JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

Cite This: *J. Am. Chem. Soc.* 2020, 142, 1475−1481

pubs.acs.org/JACS

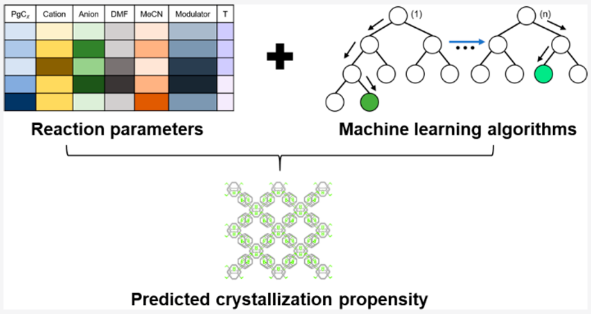
Machine Learning Assisted Synthesis of Metal−Organic Nanocapsules

Yunchao Xie,[†,#] Chen Zhang,[‡,#] Xiangquan Hu,[‡] Chi Zhang,[†] Steven P. Kelley,[‡] Jerry L. Atwood,[*,‡] and Jian Lin[*,†,§,∥]

[†] Department of Mechanical and Aerospace Engineering, University of Missouri, Columbia, Missouri 65211, United States
[‡] Department of Chemistry, University of Missouri, Columbia, Missouri 65211, United States
[§] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri 65211, United States
[∥] Department of Physics and Astronomy, University of Missouri, Columbia, Missouri 65211, United States

Supporting Information

ABSTRACT: Herein, we report machine learning algorithms by training data sets from a set of both successful and failed experiments for studying the crystallization propensity of metal−organic nanocapsules (MONCs). Among a variety of studied machine learning algorithms, XGBoost affords the highest prediction accuracy of >90%. The derived chemical feature scores that determine importance of reaction parameters from the XGBoost model assist to identify synthesis parameters for successfully synthesizing new hierarchical structures of MONCs, showing superior performance to a well-trained chemist. This work demonstrates that the machine learning algorithms can assist the chemists to faster search for the optimal reaction parameters from many experimental variables, whose features are usually hidden in the high-dimensional space.


PgCx
Cation
Anion
DMF
MeCN
Modulator
T
Reaction parameters
(1)
(n)
Machine learning algorithms
Predicted crystallization propensity

INTRODUCTION

Metal−organic nanocapsules (MONCs) have aroused a surge of interest due to their potential applications in many different fields including catalysis,[1,2] gas adsorption and separation,[3−6] and sensing.[7] These MONCs can further self-assemble into hierarchical structures.[8,9] In our previous studies, we have successfully synthesized various dimeric (M_8L_2) and hexameric ($M_{24}L_6$ or $M_{12}L_6$) MONCs by utilizing different types of metal ions and *C*-alkylpyrogallol[4]arenes (PgC_x) or *C*-alkylpyrogallol[3]resorcin[1]arene ($P_3R_1C_x$),[10−16] where *x* is the number of carbon atoms in the alkyl tails. These MONCs were synthesized by the solvothermal crystallization method, which has also been widely utilized for synthesis of inorganic−organic hybrid materials such as organohalide perovskites[17] and metal−organic frameworks (MOFs).[18] The solvothermal crystallization of the MONCs is primarily an exploratory process. It involves three major steps. First, a chemical space containing all reaction parameters for the synthesis such as metal ions, organic ligands, solvents, and temperature is created. Second, through human experience and intuition, possible synthesis parameters from the chemical space are identified and selected for experiments. Third, after a trial-and-error synthesis process, the final reaction parameters that lead to desired products are tested and reported. In the process, individuals obtain chemical intuition and knowledge from both successful and failed experiments. The whole process requires tremendous effort and resources, and the success in the synthesis of desired products heavily relies on the individuals involved. The miserable failure of human intuition in solving problems involving high complexity leads to the difficulty in analysis of all the experimental variables, which makes it almost impossible to obtain the optimum reaction parameters. Although genetic algorithms have been explored to search the complex chemical space,[19,20] their size is too overwhelming or costly to be comprehensively researched and tested. Thus, smart and cost-effective navigation based on surrogate models is quite desired.

In recent years, machine learning, which enables the provision of surrogate algorithms for material development, has gained enormous attention for effectively predicting the physical and chemical properties,[21] establishing the structure−property relationships,[22,23] and navigating the chemical space for guiding chemical synthesis.[24−29] For instance, Raccuglia et al. reported on applying a support vector machine (SVM) algorithm to exploit the chemical space from historically successful and failed experiments for elucidating factors that govern reaction outcomes.[24] Doyle et al. demonstrated the successful application of random forest (RF) regression algorithm to predict high-yielding conditions for untested substrates,[25] showing the result of a coefficient of determination of an R^2 value of 0.92. In Cronin and his colleagues'

Received: October 27, 2019
Published: December 24, 2019

DOI: 10.1021/jacs.9b11569

recent work, the well-trained neural networks could predict the reactivity of more than 1000 reaction combinations with an accuracy of >80%.[26] Despite this progress, application of the machine learning algorithms to guide the synthesis of inorganic−organic hybrid materials has still been quite limited.[24,27] So far, to the best of our knowledge, exploiting machine learning algorithms for MONCs synthesis has not yet been reported. In addition, the afforded chemical insights from these reported machine learning models, such as interpretable hypotheses and feature importance of the reaction parameters, are still quite limited.

Herein, we introduced machine learning models to the traditional trial−error synthesis process of MONCs single crystals for predicting their crystallization propensity. The best tested model, the XGBoost model, afforded the highest prediction accuracy of >90%. The feature importance and chemical hypotheses derived from the XGBoost model assist in identifying successful synthesis parameters for new MONC crystals, showing superior performance to a well-trained chemist. The extracted hidden information demonstrates that the XGBoost model afforded a simple and straightforward way of quantifying the chemical intuition, which is very hard to be realized by human or other traditional analysis methods. Thus, the machine learning shows great potential in guiding chemists, especially new entrants, to screen the reaction parameters for synthesizing new materials beyond MONCs.

■ EXPERIMENTAL SECTION

Synthesis of C-Alkylpyrogallol[4]arene (PgC$_x$). PgC$_x$ (x = 1−9) and PgC$_3$OH were synthesized using a previously reported condensation reaction.[30] Taking PgC$_3$OH as an example, 2,3-dihydrofuran (6.05 mL, 0.08 mol) and pyrogallol (0.08 mmol, 10 g) were mixed in 30 machine learnings of 95% (v/v) ethanol with the addition of 3.5 mL of concentrated HCl. Thereafter, the mixture was refluxed at 110 °C for 24 h. After cooling down, the precipitate was filtered, washed with cold 95% (v/v) ethanol and dried in vacuum. 5.4 g of white solid was prepared as the final product, PgC$_3$OH. Yield was 34.8%. Note: As for PgC$_x$ (x = 1−9), C$_{x+1}$ aldehydes including acetaldehyde (PgC$_1$), propionaldehyde (PgC$_2$), butyraldehyde (PgC$_3$), pentanal (PgC$_4$), hexanal (PgC$_5$), heptanal (PgC$_6$), octanal (PgC$_7$), nonanal (PgC$_8$), and decanal (PgC$_9$) were used for the reactions, which were conducted in either ethyl acetate (PgC$_1$ and PgC$_2$) or methanol (PgC$_x$, x = 3−9).

General Procedure for Solvothermal Crystallization of MONCs. Synthesis parameters consist of the choice of PgC$_x$, metal salts, solvents, modulators, and temperature, and the detailed information can be found in Table S1 of the Supporting Information (SI). In a typical synthesis, the metal salts (nitrates or chlorides), PgC$_x$/PgC$_3$OH, and modulators were added into a mixture of *N,N*-dimethylformamide (DMF)/acetonitrile (MeCN)/H$_2$O in a 4 mL glass vial. The mixture was sonicated for 5 min and heated overnight at various temperatures in an oven. In addition, 20 new experiments were conducted to validate the developed XGBoost model.

Synthesis of SCP-4. *C*-Propan-3-olpyrogallol[4]arene (PgC$_3$OH, 0.1 mmol, 78.4 mg), Mg(NO$_3$)$_2$·6H$_2$O (0.4 mmol, 116.4 mg), and benzoic acid (0.3 mmol, 36.6 mg) were dissolved in the mixture of 1.0 mL DMF and 2.0 mL MeCN with the addition of 0.1 mL water in a 4 mL glass vial. The mixture was sonicated for 5 min to yield a dark green solution and then heated at 130 °C overnight. Finally, green crystals were formed and collected for single crystal X-ray diffraction analysis.

XRD Characterization. The single crystal X-ray diffraction data were collected on a Bruker Apex II diffractometer at a temperature of 100 (2) K using Cu$K\alpha$ (λ = 1.54056 Å) radiation incotec Microfocus II. The XRD data were able to be collected at a resolution of 1.00 Å, and it would allow for the isotropic refinement of all non-hydrogen molar positions corresponding to the pyrogallol skeleton and metal atoms. However, a full anisotropic refinement of all positions was not performed when encoding the structure of SCP-4.

Machine Learning Models. A total of nine different machine learning algorithms, i.e., Logistic Regression (LR),[31] Gaussian Naïve Bayes (GNB),[32] k-Nearest Neighbors (KNN),[33] Support Vector Machine (SVM),[34] Decision Tree (DT),[35] Random Forest (RF),[36] Adaptive Boosting (ADA),[37] eXtreme Gradient Boosting (XGBoost),[38] and Multilayer Perceptron (MLP),[39] are trained for predicting the crystallization propensity of MONCs. All models were directly programmed using Python with the scikit-learn package.[36] A 5-fold grid-search cross-validation (5-fold GridSearchCV) strategy was implemented in our case to optimize the choice of hyper-parameters as well as preventing overfitting. We also repeated this five times by randomly splitting training and testing data sets to avoid sampling bias and reported average values of their prediction results in the evaluation metrics (Figure S1).

■ RESULTS AND DISCUSSION

Figure 1 shows the flow of predicting the crystallization propensity of MONCs with the assistance of machine learning models. First, all synthesis parameters from a total of 486 reactions and their reaction outcomes of crystallization or noncrystallization were collected from archived laboratory notebooks, and established as input and output data sets for the machine learning models. The experiments are categorized into two classes according to their reaction outcomes. Class "0" indicates the reaction outcomes of nonsingle crystals (293) while Class "1" indicates the reaction outcome of single crystals (193) at given input reaction parameters.

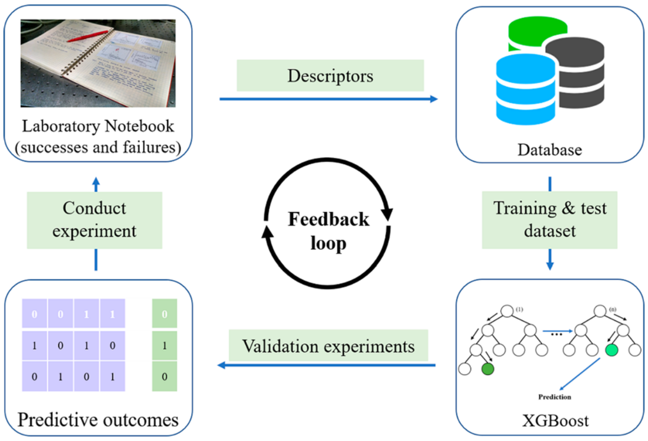


Figure 1. Schematic representation of working flow when machine learning models are incorporated into the prediction of crystallization propensity of MONCs.

We first identified a total of 17 descriptors that may govern the crystallization propensity of MONCs (Table S1). They indicate the properties of the organic ligands (molar mass, carbon length, and hydroxyl groups), the inorganic metal salts (molar mass and charge of anions, radius of cations and anions), modulators (molar mass, pK_a, and moles), and reaction conditions (temperature and solvent volume). These descriptors were developed based on our experience and chemical intuition. For example, the molar mass, carbon length, and hydroxyl group of the PgC$_x$ were chosen since the lengths of the alkyl chains and hydroxyl groups were believed to greatly affect the hydrophobicity and solubility of MONCs in organic solvents. The molar mass, charge, and radius of metal ions were selected since they affect the coordination degrees. We considered molar mass, pK_a value, and mole

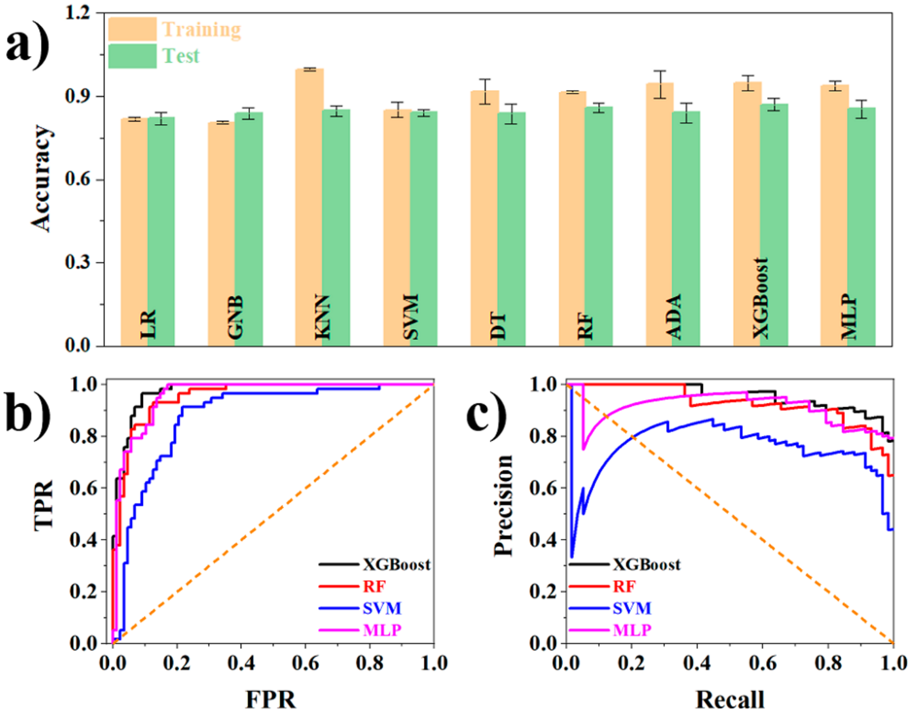


Figure 2. (a) Training and test accuracy of various machine learning models. LR: Logistic Regression; GNB: Gaussian Naïve Bayesian; KNN: k-Nearest Neighbors; SVM: Support Vector Machine; DT: Decision Tree; RF: Random Forest; ADA: AdaBoost; XGBoost: eXtreme Gradient Boosting; and MLP: Multilayer Perceptron. (b) ROC curves and (c) precision-recall curves calculated from SVM, RF, XGBoost, and MLP models.

number as the descriptors for the modulators because they can tune the deprotonation capability of PgC_x. Then, the data sets consisting of total 486 reactions with 17 descriptors were shuffled and split into training (70%) and test data sets (30%). The ratio of MONC single-crystals to nonsingle-crystal samples was equally distributed in both training and testing data sets.

After the database was established, a set of nine machine learning models, including Logistic Regression (LR),[31] Gaussian Naïve Bayes (GNB),[32] k-Nearest Neighbors (KNN),[33] Support Vector Machine (SVM),[34] Decision Tree (DT),[35] Random Forests (RF),[36] Adaptive Boosting (ADA),[37] eXtreme Gradient Boosting (XGBoost),[38] and Multilayer Perceptron (MLP)[39] were trained by a grid-search cross-validation (5-fold GridSearchCV) method and the hyper-parameters of a single-shot trial was summarized in Table S2. The evaluation metrics including accuracy, precision, recall, F_1, and receiver operating characteristic (ROC) curves were obtained by comparing the predicted results and the ground truths (Table S3). Although these machine learning models offer individual advantages, such as high accuracy for classification, ease of operation, or good interpretability, they must be weighed carefully for a new application. We evaluated their performance with a goal of finding the one that is both highly accurate and interpretable.[21] It can be seen from Figure 2a and Table S3 that all of the nine machine learning models can reach accuracy of >82% and F1 score of >81%. Among them, the XGBoost model exhibits the best performance with the highest prediction accuracy of 91% and average F1 score of 87%.

Four representative models including XGBoost and the other three models (SVM, RF, and MLP) were further compared in detail. The ROC curve indicates the relationship of the true positive rate (TPR) and false positive rate (FPR) (Figure 2b). It takes the uncertainty of each prediction into account when evaluating the performance of a machine learning model.[40,41] More deviation of an ROC curve toward the top left corner from the randomly guessing baseline (orange dash line) indicates that a machine learning model obtains a higher prediction accuracy. XGBoost shows the highest deviated ROC curve compared to those of SVM, RF, and MLP, indicating its highest prediction accuracy. AUC is the area under the ROC curve and is equal to the probability that a classifier sorts a randomly selected positive sample higher than a randomly selected negative one.[42] XGBoost exhibits the highest AUC value of 0.97 in comparison to the SVM (0.88), RF (0.96), and MLP (0.96) models. The precision-recall (PR) curves for XGBoost, SVM, RF, and MLP were employed as an additional indicator in evaluating their prediction performance (Figure 2c). Precision shows the ratio of the correctly predicted true positive numbers to the total predicted positive numbers, while the recall indicates the fraction of correctly predicted true positives in the total real positive numbers.[43] XGBoost achieved a precision of 0.9 at the recall of 0.9, which is much higher than those from SVM (0.73), RF (0.84), and MLP (0.83) at the same recall value. As shown in the confusion matrix (Table S4), the XGBoost model shows the highest recall of 0.931 among four representative machine learning models (SVC: 0.862, RF: 0.914, MLP: 0.741), which indicates the highest true positive numbers.

Different machine learning models present prediction results according to the built-in algorithms. Thus, knowing the differences among them helps us to choose the best one suitable for a real application. The relatively high flexibility of SVM usually results in limited and uncontrolled performance. Furthermore, extensive experience is needed to appropriately tune the hyperparameters when training a successful SVM model. Both RF and XGBoost are based on the decision tree

DOI: 10.1021/jacs.9b11569

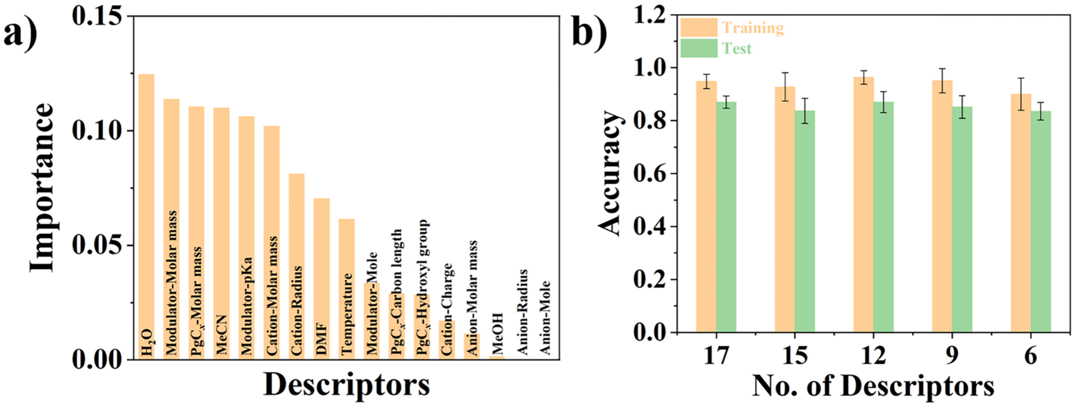


Figure 3. (a) Importance scores of descriptors derived from the XGBoost model. (b) Comparison of prediction accuracy from models trained with varied number of descriptors identified from the results shown in (a): total 17, top 15, top 12, top 9, and top 6 descriptors.

model. They are an ensemble of multiple decision trees and are proved to be effective for solving problems with high-dimensional data. Moreover, they usually present a satisfactory prediction performance even trained with default hyper-parameters. However, in contrast to RF, which makes a final decision according to the final majority vote of each classifier tree, XGBoost is a gradient boosting model that builds each classifier tree sequentially to iteratively reduce the error of the established trees. Hence, it has become one of the most widely used machine learning algorithms since introduced in 2016.[38] In our case, XGBoost affords the highest prediction accuracy among the nine tested machine learning models (Figure 2a), and is thereby selected for further analysis.

Most machine learning algorithms, especially the neural networks, have proven challenging to offer an explanation of the predicted results due to their so-called "black-box" nature. They work by fitting unknown functions via input and output data sets. The XGBoost model not only delivers the highest prediction accuracy, but also provides an out-of-the-box method to quantify the significance of the features or descriptors in making decisions. In this case, the feature importance scores calculated from the XGBoost model allow us to rank the reaction parameters that affect the crystallization propensity of MONCs, thus assisting in studying the reaction mechanism and accelerating the discovery of new MONCs crystals. The scores for the 17 total descriptors are shown in Figure 3a. It shows that the solvents (H_2O, DMF, and MeCN), organic ligands (PgC_x), modulators (molar mass, pK_a, and mole), and cations (molar mass and radius) are the dominant factors in the formation of single-crystal MONCs. Among them, water is the most significant one since it tunes the solvent polarity and is involved in the coordination of metal ions for promoting crystallization. Properties of the modulators such as molar mass and pK_a values indicate the deprotonation capability of PgC_x, making it a secondary factor. The model also shows that as an unfavorable solvent when mixed with a favorable solvent such as DMF, MeCN plays a significant role in determining the crystallization propensity. In addition, the length of the alkyl chains indicated by the molar mass greatly affects the hydrophobicity and solubility of MONCs in solvents, leading to various crystallization behaviors. Cations with different molar mass and radii display various coordination capability and affect the solubility of the MONCs. However, they are less significant than the ligands and modulators in affecting the crystallization of the MONCs.

The relative importance of these reaction parameters for synthesizing the MONC crystals agree well with a well-trained chemist's intuition. However, they are very challenging to quantify by human or traditional analysis methods. XGBoost affords a simple but straightforward way of achieving it. To further investigate whether the number of descriptors affects the prediction performance, the XGBoost models based on the top 15, 12, 9, and 6 descriptors as indicated in Figure 3a were also trained. The predictive accuracy from each model was compared and shown in Figure 3b. They are almost constant with very little variance as the number of descriptors decreases from 17 to 6. This result shows that even with the top 6 descriptors including volume of water, molar mass of modulators, molar mass of PgC_x, volume of acetonitrile, the pK_a value of modulators, and molar mass of cations, the XGBoost model is robust enough to afford satisfactory prediction results. In addition, comparison on the prediction accuracy of all machine learning models trained with the top 6 descriptors was shown in Figure S2. It is found that as the number of the descriptors is reduced to six, the predictive accuracy afforded from all nine machine learning models is more or less decreased. It is also shown that the decision tree ensemble methods (RF, ADA, and XGBoost) exhibited higher prediction accuracy (>0.83) than the remaining machine learning models. Even though XGBoost and ADA iteratively increased the accuracy by building each classifier tree in theory, the further increase became difficult because the models lack sufficient information for learning from the reduced number of the descriptors.

In order to gain more chemical insight, a flowchart was derived from the XGBoost model as shown in Figure 4a. It exhibits how the decision is made in classifying the reaction outcomes according to the input reaction parameters.[24] The tree was first divided by valence of the cations into two branches. The left branch has valence of <3, and the right one has a valence of ≥3. From the right branch, the reaction outcomes are then decided by the radii of cations (shown in green). However, the left branch with valence of <3 can be further divided according to volume of MeCN and radii of cations. The reactions with MeCN of ≥0.75 and cation radius of <0.725 nm were subsequently determined by the molar mass of PgC_x and the pK_a value of modulators (shown in pale blue), while the reactions involving cations with molar masses between 53.47 g/mol and 61.24 g/mol tend to form MONCs crystals (shown in orange). From this decision tree, one can

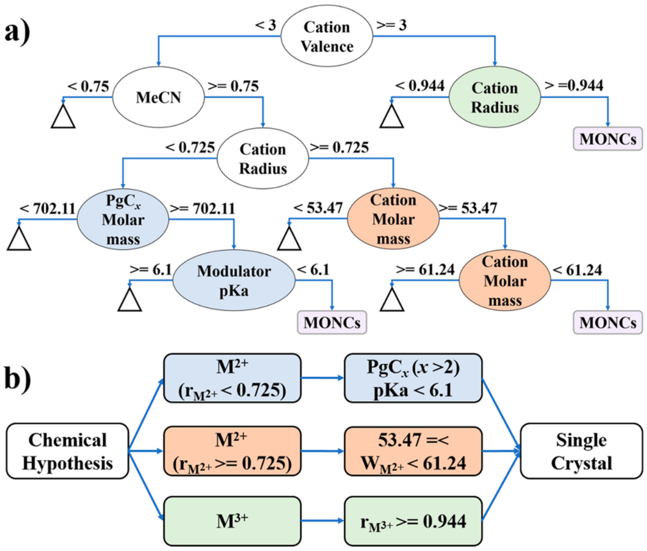


Figure 4. (a) Visualization of a decision tree from the XGBoost model for classifying single-crystals and nonsingle-crystals of MONCs. Ovals show decision nodes, rectangles show result bins, and triangles show excised subtrees. (b) Graphical representation of three hypotheses generated from the XGBoost model.

extract chemical hypotheses that include some important criteria for guiding the synthesis of the MONC single crystals. As an example shown in Figure 4b, one can deduce that the valence of the metal ions is important in the final reaction outcomes. If MONCs are crystallized from M^{2+} metal ions with radii of <0.725 nm (specifically here Ni^{2+} and Mg^{2+}), then PgC_x with x larger than 2 and modulators with pK_a of <6.1 should be provided. If the radii of the M^{2+} cations (e.g., Co^{2+} and Mn^{2+}) increase, e.g. $\geq$ 0.725 nm, then their molar mass should be between 53.47 g/mol and 61.24 g/mol to promote the crystallization. If the valence of the cations increases to 3 (M^{3+}), then they should have much larger radii (e.g., $\geq$ 0.944 nm, here Sm^{3+}) to obtain a better crystallization propensity. This new hidden information extracted from the XGBoost model is very valuable. It can assist the chemists to faster search for the optimal reaction parameters from many experimental variables, whose features can be hidden in the high-dimensional space.

Recently, Moghadam et al. introduced artificial neural networks (ANN) to rapidly predict the bulk modulus of MOFs and to illustrate effect of the structural information on the mechanical stability, which has the potential to vastly accelerate the industrial applications of MOFs in the coming years.[44] Distinguished from the Moghadam's computational work, a work conducted by Moosavi and his colleagues reported successful synthesis of MOFs (HKUST-1) with the highest surface area with the aid of chemical intuition captured from the machine learning models trained by a set of partially failed experiments.[27] Hence, it becomes practical that machine learning helps to guide the experimental synthesis and accelerate the discovery of MONCs. To compare the performance of the XGBoost model with a well-trained chemist in predicting crystallization propensity of the MONCs, 20 new validation experiments, which do not appear in the training or testing data sets, were conceived and implemented (Table S5). These 20 experiments can be categorized into three classes according to the above proposed chemical hypotheses: (i) Ni^{2+}/Mg^{2+}, (ii) Co^{2+}/Mn^{2+}, and (iii) Zn^{2+}. In the first class (Ni^{2+}/Mg^{2+}) of experiments, all synthetic parameters were designed to meet the requirement for the formation of MONC single crystals ($r_M{}^{2+}$ < 0.725 nm, PgC_x with x > 2, and pK_a of modulator <6.1). In the second class (Co^{2+}/Mn^{2+}), three experiments (No. 13, 15 and 19) were designed to confirm the importance of acetonitrile. In the last class, i.e., Zn^{2+}, the experiments were designed to fail since no recommended cations or specific experimental conditions were included. The synthesis parameters of these 20 experiments were first presented to both a skilled chemist and XGBoost for predicting reaction outcomes. Then the experiments were conducted by a chemist. The final reaction outcomes served as the benchmark to evaluate the prediction accuracy made by the chemist and XGBoost. The XGBoost model predicted the outcomes with an accuracy of 80%, which

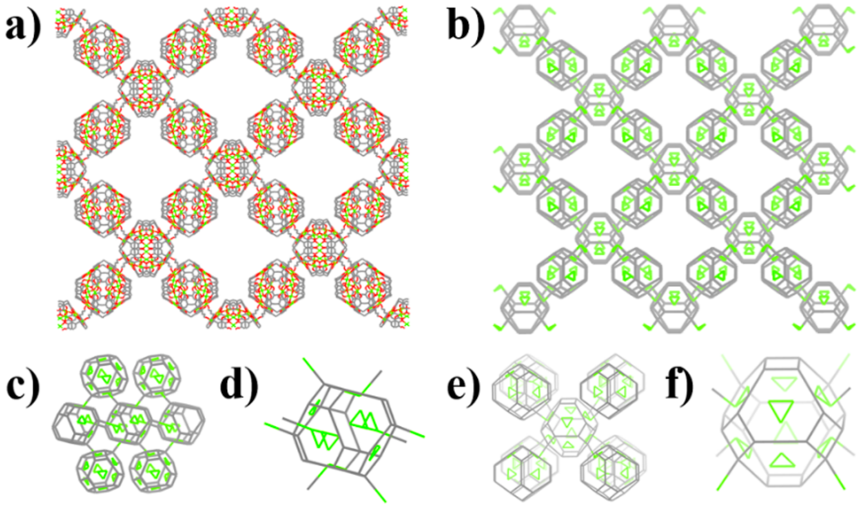


Figure 5. (a) Crystal structure and (b) network topology of SCP-4viewed along [001] direction. (c) Connection and (d) coordination mode of Type-A nanocapsules viewed along [110] direction. (e) Connection and (f) coordination mode of Type-B nanocapsules viewed along [001] direction. All hydrogen atoms and alkyl tails that do not participate in linking the nanocapsules have been omitted. Axial water molecules that coordinate to metal ions are also removed. Color codes: carbon, gray; oxygen, red; nitrogen, blue; and metal, green.

DOI: 10.1021/jacs.9b11569

is higher than that of the chemist (75%). Four unexpected failure (No. 6, 8, 9, and 17) were identified. This relatively lower accuracy of the XGBoost model compared to the one when trained with historical experiments could be due to its insufficient generalization, which is more or less influenced by a few potential exceptions (e.g., rare items in a majority of data),[45,46] or it could be due to unexpected experimental failures. Nevertheless, we believe that as the first proof-of-concept for predicting the crystallization propensity of MONCs, the XGBoost model shows great potential in guiding chemists, especially new entrants, to screen the reaction parameters for synthesizing new MONCs single-crystals.

Among the reactions that produced single crystals, a new compound SCP-4 was found (No. 2 in Table S5). This structural analysis reveals that SCP-4 is a 3D assembly of $Mg_{24}L_6$ nanocapsules (Figure 5). SCP-4 consists of two types of nanocapsules within the framework (Figure 5a,b). Along the [110] direction, each Type-A nanocapsule is connected to four Type-B nanocapsules via single alkyl chains and two Type-A nanocapsules via double alkyl chains at the (10) plane (Figure 5c). Viewed from the [001] direction, we can observe that each Type-B nanocapsule is linked with eight Type-A nanocapsules via single alkyl chains (Figure 5e). Both Type-A and Type-B nanocapsules provide 4 metal sites and 4 alkyl chains for linking, employing a “4 in 4 out” coordination mode (Figure 5d,f). Along with other supramolecular coordination polymers composed of giant $M_{24}L_6$ as building blocks,[8,9] SCP-4 exhibits the versatility of using MONCs to construct hierarchical supramolecular structures. Although the machine learning models have not yet enabled the prediction of detailed structures, they provide a tool for initially screening the possibility of crystallization, thereby offering a major advance in the synthesis of MONCs.

■ CONCLUSIONS

In summary, for the first time, this work reports a machine-learning assisted method to predict the crystallization propensity of MONCs using experimental data. The highest prediction accuracy using the XGBoost model was >91%. In addition, with the assistance of the derived important features and chemical hypotheses from the XGBoost model, we successfully synthesized a set of new crystalline MONCs. This work will shed light on the discovery of new crystalline materials by integrating human intuition and machine learning techniques. The extension of these machine learning models to other organic and inorganic materials is anticipated by developing the corresponding descriptors and fine-tuning the hyperparameters. Finally, integrating the developed machine learning models with high-throughput synthesis (i.e., robotic synthesis platforms) would greatly accelerate the development of inorganic−organic hybrid materials.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.9b11569.

Crystallographic data (CIF)

Figure S1, scheme of dataset splitting through randomly training/test splitting and hyperparameter tuning by using 5-fold GridSearchCV procedure; Figure S2, comparison of predictive accuracy of various machine learning models trained with the top 6 descriptors determined by XGBoost; Table S1, detailed chemical descriptors for machine learning models; Table S2, hyperparameters machine learning models in a single-shot trial; Table S3, comparison of evaluation metrics of nine tested machine learning models; Table S4, confusion matrix of SVM, RF, XGBoost, and MLP models on training and test datasets; and Table S5, comparison of out-of-sample prediction results made by a skillful chemist and the XGBoost model as well as actual reaction outcomes (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*atwoodj@missouri.edu
*linjian@missouri.edu

ORCID

Yunchao Xie: 0000-0001-6216-1211
Chen Zhang: 0000-0001-5552-1960
Xiangquan Hu: 0000-0001-7068-8558
Steven P. Kelley: 0000-0001-6755-4495
Jerry L. Atwood: 0000-0002-3350-9618
Jian Lin: 0000-0002-4675-2529

Author Contributions

#Authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by grants from National Science Foundation (award numbers: 1825352 and 1933861), United States Department of Energy National Energy Technology Laboratory (award number: DE-FE0031645), and Unite Sates Department of Agriculture (award number: 2018-67017-27880). The authors also thank the University of Missouri for financial and research facility support of this work.

■ REFERENCES

(1) Kaphan, D. M.; Levin, M. D.; Bergman, R. G.; Raymond, K. N.; Toste, F. D. A supramolecular microenvironment strategy for transition metal catalysis. *Science* **2015**, *350*, 1235−1238.

(2) Zhang, C.; Zhang, C.; Xie, Y.; Su, J.-W.; He, X.; Demaree, J. D.; Griep, M. H.; Atwood, J. L.; Lin, J. A Supramolecular Coordination-Polymer-Derived Electrocatalyst for the Oxygen Evolution Reaction. *Chem. - Eur. J.* **2019**, *25*, 4036−4039.

(3) Furukawa, H.; Cordova, K. E.; O’Keeffe, M.; Yaghi, O. M. The Chemistry and Applications of Metal-Organic Frameworks. *Science* **2013**, *341*, 1230444.

(4) Liu, T.-F.; Chen, Y.-P.; Yakovenko, A. A.; Zhou, H.-C. Interconversion between Discrete and a Chain of Nanocages: Self-Assembly *via* a Solvent-Driven, Dimension-Augmentation Strategy. *J. Am. Chem. Soc.* **2012**, *134*, 17358−17361.

(5) Liu, C.; Luo, T.-Y.; Feura, E. S.; Zhang, C.; Rosi, N. L. Orthogonal Ternary Functionalization of a Mesoporous Metal−Organic Framework *via* Sequential Postsynthetic Ligand Exchange. *J. Am. Chem. Soc.* **2015**, *137*, 10508−10511.

(6) Patil, R. S.; Banerjee, D.; Zhang, C.; Thallapally, P. K.; Atwood, J. L. Selective CO_2 Adsorption in a Supramolecular Organic Framework. *Angew. Chem., Int. Ed.* **2016**, *55*, 4523−4526.

(7) Zhang, M.; Feng, G.; Song, Z.; Zhou, Y.-P.; Chao, H.-Y.; Yuan, D.; Tan, T. T. Y.; Guo, Z.; Hu, Z.; Tang, B. Z.; Liu, B.; Zhao, D. Two-Dimensional Metal−Organic Framework with Wide Channels and Responsive Turn-On Fluorescence for the Chemical Sensing of

Volatile Organic Compounds. *J. Am. Chem. Soc.* **2014**, *136*, 7241−7244.

(8) Zhang, C.; Wang, F.; Patil, R. S.; Barnes, C. L.; Li, T.; Atwood, J. L. Hierarchical Self-Assembly of Supramolecular Coordination Polymers Using Giant Metal−Organic Nanocapsules as Building Blocks. *Chem. - Eur. J.* **2018**, *24*, 14335−14340.

(9) Zhang, C.; Patil, R. S.; Liu, C.; Barnes, C. L.; Atwood, J. L. Controlled 2D Assembly of Nickel-Seamed Hexameric Pyrogallol[4]-arene Nanocapsules. *J. Am. Chem. Soc.* **2017**, *139*, 2920−2923.

(10) Zhang, C.; Patil, R. S.; Li, T.; Barnes, C. L.; Atwood, J. L. Self-assembly of magnesium-seamed hexameric pyrogallol[4]arene nanocapsules. *Chem. Commun.* **2017**, *53*, 4312−4314.

(11) Kumari, H.; Dennis, C. L.; Mossine, A. V.; Deakyne, C. A.; Atwood, J. L. Exploring the Magnetic Behavior of Nickel-Coordinated Pyrogallol[4]arene Nanocapsules. *ACS Nano* **2012**, *6*, 272−275.

(12) McKinlay, R. M.; Cave, G. W. V.; Atwood, J. L. Supramolecular blueprint approach to metal-coordinated capsules. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 5944−5948.

(13) Power, N. P.; Dalgarno, S. J.; Atwood, J. L. Guest and Ligand Behavior in Zinc-Seamed Pyrogallol[4]arene Molecular Capsules. *Angew. Chem., Int. Ed.* **2007**, *46*, 8601−8604.

(14) Zhang, C.; Sikligar, K.; Patil, R. S.; Barnes, C. L.; Baker, G. A.; Atwood, J. L. A $M_{18}L_6$ metal-organic nanocapsule with open windows using mixed macrocycles. *Chem. Commun.* **2018**, *54*, 635−637.

(15) Fowler, D. A.; Rathnayake, A. S.; Kennedy, S.; Kumari, H.; Beavers, C. M.; Teat, S. J.; Atwood, J. L. Introducing Defects into Metal-Seamed Nanocapsules Using Mixed Macrocycles. *J. Am. Chem. Soc.* **2013**, *135*, 12184−12187.

(16) Zhang, C.; Patil, R. S.; Atwood, J. L. Chapter Five—Metallosupramolecular Complexes Based on Pyrogallol[4]arenes. In *Advances in Inorganic Chemistry*; van Eldik, R., Puchta, R., Eds.; Academic Press: 2018; Vol. *71*, pp 247−276.

(17) Stranks, S. D.; Snaith, H. J. Metal-halide perovskites for photovoltaic and light-emitting devices. *Nat. Nanotechnol.* **2015**, *10*, 391.

(18) Zhou, H.-C.; Long, J. R.; Yaghi, O. M. Introduction to Metal−Organic Frameworks. *Chem. Rev.* **2012**, *112*, 673−674.

(19) Le, T. C.; Winkler, D. A. Discovery and Optimization of Materials Using Evolutionary Approaches. *Chem. Rev.* **2016**, *116*, 6107−6132.

(20) Henson, A. B.; Gromski, P. S.; Cronin, L. Designing Algorithms To Aid Discovery by Chemical Robots. *ACS Cent. Sci.* **2018**, *4*, 793−804.

(21) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2016**, *2*, 16028.

(22) Dong, Y.; Wu, C.; Zhang, C.; Liu, Y.; Cheng, J.; Lin, J. Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride. *npj Comput. Mater.* **2019**, *5*, 26.

(23) Oliynyk, A. O.; Adutwum, L. A.; Rudyk, B. W.; Pisavadia, H.; Lotfi, S.; Hlukhyy, V.; Harynuk, J. J.; Mar, A.; Brgoch, J. Disentangling Structural Confusion through Machine Learning: Structure Prediction and Polymorphism of Equiatomic Ternary Phases ABC. *J. Am. Chem. Soc.* **2017**, *139*, 17870−17881.

(24) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73−76.

(25) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C−N cross-coupling using machine learning. *Science* **2018**, *360*, 186−190.

(26) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, *559*, 377−381.

(27) Moosavi, S. M.; Chidambaram, A.; Talirz, L.; Haranczyk, M.; Stylianou, K. C.; Smit, B. Capturing chemical intuition in synthesis of metal-organic frameworks. *Nat. Commun.* **2019**, *10*, 539.

(28) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.

(29) Voznyy, O.; Levina, L.; Fan, J. Z.; Askerka, M.; Jain, A.; Choi, M.-J.; Ouellette, O.; Todorović, P.; Sagar, L. K.; Sargent, E. H. Machine Learning Accelerates Discovery of Optimal Colloidal Quantum Dot Synthesis. *ACS Nano* **2019**, *13*, 11122−11128.

(30) Zhang, C.; Patil, R. S.; Li, T.; Barnes, C. L.; Teat, S. J.; Atwood, J. L. Preparation of Magnesium-Seamed C-Alkylpyrogallol[4]arene Nanocapsules with Varying Chain Lengths. *Chem. - Eur. J.* **2017**, *23*, 8520−8524.

(31) Schultz, C.; Alegría, A. C.; Cornelis, J.; Sahli, H. Comparison of spatial and aspatial logistic regression models for landmine risk mapping. *Appl. Geogr.* **2016**, *66*, 52−63.

(32) Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian Network Classifiers. *Mach. Learn.* **1997**, *29*, 131−163.

(33) Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21−27.

(34) Burges, C. J. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discovery* **1998**, *2*, 121−167.

(35) Ture, M.; Tokatli, F.; Kurt, I. Using Kaplan−Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. *Expert Syst. Appl.* **2009**, *36*, 2017−2026.

(36) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(37) Freund, Y.; Schapire, R.; Abe, N. A short introduction to boosting. *Jpn. Soc. Aritif. Intell.* **1999**, *37*, 277−296.

(38) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: San Francisco, California, 2016; pp 785−794.

(39) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.

(40) Ren, F.; Ward, L.; Williams, T.; Laws, K. J.; Wolverton, C.; Hattrick-Simpers, J.; Mehta, A. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **2018**, *4*, No. eaaq1566.

(41) Obuchowski, N. A. Receiver Operating Characteristic Curves and Their Use in Radiology. *Radiology* **2003**, *229*, 3−8.

(42) Fawcett, T. An introduction to ROC analysis. *Pattern Recog. Lett.* **2006**, *27*, 861−874.

(43) Li, F.; Han, J.; Cao, T.; Lam, W.; Fan, B.; Tang, W.; Chen, S.; Fok, K. L.; Li, L. Design of self-assembly dipeptide hydrogels and machine learning via their chemical features. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 11259−11264.

(44) Moghadam, P. Z.; Rogge, S. M. J.; Li, A.; Chow, C.-M.; Wieme, J.; Moharrami, N.; Aragones-Anglada, M.; Conduit, G.; Gomez-Gualdron, D. A.; Van Speybroeck, V.; Fairen-Jimenez, D. Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. *Matter* **2019**, *1*, 219−234.

(45) Lee, M.-H. Insights from Machine Learning Techniques for Predicting the Efficiency of Fullerene Derivatives-Based Ternary Organic Solar Cells at Ternary Blend Design. *Adv. Energy Mater.* **2019**, *9*, 1900891.

(46) Li, Q.; Fan, S.; Chen, C. An Intelligent Segmentation and Diagnosis Method for Diabetic Retinopathy Based on Improved U-NET Network. *J. Med. System.* **2019**, *43*, 304.

DOI: 10.1021/jacs.9b11569