stacking.[25,26] RF algorithm is a tree-based ensemble method with excellent performance, which is often used to solve regression and classification problems.[27] In the process of RF modeling, the bagging method is used for random sampling in place to build multiple regression trees (decision trees). At the node of each tree, a certain number of variables are selected and pruning operations are performed. When the data set contains too many secondary or redundant variables, the prediction performance of this tree will be degraded. If there are many such trees in the RF model, the performance of the ensemble model will be degraded. This is because the final result of the RF model comes from the simple average of the results of these regression trees.[28] Therefore, how to ensure that the contribution of variables to the RF model in an acceptable range is a key point worth considering when building a QSAR model based on RF.[29]

The subset generation method and the ensemble strategy of subset model prediction results are the focus of research in the modeling process of ensemble model construction.[30,31] However, in the QSAR/QSPR modeling task, the sample size is usually small, which will have serious effects on the prediction performance of the ensemble model, because it will restrict the diversity of subsets of the ensemble model.[32] MC is a method based on random sampling, which can be used to generate a group of random numbers, and can be combined with the task to be analyzed to generate a new sample and combination.[33] In recent years, MC has been employed for the construction of multiple regression models, including cross validation,[34,35] subset generation,[36] feature variable screening,[37,38] and so forth. MC is used to generate multiple groups of different subsets for the construction of the RF model, and a certain ensemble strategy is used to further integrate the output, which can effectively improve the stability of the prediction performance of the RF model, especially in small sample analysis tasks such as QSPR data.

To solve the mentioned problems and obtain a QSPR quantitative analysis model of detonation performance of energetic materials with more excellent prediction performance, an ensemble modeling strategy of MC-variable importance measurement (VIM)-random forest (RF) (MCVIMRF_Med) was proposed in this work. First, molecular descriptors and density indexes of 162 energetic compounds were obtained, and the QSPR descriptors were divided into calibration set and prediction set by the SPXY method. Then, the MC method is used to generate a subset of the ensemble model; the MCVIMRF_Med ensemble model will be constructed based on these subsets. To further optimize prediction performance of the MCVIMRF_Med ensemble model, the subset size of the MCVIMRF_Med ensemble model and the number of model iterations are optimized based on MC cross validation. The final output strategy of the ensemble model is optimized based on the optimized model

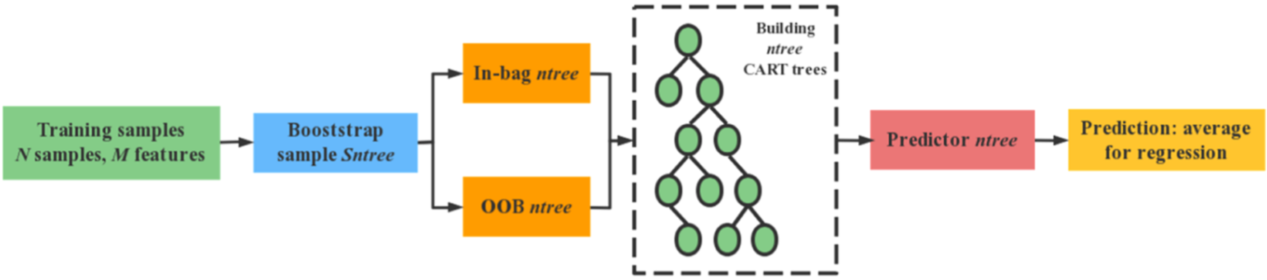


Figure 1. Schematic diagram of RF model construction.

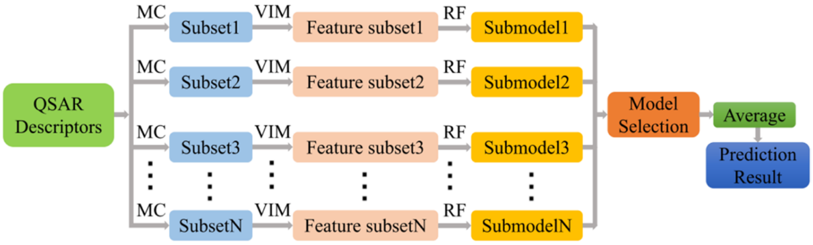


Figure 2. Schematic diagram of the MCVIMRF_Med ensemble model.

parameters, and an output optimization method based on median screening is proposed and applied for the prediction performance optimization of MCVIMRF_Med ensemble model. Finally, the MCVIMRF_Med ensemble model was built under optimized conditions to predict the density of energetic materials.

2. EXPERIMENT AND METHODS

2.1. Energetic Material Data. At present, the research of new high-energy and low-insensitive HEDM still focuses on the synthesis and preparation of C, H, O, and N energetic compounds, mainly nitrogen rich compounds. In recent years, common energetic compounds can be divided into aromatic, nitroform, nitramine, nitrate, furazan, azide, fluorine, nitrogen heterocyclic, and clathrate, according to their main skeletons. Therefore, 162 energetic compounds are involved in this study, including the mentioned nine types of nitrogenous compounds. Before calculating the molecular descriptors of energetic compounds, Gaussian software (Version 09) was utilized to optimize the molecular structures of the above energetic compounds to be analyzed at the B3LYP/6-31+G* level, and then, the molecular descriptor data were calculated based on the optimized molecular structure. The molecular descriptors of energetic materials in this study were calculated via CODESSA software (Version 3.2), mainly including the composition of nitrogen compounds, topological index, geometric configuration, electrostatic coefficient, and quantitative descriptors, the details of which can be seen in our previous research work.[13] There are 239 molecular structure descriptor data for each nitrogenous compound. The density indexes of the above energetic materials are provided by Xi'an Institute of Modern Chemistry, which is listed in Table S1 of the Supporting Information.

2.2. Chemometrics Method. *2.2.1. MC Random Sampling Method.* The MC method, which started in the 1940s, is a random simulation (or statistical simulation) method. In MC, the task to be analyzed is combined with a random probability events based on the theoretical system of statistics and probability to conduct random sampling so as to obtain approximate results of the problem. In chemometrics, the MC method has been successfully applied to variable selection, ensemble modeling subset generation, and cross validation of model performance. In this work, MC is used to generate subset samples in the iterative process. First, the MC method is used to generate a group of random sequences, and then, the calibration set samples are divided according to the above sequences, one for modeling and the other for model performance verification. Repeat the cycle until the iteration is terminated. Finally, according to a certain strategy, the models constructed from the above different subsets were ensemble (simple average, weighted average, etc.), and then, the ensemble model was applied to simulate the prediction set samples.

2.2.2. Random Forest. RF is an ensemble modeling method based on the integration of a sub model called tree, which is often used to solve classification discrimination or quantitative analysis tasks. In the regression problem, the bagging method in the RF modeling process is used to sample the samples with return, and some variables are randomly selected on each regression tree for splitting. Finally, the output of the RF model is a simple average of the outputs of all basic trees. The remarkable feature of RF model can generate multiple regression trees or decision trees with different variables. Thanks to this, RF models usually have good analysis performance, are not easy to over fit, and can be used for discrete data processing. When modeling a RF model, the Bootstrap strategy is used for resampling with placement. In this process, the selected samples are called In-bag samples, which are used to build trees in the RF, and the unselected samples are called out of bag (OOB) samples. The prediction error of OOB samples is called OOB error. The idea of RF model construction is shown in Figure 1.

VIM is a feature variable extraction method developed with RF algorithm. During RF modeling, input variables will be sorted according to their contribution to the RF model. Generally, variables with higher contribution will have a higher value of variable importance. On the contrary, variables with lower contribution will have a lower value of variable importance, even zero. VIM reserves the variables below the threshold for optimization model construction by setting a

Table 1. Prediction Performance of the MCVIMRF_Med Ensemble Model based on Different Subset Sizes

subset sizes	iterations	R^2_{CV}	RMSECV (g/cm^3)	R^2_C	RMSEC (g/cm^3)	R^2_P	RMSEP (g/cm^3)	RPD
5	500	0.6455	0.1475	0.9653	0.0785	0.7517	0.1858	1.227
10	500	0.7936	0.1183	0.9792	0.0587	0.8829	0.1467	1.582
15	500	0.8384	0.1046	0.9832	0.0502	0.9201	0.1259	1.871
20	500	0.8632	0.0948	0.9854	0.0441	0.9377	0.1086	2.157
25	500	0.8760	0.0897	0.9866	0.0406	0.9449	0.1011	2.314
30	500	0.8879	0.0848	0.9871	0.0383	0.9526	0.0925	2.526
35	500	0.8946	0.0812	0.9881	0.0362	0.9584	0.0871	2.692
40	500	0.9058	0.0776	0.9881	0.0349	0.9614	0.0830	2.817
45	500	0.9080	0.0754	0.9885	0.0338	0.9642	0.0788	2.949
50	500	0.9173	0.0721	0.9890	0.0326	0.9670	0.0752	3.080
55	500	0.9231	0.0691	0.9891	0.0320	0.9696	0.0718	3.227
60	500	0.9266	0.0670	0.9893	0.0312	0.9708	0.0692	3.348
65	500	0.9278	0.0662	0.9896	0.0304	0.9714	0.0679	3.411
70	500	0.9312	0.0640	0.9899	0.0297	0.9729	0.0657	3.524
75	500	0.9386	0.0605	0.9901	0.0293	0.9744	0.0630	3.645
80	500	0.9417	0.0583	0.9902	0.0287	0.9749	0.0617	3.724
85	500	0.9469	0.0556	0.9903	0.0283	0.9755	0.0605	3.788
90	500	0.9578	0.0491	0.9906	0.0278	0.9765	0.0587	3.882
95	500	0.9577	0.0442	0.9909	0.0272	0.9767	0.0578	3.947

certain threshold, which are called feature variables. The other part of the variables with lower variable importance values are usually deleted as interference variables. The accuracy and modeling efficiency of the RF model can be effectively improved by optimizing the input variables of the RF model through VIM.[39]

2.2.3. MCVIMRF_Med Method. As RF stipulated in the modeling process that it was unnecessary to prune the tree, the prediction performance of the RF model would be interfered by some irrelevant variables or noise information. In this paper, a MCVIMRF_Med ensemble modeling method is proposed on the basis of MC and RF algorithm to predict detonation performance of energetic materials. The construction process of the MCVIMRF_Med ensemble model is shown in Figure 2. The MCVIMRF_Med ensemble modeling process is as follows.

(1) SPXY method was used to divide the obtained data set.[40] In this study, descriptors of 162 nitrogen-containing compounds are divided into a calibration set (130 samples) and prediction set (32 samples).

(2) MC is utilized to randomly sample the training data, among which the selected samples are used for the model construction and others are for the model performance verification.[41]

(3) A RF calibration model is established on the basis of the selected variable subset to gain the variable importance value of the subset variable, and then, the input variables of the RF model were screened based on VIM method. In this study, due to the high cost of ensemble process operation, when screening subset feature variables, the threshold of VIM is set to 0 for feature variable screening.

(4) The number of iterations and subset variables of the MCVIMRF_Med ensemble modeling process are optimized based on MC cross validation, respectively.

(5) The output results of all subset construction models are calculated as the median, and those models whose prediction results are lower than the median are considered as invalid models and deleted. The output results of the remaining models are given the same weight value and averaged as the final output of MCVIMRF_Med ensemble modeling.

In the above process, the evaluation indicators of model performance mainly include the determination coefficient (R^2), root mean square error (RMSE), and relative analysis deviation of prediction set (RPD). When the R^2 is between 0.66 and 0.80, the model is generally considered to have a certain prediction effect. When R^2 is between 0.81 and 0.90, the model is considered to have good prediction performance. When R^2 is greater than 0.90, the model performance is considered excellent. The smaller the RMSE value, the better the performance of the prediction model. In the above two indicators, if they are used for cross validation, the corresponding calculation results are R^2_{CV}, RMSECV. If it is used for internal self verification of the model, the corresponding calculation result is R^2_C, RMSEC. If it is used for model prediction results, the corresponding calculation results are R^2_P and RMSEP. RPD can be obtained by the ratio of the standard deviation of the sample to be analyzed to the RMSE of prediction set. Generally, when the RPD is less than 2.5, the model is considered unable to be used for prediction tasks. When the RPD is between 2.5 and 3, the model is considered to have certain prediction performance, but the prediction results are not necessarily reliable. When the RPD is greater than 3, it is considered that the model could achieve good performance and be used for the prediction of other samples. All data calculation processes in this study were completed under the environment of MATLAB (Version 2016a).

3. RESULTS AND DISCUSSION

3.1. Model Subset Size Optimization. The sample diversity of the subset model is one of the critical factors to ensure the performance of ensemble modeling. Therefore, during the construction of the ensemble model, the input subset variables of its sub model are often optimized through certain strategies to ensure the diversity of the model as much as possible. In this study, the MC method is used to generate the sample subset, that is, each time, randomly select some samples from the calibration set as a new subset to construct

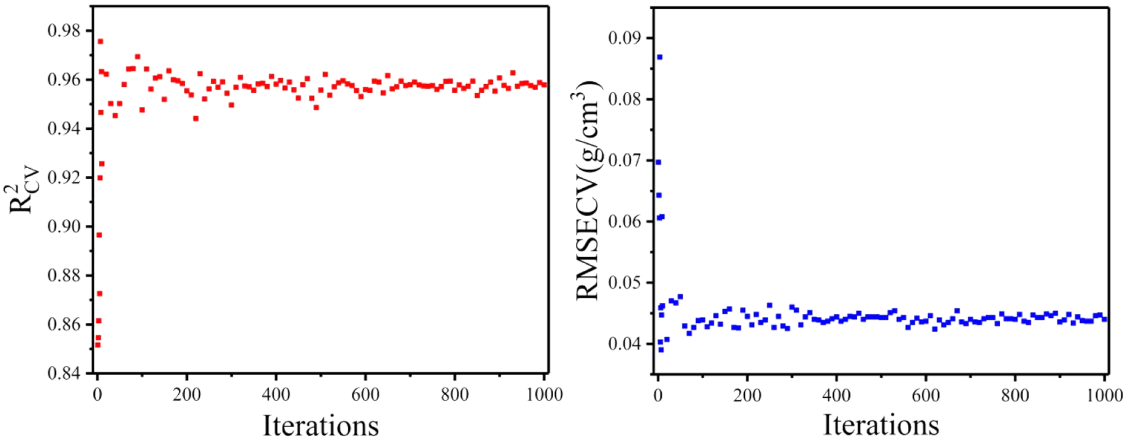


Figure 3. Prediction performance of MCVIMRF_Med ensemble modeling based on different iterations.

the sub model of the ensemble model. To fully utilize the modeling sample data set and ensure the diversity of the model subset, it is particularly important to optimize the subset size of the MC method, which is because the size of the model subset will affect the modeling speed and robustness of the ensemble model. The size of the modeling subset usually accounts for 0 to 100 of the model calibration set. Among them, 0 means that no variable is selected, and 100 means that all samples are selected (the diversity of the sub model cannot be guaranteed in the case of 100). Neither of these two conditions are meaningful in the actual modeling process. Therefore, this study mainly focuses on the prediction performance of MCVIMRF_Med ensemble modeling with a subset size of 5−95 (the number of iterations is set to 500). Table 1 shows the performance of the MCVIMRF_Med ensemble model built with a subset size of 5−95 (the interval is 5). As shown in Table 1 that with the increasing scale of the modeling subset, the prediction performance of the MCVIMRF_Med ensemble model is also improving. Compared with the model when the subset proportion is 5, the prediction performance of MCVIMRF_Med ensemble modeling has been greatly improved when the subset proportion is 95, with R^2_{CV} increasing from 0.6455 to 0.9577, RMSECV decreasing from 0.1475 to 0.0442 g/cm^3, R^2_C was raised from 0.9653 to 0.9909, and RMSEC was reduced from 0.0785 to 0.0272 g/cm^3. From the analysis results, when the subset proportion of MCVIMRF_Med ensemble modeling is 95, better prediction performance can be obtained. Therefore, the subset size of MCVIMRF_Med ensemble modeling is set to 95 in the subsequent model optimization process.

3.2. Iteration Number Optimization. The number of iterations is another key parameter of the proposed ensemble model. Proper selection of the number of iterations will improve the modeling speed and prediction accuracy of the ensemble model. However, if the number of iterations is too large, good prediction results can also be obtained, but in terms of modeling time consumption, it will generate a huge amount of computation. In contrast, when the number of iterations in modeling is too small, the modeling time is shortened, but the models built usually show some deficiencies in prediction accuracy and stability. Therefore, optimizing the appropriate number of modeling iterations is a key step in building an ensemble model. In this study, the number of iterations of the model will be further optimized on the premise of the subset size optimized in the previous step. Considering the size of the sample data set, the range of iteration number optimization is 0−1000 (the interval is 10). Figure 3 shows the prediction performance based on different iterations of MCVIMRF_Med ensemble modeling (R^2_{CV} and RMSECV). Figure 3 shows that the prediction performance of the MCVIMRF_Med ensemble model has improved to a certain extent with the increasing number of iterations both for R^2_{CV} and RMSECV. From the perspective of the stability of the model prediction performance, the stability has improved significantly with the increase of the number of iterations. As shown in Figure 3, the performance of the ensemble model basically shows a documented trend when the number of iterations is around 400. Although R^2_{CV} tends to be stable when the number of iterations is around 700 with the further increase of the number of iterations, RMSECV shows a certain fluctuation when the number of iterations is around 700. As the number of modeling iterations increases, the diversity of sub models is also constantly improved. Therefore, it can be seen that when the number of model iterations approaches 400, the MCVIMRF_Med ensemble model achieved good prediction performance. When the number of iterations increases further, the modeling and calculation consumption increases continuously, which has little effect on the improvement of model performance. Therefore, considering the modeling speed and prediction accuracy when building the MCVIMRF_Med ensemble model, the number of iterations is 410. At this time, R^2_{CV} is 0.9596, RMSECV is 0.0437 g/cm^3, R^2_C is 0.9909, and RMSEC is 0.0273 g/cm^3. The MCVIMRF_Med ensemble model is built based on the optimized iterations (VIM thresholds are all set to 0). After 410 iterations, almost every descriptor numbered 16−21 and 127−138 has been deleted, which indicates that these variables are meaningless in the process of model construction. In addition, a few other variables are eliminated, which will help prediction performance optimization of the MCVIMRF_Med ensemble model.

3.3. Ensemble Strategy Optimization. The output of the ensemble model usually combines the prediction results of all sub models with a certain strategy. Generally, when building an ensemble model, the final output results will be obtained by simple average or weighted average. During the construction of the ensemble model, a large number of sub models will be constructed, which increases the stability of the ensemble model to a certain extent. However, considering that the final result of the ensemble model comes from the joint output of all sub model, if there are some outputs with poor results in the sub model, the performance of the ensemble model will become worse. Thus, it is an essential assignment to filter the sub models to a certain extent. In this study, the RMSECV value is used as a reference. First, the output results of all sub models are sorted to obtain the median, and the sub model below the median is rounded off. The remaining sub models are ensemble based on a simple averaging method, which is the final output of the MCVIMRF_Med ensemble model. As shown in Figure 4, the performance of the MCVIMRF ensemble model is based on different ensemble strategies [the MCVIMRF ensemble model using the direct simple average (MCVIMRF) and the ensemble idea proposed in this work (MCVIMRF_Med)]. Figure 4 shows that compared with the ensemble model based on direct simple average, the MCVIMRF_Med ensemble model can achieve more excellent performance, of which the R^2_{CV} increases from 0.9042 to 0.9596, and RMSECV decreases from 0.0792 to 0.0437 g/cm^3. R^2_C was raised from 0.9888 to 0.9909, and RMSEC was reduced from 0.0318 to 0.0273 g/cm^3. Therefore, the ensemble idea based on median screening proposed in this work is an efficacious approach to further enhance the estimation performance of the MCVIMRF ensemble model.

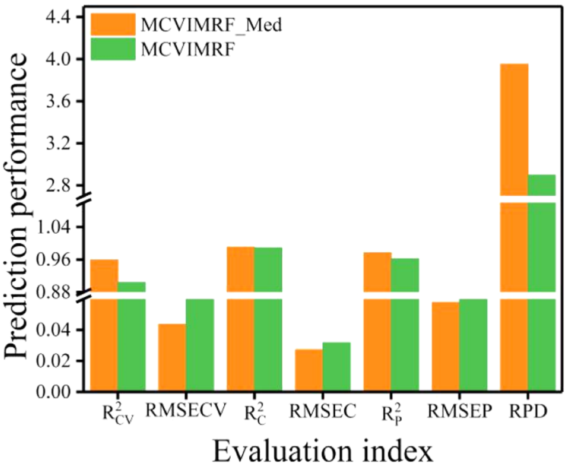


Figure 4. Prediction performance of MCVIMRF ensemble model via different ensemble strategies.

3.4. Model Performance Comparison. Due to the randomness of subset selection, the performance of the ensemble model will have some deviation, so the stability of the MCVIMRF_Med model is one of its performance evaluation indicators. Figure 5 shows the performance of the MCVIMRF_Med ensemble model, in which the deviation of the model performance obtained by repeated experiments (50 times) is highlighted. As shown in Figure 5 that the performance of the MCVIMRF_Med ensemble model is very stable, the error bar of all evaluation index is relatively small. Thus, it can therefore draw a conclusion that the MCVIMRF_Med ensemble model has excellent stability and robustness.

To further verify the prediction performance of the MCVIMRF_Med ensemble model for the density of energetic materials, it is compared with that of many other calibration models, including PLS, RF, VIMRF, and MCVIMRF ensemble models based on simple direct average. When building a PLS model, the latent variables are optimized to 10. When constructing the RF calibration model, the RF default parameters are selected, that is, the n_{tree} is 500 and m_{try} is $P/3$ (P is the number of molecular descriptors of energetic compounds). When building the VIMRF model, the threshold is set to 0. When building the MCVIMRF ensemble model, the model subset size is 50 and the number of iterations is set to 500. Table 2 shows the prediction performance of the above different calibration models for the density index of energetic materials. As shown in Table 2 that although the RF model shows better prediction results in the prediction set, its cross validation and self prediction of the calibration set are poor, so it is considered that the RF model has an over fitting phenomenon. Compared with other calibration models mentioned above, the MCVIMRF_Med ensemble model proposed in this study achieves better prediction performance at three levels of model cross validation results, internal self prediction and prediction set, with R^2_{CV} of 0.9596, RMSECV of 0.0437 g/cm^3, R^2_C of 0.9909, RMSEC of 0.0273 g/cm^3, R^2_P of 0.9768, RMSEP of 0.0578 g/cm^3, and RPD of 3.951. This result shows that the MCVIMRF_Med ensemble model can effectively improve the performance of the single RF calibration model for the density prediction of energetic materials.

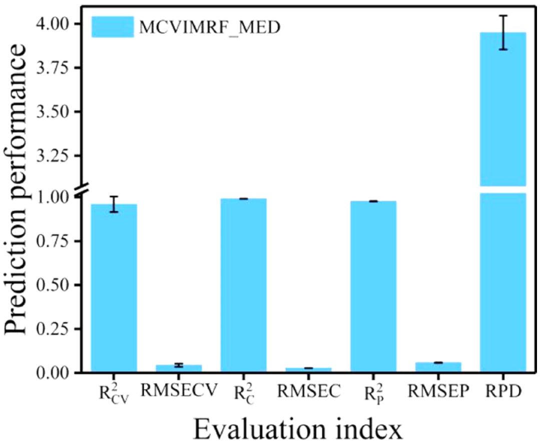


Figure 5. Prediction performance of MCVIMRF_Med ensemble model.

4. CONCLUSIONS

In this study, a MCVIMRF_Med ensemble modeling strategy combined with QSPR is proposed to successfully predict the density index of energetic materials. First, the structure of 162 energetic compounds was optimized by Gaussian software, and the molecular descriptor data were calculated by CODESSA software based on the optimized molecular structure. Then, based on the above molecular descriptor data and the corresponding energetic compound density index, the MCVIMRF_Med ensemble model was constructed. When modeling, the SPXY method is first used to divide the data set, and then, MC is used to further divide the calibration set data into multiple subset data sets for the construction of the ensemble model. The subset size and the number of iterations of the MCVIMRF_Med ensemble model are optimized based on MC cross validation (the optimized subset size is 95 and the number of iterations is 410). Then, the final output strategy of the ensemble model is optimized, and an output optimization method based on median screening is proposed and successfully applied to optimize of the MCVIMRF_Med ensemble model. To further investigate the performance of the

Table 2. Prediction Performance of Density Index of Energetic Materials based on Different Calibration Models

model	R^2_{CV}	RMSECV (g/cm^3)	R^2_C	RMSEC (g/cm^3)	R^2_P	RMSEP (g/cm^3)	RPD
PLS	0.5902	0.1236	0.7151	0.0969	0.8228	0.0917	2.345
RF	0.8527	0.0694	0.9816	0.0269	0.9544	0.0563	4.087
VIMRF	0.8516	0.0697	0.9833	0.0262	0.9550	0.0573	3.969
MCVIMRF	0.9042	0.0792	0.9888	0.0318	0.9622	0.0802	2.897
MCVIMRF_Med	0.9596	0.0437	0.9909	0.0273	0.9768	0.0578	3.951

MCVIMRF_Med ensemble model, the performance of it was compared with PLS, RF, VIMRF, and MCVIMRF calibration models. The results show that the MCVIMRF_Med ensemble model can gain a better result for the density prediction of energetic materials, with an R^2_{CV} of 0.9596, RMSECV of 0.0437 g/cm^3, R^2_C of 0.9909, RMSEC of 0.0273 g/cm^3, R^2_P of 0.9768, RMSEP of 0.0578 g/cm^3, and RPD of 3.951. Therefore, the MCVIMRF_Med ensemble modeling strategy combined with QSPR proposed in this work is an effective prediction method for density of energetic materials.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c07436.

Density index obtained from experiment of the 162 nitro energetic compounds (PDF)

AUTHOR INFORMATION

Corresponding Authors

Tianlong Zhang − *Key Laboratory of Synthetic and Natural Functional Molecule of the Ministry of Education, College of Chemistry & Materials Science, Northwest University, Xi'an 710127, China;* orcid.org/0000-0003-4289-5052; Email: tlzhang@nwu.edu.cn

Hua Li − *Key Laboratory of Synthetic and Natural Functional Molecule of the Ministry of Education, College of Chemistry & Materials Science, Northwest University, Xi'an 710127, China; College of Chemistry and Chemical Engineering, Xi'an Shiyou University, Xi'an 710065, China;* orcid.org/0000-0001-6618-7216; Email: huali@nwu.edu.cn

Authors

Maogang Li − *Key Laboratory of Synthetic and Natural Functional Molecule of the Ministry of Education, College of Chemistry & Materials Science, Northwest University, Xi'an 710127, China*

Weipeng Lai − *Xi'an Modern Chemistry Research Institute, Xi'an 710065, China*

Ruirui Li − *Guangzhou University of Chinese Medicine, Guangzhou 510006, China*

Jiajun Zhou − *Key Laboratory of Synthetic and Natural Functional Molecule of the Ministry of Education, College of Chemistry & Materials Science, Northwest University, Xi'an 710127, China*

Yingzhe Liu − *Xi'an Modern Chemistry Research Institute, Xi'an 710065, China;* orcid.org/0000-0003-4150-1111

Tao Yu − *Xi'an Modern Chemistry Research Institute, Xi'an 710065, China*

Hongsheng Tang − *Key Laboratory of Synthetic and Natural Functional Molecule of the Ministry of Education, College of Chemistry & Materials Science, Northwest University, Xi'an 710127, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c07436

Author Contributions

M.L. and W.L. contributed equally to this work. M.L.: investigation, methodology, and writingoriginal draft preparation. W.L.: sample, descriptor calculation, and data collection. R.L.: descriptor calculation and data collection. J.Z.: sample, software, and data collection. Y.L.: sample, experiment, and data collection. T.Y.: sample, experiment, and data collection. T.Z.: funding acquisition, supervision, writingreview and editing. H.T.: supervision. H.L.: funding acquisition and supervision.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful of the support of the National Natural Science Foundation of China (nos. 22173071 and 21873076).

REFERENCES

(1) Choi, C.; Ashby, D. S.; Butts, D. M.; DeBlock, R. H.; Wei, Q.; Lau, J.; Dunn, B. Achieving High Energy Density and High Power Density with Pseudocapacitive Materials. *Nat. Rev. Mater.* **2020**, *5*, 5−19.

(2) Yan, Q.; Zhao, F.; Kuo, K. K.; Zhang, X.; Zeman, S.; DeLuca, L. T. Catalytic Effects of Nano Additives on Decomposition and Combustion of RDX-, HMX-, and AP-Based Energetic Compositions. *Prog. Energy Combust.* **2016**, *57*, 75−136.

(3) Yin, P.; Zhang, Q.; Shreeve, J. M. Dancing with Energetic Nitrogen Atoms: Versatile n-Functionalization Strategies for N-Heterocyclic Frameworks in High Energy Density Materials. *Acc. Chem. Res.* **2016**, *49*, 4−16.

(4) Gałuszka, A.; Migaszewski, Z.; Namieśnik, J. The 12 Principles of Green Analytical Chemistry and The SIGNIFICANCE Mnemonic of Green Analytical Practices. *TrAC, Trends Anal. Chem.* **2013**, *50*, 78−84.

(5) Erythropel, H. C.; Zimmerman, J. B.; de Winter, T. M.; Petitjean, L.; Melnikov, F.; Lam, C. H.; Lounsbury, A. W.; Mellor, K. E.; Janković, N. Z.; Tu, Q.; Pincus, L. N.; Falinski, M. M.; Shi, W.; Coish, P.; Plata, D. L.; Anastas, P. T. The Green ChemisTREE: 20 Years After Taking Root With The 12 Principles. *Green Chem.* **2018**, *20*, 1929−1961.

(6) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.

(7) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A Mobile Robotic Chemist. *Nature* **2020**, *583*, 237−241.

(8) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR Without Borders. *Chem. Soc. Rev.* **2020**, *49*, 3525−3564.

(9) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure-Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112*, 2889−2919.

(10) Burello, E.; Worth, A. P. QSAR Modeling of Nanomaterials. *WIRES Nanomed. Nanobi.* **2011**, *3*, 298−306.

(11) Ambure, P.; Aher, R. B.; Gajewicz, A.; Puzyn, T.; Roy, K. "NanoBRIDGES" software: Open access tools to perform QSAR and nano-QSAR modeling. *Chemom. Intell. Lab. Syst.* **2015**, *147*, 1−13.

(12) Pontiki, E.; Hadjipavlou-Litina, D. Lipoxygenase Inhibitors: A Comparative QSAR Study Review and Evaluation of New QSARs. *Med. Res. Rev.* **2008**, *28*, 39−117.

(13) He, T.; Lai, W.; Li, M.; Feng, Y.; Liu, Y.; Yu, T.; Tang, H.; Zhang, T.; Li, H. The Detonation Heat Prediction of Nitrogen-Containing Compounds Based on Quantitative Structure-Activity Relationship (QSAR) Combined with Random Forest (RF). *Chemom. Intell. Lab. Syst.* **2021**, *213*, 104249.

(14) Roy, K.; Kar, S.; Ambure, P. On A Simple Approach for Determining Applicability Domain of QSAR Models. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 22−29.

(15) Xu, X.; Zhang, W.; Huang, C.; Li, Y.; Yu, H.; Wang, Y.; Duan, J.; Ling, Y. A Novel Chemometric Method for The Prediction of Human Oral Bioavailability. *Int. J. Mol. Sci.* **2012**, *13*, 6964−6982.

(16) Islam, R.; Parves, M. R.; Paul, A. S.; Uddin, N.; Rahman, M. S.; Mamun, A. A.; Hossain, M. N.; Ali, M. A.; Halim, M. A. A Molecular Modeling Approach to Identify Effective Antiviral Phytochemicals Against the Main Protease of SARS-CoV-2. *J. Biomol. Struct. Dyn.* **2021**, *39*, 3213−3224.

(17) Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(18) Mansouri, K.; Ringsted, T.; Ballabio, D.; Todeschini, R.; Consonni, V. Quantitative Structure-Activity Relationship Models for Ready Biodegradability of Chemicals. *J. Chem. Inf. Model.* **2013**, *53*, 867−878.

(19) Burden, F. R.; Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, *42*, 3183−3187.

(20) Burden, F. R.; Ford, M. G.; Whitley, D. C.; Winkler, D. A. Use of Automatic Relevance Determination in QSAR Studies Using Bayesian Neural Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423−1430.

(21) Winkler, D. A. The Impact of Machine Learning on Future Tuberculosis Drug Discovery. *Expert Opin. Drug Discovery* **2022**, *17*, 925.

(22) Li, J.; Liu, H.; Yao, X.; Liu, M.; Hu, Z.; Fan, B. Quantitative Structure-Activity Relationship Study of Acyl Ureas as Inhibitors of Human Liver Glycogen Phosphorylase Using Least Squares Support Vector Machines. *Chemom. Intell. Lab. Syst.* **2007**, *87*, 139−146.

(23) Wang, K.; Bian, X.; Tan, X.; Wang, H.; Li, Y. A New Ensemble Modeling Method for Multivariate Calibration of Near Infrared Spectra. *Anal. Methods* **2021**, *13*, 1374−1380.

(24) Liu, G.; Tang, Z.; Qin, H.; Liu, S.; Shen, Q.; Qu, Y.; Zhou, J. Short-Term Runoff Prediction Using Deep Learning Multi-Dimensional Ensemble Method. *J. Hydrol.* **2022**, *609*, 127762.

(25) Sagi, O.; Rokach, L. Ensemble Learning: A survey. *Wires Data Min. Knowl.* **2018**, *8*, No. e1249.

(26) Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A Survey on Ensemble Learning. *Front. Comput. Sci. China* **2020**, *14*, 241−258.

(27) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(28) Bin, J.; Ai, F.; Fan, W.; Zhou, J.; Yun, Y.; Liang, Y. A Modified Random Forest Approach to Improve Multi-Class Classification Performance of Tobacco Leaf Grades Coupled with NIR Spectroscopy. *RSC Adv.* **2016**, *6*, 30353−30361.

(29) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(30) Lv, Y.; Liu, J.; Yang, T.; Zeng, D. A Novel Least Squares Support Vector Machine Ensemble Model for NOx Emission Prediction of A Coal-Fired Boiler. *Energy* **2013**, *55*, 319−329.

(31) Lee, M. C.; Boroczky, L.; Sungur-Stasik, K.; Cann, A. D.; Borczuk, A. C.; Kawut, S. M.; Powell, C. A. Computer-Aided Diagnosis of Pulmonary Nodules Using A Two-Step Approach for Feature Selection and Classifier Ensemble Construction. *Artif. Intell. Med.* **2010**, *50*, 43−53.

(32) Bian, X.; Li, S.; Lin, L.; Tan, X.; Fan, Q.; Li, M. High and Low Frequency Unfolded Partial Least Squares Regression Based on Empirical Mode Decomposition for Quantitative Analysis of Fuel Oil Samples. *Anal. Chim. Acta* **2016**, *925*, 16−22.

(33) Miller, B. R., III; McGee, T. D., Jr; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8*, 3314−3321.

(34) Li, H.; Liang, Y.; Xu, Q.; Cao, D. Key Wavelengths Screening Using Competitive Adaptive Reweighted Sampling Method for Multivariate Calibration. *Anal. Chim. Acta* **2009**, *648*, 77−84.

(35) Xu, Q.; Liang, Y. Monte Carlo Cross Validation. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1−11.

(36) Wang, K.; Bian, X.; Zheng, M.; Liu, P.; Lin, L.; Tan, X. Rapid Determination of Hemoglobin Concentration by A Novel Ensemble Extreme Learning Machine Method Combined with Near-Infrared Spectroscopy. *Spectrochim. Acta A* **2021**, *263*, 120138.

(37) Cai, W.; Li, Y.; Shao, X. A Variable Selection Method Based on Uninformative Variable Elimination for Multivariate Calibration of Near-Infrared Spectra. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 188−194.

(38) Anders, U.; Korn, O. Model Selection in Neural Networks. *Neural Network.* **1999**, *12*, 309−323.

(39) Li, M.; Ruan, F.; Li, R.; Zhou, J.; Zhang, T.; Tang, H.; Li, H. In Situ Simultaneous Quantitative Analysis Multi-Elements of Archaeological Ceramics via Laser-Induced Breakdown Spectroscopy Combined with Machine Learning Strategy. *Microchem. J.* **2022**, *182*, 107928.

(40) Galvão, R. K. H.; Araujo, M. C. U.; José, 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) Veselinović, A. M.; Velimorović, D.; Kaličanin, B.; Toropova, A.; Toropov, A.; Veselinović, J. Prediction of Gas Chromatographic Retention Indices Based on Monte Carlo Method. *Talanta* **2017**, *168*, 257−262.