Supporting Information for

Machine Learning Assisted Synthesis of Metal-Organic Nanocapsules

Yunchao Xie,[1] Chen Zhang,[2] Xiangquan Hu,[2] Chi Zhang,[1] Steven P Kelley,[2] Jerry L. Atwood,[2,*] and Jian Lin[1,3,4,*]

[1] Department of Mechanical Engineering, University of Missouri, Columbia MO 65211 USA

[2] Department of Chemistry, University of Missouri, Columbia MO 65211 USA

[3]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia MO 65211 USA

[4]Department of Physics and Astronomy, University of Missouri, Columbia MO 65211 USA

Supplementary Figures and Tables

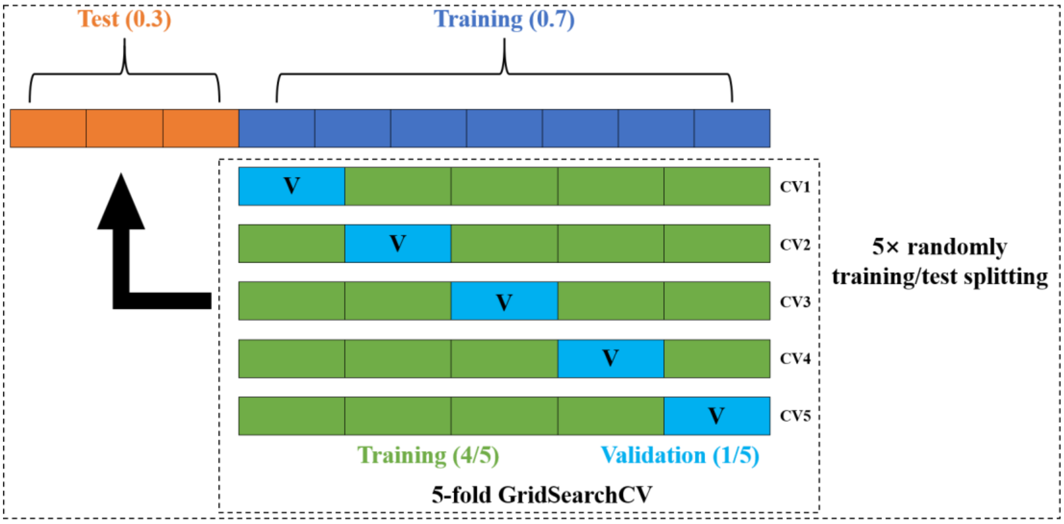


Figure S1. Scheme of dataset splitting through randomly training/test splitting and hyperparameter tuning *via* 5-fold GridSearchCV. This process was repeated 5 times through changing random seed and the average of evaluation metrics was reported.

Table S1. Detailed chemical descriptors (features) for ML models.

Parameters	PgC_x	Metal salts		Solvents	Modulators	Temperature
		Cations	Anions			
Examples	PgC_1, PgC_2, PgC_3, PgC_4, PgC_5, PgC_6, PgC_7, PgC_8, PgC_9, PgC_3OH (x represents the number of carbon atoms in the alkyl tail)	Al^{3+}, Ca^{2+}, Cd^{2+}, Co^{2+}, Cr^{2+}, Cr^{3+}, Cu^{2+}, Er^{3+}, Fe^{2+}, Fe^{3+}, Ga^{3+}, Gd^{3+}, Lu^{3+}, Mg^{2+}, Mn^{2+}, Ni^{2+}, Sm^{3+}, Sr^{2+}, Tb^{3+}, V^{3+}, Zn^{2+}	NO_3^-, Cl^-	(10) DMF: dimethyl formamide (11) MeOH: methanol (12) MeCN: Acetonitrile (13) H_2O	Arginine, Aspartic acid, Benzoic acid, Cysteine, Glycine, Histidine, Imidazole, Lysine, Proline, Pyridine, Serine, Sodium ethoxide, Triethylamine, Tyrosine	100 °C, 110 °C, 120 °C, 130 °C
Chemical descriptors	1.Molar mass 2.Carbon length 3.Hydroxyl group	4.Molar mass 5.Charge 6.Radius	7.Molar mass 8.Charge 9.Radius	10/11/12/13. Solvent volume	14.Molar mass 15.pKa 16.Mole	17. Temperature

Note:

Before these datasets are fed into training machine learning model, the values of chemical descriptors were first scaled within 0 and 1 to make sure chemical descriptors have the same numeric scale and can be equally treated. The scaled procedure is calculated according to the following formula:

$$x' = \frac{x_i - x_{\min,i}}{x_{\max,i} - x_{\min,i}}$$

where the x', x_i, $x_{\max,i}$, and $x_{\max,i}$ refer to scaled, original, maximum and minimum value of a selected chemical descriptor, respectively.

For example, the molar mass of PgC_x has the maximum and minimum value of 608.6 and 1057.5, respectively. If we want to do the scaler on a given value, i.e., 720.8, we just substitute the value into the above formula and get 0.250. Following is a list showing how the values changed after scaling process.

PgC_x	PgC_1	PgC_2	PgC_3	PgC_3OH	PgC_4	PgC_5	PgC_6	PgC_7	PgC_8	PgC_9
Molar mass	608.6	664.7	720.8	776.9	784.8	833.0	889.1	945.2	1001.3	1057.5
Scaled	0	0.125	0.250	0.375	0.393	0.500	0.625	0.750	0.875	1

Table S2. Hyperparameters for four machine learning models in a single-shot trial.

Model	Full Name	Hyperparameters
LR	Logistic Regression	-
GNB	Gaussian Naïve Bayes	-
DT	Decision Tree	max_depth = 5
SVM	Support Vector Machine	C = 1000, kernel = 'linear'
RF	Random Forest	n_estimators = 200, max_depth = 7
KNN	k-Nearest Neighbors	n_neighbors = 1
ADA	Adaptive Boosting	DecisionTreeClassifier(max_depth = 5), n_estimators = 200, learning_rate = 0.001
XGB	eXtreme Gradient Boosting	max_depth= 5, learning_rate=0.1, n_estimators = 200, subsample = 0.7
MLP	Deep neural networks	3 layers, dense = 40/40/2, input_dim = 17, activation = ‘relu’/’relu’/’softmax’, loss = ‘categorical_crossentropy’, optimizer = ‘adam’

Note: the hyperparameters of each model not explicitly listed above were set to its defaults in the scikit-learn package.

Table S3. Comparison of evaluation metrics of the four machine learning models.

Model	Accuracy		Test			
	Training	Test	AUC	Precision	Recall	F_1
LR	0.87	0.83	0.87	0.81	0.82	0.81
GNB	0.78	0.82	0.86	0.83	0.84	0.81
KNN	0.996	0.84	0.87	0.84	0.85	0.84
SVM	0.83	0.82	0.88	0.83	0.84	0.82
DT	0.90	0.84	0.91	0.83	0.84	0.83
RF	0.90	0.85	0.96	0.85	0.87	0.85
ADA	0.93	0.84	0.95	0.83	0.84	0.83
XGB	**0.95**	**0.87**	**0.97**	**0.87**	**0.87**	**0.87**
MLP	0.94	0.85	0.96	0.85	0.84	0.85

It includes accuracy scores of training and test datasets. Area under receiver operating characteristic (AUC), precision, recall and F_1 scores of test datasets. All values were averaged of multiple random splitting with GridSearchCV.

Note:

Several metrics including accuracy, precision, recall, and F1 score are calculated to evaluate the performance of machine learning models and shown in **Equation** 1 – 4.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

where TP, FP, TN, and FN represent the number of true positives, the number of false positives, the number of true negatives, and the number of false negatives. *F_1* score represents the harmonic mean of precision and recall.

The receiver operating characteristic (ROC) curves were plotted with true positive rate (TPR, **Equation** 5) against false positive rate (FPR, **Equation** 6) at various decision thresholds. The area under a ROC curve (AUC) was also calculated, which indicates the ability of ML models to separate different classes. The precision-recall (PR) curves were show the plot of precision (**Equation** 2) versus recall (**Equation** 3) at different threshold settings.

$$TPR = \frac{TP}{TP + FN} \quad (5)$$

$$FPR = \frac{FP}{FP + TN} \quad (6)$$

Table S4. Confusion matrix of SVM, RF, XGB and MLP models on training and test datasets. (SC: Single Crystal, NSC: Non-Single Crystal)

SVM		Predicted			
		Training		Test	
		NSC	SC	NSC	SC
Actual	NSC	TN 156	FP 49	TN 70	FP 18
	SC	FN 5	TP 130	FN 8	TP 50

RF		Predicted			
		Training		Test	
		NSC	SC	NSC	SC
Actual	NSC	TN 171	FP 34	TN 77	FP 11
	SC	FN 1	TP 134	FN 5	TP 53

XGB		Predicted			
		Training		Test	
		NSC	SC	NSC	SC
Actual	NSC	TN 192	FP 13	TN 77	FP 11
	SC	FN 7	TP 128	FN 4	TP 54

MLP		Predicted			
		Training		Test	
		NSC	SC	NSC	SC
Actual	NSC	TN 200	FP 5	TN 85	FP 3
	SC	FN 7	TP 128	FN 15	TP 43

Table S5. Comparison of out-of-sample prediction results made by a skillful chemist and XGB model as well as actual reaction outcomes.

Class	#	PgC_x	Metal salts		DMF	MeCN	H_2O	Modulator	T	Prediction		Results
			Cation	Anion	(mL)	(mL)	(mL)		(ºC)	Chemist	XGB	
1	1	PgC_3OH	Mg^{2+}	NO_3^-	0.5	2.0	1.0	BC	130	1	1	1
	2	PgC_3OH	Mg^{2+}	NO_3^-	1.0	2.0	0.1	BC	130	1	1	1
	3	PgC_3OH	Mg^{2+}	NO_3^-	1.0	2.0	0.5	BC	130	1	1	1
	4	PgC_3OH	Mg^{2+}	NO_3^-	2.0	1.0	0.5	BC	130	1	1	1
	5	PgC_3OH	Mg^{2+}	NO_3^-	2.0	1.0	0.5	BC	100	0	**1**	0
	6	PgC_4	Mg^{2+}	NO_3^-	2.0	1.0	0.1	BC	130	**0**	1	1
	7	PgC_3OH	Mg^{2+}	Cl^-	1.0	2.0	0.5	BC	100	**0**	1	1
	8	PgC_3OH	Mg^{2+}	Cl^-	2.0	1.0	0.5	BC	130	1	**1**	0
	9	PgC_4	Mg^{2+}	Cl^-	2.0	1.0	0.1	BC	100	**0**	**1**	0
	10	PgC_3OH	Ni^{2+}	NO_3^-	2.0	1.0	0.1	BC	130	1	1	1
	11	PgC_3OH	Ni^{2+}	NO_3^-	2.0	1.0	0.5	BC	130	1	1	1
	12	PgC_4	Ni^{2+}	NO_3^-	2.0	1.0	0.5	PY	100	1	1	1
2	13	PgC_3OH	Co^{2+}	NO_3^-	2.0	0.5	1.0	PY	130	0	0	0
	14	PgC_4	Co^{2+}	NO_3^-	1.0	2.0	0.1	PY	100	1	1	1
	15	PgC_4	Co^{2+}	NO_3^-	2.0	0.5	1.0	PY	130	0	0	0
	16	PgC_4	Co^{2+}	NO_3^-	2.0	1.0	0.1	PY	130	**0**	1	1
	17	PgC_4	Co^{2+}	NO_3^-	2.0	1.0	0.5	PY	100	**0**	**1**	0
	18	PgC_3OH	Mn^{2+}	NO_3^-	1.0	2.0	0.5	BC	130	**0**	1	1
	19	PgC_4	Mn^{2+}	NO_3^-	2.0	0.5	1.0	PY	130	0	0	0
3	20	PgC_3OH	Zn^{2+}	NO_3^-	2.0	0.5	1.0	PY	130	0	0	0
	Predictive accuracy									75 %	80 %	

Note: For all 20 experiments, modulators including benzoic acid (BC) and pyridine (PY) were added and the mole of BC and PY were 0.3 mmol and 1.2 mmol, respectively.

Machine learning Assisted Synthesis of Metal-Organic Nanocapsules

Yunchao Xie,[1] Chen Zhang,[2] Xiangquan Hu,[2] Chi Zhang,[1] Steven P Kelley,[2] Jerry L. Atwood,[2,*] and Jian Lin[1,3,4,*]

[1] Department of Mechanical Engineering, University of Missouri, Columbia MO 65211 USA

[2] Department of Chemistry, University of Missouri, Columbia MO 65211 USA

[3]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia MO 65211 USA

[4]Department of Physics and Astronomy, University of Missouri, Columbia MO 65211 USA

ABSTRACT

Herein, we report the successful discovery of a new hierarchical structure of metal-organic nanocapsules (MONCs) by integrating chemical intuition and machine learning algorithms. By training datasets from a set of both succeeded and failed experiments, we studied the crystallization propensity of metal-organic nanocapsules (MONCs). Among four machine learning models, XGB model affords the highest prediction accuracy of 91%. The derived chemical feature scores and chemical hypothesis from the XGB model assist to identify proper synthesis parameters showing superior performance to a well-trained chemist. This paper will shed light on the discovery of new crystalline inorganic-organic hybrid materials guided by machine learning algorithms.

Keywords: machine learning, metal-organic nanocapsules, crystallization propensity, XGB

1. INTRODUCTION

Metal-organic nanocapsules (MONCs) have aroused a surge of interests due to their potential applications in many different fields including catalysis,[1] gas adsorption and separation,[2-5] and sensing.[6] These MONCs can further self-assemble into hierarchical structures.[7-8] 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 ($P_3R_1C_x$),[9-15] where *x* is the number of carbon atoms in the alkyl tail. 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[16] and metal-organic frameworks (MOFs).[17] The solvothermal crystallization of the MONCs is primarily an exploratory process. It involves three major steps. First, chemical space containing all synthesis parameters for the synthesis such as metal ions, organic ligands, solvents, and temperature is included. 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 synthesis 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 individuals. Although genetic algorithms have been explored to search the chemical space,[18-19] its size is too overwhelming to be all researched and tested. Thus, smart navigation based on surrogate models is quite desired.

In recent years, machine learning, which enables to provide surrogate algorithms for material development, has gained enormous attention for effectively predicting the physiochemical properties,[20] establishing the structure-property relationships,[21-22] and navigating chemical space for guiding chemical synthesis.[23-28] For instance, Raccuglia et al. reported applying support vector machine algorithm to exploit chemical space from historical successful and failed experiments for elucidating factors that govern reaction outcomes.[23] Doyle et al. demonstrated the successful application of random forest regression algorithm to predict high-yielding conditions for untested substrates,[24] showing a result of a coefficient of determination R^2 value of 0.92. In Cronin and his colleagues' recent work, the well-trained neural networks could predict the reactivity of more than 1000 reaction combinations with accuracy of greater than 80%.[25] Despite such progress, application of the machine learning algorithms to guide synthesis of inorganic-organic hybrid materials has still been quite limited.[23, 26] So far, to the best of our knowledge, exploiting machine learning algorithms for MONCs synthesis has not yet been reported. The miserable failure of human intuition in high-dimensional problems lead to the difficulty in analysis of high-dimensional parameters, which makes it impossible in optimization of synthesis parameters. In addition, afforded chemical insights such as interpretable hypotheses and feature importance of the synthesis parameters from these reported machine learning models are still quite limited.

Herein, we propose to introduce well-trained machine learning models (e.g., Support Vector Machine, Random Forest, eXtreme Gradient Boosting, Multilayer Perceptron) to the traditional trial-error process of MONC synthesis and afford the highest prediction accuracy of 91% when predicting the crystallization propensity. The feature importance and chemical hypothesis derived from the XGB model assist to identify successful synthesis parameters for MONCs, showing a higher prediction accuracy than a well-trained chemist. The machine learning algorithm could learn the hidden discipline automatically from feeding descriptors, which provides a proper approach to accelerate the discovery of MONCs single-crystal. To the best of our knowledge, it is the first time to introduce machine learning algorithm into the specific field of MONCs for predicting the crystallization propensity. The developed machine learning models are envisioned to be easily extended to other synthesis systems.

2. EXPERIMENTAL SECTION

2.1 Synthesis of C-alkylpyrogallol[4]arene (PgC*x*). PgC$_x$ (x = 1-9) and PgC$_3$OH were synthesized using previously reported condensation reaction.[29] Taking PgC$_3$OH as an example, **2**,3-dihydrofuran (6.05 machine learning, 0.08 mol), and pyrogallol (0.08 mmol, 10 g) were mixed in 30 machine learning 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 hours. 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).

2.2 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. In a typical synthesis, the metal salts (nitrates or chlorides), PgC_x, and modulators were added into a mixture of *N*,*N*-dimethylformamide (DMF) / acetonitrile (MeCN) / H_2O in a 4 mL glass vial. The mixture was sonicated for 5 minutes and heated overnight at various temperatures in an oven. In addition, 20 new-designed experiments were conducted to validate the robustness of the machine learning models, which are listed in Table S5.

2.3 Synthesis of SCP-4 from reaction (No. 2 in Table S5). *C*-propan-3-olpyrogallol[4]arene (PgC_3OH, 0.1 mmol, 78.4 mg), $Mg(NO_3)_2 \cdot 6H_2O$ (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.

2.4 XRD characterization. The single crystal X-ray diffraction data was collected on a Brucker Apex II diffractometer at a temperature of 100 (2) K using CuKα (λ = 1.54056Å) radiation incotec Microfocus II.

2.5 Machine learning models. Total nine different machine learning algorithms, i.e., Logistic Regression (LR),[30] Gaussian Naïve Bayes (GNB),[31] k-Nearest Neighbors (KNN),[32]Support Vector Machine (SVM),[33] Decision Tree (DT),[34] Random Forest (RF),[35] Adaptive Boosting (ADA),[36] eXtreme Gradient Boosting (XGB),[37] and Multilayer Perceptron (MLP)[38] are trained on historical datasets for predicting crystallization propensity of MONCs. All machine learning models were directly programmed using Python with scikit-learn package.[35] We repeated five times random test/training splits to avoid sampling bias and the average of evaluation metrics was reported.

3. RESULTS AND DISCUSSION

Figure 1 shows the flow chart of predicting the crystallization propensity of MONCs with assistance of machine learning models. First, historical synthesis parameters from a total of 486 reactions including both successes and failures as well as reaction outcomes were collected from archived laboratory notebooks, and established as input and output for the machine learning models. 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, radius of cations and anions, valence of cations, and moles of anions), modulators (molar mass, pKa and moles), and experimental 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 length 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, pKa value, and mole number as the descriptors for the modulators because they can tune the deprotonation capability of PgC_x. The datasets are from total 486 experiments. They are categorized into two classes according to their reaction outcomes. Class "0" indicates the reaction outcomes of non-single crystals (293) while Class "1" indicates the reaction outcome of single crystals (193) at given input reaction parameters. Then, these datasets consisting of total 486 reactions with 17 descriptors were shuffled and split into training (70%) and test datasets (30%). The ratio of MONCs single-crystal to non-single-crystal samples was equally distributed in both training and test datasets. Five times random training/test splits were conducted to reduce sampling bias and the average of evaluation metrics were reported (Figure S1).

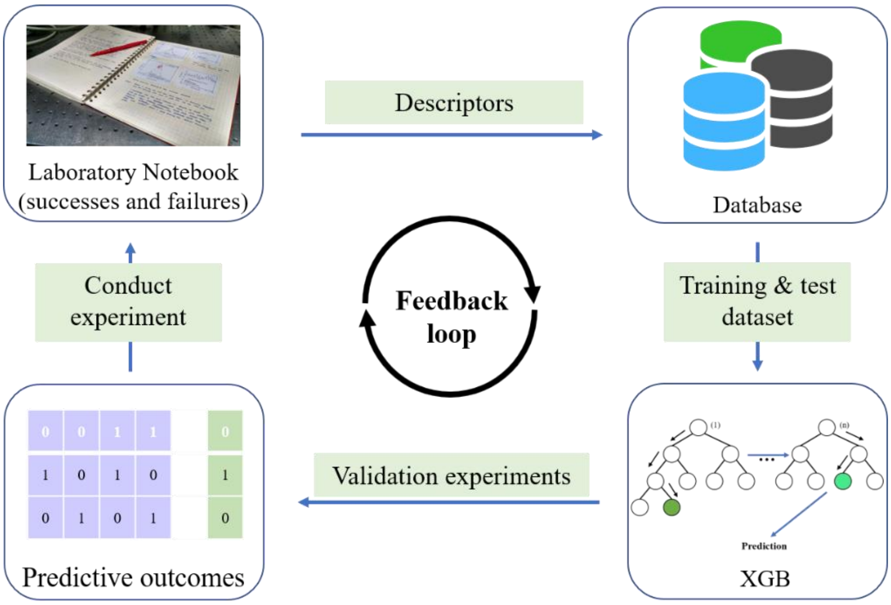


Figure 1. Schematic representation of working flow when machine learning models are incorporated in the prediction of crystallization propensity of MONCs.

After the database was established, a broad set of nine machine learning models, including Logistic Regression (LR),[30] Gaussian Naïve Bayes (GNB),[31] k-Nearest Neighbors (KNN),[32] Support Vector Machine (SVM),[33] Decision Tree (DT),[34] Random Forests (RF),[35] Adaptive Boosting (ADA),[36] eXtreme Gradient Boosting (XGB)[37] and Multilayer Perceptron (MLP)[38] were trained by a grid-search cross-validation (5-fold GridSearchCV) method and the hyperparameters of a single-shot trial was summarized in Table S2. The evaluation metrics including accuracy, precision, recall, F_1, receiver operating characteristic (ROC) curve were obtained by comparing the predicted results and the ground truths (Table S3). Finally, the XGB model with the highest prediction performance were used to predict reaction outcomes of new experiments. These well-trained machine learning models enable the extraction of important features that decide the reaction outcomes for recommending new synthesis parameters for the next experimental cycle, thus helping to generate human-interpretable hypotheses in the formation of single crystal MONCs.

Although these machine learning models offer individual advantages, such as high accuracy for classification, easiness to operate or good interpretability, they must be weighed carefully for a new application. We evaluated their performance with a goal of finding the one that shows both high prediction accuracy and easy interpretation.[20] It can be seen from Figure 2a and Table S3 that all of nine machine learning models can reach accuracy of > 82% and F1 score of > 81%. Among these machine learning models, XGB exhibited the superior performance, showing the highest accuracy of 91% with an average of 87% and average F1 score of 87%. As shown in the confusion matrix (Table S4), the XGB model shows the highest recall of 0.931

among four machine learning models (SVC: 0.862, RF: 0.914, MLP: 0.741), which indicates the highest true positive numbers.

Four representative models including XGB and other three models (SVM, RF, and MLP) were discussed in detail due to their prediction accuracy, easy to operate and good interpretability. A ROC curve indicates the relationship of 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.[39-40] More deviation of a 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. XGB shows the most deviated ROC curve compared to those of SVM, RF and MLP, indicating the 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.[41] XGB 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 XGB, SVM, RF and MLP were employed as an additional indicator in evaluating prediction performance (Figure 2c). Precision shows the ratio of correctly predicted true positive numbers to total predicted positive numbers, while recall indicates the fraction of correctly predicted true positive in the total real positive numbers.[42] XGB achieved precision of 0.9 at the recall of 0.9, which is much higher than to those for SVM (0.73), RF (0.84) and MLP (0.83) at the same recall value.

Different machine learning models present prediction results according to the built-in algorithms. Knowing the difference among them helps us to choose the best one suitable for our application. The relatively high flexibility of SVM usually results in limited and uncontrolled performance. Furthermore, SVM requires extensive experience to appropriately tune the hyperparameters. Both RF and XGB are based on the decision tree. 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 hyperparameters. However, in contrast to RF which makes final decision according to the final majority vote of each classifier tree, XGB is a gradient boosting model which builds each classifier tree sequentially to iteratively reduce the error of the established classifier trees. Hence, it has become one of the most widely used machine learning algorithms since introduced in 2016.[37] In our case, XGB affords the highest prediction accuracy among nine tested machine learning models (Figure 2a), thereby is selected for further analysis.[37, 43-44]

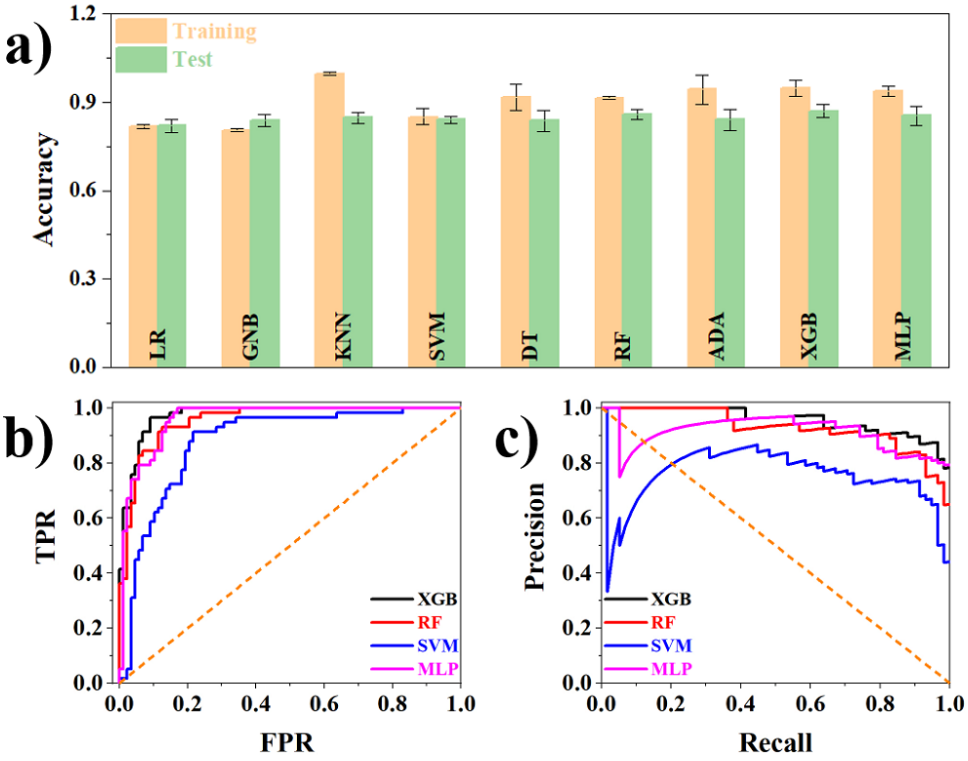


Figure 2. (a) Training and test accuracy of various machine learning models. LR: Logistic Regression; GNB: Gaussian Naïve Bayesian; KNN: k-Nearest Neighbors; SVM: Support Vector Machine; DT: Decision Tree; RF: Random Forest; ADA: AdaBoost; XGB: eXtreme Gradient Boosting; MLP: Multilayer Perceptron. (b)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

Most machine learning algorithms, especially the neural networks, are proved challenging to offer explanation of the predicted results due to their so-called “black-box” nature. They work by fitting unknown functions *via* input and output datasets. The XGB model not only delivers the highest prediction accuracy, but also provides an out-of-the-box method to quantify significance of the features or descriptors in making decisions. In this case, the feature importance scores calculated from the XGB model allow us to rank the reaction parameters that affect the crystallization propensity of MONCs, thus assisting in studying reaction mechanism and accelerating discovery of new MONCs crystals. The scores for the total 17 descriptors were shown in Figure 3. It shows that solvents (H_2O, DMF, and MeCN), the organic ligands (PgC_x), modulators (molar mass, pKa, and mole), and cation (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 involved in the coordination of metal ions for promoting crystallization. Properties of the modulators such as molar mass and pKa values indicate the deprotonation capability of $PgC_{x,}$ thus making it a secondary factor. The model also shows that as an unfavorable solvent when mixed with 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' intuition. However, they are very challenging to be quantified by human or other traditional analysis methods. The XGB algorithm affords a simple but straightforward way of achieving it. To further investigate whether the number of descriptors affects the prediction performance, the XGB models based on 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. Interestingly, the prediction accuracies 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 PgCx, volume of acetonitrile, the pKa value of modulators, and molar mass of cations, the XGB model is robust enough to afford satisfactory prediction results.

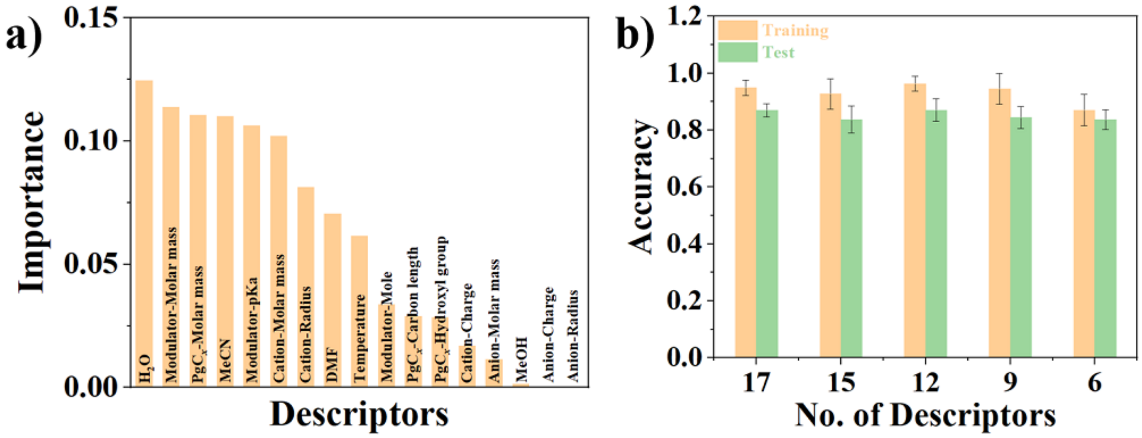


Figure 3. (a) Importance scores of descriptors derived from the XGB model. (b) Comparison of prediction accuracy from models trained with varied number of descriptors identified from the results shown in Figure 3a: total 17, top 15, top 12, top 9 and top 6 descriptors.

In order to gain more chemical insight, a flow chart was derived from the XGB model as shown in Figure 4a. It exhibits how the decision is made in classifying the reaction outcomes according to the input reaction parameters.[23] The tree was first divided by valence of the cations into two branches. The left branch has valence of < 3, and right one has 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 pKa value of modulators (shown in pale blue), while the reactions involving cations with molar mass between 53.47 g/mol and 61.24 g/mol tend to form MONCs crystals (shown in orange). From this decision tree, one can extract chemical hypothesis that includes 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+}), PgC_x with x larger than 2 and modulators with pKa of < 6.1 should be provided. If the radii of the M^{2+} cations (e.g. Co^{2+} and Mn^{2+}) increase, e.g. >= 0.725 nm, their molar mass should be between 53.47 g/mol and 61.24 g/mol in order to promote the crystallization. If the valence of the cations increases to 3 (M^{3+}), they should have much larger radii (e.g. >= 0.944 nm, here Sm^{3+}) in order to obtain a better crystallization propensity. These new hidden information extracted from the XGB 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.

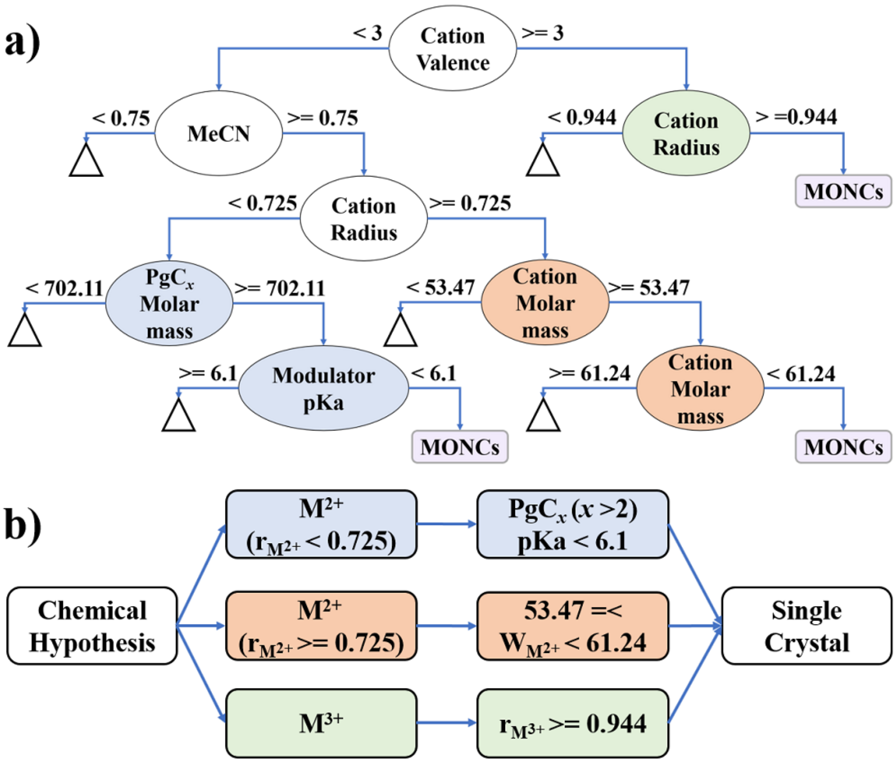


Figure 4. (a) Visualization of a decision tree from XGB model for classifying single-crystal and non-single-crystal of MONCs. Ovals show decision nodes, rectangles show result bins and

triangles show excised subtrees. (b) Graphical representation of three hypotheses generated from the XGB model.

To compare the performance of the XGB 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 datasets, were conceived and implemented (Table S5). These 20 experiments can be categorized into three classes according to the above proposed chemical hypothesis: (i) Ni^{2+}/Mg^{2+}, (ii) Co^{2+}/Mn^{2+}, (iii) Zn^{2+}. For the first class (Ni^{2+}/Mg^{2+}), all synthetic parameters were designed to meet the requirements for the formation of MONC single crystals ($r_{M^{2+}} < 0.725$ nm, PgC_x with $x > 2$, and pKa of modulator < 6.1). For the second class (Co^{2+}/Mn^{2+}), three experiments (No. **13**, **15** and **19**) were designed to confirm the importance of acetonitrile. For the last class, i.e., Zn^{2+}, the experiment was designed to be a failed experiment since no recommended cations or specific experiment conditions were included. The synthesis parameters of these 20 experiments were first presented to both the chemist and XGB for predicting reaction outcomes. Then the validation experiments were conducted by the chemist. The final reaction outcomes serve as the benchmark to evaluate the prediction accuracy by chemist and XGB model (Table S5). The XGB model successfully predicted the outcomes with accuracy of 80%, which is higher than that predicted by the skilled chemist (75%). Four unexpected failure (No. **6**, **8**, **9** and **17**) were found. It is proposed that the lower accuracy of the XGB model when predicting 20 new validation experiments may be due to insufficient generalization of the developed XGB model which is more or less influenced by a few potential exceptions (e.g., rare items in a majority of data).[45-46] Nevertheless, we believe that as the first proof-of-concept for predicting the crystallization propensity of MONCs the XGB model shows great potentials in guiding chemists, especially new entrants, to screen the reaction parameters for synthesizing new MONCs single-crystal.

Among the reactions that produced single crystals, a new compound **SCP-4** was found (No. **2** in Table S5). Single crystal X-ray diffraction data from a crystal of **SCP-4** was 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. A full anisotropic refinement of all positions was not possible, but because the packing of the structure can be inferred solely from the metal atom-to-metal atom vectors which are much longer than the resolution of the data, we consider the analysis of the packing reliable. This structural analysis reveals that **SCP-4** is 3D assembly of $Mg_{24}L_6$ nanocapsules (Figure 5sss). **SCP-4** consists of two types of nanocapsules within the framework (Figure 5a and 5b). Along [110] direction, each Type-A nanocapsules is connected to four Type-B nanocapsules *via* single alkyl chains and two Type-A nanocapsules *via* double alkyl chains at $(1\,\overline{1}\,0)$ plane (Figure 5c). Viewed from [001] direction, we can observe that each Type-B nanocapsules 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 and f). Along with other supramolecular coordination polymers composed of giant $M_{24}L_6$ as building blocks,[7-8] **SCP-4** exhibits the versatility of using MONCs to construct hierarchical supramolecular structures. Although the machine learning models have not yet enabled 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.

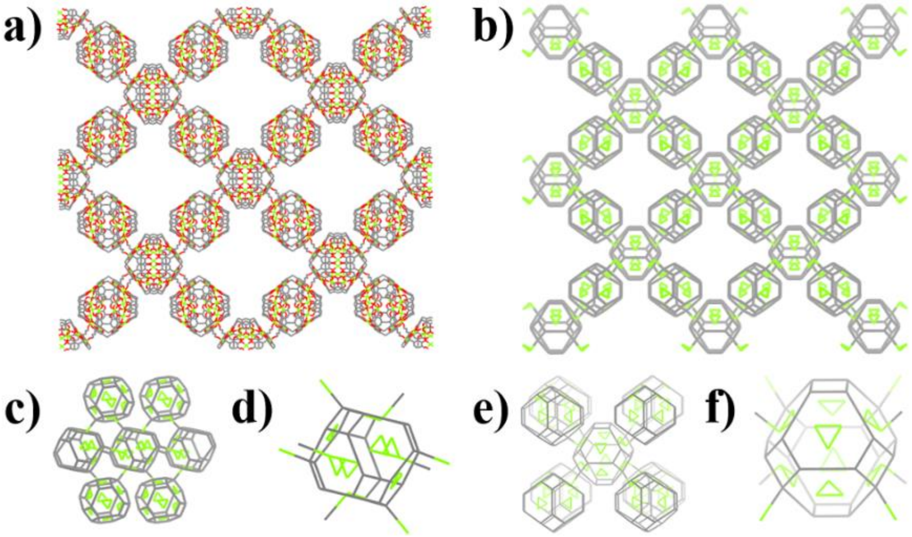


Figure 5. (a) Crystal structure and (b) network topology of **SCP-4** viewed along [001] direction. (c) Connection and (d) coordination mode of Type-A nanocapsules viewed along [110] direction. (e) Connection and (f) coordination mode of Type-B nanocapsules viewed along [001] direction. All hydrogen 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, grey; oxygen, red; nitrogen, blue; metal, green.

3. CONCLUSION

In summary, for the first time, this paper reports the machine-learning assisted method to predict the crystallization propensity of MONCs using historical successful/failed data. The highest prediction accuracy using the XGB model reaches 91% (averagely 87%). In addition, guided by the XGB model, we successfully discovered a new crystalline compound **SCP-4**.

This work will shed light on the discovery of new crystalline materials by integrating human intuition and machine learning techniques. The extension of our models to other organic synthesis systems is anticipated by substituting the corresponding descriptors into the machine learning models and fine-tuning the hyperparameters correspondingly. Finally, integrating the developed machine learning models with high-throughput synthesis (i.e., robotic synthesis platforms) would greatly accelerate development of inorganic-organic hybrid materials.

ASSOCIATED CONTENT

Supporting Information

All data are available in the manuscript and supplementary information or requesting from the corresponding author. All Python script for this project is available at https://github.com/linresearchgroup/MONCs.

AUTHOR INFORMATION

Corresponding Authors

*linjian@missouri.edu (J.L.)

*atwoodj@missouri.edu (J.L.A)

Author Contributions

#Authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKOWLEDEGEMENTS

This work was supported by grants from National Science Foundation (Award number: 1825352) with Program Manager Khershed P. Cooper, United States Department of Energy National Energy Technology Laboratory (Award number: DE-FE0031645) with Program Manager Karol K. Schrems, and Unite Sates Department of Agriculture (Award number: 2018-67017-27880). 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* (6265), 1235-1238.
2. Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M., The Chemistry and Applications of Metal-Organic Frameworks. *Science* **2013,** *341* (6149).
3. 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. *Journal of the American Chemical Society* **2012,** *134* (42), 17358-17361.
4. 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. *Journal of the American Chemical Society* **2015,** *137* (33), 10508-10511.
5. Patil, R. S.; Banerjee, D.; Zhang, C.; Thallapally, P. K.; Atwood, J. L., Selective CO_2 Adsorption in a Supramolecular Organic Framework. *Angewandte Chemie International Edition* **2016,** *55* (14), 4523-4526.
6. 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. *Journal of the American Chemical Society* **2014,** *136* (20), 7241-7244.
7. 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. *Chemistry – A European Journal* **2018,** *24* (54), 14335-14340.

8. Zhang, C.; Patil, R. S.; Liu, C.; Barnes, C. L.; Atwood, J. L., Controlled 2D Assembly of Nickel-Seamed Hexameric Pyrogallol[4]arene Nanocapsules. *Journal of the American Chemical Society* **2017,** *139* (8), 2920-2923.

9. Zhang, C.; Patil, R. S.; Li, T.; Barnes, C. L.; Atwood, J. L., Self-assembly of magnesium-seamed hexameric pyrogallol[4]arene nanocapsules. *Chemical Communications* **2017,** *53* (31), 4312-4314.

10. 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* **2011,** *6* (1), 272-275.

11. McKinlay, R. M.; Cave, G. W. V.; Atwood, J. L., Supramolecular blueprint approach to metal-coordinated capsules. *Proceedings of the National Academy of Sciences* **2005,** *102* (17), 5944-5948.

12. Power, N. P.; Dalgarno, S. J.; Atwood, J. L., Guest and Ligand Behavior in Zinc-Seamed Pyrogallol[4]arene Molecular Capsules. *Angewandte Chemie International Edition* **2007,** *46* (45), 8601-8604.

13. 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. *Chemical Communications* **2018,** *54* (6), 635-637.

14. 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. *Journal of the American Chemical Society* **2013,** *135* (33), 12184-12187.

15. 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.

16. Stranks, S. D.; Snaith, H. J., Metal-halide perovskites for photovoltaic and light-emitting devices. *Nature Nanotechnology* **2015,** *10*, 391.

17. Zhou, H.-C.; Long, J. R.; Yaghi, O. M., Introduction to Metal–Organic Frameworks. *Chemical Reviews* **2012,** *112* (2), 673-674.

18. Le, T. C.; Winkler, D. A., Discovery and Optimization of Materials Using Evolutionary Approaches. *Chemical Reviews* **2016,** *116* (10), 6107-6132.

19. Henson, A. B.; Gromski, P. S.; Cronin, L., Designing Algorithms To Aid Discovery by Chemical Robots. *ACS Central Science* **2018,** *4* (7), 793-804.

20. Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C., A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Computational Materials* **2016,** *2*, 16028.

21. 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 Computational Materials* **2019,** *5* (1), 26.

22. 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. *Journal of the American Chemical Society* **2017,** *139* (49), 17870-17881.

23. 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.

24. 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* (6385), 186-190.

25. 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* (7714), 377-381.

26. Moosavi, S. M.; Chidambaram, A.; Talirz, L.; Haranczyk, M.; Stylianou, K. C.; Smit, B., Capturing chemical intuition in synthesis of metal-organic frameworks. *Nature Communications* **2019,** *10* (1), 539.

27. Coley, C. W.; Green, W. H.; Jensen, K. F., Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research* **2018,** *51* (5), 1281-1289.

28. 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**.

29. 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. *Chemistry – A European Journal* **2017,** *23* (35), 8520-8524.

30. Schultz, C.; Alegría, A. C.; Cornelis, J.; Sahli, H., Comparison of spatial and aspatial logistic regression models for landmine risk mapping. *Applied Geography* **2016,** *66*, 52-63.

31. Duda, R. O.; Hart, P. E., *Pattern classification and scene analysis*. Wiley New York: 1973; Vol. 3.

32. Dasarathy, B. V., Nearest neighbor (NN) norms: NN pattern classification techniques. **1991**.

33. Burges, C. J., A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery* **1998,** *2* (2), 121-167.

34. 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 Systems with Applications* **2009,** *36* (2, Part 1), 2017-2026.

35. 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. *Journal of machine learning research* **2011,** *12* (Oct), 2825-2830.

36. Freund, Y.; Schapire, R.; Abe, N., A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence* **1999,** *14* (771-780), 1612.

37. 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, USA, 2016; pp 785-794.

38. Jaeger, S.; Fulle, S.; Turk, S., Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *Journal of Chemical Information and Modeling* **2018,** *58* (1), 27-35.

39. 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* (4), eaaq1566.

40. Obuchowski, N. A., Receiver Operating Characteristic Curves and Their Use in Radiology. *Radiology* **2003,** *229* (1), 3-8.

41. Fawcett, T., An introduction to ROC analysis. *Pattern Recogn. Lett.* **2006,** *27* (8), 861-874.

42. 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. *Proceedings of the National Academy of Sciences* **2019,** *116* (23), 11259-11264.

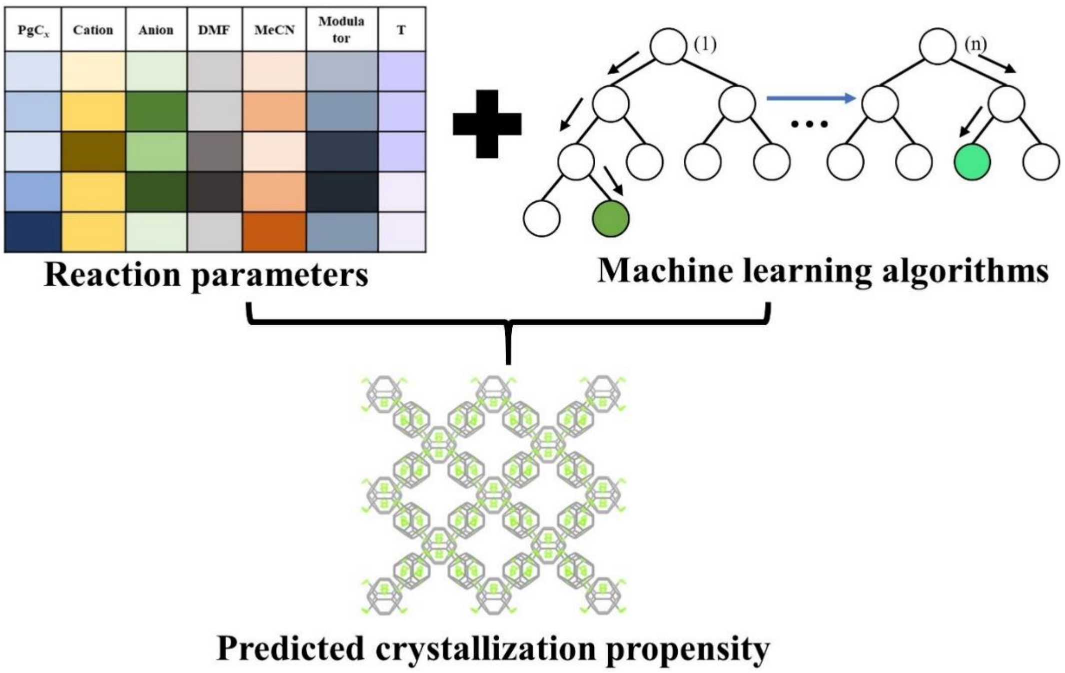
43. Lei, T.; Chen, F.; Liu, H.; Sun, H.; Kang, Y.; Li, D.; Li, Y.; Hou, T., ADMET Evaluation in Drug Discovery. Part 17: Development of Quantitative and Qualitative Prediction Models for Chemical-Induced Respiratory Toxicity. *Molecular Pharmaceutics* **2017,** *14* (7), 2407-2421.

44. Wu, Z.; Ramsundar, B.; Feinberg, Evan N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V., MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* **2018,** *9* (2), 513-530.

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. *Advanced Energy Materials* **2019,** *9* (26), 1900891.

46. Li, Q.; Fan, S.; Chen, C., An Intelligent Segmentation and Diagnosis Method for Diabetic Retinopathy Based on Improved U-NET Network. *Journal of Medical Systems* **2019,** *43* (9), 304.

TOC Figure


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