Chemical Engineering Journal 473 (2023) 145216

Contents lists available at ScienceDirect

Chemical Engineering Journal

journal homepage: www.elsevier.com/locate/cej

Machine learning assisted phase and size-controlled synthesis of iron oxide particles

Juejing Liu [a,b,c,1], Zimeng Zhang [a,d,1], Xiaoxu Li [a,1], Meirong Zong [a], Yining Wang [a], Suyun Wang [a], Ping Chen [a], Zaoyan Wan [a], Lili Liu [a], Yangang Liang [e], Wei Wang [e], Shiren Wang [d], Xiaofeng Guo [b], Emily G. Saldanha [f,*], Kevin M. Rosso [a,*], Xin Zhang [a,*]

[a] Physical and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA 99354, USA
[b] Department of Chemistry, Washington State University, Pullman, WA 99164, USA
[c] Materials Science and Engineering Program, Washington State University, Pullman, WA 99164, USA
[d] Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843, USA
[e] Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, WA 99354, USA
[f] National Security Directorate, Pacific Northwest National Laboratory, Richland, WA 99354, USA

ARTICLE INFO

Keywords:
Iron oxides synthesis
Machine learning
Prediction
Particle size
Phase
Hematite

ABSTRACT

Synthesis of iron oxides with specific phases and particle sizes is a crucial challenge in various fields, including materials science, energy storage, biomedical applications, environmental science, and earth science. However, despite significant advances in this area, much of the current palette of particle outcomes has been based on time-consuming trial-and-error exploration of synthesis conditions. The present study was designed to explore a very different approach to 1) predict the outcome of synthesis from specified reaction parameters based on using machine learning (ML) techniques, and 2) correlate sets of parameters to obtain products with desired outcomes by a newly designed recommendation algorithm. To achieve this, four ML algorithms were tested, namely random forest, logistic regression, support vector machine, and k-nearest neighbor. Among the models, random forest outperformed the others, attaining 96% and 81% accuracy when predicting the phase and size of iron oxide particles in the test dataset. Surprisingly, the permutation feature importance analysis revealed that volume, which may strongly relate to pressure, was one of the important features, along with precursor concentration, pH, temperature, and time, influencing the phase and size of iron oxide particles during synthesis. To verify the robustness of the random forest models, prediction and experimental results were compared based on 24 randomly generated methods in additive and non-additive systems not included in the datasets. The predictions of product phase and particle size from the models agreed well with the experimental results. Furthermore, a searching and ranking algorithm was developed to recommend potential synthesis parameters for obtaining iron oxide products with the desired phase and particle size from previous studies in the dataset. This study lays the foundation for a closed-loop approach in materials synthesis and preparation, beginning with suggesting potential reaction parameters from the dataset and predicting potential outcomes, followed by conducting experiments and analyses, and ultimately enriching the dataset.

1. Introduction

Controlled synthesis of nanomaterials with precisely defined size and phase has received significant attention for diverse applications in catalysis, energy storage, biomedicine, and environmental remediation. The size and phase of nanomaterials are critical factors that affect their optical, electronic, magnetic, and catalytic properties, and consequently, their performance [1–4]. For instance, size-dependent behavior, such as the quantum confinement effect in semiconductor nanocrystals, can be harnessed in applications like solar cells, light-emitting diodes, and other optoelectronic devices [5–7]. Smaller nanoparticles also have a higher surface-to-volume ratio leading to a higher number of surface atoms that define particle reactivity and catalytic activity [8–10]. Furthermore, the phase of a material profoundly

* Corresponding authors.
E-mail addresses: emily.saldanha@pnnl.gov (E.G. Saldanha), kevin.rosso@pnnl.gov (K.M. Rosso), xin.zhang@pnnl.gov (X. Zhang).
[1] These three authors contribute equally to this paper.

https://doi.org/10.1016/j.cej.2023.145216
Received 13 May 2023; Received in revised form 29 July 2023; Accepted 2 August 2023
Available online 5 August 2023

influences its catalytic activity and selectivity, as different crystal structures can expose unique active sites and alter the adsorption and desorption kinetics of reactants and products [1,9,11,12]. Phase-dependent properties such as electrical conductivity, magnetism, and mechanical strength significantly impact the performance of a material in applications such as energy storage, sensing, and drug delivery [13,14]. Complicating matters is the fact that particle size can affect relative phase stabilities through the surface free energy contribution to the total free energy of the nanomaterial [15,16]. Therefore, achieving control over both the size and phase of nanomaterials is fundamentally important, but is also a complex endeavor that has traditionally relied on empirical trial-and-error approaches to achieve a desired result.

Controlling the properties of iron oxide particles is a prominent case in point. Tailored synthesis of iron oxide particles enables exploitation of their full range of remarkable magnetic, electrical, and catalytic properties [17]. Among the different iron oxide phases, hematite (α-Fe_2O_3) has been extensively studied due to its stability, nontoxicity, photoelectric properties, and potential applications in photocatalysis, gas sensing, antibacterial, and nanofluid applications [18–20]. However, achieving precise phase control to yield pure hematite particles as opposed to other iron oxide phases [e.g., maghemite (γ-Fe_2O_3), goethite (α-FeOOH), akageneite (β-FeOOH), lepidocrocite (γ-FeOOH), magnetite (Fe_3O_4), and ferrihydrite], while also controlling particle size and size distribution, remains challenging [21]. In recent years, significant progress has been made in developing various synthesis strategies for iron oxides, including hydrothermal synthesis, sol–gel method, and co-precipitation [22–24]. Nonetheless, because this has relied largely on empirical synthesis strategies, progress has not only been gradual and time-consuming but has also left uncertainty regarding the extent to which the full palette of particle properties has been sampled. A robust and accurate predictive model for optimizing the synthesis protocol of iron oxide nanoparticles to obtain the desired phase and particle size is still lacking. Consequently so is our fundamental understanding of the relevant nucleation and growth pathways that control particle outcomes, and the range of currently untapped additional possibilities.

Machine learning (ML) has emerged as a promising approach to address the challenges of predicting synthesis-structure-property relationships of nanomaterials. ML utilizes computer algorithms and mathematical models to uncover the underlying relationships between features and labels, such as synthesis conditions and morphological parameters [25,26]. This predictive capability is particularly useful when the relationships between variables and outcomes are complicated or unknown [27,28]. With the advent of high-throughput experimental setups and an increase in experimental data sources, ML models can take advantage of the large amount of available data [29–31]. Various ML models, such as logistic regression [32], random forest [33], k-nearest neighbor [34], Gaussian processes [35], support vector machines [36], deep neural network [37], and Bayesian optimization [38], have been proposed for the analysis of different nanomaterials and their specific properties. For instance, Sun et al. combined various ML models and physics-based simulations to achieve efficient and accurate high-throughput production of silver nanoparticles, among studies of other metal participles [39]. Wang et al. developed an unsupervised ML model for transmission electron microscopy (TEM) image analysis and classification of the gold nanoparticles [40]. Lee et al. used ML to quantitatively analyze the morphology of gold nanoparticles via TEM images and achieved high precision [41]. Pellegrino et al. studied the performance of ML models for predicting the morphology of TiO_2 nanoparticles obtained with different synthesis parameters [42]. Recently, Lu et al. and Wu et al. successfully trained models with various algorithm to predict the structure of 2D materials based on reaction conditions and to determine the stability of 2D materials from their crystallography structures [43,44]. These studies demonstrate the potential of ML in revealing the underlying relationships between synthesis conditions and the properties of iron oxide nanomaterials.

Another common question in materials synthesis and preparation is how to choose the right experimental conditions. The selection of appropriate synthesis parameters is key to obtaining materials with desired properties. Currently, researchers need to extract and summarize the optimal sets of synthesis parameters manually. This is a time-consuming and error-prone process. Recently, advanced ML tools have been used to efficiently gather synthesis parameters from thousands of studies [45]. However, such large datasets make manual verification virtually impossible. Therefore, it is important to develop an objective method to evaluate, rank, and recommend conditions with a higher likelihood of synthesizing desired products. To our knowledge, only a few studies have discussed the recommendation of synthesis parameters from the existing datasets [46,47].

In this study, we utilized an iron oxide synthesis dataset, which we collected from published sources and unpublished data from our laboratory, to explore potential solutions for two important tasks: (i) predicting the outcomes of different synthesis parameters, and (ii) determining the optimal parameters for a desired product. To address the first task, we trained and tested various ML models, including logistic regression, random forest, k-nearest neighbor, and support vector machine. These models predicted the synthesis outcomes, such as the formation of hematite and the size of particles, based on experimental parameters. We also conducted permutation feature importance analysis to identify the most critical features that the models relied on to predict particle size and phase from synthesis conditions. Additionally, we conducted correlation analysis to examine the relationships between important features. To verify the accuracy of our random forest model, we compared its predictions with experimental results from 24 randomly generated methods in both additive-added and additive-free systems, which were not included in the original dataset. The model's predictions of product phase and particle size aligned well with the experimental results. For the second task, we developed a searching and ranking algorithm that can identify potential synthesis parameters for iron oxides with desired phase and particle size. The combination of these two solutions has the potential to form a "closed-loop" method that integrates parameter selection from the dataset, outcome prediction, experiment execution, and dataset enrichment.

2. Experiments and methods

2.1. Chemicals and materials

Ferric chloride anhydrous, sodium dodecyl sulphate (SDS), and sodium citrate, and sodium hydroxide were purchased from Sigma-Aldrich Chemical Reagent Co., Ltd. All chemicals are analytical purity and can be used directly without any further treatment. Deionized (DI) water used in this work was prepared using a Barnstead water purification system.

2.2. Synthesis of iron oxides

In a typical procedure, anhydrous ferric chloride and additives were dissolved in 15 ml of DI water under magnetic stirring at room temperature. The pH of the solution was then adjusted to the desired value by slowly adding 5 M NaOH and monitoring with a Thermo Scientific Orion Star A221 pH meter. The resulting solution was transferred into a 20 ml Teflon liner stainless steel autoclave and kept at the specific temperature and time. The products were washed three times by DI water and the final products were collected by centrifugation at 8000 rpm.

2.3. Characterization

Powder XRD of all as-prepared samples was performed on a Philips X'pert Multi-Purpose Diffractometer (MPD) (PANAlytical) equipped with Cu Kα radiation operating at 50 kV and 40 mA. Scanning Electron Microscopy (SEM) imaging was conducted on the FEI Helios NanoLab

600i dual-beam focused ion beam precision manufacturing instrument operating at 5 kV and 86 μA. To improve the electronic conductivity of the samples before SEM imaging, a thin carbon layer (about 5 nm) was deposited on the particle surface by using a carbon coater (208C; Ted Pella, Inc.).

2.4. Data set characterization and collection

Data on the synthesis methods, phase and particle size of iron oxide nanomaterials have been collected from previously reported studies as well as experimental data collected in the laboratory at Pacific Northwest National Laboratory, USA. The dataset includes 780 pieces of data corresponding to iron oxides synthesis in different sets of conditions, including precursors, additives, solvents, concentrations of each ingredient and temperature. The corresponding phases of the iron oxides such as hematite (α-Fe_2O_3), maghemite (γ-Fe_2O_3), magnetite (Fe_3O_4). Goethite (α-FeOOH), akageneite (β-FeOOH), lepidocrocite (γ-FeOOH), ferrihydrite and the size of the nanoparticles were also included in the dataset. The dataset was consisted of about 729 sets of data from previous publications reporting the iron oxides synthesis from 2010 to 2020 as well as the 51 sets of laboratory experimental data.

2.5. Software libraries

All software libraries used in this study are listed below: Pandas [48], NumPy [49], scikit-learn [50], openpyxl [51], and Jupyter Notebook [52]. Pandas, NumPy, and openpyxl were employed for the purpose of importing, cleansing, and preprocessing data. The scikit-learn library and Jupyter Notebook were employed to train models based on different algorithms in an interactive Python environment. Data visualization and cluster analysis were conducted using Seaborn and Matplotlib [53,54].

2.6. Feature analytics and selection

Eleven features were selected based on our experimental experience as having potential to significantly impact the synthesis of iron oxides. These features include precursor species and concentration, surfactant species and concentration, hydrogen ion concentration, temperature, reaction time, solvent, and solvent volume. Among the features selected for this study, precursor and surfactant species, as well as solvent, were categorical variables. To prepare the data for training the machine learning models, one-hot encoding was applied to each categorical variable. This process created additional feature columns from each categorical variable, with each unique value represented as a binary feature. For instance, the precursor variable contained several unique values, including $FeCl_3$, $Fe(NO_3)_3$, $FeSO_4$, FeC_2O_4, $K_3[Fe(CN)_6]$, etc. Consequently, 10 additional feature columns were generated to replace the original precursor feature column. A value of 1 in a particular precursor column indicated the use of that precursor in an experiment, while 0 indicated the use of a different precursor. Normalization was performed on the training set using the StandardScaler class from scikit-learn. The resulting standard scaler was then applied to both the training and test sets. The purpose of this normalization was to minimize any potential bias towards the test set, which should be considered as unknown during the training phase. By normalizing both the training and test sets in the same way, we aimed to ensure that the machine learning models could make accurate predictions on previously unseen data.

2.7. Stratification, sampling, and training

In our dataset, hematite was the primary synthesis product. However, to address the issue of dataset imbalance, we used stratified sampling to ensure a relatively equal distribution of hematite and non-hematite samples in both the training and testing subsets. Specifically, we allocated 80% of the data to the training subset and 20% to the testing subset. This approach aimed to improve the robustness of the machine learning models by training them on a representative sample of the data and testing their generalization ability on previously unseen data.

We trained models based on four ML algorithms, namely k-nearest neighbor (KNN), logistic regression (LR), support vector machine (SVM), and random forest (RF) [32–34,36]. We converted the prediction of whether the experimental conditions led to the formation of hematite or not into a binary classification. Two methods were employed to predict the size of nanoparticles, with regression based solely on the RF algorithm and classification based on KNN, SVM, and RF. In the classification method, we sorted the nanoparticle sizes into three categories: nano (less than 100 nm), sub-micron (between 100 nm and 1000 nm), and micron (greater than 1000 nm). We excluded LR from the classification method. LR is mainly a binary classification method, while three labels were presented in the particle size prediction experiment.

To achieve high accuracy in our analysis, k-fold cross-validation was then used in this study. We split the training dataset into k (in this study, $k = 5$) equally sized subgroups. In each validation subgroup, we used the confusion matrix to measure accuracy, while the remaining four subgroups were used for training the machine learning models. The k-fold cross-validation process was repeated five times for each algorithm, with each subgroup used only once for validation. The mean value across the k-folds was calculated, and the cross-validation process was iterated for various combinations of hyperparameters. During the training of the machine learning models for all four algorithms, we performed grid search cross-validation to evaluate multiple models with different training parameters. We then selected the best model for each algorithm and compared their performance (see Table S1 for initial sets of parameters and the best parameters). To evaluate the performance of the models, we used the testing dataset, which was not used during the training process. We measured the prediction performance using accuracy, which is defined as the ratio of correctly predicted data to the total testing data.

To obtain lists of feature importance for each algorithm after training, we utilized the permutation method available in scikit-learn. This method involves shuffling each feature per epoch and evaluating the resulting impact on model accuracy. Features that have a significant effect on model accuracy when shuffled are considered to be of high importance, while those that do not significantly affect model accuracy are considered to be of lower importance. We then performed a correlation analysis on the five most important features (temperature, precursor concentration, pH, time, and solution volume) using Pearson's correlation coefficients. Additionally, a hierarchically clustered heatmap was generated using seaborn to observe the relationship between these five conditions and the phase of the iron oxide products. We used all available features to train all models presented in this study. No features were removed based on the results of premutation and correlation analysis (see Table S3 to S7).

2.8. Note for correlation analysis

The correlation analysis was performed by using seaborn package [53]. The Pearson correlation coefficient method was used to obtain the dendrogram and reveal the relationship among important features. Two hierarchically clustered heatmaps were obtained to study the influence of important features on the phase and size of iron oxide particles. We noticed that the range of values for several features, including volume, precursor concentration, surfactant concentration, and time, may vary in a wide range (a magnitude of 3). For better comparison, we took the logarithm with a base of 10 for these four features. We also shift the surfactant concentration upward with +1 mM to prevent calculating the logarithm of 0.

2.9. Design of searching and ranking recommendation algorithm

The recommendation algorithm took the desired phase and particle

size of iron oxide as input. The ranking algorithm allowed us to quantify the degree of deviation of each parameter from its corresponding average value, using the standard deviation as a measure of variability, to recommend the most suitable conditions for achieving the desired phase and particle size of iron oxide. It then searched the entire dataset and selected all possible sets of conditions that met the desired criteria. We calculated the average (avg.) and standard deviation (std.) of the major features, including time, temperature, pH, precursor concentration, and volume, from the selected sets of parameters. To evaluate how a specific parameter of a specific set differs from the averaged value, we used a ranking algorithm as shown in Eq. (1):

$$s_{feature} = \frac{\sigma_{feature}}{|V_{feature-avg_{feature}}|} \quad (1)$$

Where $s_{feature}$ is the score of the specific parameter in the specific set, $V_{feature}$ is the raw value of the feature, $avg_{feature}$ and $\sigma_{feature}$ is the average and standard derivation of the feature. The scores from every parameter in the set were then added up and formed the final score of the set of features:

$$S_{set} = \sum_{i} s_{feature,i} \quad (2)$$

A high score (S_{set}) indicates that the parameters in this set are generally closer to the averaged value and more previous studies used similar parameters to synthesize the desired products. In contrast, a low score suggests that the parameters are farther from the averaged values and less studies used similar parameters to synthesize the desired products.

3. Results and discussion

3.1. Machine learning models

We collected data from 780 iron oxide synthesis experiments (as shown in Fig. 1a), with each set of data comprising experiment conditions (features), e.g., precursor concentration, pH, reaction temperature, etc., and corresponding resultant parameters (labels), such as particle phase and size. Given that training ML models requires a substantial amount of data, we had to be mindful of our dataset's small size, which precluded the use of more sophisticated algorithms like neural networks [55–58]. Consequently, we opted for four algorithms suitable for a small dataset as illustrated in Fig. 1b: KNN, LR, SVM, and RF [32–34,36]. KNN is an instance-based classification algorithm that assigns labels to unseen features based on their similarity to known features with certain labels. LR is a generalized linear model that exclusively handles binary classification. It builds a function to calculate the probability of features belonging to one of the two bins and uses this to categorize unseen features. SVM is a kernel-based algorithm that maps features to a higher dimension and generates rules to classify them. We used a radial kernel for our SVM model. RF models comprise many decision trees, with their results combined to make the final prediction. To test the performance of these models, we first assessed their accuracy in binary classification, determining whether experimental conditions led to hematite formation or not.

The Random Forest model outperformed all three other models for this binary classification question, with an accuracy of 96% (see Table S2). The KNN model predicted the formation of non-hematite compounds with 83% recall, while the recall for predicting the formation of hematite compounds was 90% (Fig. 1c). The LR model had more divergent prediction accuracies, with a recall of only 60% for detecting the formation of non-hematite (Fig. 1d). In contrast, the model achieved

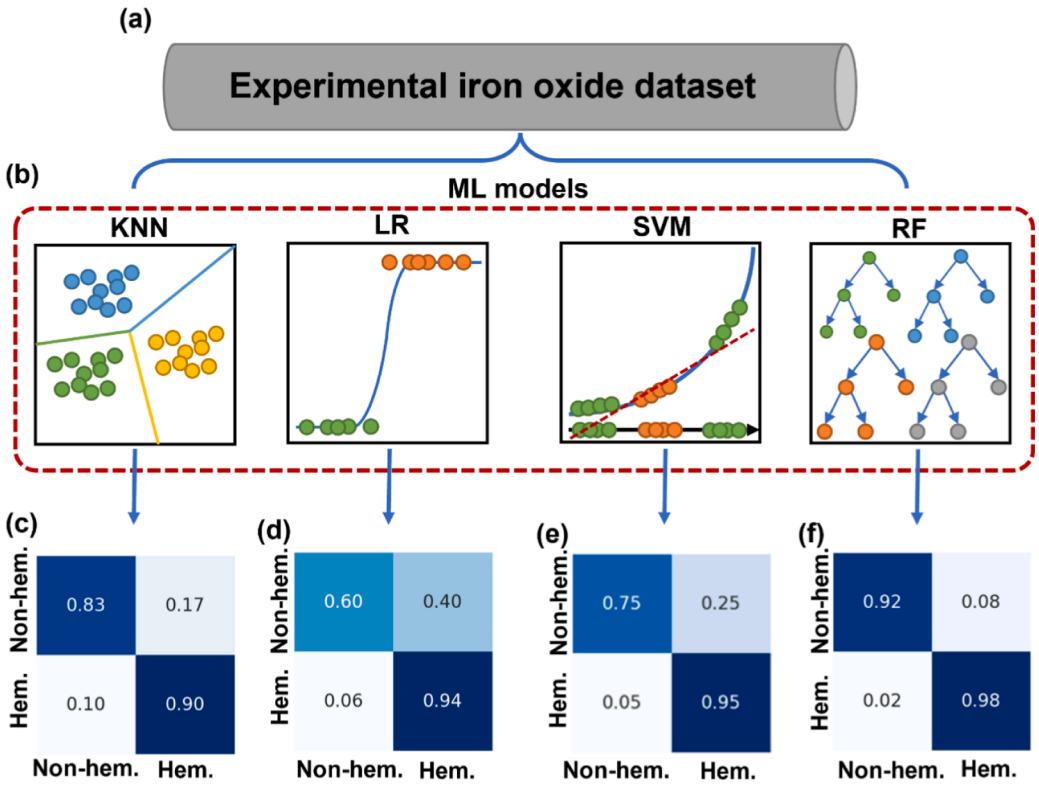

Fig. 1. Prediction of formation of hematite based on experimental conditions by different ML methods. (a) A dataset of iron oxide synthesis conditions obtained from previous studies and our experimental results. (b) ML algorithms used in this study, including k-nearest neighborhood (KNN), logistic regression (LR), support vector machine (SVM), and random forest (RF). (c) to (f) Confusion matrices for models from KNN (c), LR (d), SVM (e), RF (f) to predict the formation of hematite.

94% recall in predicting the formation of hematite compounds. The SVM model achieved 75% recall in detecting the formation of non-hematite and 95% recall in predicting the formation of hematite (Fig. 1e). The RF model demonstrated the highest accuracy, with an recall of 92% for predicting non-hematite formation and 98% recall for predicting hematite formation. The overall accuracy of the RF model was 96% (Fig. 1f).

We employed the permutation method to derive feature importance rankings for the four models (Fig. 2). This technique entails shuffling the values of a single feature while keeping the others constant, and then observing the resulting impact on model accuracy. Features that exhibit a significant effect on model accuracy are deemed highly important, whereas those with minimal effect are considered less important. As a result, we are able to uncover the relationships between features (experimental conditions) and labels (phases of products). This approach proves valuable in determining such relationships, especially considering the "black box" nature of the machine learning models employed in this study. Despite the relative simplicity of the models used, they rely on thousands, if not tens of thousands, of internal parameters to establish functions that connect features and labels [59]. It may also reveal previously ignored relationships between features (precursors, additives, and conditions) and labels (phase of iron oxide).

The feature importance analysis reveals that temperature, pH, precursor concentration, and time are crucial features for predicting the formation of hematite for most models. In the KNN model (Fig. 2a), temperature is the most important feature, followed by pH, volume, precursor concentration, and time. Interestingly, the model considers volume an important feature, despite the presence of precursor concentration. A possible explanation is that the majority of the dataset involved hydrothermal synthesis using autoclaves, where the solution-to-autoclave volume ratio determines the pressure inside the autoclave [60]. Given that pressure is a key factor in hydrothermal reactions, the algorithm chose volume as a feature for predicting hematite formation [61].

The LR model is the only model that shows a different set of feature importance ranks than those of the other models. Although temperature is still the most important feature, the ranking of the other features is less intuitive. After temperature, this model relies on whether water and ethanol are used as solvents, or whether $Fe(NO_3)_3$ and polyvinylpyrrolidone (PVP) are used during the synthesis. In addition, the LR model relies less on these features to predict hematite formation. The value of the importance of temperature is about two thirds than that of the KNN models (0.09 vs. 0.14). Hence, we believe that the LR model does not fully recognize the pattern of iron-oxide synthesis, especially for the non-hematite part (see Fig. 1d and Table S2).

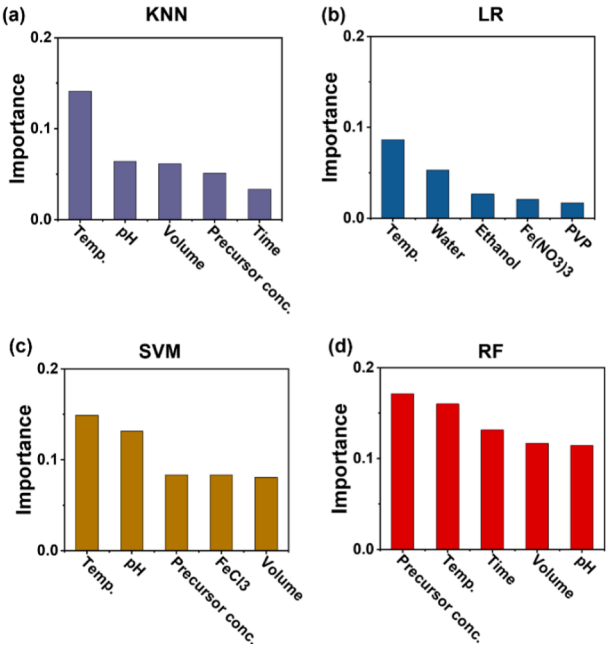

Fig. 2. Premutation feature importance of four binary classification models. (a) to (d) Important features in of k-nearest neighborhood (a), logistic regression (b), support vector machine (c), and random forest (d) model.

The SVM model relies on similar features to the KNN model to predict the outcome of iron oxide synthesis, as shown in Fig. 2c. However, the presence of $FeCl_3$ is more important than reaction time. The SVM model also relies more on the first five features than KNN to predict the phase of iron oxide, as the values of importance of the first five features are higher in the SVM model. As a result, the SVM model shows a higher accuracy in terms of binary phase prediction (see Table S2).

The RF model which preforms the best accuracy on phase identification, uses the commonly known important features (precursor concentration, temperature, time, volume, and pH) equally, with the highest and lowest feature importance scores of 0.17 and 0.11, respectively. Shuffling any of these five features has a similar effect on the accuracy of the model. The rest features show much lower important (<0.05, see Table S5). This is consisted with the principle of classical nucleation theory (CNT). According to CNT, the phase with lowest nucleation barrier will form fist and then consume the precursors to avoid the immediate formation of more stable phases. The steady-state nucleation rate (J) can be expressed as:

$$J = J_0 exp(-\Delta G^*/k_B T) \quad (3)$$

$$\Delta G^* = 16\pi V_m^2 \gamma^3 / \Delta G^2 \quad (4)$$

in which, J_0 is per-factor and typically negligible compared to the difference in the exponential term, ΔG^* is the nucleation barrier to form a spherical critical nucleus; k_B is Boltzmann constant; T is the temperature γ is the Gibbs surface free energy of nucleus (J/m^2); V_m is the molar volume (cm^3/mol) and ΔG is the thermodynamic driving force of phase change. As shown in equation (3) and (4), the precursor concentration, temperature, and pH are the dominating features to affect the ΔG, and thereby the phase. And the other less important features such as the type and concentration of additives has negligible impact on the driving force term, showing the good agreement between RF model and CNT. Interestingly, the additives like surfactants can also potentially alter the final products by changing the surface energy, however RF model didn't rely heavily on this term to predict the phase. This may be due to the limited availability of data containing specific additives, particularly surfactants, the permutation algorithm may be misled as shuffling the column of such categorical features does not result in substantial changes.

We performed correlation analysis using the Pearson correlation coefficient for the five main features determined by the premutation method (temperature, volume, precursor concentration, pH, and time) [62]. The results show that most of the features are independent from each other (Fig. 3a), although a dendrogram was still obtained. Most of the correlation coefficients are smaller than 0.1, indicating a negligible correlation. This is understandable as iron oxides can be synthesized in many different combinations of conditions [63]. One exception is temperature, which has a relatively strong negative correlation with the other features. One possible explanation for the correlation between temperature and solution volume could be related to the synthesis method used in the majority of the experiments in our dataset. Specifically, most of the syntheses used a hydrothermal method with autoclaves. In this method, high temperatures can lead to a significant increase of pressure inside the autoclave. If the volume of the solution is also high, this can exacerbate the pressure increase and may even cause leaking. As a result, the experimental conditions used in these experiments may have been limited by the maximum allowable temperature and volume for the given autoclave setup. Further experiments are needed to confirm this hypothesis [60]. Additionally, the negative

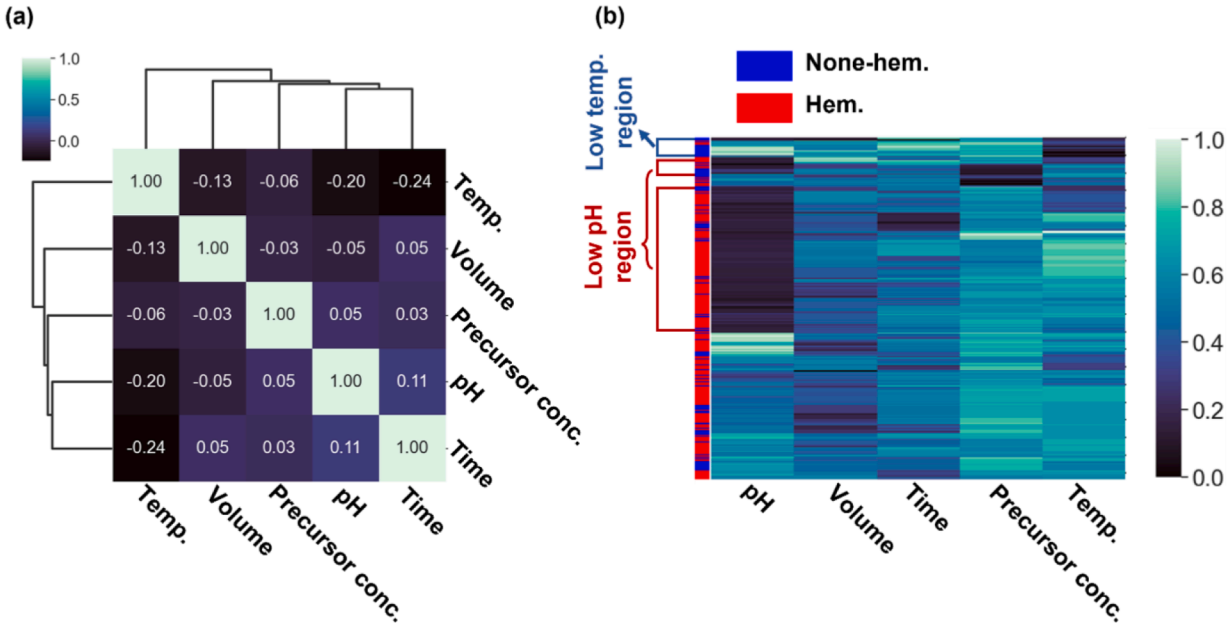

Fig. 3. Correlation analysis of important features, including temperature, volume, precursor concentration, pH value, and time. The dendrogram with Pearson correlation coefficient (a) and cluster map (b) were generated based on the whole iron oxide synthesis dataset.

correlation between temperature and time could likely be explained by the reaction kinetics of iron oxide synthesis. Increasing the temperature during the synthesis process is generally favorable for the formation of iron oxide crystals, which can reduce the reaction time required to achieve a desired outcome.

A hierarchically clustered heatmap (see Fig. 3b) provided insight into the relationship among pH, temperature, and the formation of hematite or other iron oxide phases, consistent with the principle of classical nucleation theory [21]. Our results show that, under low pH (below 4.0), most of the synthesized products were hematite (indicated in red). Conversely, at higher pH levels (above 4.0), the proportion of non-hematite products (indicated in blue) increased. The preference for acid solutions in hematite synthesis is due to the thermodynamic stability of hematite under low pH conditions. In addition, acidic conditions dissolve iron-containing precursors and promote the nucleation and growth of hematite crystals. However, the oversaturation state of the solution decreases with decreasing pH. A higher temperature is desired to overcome the thermodynamical barrier of nucleation to form the stable hematite phase [21,64]. In the higher pH range, the solution is oversaturated in terms of all ferric oxides. In such case, the surface energy will have a higher impact on the phase selection. Because the metastable phase has lower surface energies than thermodynamically stable phase hematite, the barrier to form a critical nucleus of metastable phase is lower. The metastable phase will form first and simultaneously depletes the concentration of precursor, which hinders the formation of hematite at relatively high pH conditions.

We employed permutation analysis and correlation analysis to explore previously undiscovered relationships between reaction conditions and the resulting phases of iron oxide. This approach allowed us to gain valuable insights into the mechanisms underlying iron oxide formation. Despite certain features, such as the type of surfactant, being considered unimportant, we still believe they may have a role in the process of iron oxide formation. This apparent lack of importance could be attributed to the limited size of our training dataset, which might have led to some surfactants being underrepresented. We are cautious about removing these features from the model, as doing so could potentially compromise the accuracy of our results. Therefore, to maintain the integrity of our analysis and to avoid any biases in feature selection, we chose not to perform any screening or elimination of features based on permutation and correlation analysis. Instead, we utilized all available features to train our models comprehensively for this study. This decision ensures that we do not overlook any potential relationships between the features and the phases of iron oxide, and it allows us to draw more robust conclusions from our analysis [62,65].

Particle size significantly influences the catalytic performance of iron oxide products, so it is important to know the size of the particles before conducting experiments [24,66–68]. Predicting the particle size of synthesized iron oxides based on the reaction conditions is of great interest. In this study, we initially attempted to use a random forest regression algorithm to train models for predicting the exact particle size based on the experimental conditions. The features used to train the particle size prediction models are exactly the same as the phase prediction. However, as shown in Fig. S1, the random forest regression model was unable to accurately predict the particle size of iron oxide particles from the test dataset. We hypothesized that the small size of our dataset may have contributed to the low accuracy of the random forest regression model, as the insufficient information prevented training the model effectively. To address this issue, we converted the prediction of particle size from a regression question into a classification question. By sorting the particle size into three categories: nano (less than 100 nm), submicron (100 nm to 1000 nm), and micron (greater than 1000 nm), we were able to use three machine learning algorithms to train models for predicting particle size based on the experimental conditions, including SVM, KNN, and RF (Fig. 4a). LR was excluded from the analysis for two reasons. Firstly, the performance of the LR model for binary phase classification was poor. Secondly, the LR model is generally not suitable for multi-class classification without employing several tricks, such as transforming the multi-class problem into multiple binary classification problems.

During the training process, the SVM algorithm failed to converge and produced no model (Fig. 4b). The KNN algorithm converged during training, but the resulting model has relatively low overall accuracy (62%), with recalls of 61% (nano), 63% (sub-micron), and 67% (micron) for predicting particle sizes from conditions in the test dataset (see Fig. 4c). In contrast, the RF-based model demonstrated the highest accuracy among the three algorithms, with an overall accuracy of 81%.

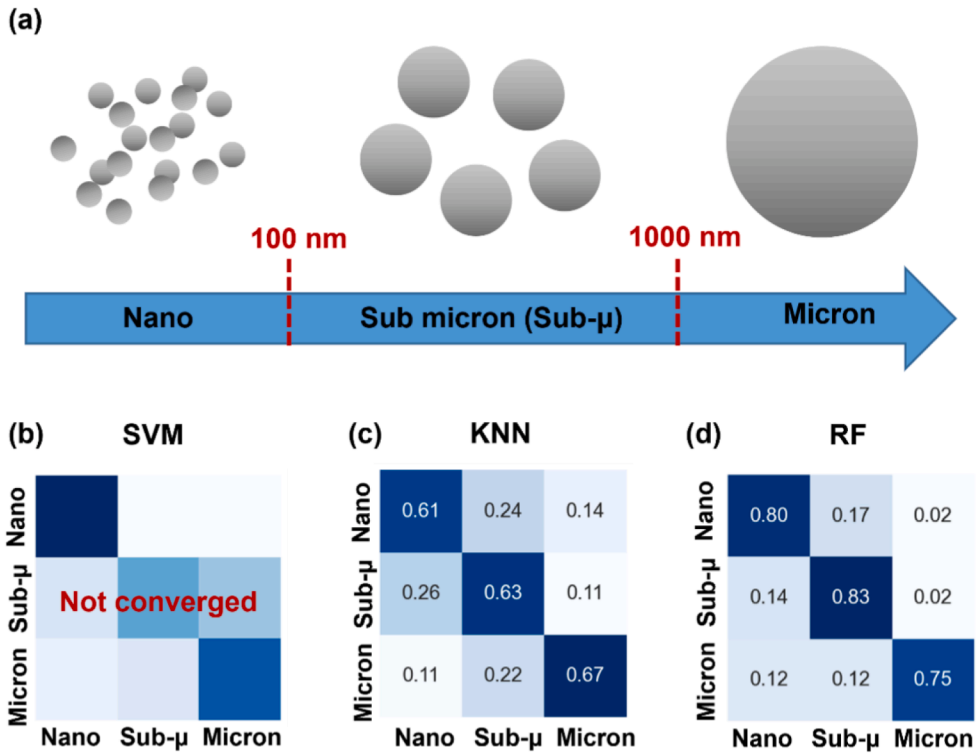

Fig. 4. Prediction of particle size of products based on conditions of reactions. (a) Dividing product particles into three categories based on size. (b) to (d) Prediction results based on models from SVM (b), KNN (c), and RF (d). Training SVM models were unsuccessful since fitting was not converged.

The RF-based model achieved recalls of 80%, 83%, and 75% for predicting nano, sub-micron, and micron particles, respectively (see Fig. 4d).

The feature importance analysis of the RF-based model showed that the pH, time, precursor concentration, volume, and temperature are the key features contributing to the categorization of the size of particles (see Fig. S2a). The pH has the highest importance (0.18), followed by time (0.17), precursor concentration (0.16), volume (0.13), and temperature (0.11). The pH is well-known to affect the phase of iron oxide, which, in turn, influences the size of particles. A high concentration of Fe^{3+} and long reaction time favor the growth of iron oxide particles with a large diameter, vice versa [69,70]. Similar to pH, the main effect of temperature is changing the possible phase of the products and therefore influences the size of particles. The impacts of volume (corresponding to pressure) during the hydrothermal synthesis are less studied than the rest of the important features. As the model heavily relies on this feature to predict the diameter of the particles, more experimental studies are needed to understand the relationship between pressure during the synthesis and the size of iron oxide particles. Similar to the RF based phase prediction model, this particle size prediction also utilized additive concentrations as the 6th important feature (see Table S7).

The model also used the concentration of surfactants to predict the size of the iron oxide particles, but the importance of this feature (0.04) was much lower than the others. One plausible reason is that the majority of parameter sets were surfactant-free iron oxide synthesis hoping to avoid the contamination of the iron oxide particles by the organic molecules. Thus, the number of sets of parameters with surfactant may not have been enough to unveil the relationship between surfactants and particle size, and the model is not able to fully utilize this feature to predict the size of particles.

We also tried using a hierarchical cluster map to investigate whether any of the key features may influence the particle size, in a way similar to how temperature and pH value affect the formation of hematite. However, such an influence was not able to be revealed, even when we narrowed down of the dataset and excluded all reactions that produced non-hematite particles. It seems that the control of particle size is achieved through a complex combination of all the features, and the cluster map method may not be capable of revealing such influences.

3.2. Experimental validation of ML models

To further evaluate the performance of the RF-based binary classification model and particle size prediction model, we synthesized iron oxides with and without adding surfactants (see Table 1 and 2). We prepared 18 surfactant-free samples, which are shown in Table 1, Figs. S3 and S4. The binary phase classification model correctly identified the formation of hematite from 15 sets of conditions, resulting in an accuracy of 83%. The model incorrectly classified three sets of conditions, which were the combination of high pH with high temperature (Exp. 3) and low pH with low temperature (Exp. 4 and 5).

SEM observations suggest that the RF model for predicting particle size based on experimental conditions is less accurate than the binary phase classification model. The RF model accurately predicted the size range of particles in 11 out of 18 samples (Sample 1, 2, 3, 6, 8, 9, 10, 12, 14, 15, and 16). For Samples 7 and 13, the model's predictions were mixed, with SEM images showing that Sample 7 contained both nano and sub-micron particles, while Sample 13 was made up of sub-micron and micron particles. The model also incorrectly predicted the particle size of Sample 4, 5, 11, 17, and 18.

The RF model accurately predicted the phase and size of iron oxide particles synthesized with sodium dodecyl sulfate (SDS) while the predictions for all experiments involving sodium citrate (SC) were inaccurate. Closer examination of the experimental conditions for iron oxide synthesis with SDS (Exp. 19–21) revealed that the additive may alter the morphology of the iron oxide particles but not the phase type. The synthesis conditions of Exp. 19 and 21 were acidic or weak basic environments (pH of 2.0 and 8.0 respectively) with relatively high reaction temperature (180 °C). The combination of these two conditions is

Table 1
Utilizing random forest-based models to predict phase and size of iron oxide products on random generated experiments. Volume of solvent and reaction time were fixed at 15 ml and 16 h, respectively. XRD was used to identify phases (Figs. S3 and S4), and SEM was utilized to identify size of particles (Fig. S5).

No.	$FeCl_3$ conc. (mM)	pH	Temp. (°C)	Is hem. Predict	Predict size	Exp. phase *	Exp. size
Exp 1	500	1.1	180	True	Micron	Hem	Micron
Exp 2	428	2.0	180	True	Micron	Hem	Micron
Exp 3	372	12.6	180	True	Nano	Gt	Nano
Exp 4	500	1.1	80	True	Micron	Gt	Sub-μ
Exp 5	428	2.0	80	True	Micron	Gt	Nano
Exp 6	372	12.6	80	False	Nano	Gt	Nano
Exp 7	100	1.7	180	True	Sub-μ	Hem	Sub-μ & Micron
Exp 8	98	2.1	180	True	Sub-μ	Hem	Sub-μ
Exp 9	90	12.8	180	False	Sub-μ	Gt	Sub-μ
Exp 10	100	1.7	80	False	Sub-μ	Aka	Sub-μ
Exp 11	98	2.1	80	False	Sub-μ	Aka	Nano
Exp 12	90	12.8	80	False	Nano	Gt	Nano
Exp 13	10	2.4	180	True	Sub-μ	Hem	Nano & Sub-μ
Exp 14	10	11.9	180	False	Sub-μ	Gt	Sub-μ
Exp 15	10	3.7	180	True	Sub-μ	Hem	Sub-μ
Exp 16	10	2.4	80	False	Sub-μ	Aka	Sub-μ
Exp 17	10	11.9	80	False	Nano	Gt	Sub-μ
Exp 18	10	3.7	80	False	Sub-μ	Aka	Nano

* Hem.: Hematite, Gt.: Goethite, Aka.: Akageneite.

known to cause the formation of hematite. In Exp. 20, the reaction temperature was only 110 °C, favoring the formation of non-hematite products, such as akageneite in this case. On the other hand, the introduction of SC significantly altered the phase of the iron oxide products, even when the experimental conditions clearly favored the formation of hematite. For instance, in Exp. 22, we observed the formation of magnetite (Fe_3O_4) at high reaction temperature and low pH value, while ferrihydrite was formed in the low-temperature reaction. The reason SC significantly altered the formation of phases is that this additive can cause the reduction of Fe^{3+} ions into Fe^{2+}. In contrast, SDS additive does not react with Fe^{3+} ions. Overall, the RF model exhibited good predicting accuracy when the additive was not reactive with Fe ions.

The findings from the feature importance analysis (Fig. 2d), correlation analysis (Fig. 3b), and the comparison of model predictions versus experimental results (see Tables 1 and 2) indicate that the RF-based model is most likely to predict the formation of hematite when either of the following conditions is met: low pH or high reaction temperature. This observation aligns with the thermodynamic perspective on the preferred conditions that induce the formation of various iron oxide phases, as discussed previously [21,64]. Notably, the RF model also takes into account precursor concentration, volume of solvent (may be corresponded to pressure), and reaction time to determine the likelihood of hematite formation (Fig. 2d). However, no evident relationship between the formation of hematite and these features was observed during the correlation analysis (Fig. 3d). It is believed that these three features exert a more subtle influence on the formation of hematite compared to pH and reaction temperature, warranting further investigation.

3.3. Optimizing synthesis algorithm based on dataset

We have devised a search and ranking algorithm for suggesting relevant previous studies from our dataset, aiding in the synthesis of iron oxide particles with specific properties, including phase and particle size. This algorithm presents the parameters employed in previous studies, along with their corresponding digital object identifiers (DOIs) (refer to Fig. 5). The algorithm operates on the principle that values frequently employed in a particular feature to synthesize the desired product are more likely to yield successful outcomes compared to infrequently used values. Therefore, if a previous study employs values that are consistently employed to synthesize the desired products, it receives a high recommendation, and vice versa.

To implement this idea, we retrieved all possible sets of reaction conditions from the iron oxide dataset that could produce the desired product and calculated the mean and standard deviation of the features from the resulting sub-dataset (Fig. 5a and 5b). We then calculated a ranking score for each set of parameters using a detailed algorithm outlined in the methods section. The algorithm considers the distance between each feature value and its corresponding average value in the sub-dataset (Fig. 5c). Sets of parameters with higher ranking scores are more likely to have been used in previous studies to synthesize the desired product. Conversely, sets of parameters with lower ranking scores are less likely to have been used. The output of the algorithm is a list of recommended sets of parameters for synthesizing the desired product in the previous studies and DOIs of the studies using these parameters. This algorithm can significantly reduce the time and effort required to search for specific iron oxide particles. An example of the algorithm's output for synthesizing hematite nanoparticles with diameters between 25 nm and 75 nm is shown in Table S4. It should be noted that this algorithm is not capable of suggesting synthesis parameters beyond the range the dataset.

4. Conclusion

This study addresses two significant challenges in materials synthesis: predicting the outcome of a synthesis from specified reaction

Table 2
Utilizing random forest-based models to predict phase and size of iron oxide products on random generated experiments with surfactant. Volume was fixed at 15 ml. XRD was used to identify phases (see Fig. S6), and SEM was utilized to identify size of particles (see Fig. S7).

No.	$FeCl_3$ conc. (mM)	Time (h)	pH	Additive*	Additive conc. (mM)	Temp. (°C)	Is hem. Predict	Predict size	Exp. phase**	Exp. size
Exp 19	831	7	2.0	SDS	131	180	True	Nano	Hem	Nano
Exp 20	971	5	8.0	SDS	181	110	False	Nano	Aka	Nano
Exp 21	501	29	8.0	SDS	121	180	True	Sub-μ	Hem	Sub-μ
Exp 22	461	37	2.0	SC	141	180	True	Sub-μ	Mag	Nano
Exp 23	671	41	2.0	SC	191	95	True	Micron	2L-Fh	Nano
Exp 24	471	34	10.0	SC	61	110	False	Nano	2L-Fh	Nano

* SDS: Sodium dodecyl sulfate, SC: Sodium citrate.
** Hem: Hematite, Gt: Goethite, Aka: Akaganeite, Mag: Magnetite, 2L-Fh: 2-line Ferrihydrite.

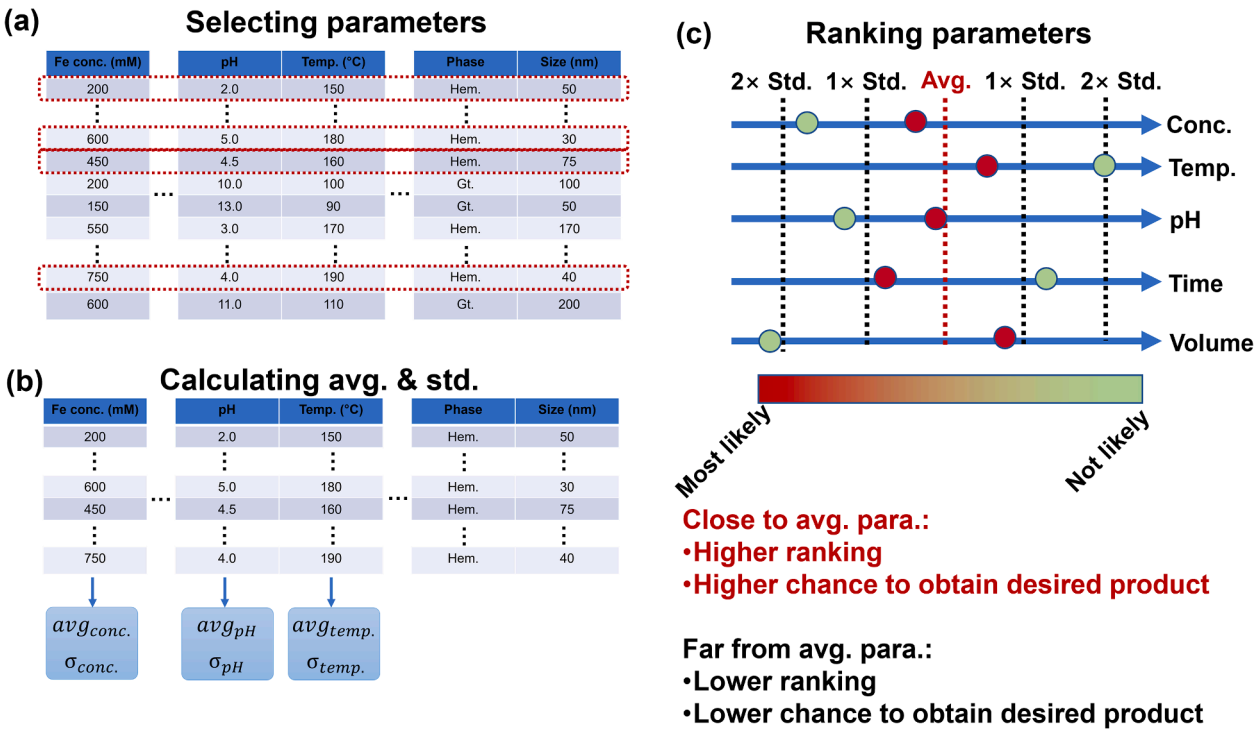

Fig. 5. Retrieving desired synthesis parameters to obtain iron oxide particles with specified phase and particle size. (a) Selecting all possible sets of parameters from dataset (e.g., hematite with 25 nm to 75 nm diameter). (b) Calculating average and standard derivation from every category of parameter. (c) Illustration of ranking parameters in every set and summarize result. The red dots show the set of parameters more likely to synthesize desired product, vice versa for the green dots.

parameters, and correlating sets of parameters to obtain products with desired outcomes. To predict experimental outcomes, we trained four machine learning algorithms, including random forest, logistic regression, support vector machine, and k-nearest neighbor, to predict the phase and particle size of iron oxide based on experimental conditions. Among these models, random forest demonstrated the best performance, achieving 96% and 81% accuracy in predicting the phase and size of iron oxides in the test dataset. The permutation feature importance analysis revealed that volume, which is plausibly correlated with pressure, exhibits a significant influence to the phase and size of iron oxide particles, along with precursor concentration, pH, temperature, and time. The random forest-based models were further evaluated by experimentally synthesizing iron oxide particles in both additive-free and additive systems, demonstrating overall good accuracy. Additionally, a searching and ranking algorithm was developed to recommend potential synthesis parameters from previous studies for obtaining iron oxide products with desired phase and particle size from previous studies in the dataset. This study lays the groundwork for a closed-loop approach to materials synthesis and preparation, from suggesting potential reaction parameters in the dataset and predicting potential outcomes, through conducting experiments and analysis, to enriching the dataset.

Author Contributions

X.Z. conceived the project. X. Z., E.G.S. and K.M.R. supervised the project. X.L., M.Z., Y.W., S.W., P.C., Y.Z., L.L. and Y.L. collected the literature and experimental data and developed the database. J.L, Z.Z., E.G.S., and X.Z. performed the data analysis and ML. J.L., Z.Z, X.L. and X.Z. wrote the manuscript with inputs from all co-authors. J.L., Z.Z., and X.L. are considered the co-first author. W.W., S.W., and X.G. helped with project design and manuscript refinement. All authors have given approval to the final version of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have made the data publicly available. The iron oxide synthesis dataset is available at https://data.pnl.gov/group/189/nodes/dataset/35215.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division through its Geosciences program at Pacific Northwest National Laboratory (PNNL) (FWP 56674). A portion of the work was performed with the user proposal 51382 using the Environmental and Molecular Sciences Laboratory (EMSL), a national scientific user facility at PNNL sponsored by the DOE's Office of Biological and Environmental Research. PNNL is a multi-program national laboratory operated by Battelle Memorial Institute under contract no. DE-AC05-76RL01830 for the DOE. E.S., Y.L., W.W. and X.Z. also acknowledge supported by Energy Storage Materials Initiative (ESMI), which is a Laboratory Directed Research and Development Project at Pacific Northwest National Laboratory (PNNL). S.W. acknowledge the support of this work by the National Science Foundation (NSF), Division of Civil, Mechanical, & Manufact Innovation, under award No. 1934120. X.G. acknowledge the support of this work by the National Science Foundation (NSF), Division of Earth Sciences, under award No. 2149848.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cej.2023.145216.

References

[1] Y. Chen, Z. Lai, X. Zhang, Z. Fan, Q. He, C. Tan, H. Zhang, Phase engineering of nanomaterials, Nat. Rev. Chem. 4 (5) (2020) 243–256, https://doi.org/10.1038/s41570-020-0173-4.

[2] X.H. Liu, L. Zhong, S. Huang, S.X. Mao, T. Zhu, J.Y. Huang, Size-dependent fracture of silicon nanoparticles during lithiation, ACS Nano 6 (2) (2012) 1522–1531.

[3] Z. Abbas, C. Labbez, S. Nordholm, E. Ahlberg, Size-dependent surface charging of nanoparticles, J. Phys. Chem. C 112 (15) (2008) 5715–5723.

[4] H. Jiang, K.-S. Moon, H. Dong, F. Hua, C. Wong, Size-dependent melting properties of tin nanoparticles, Chem. Phys. Lett. 429 (4–6) (2006) 492–496.

[5] S.R. Emory, W.E. Haskins, S. Nie, Direct observation of size-dependent optical enhancement in single metal nanoparticles, J. Am. Chem. Soc. 120 (31) (1998) 8009–8010.

[6] F.W. Wise, Lead salt quantum dots: the limit of strong quantum confinement, Acc. Chem. Res. 33 (11) (2000) 773–780.

[7] T. Takagahara, K. Takeda, Theory of the quantum confinement effect on excitons in quantum dots of indirect-gap materials, Phys. Rev. B 46 (23) (1992) 15578–15581.

[8] X. Zhou, W. Xu, G. Liu, D. Panda, P. Chen, Size-dependent catalytic activity and dynamics of gold nanoparticles at the single-molecule level, J. Am. Chem. Soc. 132 (1) (2010) 138–146.

[9] K. An, G.A. Somorjai, Size and shape control of metal nanoparticles for reaction selectivity in catalysis, ChemCatChem 4 (10) (2012) 1512–1524.

[10] W. Xu, Y. Bai, Y. Yin, Surface engineering of nanostructured energy materials, Adv. Mater. 30 (48) (2018), 1802091.

[11] Z. Fan, H. Zhang, Crystal phase-controlled synthesis, properties and applications of noble metal nanomaterials, Chem. Soc. Rev. 45 (1) (2016) 63–82.

[12] G.C. Hadjipanayis, R.W. Siegel, Nanophase Materials: Synthesis-Properties-Applications, Springer Science & Business Media, 2012.

[13] K.E. Drexler, Machine-phase nanotechnology, Sci. Am. 285 (3) (2001) 74–75.

[14] H. Li, X. Zhou, W. Zhai, S. Lu, J. Liang, Z. He, H. Long, T. Xiong, H. Sun, Q. He, Z. Fan, H. Zhang, Phase engineering of nanomaterials for clean energy and catalytic applications, Adv. Energy Mater. 10 (40) (2020), 2002019.

[15] A. Navrotsky, L. Mazeina, J. Majzlan, Size-driven structural and thermodynamic complexity in iron oxides, science 319 (5870) (2008) 1635–1638, https://doi.org/10.1126/science.1148614.

[16] J. McHale, A. Auroux, A. Perrotta, A. Navrotsky, Surface energies and thermodynamic phase stability in nanocrystalline aluminas, Science 277 (5327) (1997) 788–791.

[17] M.S. Chavali, M.P. Nikolova, Metal oxide nanoparticles and their applications in nanotechnology, SN Appl. Sci. 1 (6) (2019) 607.

[18] K. Sivula, F. Le Formal, M. Grätzel, Solar water splitting: progress using hematite (α-Fe2O3) photoelectrodes, ChemSusChem 4 (4) (2011) 432–449.

[19] S.S. Shinde, R.A. Bansode, C.H. Bhosale, K.Y. Rajpure, Physical properties of hematite α-Fe2O3 thin films: application to photoelectrochemical solar cells, J. Semicond. 32 (1) (2011), 013001.

[20] A. Rufus, N. Sreeju, D. Philip, Synthesis of biogenic hematite (α-Fe 2 O 3) nanoparticles for antibacterial and nanofluid applications, RSC Adv. 6 (96) (2016) 94206–94217.

[21] X. Li, A. Sheng, Y. Ding, J. Liu, A model towards understanding stabilities and crystallization pathways of iron (oxyhydr)oxides in redox-dynamic environments, Geochim. Cosmochim. Acta 336 (2022) 92–103, https://doi.org/10.1016/j.gca.2022.09.002.

[22] M. Zong, D. Song, X. Zhang, X. Huang, X. Lu, K.M. Rosso, Facet-dependent photodegradation of methylene blue by hematite nanoplates in visible light, Environ Sci Technol 55 (1) (2021) 677–688, https://doi.org/10.1021/acs.est.0c05592.

[23] X. Huang, Y. Chen, E. Walter, M. Zong, Y. Wang, X. Zhang, O. Qafoku, Z. Wang, K. M. Rosso, Facet-specific photocatalytic degradation of organics by heterogeneous fenton chemistry on hematite nanoparticles, Environ. Sci. Technol. 53 (17) (2019) 10197–10207, https://doi.org/10.1021/acs.est.9b02946.

[24] M. Zong, X. Zhang, Y. Wang, X. Huang, J. Zhou, Z. Wang, J.J. De Yoreo, X. Lu, K. M. Rosso, Synthesis of 2D hexagonal hematite nanosheets and the crystal growth mechanism, Inorg Chem 58 (24) (2019) 16727–16735, https://doi.org/10.1021/acs.inorgchem.9b02883.

[25] H. Tao, T. Wu, M. Aldeghi, T.C. Wu, A. Aspuru-Guzik, E. Kumacheva, Nanoparticle synthesis assisted by machine learning, Nat. Rev. Mater. 6 (8) (2021) 701–716, https://doi.org/10.1038/s41578-021-00337-5.

[26] H. Lv, X. Chen, Intelligent control of nanoparticle synthesis through machine learning, Nanoscale 14 (18) (2022) 6688–6708, https://doi.org/10.1039/d2nr00124a.

[27] J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, NPJ Comput. Mater. 5 (1) (2019) 83, https://doi.org/10.1038/s41524-019-0221-0.

[28] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (7715) (2018) 547–555, https://doi.org/10.1038/s41586-018-0337-2.

[29] D.E. Jones, H. Ghandehari, J.C. Facelli, A review of the applications of data mining and machine learning for the prediction of biomedical properties of nanoparticles, Comput. Methods Programs Biomed. 132 (2016) 93–103, https://doi.org/10.1016/j.cmpb.2016.04.025.

[30] R. Jinnouchi, R. Asahi, Predicting catalytic activity of nanoparticles by a DFT-aided machine-learning algorithm, J. Phys. Chem. Lett. 8 (17) (2017) 4279–4283, https://doi.org/10.1021/acs.jpclett.7b02010.

[31] X. Yan, A. Sedykh, W. Wang, X. Zhao, B. Yan, H. Zhu, In silico profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches, Nanoscale 11 (17) (2019) 8352–8362, https://doi.org/10.1039/c9nr00844f.

[32] D.R. Cox, The regression analysis of binary sequences, J. Roy. Stat. Soc.: Ser. B (Methodol.) 20 (2) (1958) 215–242.

[33] A. Liaw, M. Wiener, Classification and regression by randomForest, R news 2 (3) (2002) 18–22.

[34] E. Fix, J.L. Hodges Jr, Discriminatory analysis-nonparametric discrimination: Small sample performance, California Univ Berkeley, 1952.

[35] D.J. MacKay, D.J. Mac Kay, Information Theory, Inference and Learning Algorithms, Cambridge University Press, 2003.

[36] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297, https://doi.org/10.1007/BF00994018.

[37] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.

[38] J. Mockus, Bayesian Approach to Global Optimization: Theory and Applications, Springer Science & Business Media, 2012.

[39] B. Sun, A.S. Barnard, Visualising multi-dimensional structure/property relationships with machine learning, J. Phys.: Mater. 2 (3) (2019), 034003, https://doi.org/10.1088/2515-7639/ab0faa.

[40] X. Wang, J. Li, H.D. Ha, J.C. Dahl, J.C. Ondry, I. Moreno-Hernandez, T. Head-Gordon, A.P. Alivisatos, AutoDetect-mNP: an unsupervised machine learning algorithm for automated analysis of transmission electron microscope images of metal nanoparticles, JACS Au 1 (3) (2021) 316–327, https://doi.org/10.1021/jacsau.0c00030.

[41] W. Lee, H.S. Nam, Y.G. Kim, Y.J. Kim, J.H. Lee, H. Yoo, Robust autofocusing for scanning electron microscopy based on a dual deep learning network, Sci. Rep. 11 (1) (2021) 20933, https://doi.org/10.1038/s41598-021-00412-5.

[42] F. Pellegrino, R. Isopescu, L. Pellutie, F. Sordello, A.M. Rossi, E. Ortel, G. Martra, V. D. Hodoroaba, V. Maurino, Machine learning approach for elucidating and predicting the role of synthesis parameters on the shape and size of TiO2 nanoparticles, Sci. Rep. 10 (1) (2020) 18910, https://doi.org/10.1038/s41598-020-75967-w.

[43] M. Lu, H. Ji, Y. Zhao, Y. Chen, J. Tao, Y. Ou, Y. Wang, Y. Huang, J. Wang, G. Hao, Machine learning-assisted synthesis of two-dimensional materials, ACS Appl. Mater. Interfaces 15 (1) (2023) 1871–1878, https://doi.org/10.1021/acsami.2c18167.

[44] S. Wu, Z. Wang, H. Zhang, J. Cai, J. Li, Deep learning accelerates the discovery of two-dimensional catalysts for hydrogen evolution reaction, Energy & Environ. Mater. 6 (1) (2023) e12259, https://doi.org/10.1002/eem2.12259.

[45] O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, G. Ceder, Text-mined dataset of inorganic materials synthesis recipes, Sci. Data 6 (1) (2019) 203, https://doi.org/10.1038/s41597-019-0224-1.

[46] J. Savage, A. Kishimoto, B. Buesser, E. Diaz-Aviles, C. Alzate, Chemical reactant recommendation using a network of organic chemistry, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, Como, Italy; 2017.

[47] R. Shibukawa, S. Ishida, K. Yoshizoe, K. Wasa, K. Takasu, Y. Okuno, K. Terayama, K. Tsuda, CompRet: a comprehensive recommendation framework for chemical synthesis planning with algorithmic enumeration, J. Cheminf. 12 (1) (2020) 52, https://doi.org/10.1186/s13321-020-00452-5.

[48] *pandas-dev/pandas: Pandas 1.2.2*; Zenodo: 2021. https://doi.org/10.5281/zenodo.4524629 (accessed).

[49] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nature 585 (7825) (2020) 357–362.

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, D.V. Scikit-learn, Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[51] *openpyxl*; 2022. https://foss.heptapod.net/openpyxl/openpyxl (accessed).

[52] T. Kluyver, B. Ragan-Kelley, F. Pérez, B.E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J.B. Hamrick, J. Grout, S. Corlay, *Jupyter Notebooks-a publishing format for reproducible computational workflows*; 2016.

[53] M. Waskom, seaborn: statistical data visualization, J. Open Source Softw. 6 (60) (2021), https://doi.org/10.21105/joss.03021.

[54] J.D. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. 9 (3) (2007) 90–95, https://doi.org/10.1109/MCSE.2007.55.

[55] J.W. Lee, W.B. Park, J.H. Lee, S.P. Singh, K.S. Sohn, A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns, Nat. Commun. 11 (1) (2020) 86, https://doi.org/10.1038/s41467-019-13749-3.

[56] S.M. Azimi, D. Britz, M. Engstler, M. Fritz, F. Mucklich, Advanced steel microstructural classification by deep learning methods, Sci. Rep. 8 (1) (2018) 2128, https://doi.org/10.1038/s41598-018-20037-5.

[57] L. Ju, A. Lyu, H. Hao, W. Shen, H. Cui, Deep learning-assisted three-dimensional fluorescence difference spectroscopy for identification and semiquantification of illicit drugs in biofluids, Anal. Chem. 91 (15) (2019) 9343–9347, https://doi.org/10.1021/acs.analchem.9b01315.

[58] M. Ziatdinov, O. Dyck, A. Maksov, X. Li, X. Sang, K. Xiao, R.R. Unocic, R. Vasudevan, S. Jesse, S.V. Kalinin, Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations, ACS Nano 11 (12) (2017) 12742–12752, https://doi.org/10.1021/acsnano.7b07504.

[59] M. Ojala, G.C. Garriga, Permutation tests for studying classifier performance, in: *2009 Ninth IEEE International Conference on Data Mining*, 6-9 Dec. 2009, 2009; pp 908-913. DOI: 10.1109/ICDM.2009.108.

[60] R.I. Walton, Subcritical solvothermal synthesis of condensed inorganic materials, Chem. Soc. Rev. 31 (4) (2002) 230–238, https://doi.org/10.1039/b105762f From NLM PubMed-not-MEDLINE.

[61] O. Schäf, H. Ghobarkar, P. Knauth, Hydrothermal synthesis of nanomaterials, in: P. Knauth, J. Schoonman (Eds.), Electronic Materials: Science & TechnologyNanostructured Materials, Kluwer Academic Publishers, Boston, 2004, pp. 23–41.

[62] H. Zhang, H. Fu, X. He, C. Wang, L. Jiang, L.-Q. Chen, J. Xie, Dramatically enhanced combination of ultimate tensile strength and electric conductivity of alloys via machine learning screening, Acta Mater. 200 (2020) 803–810, https://doi.org/10.1016/j.actamat.2020.09.068.

[63] A. Ali, H. Zafar, M. Zia, I. Ul Haq, A.R. Phull, J.S. Ali, A. Hussain, Synthesis, characterization, applications, and challenges of iron oxide nanoparticles, Nanotechnol Sci Appl 9 (2016) 49–67, https://doi.org/10.2147/nsa.S99986 From NLM.

[64] R.M. Cornell, U. Schwertmann, The Iron Oxides: Structure, Properties, Reactions, Occurrences, And Uses, Wiley-VCH Weinheim, 2003.

[65] H.-X. Liu, Y.-F. Yang, Y.-F. Cai, C.-H. Wang, C. Lai, Y.-W. Hao, J.-S. Wang, Prediction of sintered density of binary W(Mo) alloys using machine learning, Rare Met. 42 (8) (2023) 2713–2724, https://doi.org/10.1007/s12598-022-02238-0.

[66] M. Khalil, N. Liu, R.L. Lee, Catalytic aquathermolysis of heavy crude oil using surface-modified hematite nanoparticles, Ind. Eng. Chem. Res. 56 (15) (2017) 4572–4579, https://doi.org/10.1021/acs.iecr.7b00468.

[67] M. Ashraf, I. Khan, M. Usman, A. Khan, S.S. Shah, A.Z. Khan, K. Saeed, M. Yaseen, M.F. Ehsan, M.N. Tahir, N. Ullah, Hematite and magnetite nanostructures for green and sustainable energy harnessing and environmental pollution control: A review, Chem. Res. Toxicol. 33 (6) (2020) 1292–1311.

[68] B. Hashemzadeh, H. Alamgholiloo, N. Noroozi Pesyan, E. Asgari, A. Sheikhmohammadi, J. Yeganeh, H. Hashemzadeh, Degradation of ciprofloxacin using hematite/MOF nanocomposite as a heterogeneous Fenton-like catalyst: A comparison of composite and core−shell structures, Chemosphere 281 (2021), 130970, https://doi.org/10.1016/j.chemosphere.2021.130970.

[69] A. Lassoued, B. Dkhil, A. Gadri, S. Ammar, Control of the shape and size of iron oxide (α-Fe2O3) nanoparticles synthesized through the chemical precipitation method, Results Phys. 7 (2017) 3007–3015, https://doi.org/10.1016/j.rinp.2017.07.066.

[70] M. Khalil, J. Yu, N. Liu, R.L. Lee, Hydrothermal synthesis, characterization, and growth mechanism of hematite nanoparticles, J. Nanopart. Res. 16 (4) (2014), https://doi.org/10.1007/s11051-014-2362-x.