

# Machine Learning-Assisted Carbon Dot Synthesis: Prediction of Emission Color and Wavelength

Ravithree D. Senanayake,<sup>†</sup> Xiaoxiao Yao,<sup>†</sup> Clarice E. Froehlich, Meghan S. Cahill, Trever R. Sheldon, Mary McIntire, Christy L. Haynes,\* and Rigoberto Hernandez\*



Cite This: <https://doi.org/10.1021/acs.jcim.2c01007>



Read Online

ACCESS |



Metrics & More

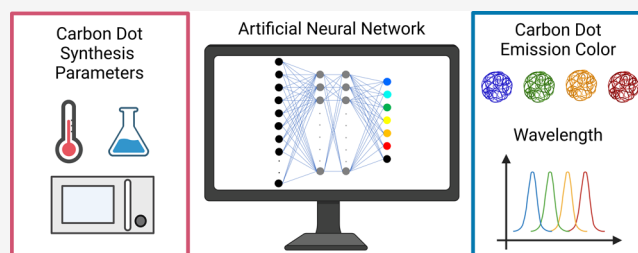


Article Recommendations



Supporting Information

**ABSTRACT:** Carbon dots (CDs) have attracted great attention in a range of applications due to their bright photoluminescence, high photostability, and good biocompatibility. However, it is challenging to design CDs with specific emission properties because the syntheses involve many parameters, and it is not clear how each parameter influences the CD properties. To help bridge this gap, machine learning, specifically an artificial neural network, is employed in this work to characterize the impact of synthesis parameters on and make predictions for the emission color and wavelength for CDs. The machine reveals that the choice of reaction method, purification method, and solvent relate more closely to CD emission characteristics than the reaction temperature or time, which are frequently tuned in experiments. After considering multiple models, the best performing machine learning classification model achieved an accuracy of 94% in predicting relative to actual color. In addition, hybrid (two-stage) models incorporating both color classification and an artificial neural network *k*-ensemble model for wavelength prediction through regression performed significantly better than either a standard artificial neural network or a single-stage artificial neural network *k*-ensemble regression model. The accuracy of the model predictions was evaluated against CD emission wavelengths measured from experiments, and the minimum mean average error is 25.8 nm. Overall, the models developed in this work can effectively predict the photoluminescence emission of CDs and help design CDs with targeted optical properties.



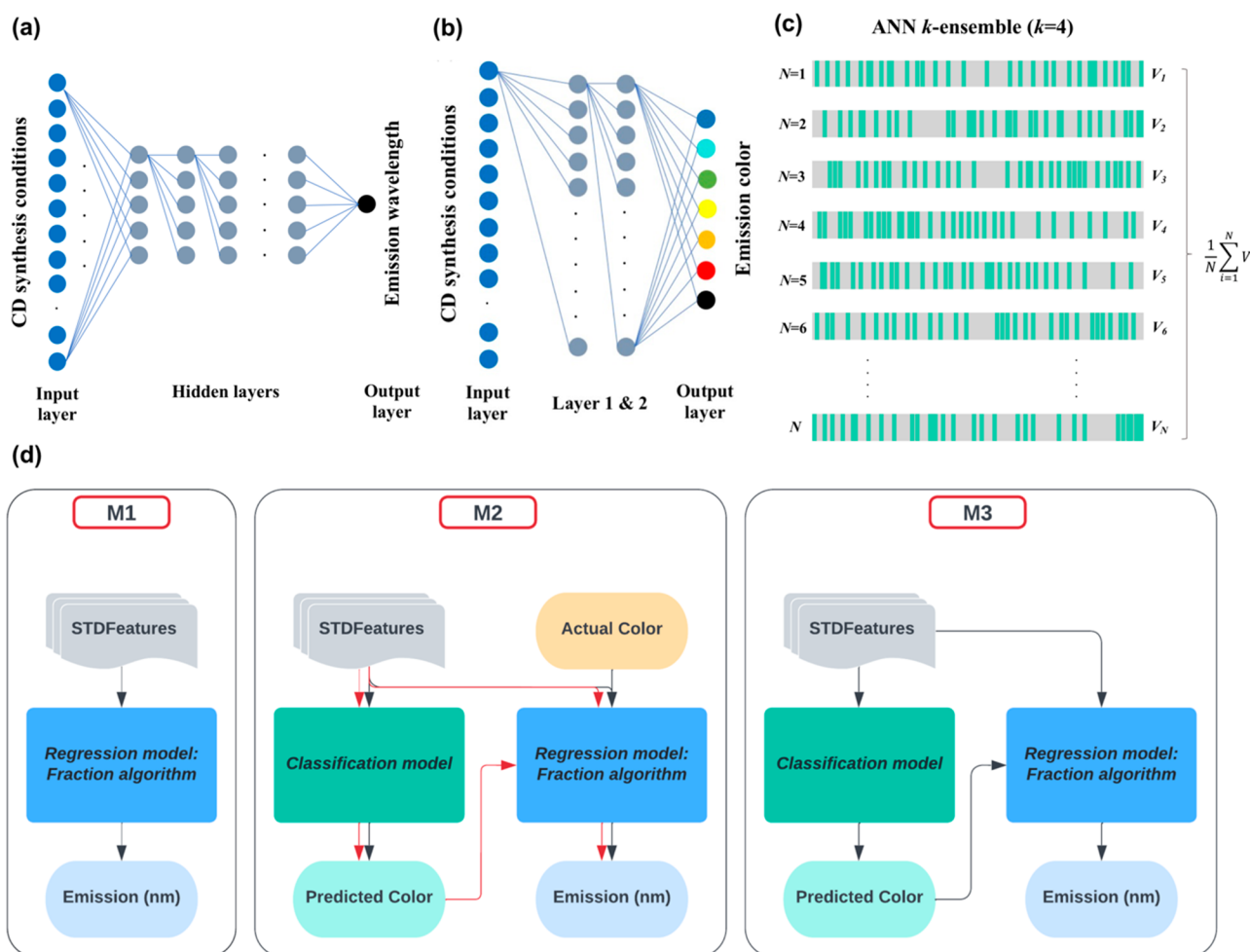
## INTRODUCTION

Carbon dots (CDs) are emerging fluorescent nanomaterials that have demonstrated great potential in a broad range of fields, such as bioimaging, sensing, and LEDs, due to their tunable fluorescence, good photostability, and low toxicity.<sup>1–3</sup> A range of precursors can be utilized to make CDs, from small molecules such as citric acid and ethylenediamine to bulk carbon materials such as graphite or biomass.<sup>4–6</sup> The synthesis methods are simple: usually, heat is employed to start the reaction using methods like hydrothermal treatment, pyrolysis, etc. Recently, room temperature syntheses have also been developed.<sup>7,8</sup> The broad range of these characteristics makes CDs excellent candidates for scale up in industrial applications. Currently, most CDs have blue or green emission characteristics, but they can also be engineered to have emission colors ranging from blue to red, covering the whole visible spectrum, or even near IR.<sup>9,10</sup> This tunable fluorescence makes CDs a competitive substitute for semiconductor quantum dots as a more environmentally friendly option available for a variety of applications. Unfortunately, despite the fact that CDs can be synthesized to have various emission characteristics, there is presently no good way to predict the emission color in the design of experiments. This is largely due to a lack of clarity regarding the formation mechanism and origin of photo-

luminescence for CDs. Thus, much laboratory effort is spent empirically tuning reaction parameters to obtain CDs with desired properties. It is especially challenging to expand the emission window to the red or near IR within such searches because most CDs exhibit blue or green fluorescence.

There are a few key factors, such as solvents and precursors, that have been reported to play a role in tuning the fluorescence properties of CDs.<sup>11</sup> Compared with water, organic solvents—such as formamide or dimethylformamide—are more likely to generate CDs with red-shifted fluorescence.<sup>12,13</sup> Precursors with aromatic structure such as phenylenediamine are likely to form extended  $\pi$  structures potentially contributing to red emission.<sup>14,15</sup> In addition, long wavelength-emissive CDs have been found to have either larger particle sizes,<sup>10</sup> higher surface oxidation degree,<sup>16</sup> or higher nitrogen doping.<sup>17</sup> Therefore, multiple parameters could influence the emission of CDs. To account for all these

Received: August 9, 2022



**Figure 1.** Machine learning models, their algorithm, and training workflows. (a) Artificial neural network (ANN) structure with input, hidden, and output layers used to predict emission wavelength through the regression model. (b) ANN structure with input, two hidden, and output layers used to predict emission color in the classification model, with only a few connections shown between nodes in ANN for simplicity. (c) ANN  $k$ -ensemble ( $k = 4$ ). (d) Training workflow of the three ML models (black arrows). M1 is the regression model, M2 is the hybrid model with classification and regression in parallel, and M3 is the hybrid model where regression model training follows classification training. The testing workflow in M1 and M3 are similar to their training workflows. In M2, the testing workflow is different from the training workflow and testing workflow as indicated by the red arrows.

parameters within an experimental design, usually a combinatorial set of reactions and conditions is performed to identify CDs with targeted performance, and this process can be very time-consuming.

A promising alternative to the slow empirical approach for tuning CD emission characteristics relies on a computational method that can incorporate multiple nonlinear and complex parameters to reveal the relationship between the synthesis parameters and CD emission wavelength. Machine learning (ML) is a subset of artificial intelligence that can make predictions for new samples based on training data. It is also an efficient way for finding relationships in a complex data set without explicit instructions. Recently, ML has been applied widely in chemistry and materials science in a variety of applications, such as in the design of retrosynthetic pathways for organic compounds and in the prediction of the likelihood of a molecule to crystallize.<sup>18</sup> Additionally, ML has been reported to assist with virtual rapid screening of pharmaceutical drug products.<sup>18</sup> Meanwhile, there is literature precedent demonstrating the feasibility of utilizing ML in the prediction of CD properties. For example, Han et al.<sup>19</sup> utilized a

regression ML model to assist in the prediction of quantum yield (QY) for *p*-dihydroxybenzene and ethylenediamine CDs. With this model, they were able to design green fluorescent CDs to achieve a QY as high as 39.3%. Similarly, Hong et al.<sup>8</sup> synthesized CDs from 400 reactions of *p*-benzoquinone and ethylenediamine at room temperature. Their use of the XGBoost<sup>20</sup> model demonstrated the best performance among other ML models and achieved a prediction coefficient of determination higher than 0.96. Additionally, Wang et al.<sup>21</sup> utilized a deep convolution neural network to predict the emission color of CDs and the nature of the emission as either excitation-dependent or not. The color prediction achieved 81% accuracy for CDs with blue and green fluorescence, while the prediction for red or multicolor CDs was less accurate. Unfortunately, the precursors used in two of the previous publications (*p*-dihydroxybenzene and *p*-benzoquinone) are relatively uncommon in the CD synthesis undertaken throughout the literature. Thus, the models or conclusions drawn from these studies are potentially limited by the absence of a full combinatorial search over the broad set of design parameters that is available through the use of ML.

Herein, we conduct a meta-analysis using data examples of various CD syntheses reported in the literature employing citric acid and urea or ethylenediamine as precursors. We restrict ourselves to these precursors because they are employed most frequently and thus, offer the most comprehensive data measured by a large number of research groups. A total of 407 data examples were collected from the literature, with 379 data examples used as the training database, while the other 28 data examples were set aside as an external test set to validate the model. In this way, we can account for different synthesis methods and errors across different laboratories. Statistical analysis of the database revealed the correlation between the features and CD emission. Artificial neural networks (ANNs) have been selected as the primary engines in the ML implementations for the prediction of CD emission because of their flexibility to address multicomponent data sets. Input features include precursor molar ratios, reaction method, solvent, purification method, pH, reaction temperature (K), and reaction time (min). We built three machines (M1–M3) based on combinations of two types of models: a classification model and a regression model. The classification model is used for categorical color prediction, and the regression model is used for numerical emission wavelength prediction. Both regression and classification models have a layered structure that delivers a complex transformation between the input and output layers through hidden layers along with nonlinearity by an activation function at each node. To evaluate the accuracy of the prediction, 16 new targeted CD synthesis reactions were performed with variation in a number of synthesis parameters, and their emission wavelengths were recorded and compared to model predictions. The color prediction from the classification model that does not include reaction temperature and time as features achieved a training accuracy value of 0.94. Interestingly, the emission wavelength prediction improved from mean average error (MAE) = 38.4 to 25.8 nm if the color from the training set was used as an input for the wavelength prediction model. This suggests that the emission prediction can be improved when both classification and regression methods are combined. The impact of including reaction temperature and time as features in the models is also assessed.

## METHODS

**Data Analysis and Feature Engineering.** The initial feature-label correlations were assessed utilizing the Pearson correlation coefficient in a linear regression: ordinary least-squares (OLS) regression and ANOVA modules available in the `statmodel` python package. Pearson correlation coefficients demonstrate the correlations between the numerical features and emission whereas the ANOVA test shows which categorical features are better correlated with emission. Feature values were normalized to have a standard deviation of 1 and a mean of 0. One-hot-encoding<sup>22,23</sup> (as implemented in the python `pandas` module) was used to convert categorical features into continuous values before passing them through to the ML model, as ANNs cannot interact with discrete data. In one-hot-encoding, a categorical feature is replaced by new numerical features created for each type in the feature category to represent the full information from the original categorical feature. The precursor molar ratios were treated in a semi-one-hot-encoded manner where we created a separate feature for each precursor (including acid and base additives) which was equal to the mole percentage of the given precursor. After

employing feature engineering with one-hot-encoding, a total of 61 and 63 numerical features were available for the cases without temperature/time and with temperature/time, respectively.

**ANN  $k$ -Model, Training, and Prediction.** To address the limitations arising from the use of a small data set in an ML model, we develop an ensemble of  $N$  ANNs each consisting of input, hidden, and output layers of nodes with a similar structure. The prediction (or label) can be an analog value such as the emission wavelength as indicated in the regression model of Figure 1a, or can be a set of one-hot encoded nodes as indicated in the classification model of Figure 1b. Each ANN is trained on a training set consisting of a randomized selection of examples. The training sets are a subset of the full database with a size equal to the  $(k - 1)/k$  fraction of the database while the remaining examples are used as the internal validation set (whose size is the  $(1/k)$  fraction of the database) for the corresponding training set. If the database were partitioned into  $k$  randomized subsets of equal size, then the ensuing  $k$  validation sets and their complementary subsets—viz training sets—would give rise to a  $k$ -bag of ANNs in bootstrap aggregation.<sup>24–31</sup> Instead, in the ANN  $k$ -ensemble, we pick only one of the members of a given  $k$ -bag, but we sample  $N$  of them, as illustrated in Figure 1c. In the limit that  $N$  is large, both ensemble methods should result in the same distribution of predictions, providing both an expectation value from its average and an estimate of the error from its root-mean-square deviation. In the present application, an emission prediction is obtained for each run, and the predictions are averaged over  $N$ . One additional layer of complexity in the model building is the use of staged ANNs, combining classification and then regression in different combinations as indicated in Figure 1d, discussed in more detail below. Together, a standard ANN model for color prediction through classification and the ANN  $k$ -ensemble for continuous label prediction optimized through regression lead to a better utilization of the available data set in training the model, and this helps to eliminate any bias.

There are, of course, alternate ML models that could have been used instead of the ANN  $k$ -ensemble model for CD property prediction—e.g., XGBoost,<sup>20</sup> K-nearest neighbor (KNN),<sup>32,33</sup> and support vector machine (SVM).<sup>34</sup> To confirm the efficacy of the current approach, such ML models were also trained from the CD data set. However, as reported in Figure S1, they were not as good as the selected approach in providing both classification and regression given the limited size of the data set.

The ANN  $k$ -ensemble was used previously by some of us to predict the viability of organisms exposed to engineered nanoparticles.<sup>30,35</sup> In that work, we found that a single ANN could, not surprisingly, overfit the data in predicting nanoparticle properties because of the size of the data set. However, use of bags of ANNs allowed us to address this problem through a sampling of an ensemble of ANNs whose relative errors could cancel out. As reported below, we found a similar reduction in error in employing the ANN  $k$ -ensemble rather than a single ANN on the CD data set consisting of only 379 examples.

From the initial feature-label correlations, the feature “color” indicated stronger correlations with the emission. However, “color” cannot be used as an input since it is an observed property, not a controlled factor in CD synthesis. Therefore, a classification model (Figure 1b) was built where we first predicted color using the standard feature set (STDFeatures)

listed in Table 1. The classification model consists of an input layer, two hidden layers, and an output layer with seven nodes

**Table 1. Numerical, Categorical, and Mixed Features Collected from the CD Synthesis Conditions, Reagents, and CD Nanomaterial Properties**

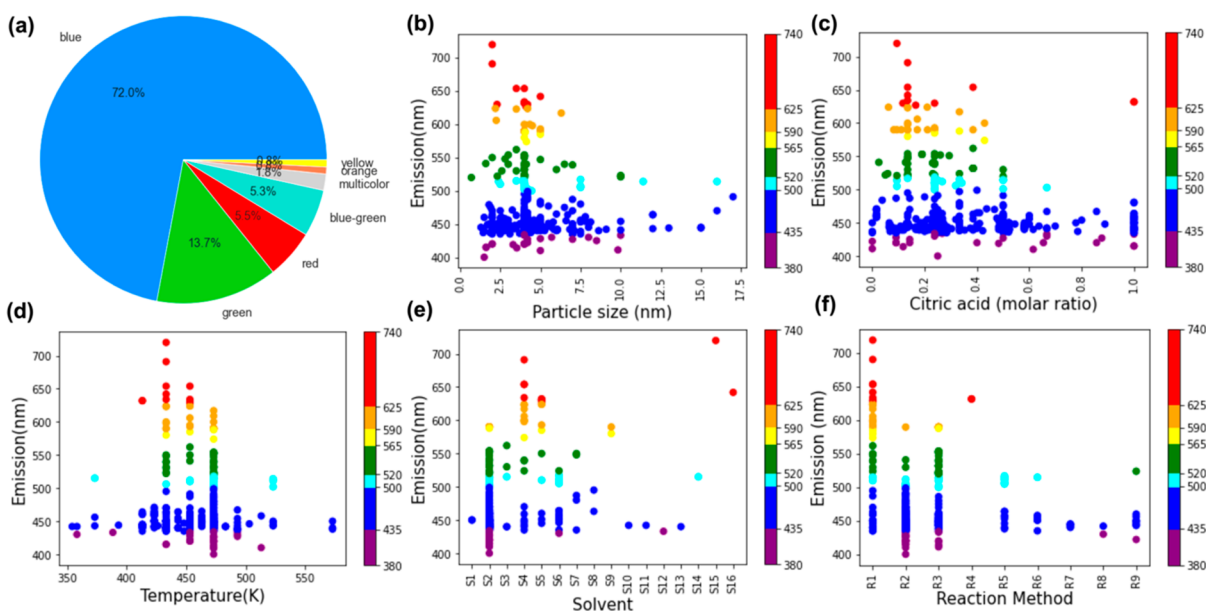
	standard features (STDFeatures)
numerical	reaction temperature (K), reaction time (min)
categorical	reaction method, solvent, pH, purification method
mixed	precursor composition (citric acid, urea, ethylenediamine, ammonium, NaOH, boric acid, sodium thiosulfate, KOH)

one-hot encoded to different output colors. Then, the predicted color from the classification model was used as a feature along with the STDFeatures to predict the emission wavelength from a regression model. This is a hybrid approach that uses both classification and regression models together to predict emission wavelengths. Our hypothesis is that using predicted color as an intermediate feature (or hidden variable) will improve the prediction of emission wavelength due to their strong correlations. The ANNs were built using Tensorflow and Keras version 2.2.0 in python.<sup>36–38</sup> We used python libraries numpy, scipy, pandas, scikit-learn, sklearn, matplotlib, and Jupyter Notebooks to preprocess, postprocess, and visualize data.<sup>36,39–45</sup>

Three ML machines were built to predict the emissions of CDs. The M1 machine was built using the ANN structure of Figure 1a and the ANN  $k$ -ensemble. In other words, M1 is simply an ANN  $k$ -ensemble that is optimized through regression. The STDFeatures (Table 1) are used in M1 to train the model and to predict emission wavelength. M2 and M3 are hybrid machines, containing both classification for color prediction and regression for wavelength prediction.

Specifically, M2 and M3 combine an ANN classification model with an ANN  $k$ -ensemble in stages. We are not aware of prior use of a hybrid machine to predict CD emission wavelength in which a classification model is used to predict the color first and then used as an input (combined with actual color in the training) to predict the emission wavelength from a regression ML model. Using color as an intermediate input feature is a unique way to increase the prediction accuracy. In M2, the ANN  $k$ -ensemble model is trained in parallel to the color prediction through classification. In M3, the color is first predicted from the classifier without optimization relative to the known color, and it is then fed into the ANN  $k$ -ensemble model along with the STDFeatures. The combined M3 model is trained only relative to the regression of the frequency prediction. For validation and prediction purposes, the emission color enters M2 in a slightly different way than how it is used in the training workflow. The reason is that the colors of the test CDs are unknown until they are predicted. Consequently, the actual color should not be used as a feature in predicting the emission for an unknown CD sample. Thus, for prediction from examples in new tests (or for a single material with novel features), M2 uses the same data workflow as in M3. That is, the predicted color from the classifier is entered as the value of the corresponding feature for the ANN  $k$ -ensemble model. Here, we implement an ensemble of ANNs, which uses an ANN  $k$ -ensemble ( $k = 4$ ), as shown in Figure 1c, to minimize biases arising from nonuniformly distributed data, since a small data set ( $\sim 379$ ) can be vulnerable to biased predictions.

**Design of Experiments.** Here, 5.16 mmol (1 unit) of urea or ethylenediamine was used in the postprediction reactions, while the amount of citric acid was randomly chosen from a generator to be from 0.1 to 1 unit, with a step size of 0.1 units.



**Figure 2.** (a) Pie chart representing the distribution of CD emission color in the training database (379 entries); scatter plots showing the relationship between emission wavelength (nm) and (b) particle size (nm), (c) citric acid (molar ratio), (d) temperature (K), (e) solvent, and (f) reaction method. The solvents in panel e are labeled S1 to S16 as follows: S1 (ethanol, formamide), S2 (water), S3 (ethanol), S4 (DMF), S5 (formamide), S6 (no solvent), S7 (glycerol), S8 (water, glycerol), S9 (glycerol, DMF), S10 (hydrogen, oxygen gas), S11 (acetonitrile), S12 (pyridine), S13 (water, formamide), S14 (toluene), S15 (DMSO), and S16 (DEF). The reaction methods in panel f are labeled R1 to R9 as follows: R1 (solvothermal treatment), R2 (hydrothermal treatment), R3 (microwave), R4 (microwave-assisted solvothermal), R5 (pyrolysis), R6 (conventional heating), R7 (microwave-assisted hydrothermal), R8 (gaseous detonation), and R9 (oven).

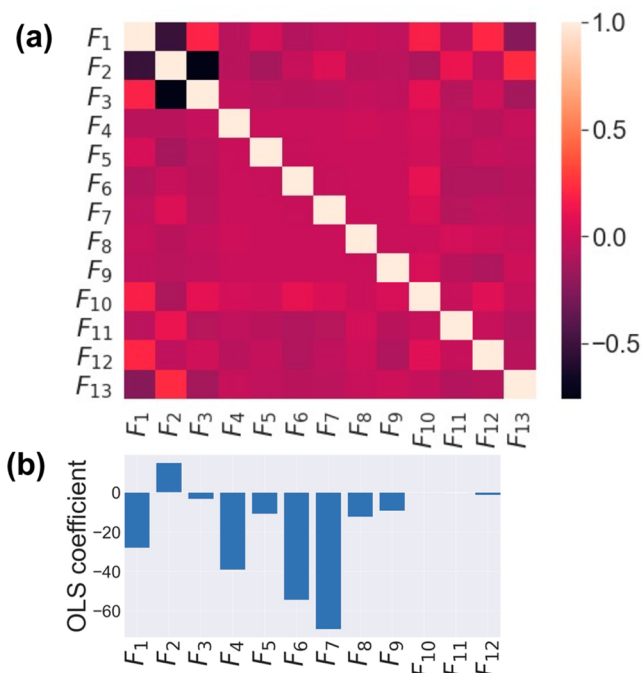
In cases where there was no urea or ethylenediamine, 1 unit of citric acid was used in the reaction. For 14 of the reactions, there was no additive, and the pH was set to be “neutral”. For the other 2 reactions, one was set as acidic, and the other was set as basic. Additives such as HCl or NaOH were added for pH adjustment. The reaction method was randomly selected to be one of the following: microwave-assisted hydrothermal or solvothermal treatment; microwave; hydrothermal or solvothermal treatment; or pyrolysis. The solvent was randomly selected from the following: water; formamide; dimethylformamide; or none. For the first set of experiments, the temperature was set as 150 °C, and the reaction time was set at 1 h. For the microwave method, the temperature cannot be controlled to be the same among different reactions. The reaction time was set as the length of time for all the water to evaporate when used as the solvent or 25 s when organic solvents were used. For the second set of experiments, the temperature was randomly chosen from 140, 160, 180, and 200 °C, and the reaction time was randomly chosen from 1, 2, 3, or 4 h, except for those reactions using the microwave method. For the microwave methods, the reaction time was randomly chosen from 20 to 80 s. If there was any precipitate observed during the initial mixing of starting materials, the reaction was not added to the database, and new conditions were randomly chosen. For example, CA:EDA = 0.5:1 in DMF generated a white precipitate, and the sample was discarded from the test set. All experimental conditions can be found in the [Supporting Information](#).

## RESULTS AND DISCUSSION

The majority of the CDs in the database (72%) are blue-colored. Green, red, and blue-green CDs represent 13.7%, 5.5%, and 5.3% of the database, respectively ([Figure 2a](#)). The database also contains small fractions of orange and yellow CDs. This color bias is unavoidable because it reflects the distributions of CDs that have been reported in the peer-reviewed literature. For multicolor CDs, the peak with the highest intensity for any excitation wavelength was used to define the main emission wavelength. The direct correlation plots reveal several notable connections. First, the low correlation of CD particle size and its emission wavelength suggests that quantum confinement and size effects do not play a major role for these types of CDs ([Figure 2b](#)). Most of the CDs have diameters between 1 and 7.5 nm, with emissions spanning from blue to red, whereas CDs outside of this diameter range mainly have blue colors. Additionally, when the molar ratio of citric acid is around 0.1 and that of urea/ethylenediamine is around 0.9, the CDs that are produced are most likely to exhibit longer emission wavelengths ([Figure 2c](#)). This suggests that nitrogen doping could be an important factor to consider when designing red-emissive CDs. It has been previously reported that nitrogen doping occasionally leads to a red shift in CD emissions,<sup>17</sup> though nitrogen doping is more frequently related to the higher fluorescence intensity/quantum yields of CDs.<sup>46,47</sup> The direct correlation plot of temperature and emission shows that blue and blue-green CDs can be synthesized within a wide range of temperatures while the green, yellow, orange, and red CDs are mostly synthesized within the temperature range of 420 to 475 K, corresponding with 147 to 202 °C ([Figure 2d](#)). Interestingly, some organic solvents such as DMF, formamide, DMSO, and DEF contribute to the formation of CDs with long wavelength emissions ([Figure 2e](#)). This observation is also supported by a

reaction method analysis as the solvothermal treatment method is more likely to generate orange and red CDs ([Figure 2f](#)). Furthermore, previous studies have also indicated that some organic solvents can contribute to the formation of CDs with long wavelength emissions.<sup>11</sup> For example, Tian et al.<sup>13</sup> synthesized CDs from citric acid and urea in three solvents: water, glycerol, and DMF. DMF had earlier been seen to lead to the formation of CDs with the longest emission wavelengths, and they concluded that this is due to a larger  $sp^2$  domain formed from dehydration and carbonization in DMF. Other direct correlations plotted with emission are shown in the [Supporting Information](#) ([Figure S2](#)).

Though the direct correlations can reveal key evidence about experimental factors that influence CD emission characteristics, they can miss more complex connections between the factors. Therefore, Pearson correlation coefficients and OLS regression were used to evaluate the numerical features of interest in predicting emission wavelength. The Pearson correlation coefficient heat map can reveal how each individual numerical feature correlates with other features, including the main emission peak. For example, the correlation between the citric acid feature and the main peak in emission feature is weakly inverse (purple colored in heat map) as indicated by the negative value close to zero ([Figure 3a](#)). The OLS establishes the relationship between a dependent variable, i.e. the main peak in emission, and independent variables (feature set) with a coefficient for each independent variable. The OLS



**Figure 3.** Numerical feature-emission correlations. (a) Pearson correlation coefficient heat map. The 0–1 values give the strength, white-magenta-black color scheme and  $\pm$  signs display the direction of the correlations. (b) Calculated coefficients from the OLS regression model, where values represent the strength, and the  $\pm$  signs give the direction of correlations. The F1–F13 denote the numerical features: F1 (citric acid), F2 (urea), F3 (ethylenediamine), F4 (ammonium), F5 (NaOH), F6 (boric acid), F7 (sodium thiosulfate), F8 (KOH), F9 (formic acid), F10 (reaction temperature (K)), F11 (reaction time (min)), F12 (particle size (TEM)), and F13 (main peak (in water)).

coefficients are calculated by minimizing the total sum of squares of the difference between the calculated and observed values of the main emission peak. All calculated coefficients for each numerical feature are shown in Figure 3b. Both Pearson and OLS coefficients indicate that reaction temperature, time, and particle size show low or no correlation with the emission, while varied precursors show some small correlation to the emission (Figure 3, Tables S1 and S2). The ANOVA test performed on the categorical features demonstrates that variations in the categorical features of reaction method, solvent, and purification method likely influence the emission as the  $p$  value was less than 0.05 in these cases (Table S2). Note that  $F$  values are inversely related to  $p$  values, so a higher  $F$  value indicates a more significant  $p$  value. For pH,  $p > 0.05$ , so pH is not likely an essential feature in predicting CD emission. Pearson correlation coefficients and OLS regression were also used to explore the correlations between all features and emission (Table S3 and Figures S3 and S4). As expected, the feature “color” has strong correlation with the emission wavelength as was found in our initial feature-label correlation evaluations. However, Pearson correlation coefficients and OLS regression may not provide correct correlations between the categorical features and emission due to the one-hot-encoding of categorical features prior to evaluations.

At first, the reaction temperature and time were excluded from the initial input feature set in the ML models because of the poor correlations found from the initial evaluation in the statistical analysis. Later, reaction temperature and time were included to assess the importance of including those features, since those two parameters are frequently varied in experiments to optimize CD emission.<sup>48–50</sup>

Without the inclusion of reaction temperature and time as features, the M1 machine obtained a training MAE of 16.3 nm with a maximum standard deviation of 17.1 nm. Prior to the validation of M1 with test sets of data examples, we used the training data set as a test set. As expected, the training set exhibited an MAE on the validation set—viz.  $17.0 \pm 20.9$  nm—close to the training error (Table 2). The errors are

**Table 2. Training and Test MAEs from M1, M2, and M3 without Including Reaction Temperature and Time as Features<sup>a</sup>**

	MAE (nm) for machines		
	M1	M2	M3
train	$16.3 \pm 17.1$	$9.8 \pm 11.1$	$9.6 \pm 10.5$
train set as test	$17.0 \pm 20.9$	$10.6 \pm 40.1$	$9.8 \pm 26.5$
test 1	$23.9 \pm 12.4$	$19.0 \pm 10.4$	$19.4 \pm 10.7$
test 2	$39.4 \pm 18.3$	$35.1 \pm 14.3$	$36.2 \pm 11.8$
test 3	$38.4 \pm 17.2$	$28.5 \pm 12.8$	$25.8 \pm 15.0$

<sup>a</sup>The test errors are maximum standard train/test error out of all data points in each train/test iteration. The average MAE of tests 1–3 is also shown.

comparable as M1 has been trained with the same data. Three different sets were assembled or constructed from experiments to test the machines: test 1 contains 28 examples from the literature while test 2 and test 3 report new experimental results obtained in this work. Most reactions in test 3 were controlled to have the same reaction temperature of 150 °C and reaction time of 1 h (except for when the microwave method was used), while reactions in test 2 all had parameters randomly assigned (details in Supporting Information).

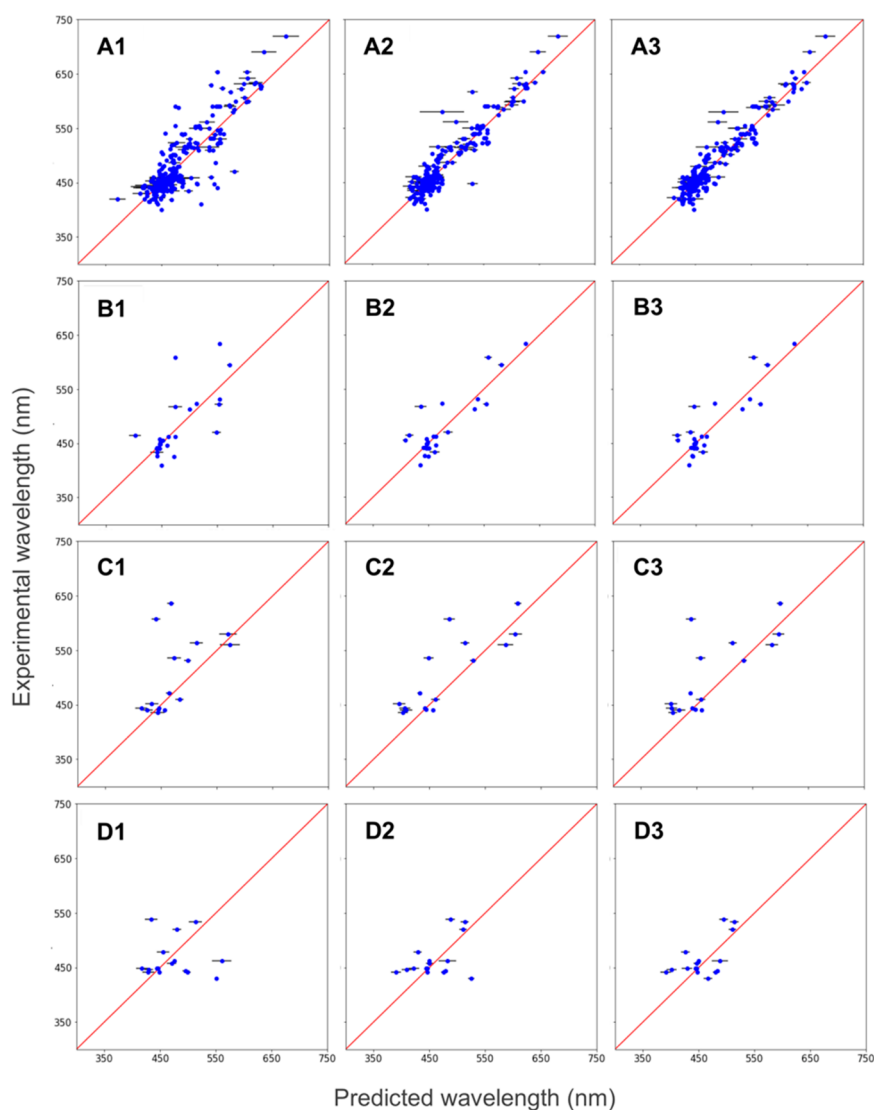
Evaluation of the M1 machine gave MAEs of  $27.5 \pm 12.4$  nm,  $39.4 \pm 18.3$  nm, and  $38.4 \pm 17.2$  nm for tests 1, 2, and 3, respectively (Table 2). The recorded errors for test 2 and test 3 were higher than those for test 1 mainly because the data examples in test 1 were statistically more similar to the training data, whereas the features in tests 2 and 3 were more dissimilar from those in the training set. Moreover, in tests 2 and 3, the distribution of examples includes CDs with the reaction method feature assigned to microwave-assisted hydrothermal and solvothermal methods in a much larger abundance than in the training set or in test 1. This could also increase MAEs.

It is useful to consider the accuracy of individual predictions from M1 (Figure 4) as MAEs for a particular test set could hide nuance in providing only the average prediction error for all the CDs in that test set. For example, such plots can reveal whether a given machine has differential accuracy in predicting the frequency of CDs in different color domains. When the training set is used as the test set in M1, most of the CD predictions lie on or near the line  $X = Y$ , which shows that most of the predictions are very close to actual emissions regardless of the color. The M1 machine predicts most of the blue and green CDs well in all three test sets (tests 1, 2, and 3) even though tests 1, 2, and 3 display different prediction MAEs. Red CDs are often not accurately predicted by the M1 machine for the test sets.

As a precursor to evaluating machines M2 and M3, two ANN ML models for predicting only color through classification was trained with the same database with and without the inclusion of temperature and time as features among the STDFeatures. The accuracy of a classification model can be assigned using several different metrics. Here we use a common choice by defining it as the fraction of predictions the model predicts correctly out of the total number of predictions made.

The ANN-based classification model in M2 was tested against the XGBoost,<sup>20</sup> KNN,<sup>32,33</sup> and SVM<sup>34</sup> models. The M2 classification model and XGBoost model were the most appropriate algorithms (Figure S1a) for the classification with similar prediction accuracies (0.94 and 0.95 respectively). The ANN  $k$ -ensemble ( $k = 4$ ) based regression model in M1 was also evaluated against XGBoost, KNN, and SVM regression models (Figure S1b). The data set without temperature and time features was used for the comparisons. It is not surprising that the ANN  $k$ -ensemble ( $k = 4$ ) based regression model and the XGBoost are the most similar in their predictions because they both benefit from being ensemble methods. Nevertheless, the ANN  $k$ -ensemble ( $k = 4$ ) based regression model was the most accurate with the maximum  $R^2$  value, and it is consequently employed in the M1, M2 and M3 models reported here.

The accuracy obtained during the training of the classification model is 0.94 and 0.93 when temperature and time are included or not, respectively, as features. Confusion matrices are used to summarize the performance of the classification model. The confusion matrices plotted for tests 1, 2, and 3 (Figure 5) illustrate how the predicted colors compare to the actual CD colors. The plots also show how the predictions change with the inclusion of temperature and time. Note that two confusion matrices were plotted together for tests 1 and 2 and for tests 1 and 3, noted as 1U2 and 1U3 respectively, to test how 1U2 predict colors compared to 1U3 as test 2 contains experiments with more time and temperature variations than in test 3. The classification model predicts most

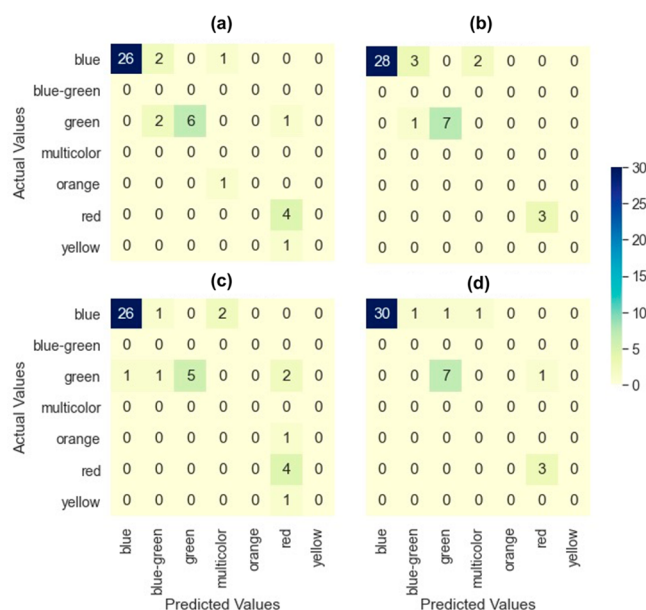


**Figure 4.** Plots of predicted versus experimental emission wavelength (nm) of CDs for test 1, test 2, test 3 obtained from M1, M2, and M3 without including reaction temperature and time as features. Each data entry represents a CD with an emission wavelength (nm) from the model prediction and an experimental measured value. The error bars show the calculated standard deviations. X = Y line (red line) M1 is the model where no color feature was used; M2 is the model where the actual color is used in the training set and the predicted color was used in test set; M3 is the model where predicted color is used in both training and test set. Test 1 is the 28 examples from literature; tests 2 and 3 are from newly synthesized CDs where test 2 has more reaction temperature and time variations. Panels in rows A, B, C, and D denote the training data set, test 1, test 2, and test 3, respectively. Panels in columns 1, 2, and 3 correspond to the M1, M2, and M3 models, respectively.

of the blue, green, and red CDs accurately (Figure 5 and Table S5) in both tests 1U2 and tests 1U3. Inclusion of temperature and time as features shifted predictions for green CDs more toward blue or red in test 1U2, and blue CD predictions improved in test 1U3. Overall, the classification model predicted test 1U3 CDs better than those in test 1U2. The classification model predictions of color are more accurate than the colors inferred from the frequency predictions of the M1 machine generated using the same STDFeatures, and primarily so for CDs with red emission. Unfortunately, the classification model can only predict a categorical color corresponding to a relatively broad range of visible emission wavelengths, while M1 predicts a specific wavelength. Therefore, we explore a pair of hybrid machines, M2 and M3, in the next section which used the classification model predictions as a feature in the staged regression model. These machines leverage the advantage of the accuracy seen in the predictions from the

classification model and have the potential to predict the emission wavelengths.

The hybrid M2 machine trains both the ANN model for color classification and the ANN *k*-ensemble model simultaneously. The hybrid M3 machine trains the classification model first using the actual color as a label, and then feeds the predicted color along with the STDFeatures in training the ANN *k*-ensemble. Thus, the main difference between M2 and M3 lies in the training of the ANN *k*-ensemble model: M2 uses the actual colors in the example while M3 utilizes the predicted colors from the classification model as the input feature. In the validation process for both M2 and M3, the predicted color from the classifier is fed into the ANN *k*-ensemble model to predict the emission wavelength. Inclusion of the color feature through these hybrid models has significantly reduced the training MAEs of M2 and M3 in comparison to M1, which were  $9.8 \pm 11.1$  nm and  $9.6 \pm 10.5$  nm, respectively (Table 2).



**Figure 5.** Confusion matrix of the classification model results for test 1, test 2, and test 3 with and without the temperature and time, comparing actual and predicted CD color. (a) Test 1 and 2 without the reaction temperature and time. (b) Test 1 and 3 without the reaction temperature and time. (c) Test 1 and 2 with reaction temperature and time. (d) Test 1 and 3 with reaction temperature and time.

M2 shows clear progress in predictions of the test sets 1, 2, and 3, including when using the training set as a test set in contrast to the M1 (Table 2). Similar to M1, M2 and M3 predict test 1 better than tests 2 and 3. Test 2 and 3 prediction errors decrease by ~3–13 nm with M2/M3 models compared to the M1 model (Table 2). Using the predicted color (M3) instead of the actual color (M2) in the training has further improved only the predictions of test 3. The significant decrease in MAEs suggests that the hybrid machines are effective by combining both classification and regression for color and wavelength prediction. However, the current M2 and M3 test errors do not vary significantly from each other, and further assessment is needed to decide if M3 is better than M2 as the current model used for color classification has room for improvement. It is also worth considering the individual predictions for tests 1, 2, and 3 using M2 and M3 in assessing the (improved) predictability of the hybrid models (Figure 4). Clearly, red CD predictions have improved in tests 1 and 2 in both M2 and M3. Also, green CDs in test 3 are predicted more accurately. The blue CD individual predictions in tests 1, 2, and 3 have deviated slightly from the  $X = Y$  line in hybrid models compared to M1 predictions. To better picture how different machines (classification and M1–M3) predict the color of each CD compared to their respective experimental colors, we color-coded the emission wavelengths based on wavelength regions in the visible spectrum (Table S6). The detailed color-coded predictions from color classification model and wavelength prediction of test 1–3 obtained from M1–M3 without considering reaction temperature and time as features are shown in Tables S7–S9.

It is also instructive to consider how the CD emission predictions in M1, M2, and M3 change with the inclusion of temperature and time as features. As summarized in the Introduction, the current dogma suggests that reaction time

and temperature can tune the optical properties of CDs. However, the analysis of the metadata reported above suggests that temperature and time do not correlate strongly with emission. There is also some literature precedent that these two parameters are relatively less important compared to other features—such as the precursors, their mass, their volume, the choice of solvents, etc.—in predicting the quantum yield and color of the CDs using a machine learning model.<sup>19,21</sup> In the Supporting Information, we report the differences in the accuracies of M1, M2, and M3 machines when they are constructed with or without temperature and time as features in predicting the emission wavelength of CDs. Inclusion of these two features increased the training and test MAE of M1 only slightly (Table S10 and Figure S5). In contrast, M2 and M3 have lower MAEs in both training and test sets. Specifically, the predictability for the CDs in tests 2 and 3 improved using the M2 and M3 machines trained with temperature and time. Although the effect is more significant in the model that predicts the color through classification than in the ANN  $k$ -ensemble model, inclusion of temperature and time as features in the hybrid machines has a combined positive effect. Test 2 has 16 experiments, which have more temperature and time variations than those in test 3. However, test 2 only shows a slight decrease in MAEs whereas test 3 shows a significant decrease in errors when temperature and time are considered. The reason could be that temperature and time only finetune the individual carbon dot emission while reaction methods, solvent, and purification methods play a major role in determining the emission colors. Further, the calculated MAEs for tests 1, 2, and 3 indicate that temperature and time could be critical for certain colored CDs.

The addition of temperature and time leads to higher error in blue and green CD predictions using the M1 machine, whereas there is no change or improvement for red CD predictions (Figures 4 and S5). The color-coded predictions of tests 1–3 from the color classifier and machines M1–3 with the inclusion of temperature and time are shown in Tables S11–S13. Further, it is shown how the individual predictions of tests 1–3 from M1–M3 machines change upon inclusion of temperature and time and where the predictions lie with respect to the experimental wavelengths (Figure S6). Use of temperature and time as features in the hybrid M2 and M3 machines generally leads to better predictions of blue CDs, except for M2 in test 2. Green CD predictions are better in the hybrid machines except in test 1, and red CD predictions improved in the hybrid machines. In the classification model within the hybrid machines, only blue CD predictions improved in tests 1 and 3 with the inclusion of temperature and time as features. Perhaps surprisingly, the model for color classification predicts the blue, green, and red CDs reasonably accurately without inclusion of temperature and time as features.

The root-mean-square error (RMSE) was calculated for the training, train set as test and test 1–3 from M1–M3 machines (Figure S7) as a comparison to the MAEs shown in Tables 2 and S10 with and without including reaction temperature and time. RMSE demonstrates how the prediction errors spread around the line of best fit. Lower RMSE values indicate better predictions. It is further established that M2 and M3 performance has improved compared to M1 in all tests conducted. Inclusion of the temperature and time has improved the predictions from all three machines (M1–M3) which is reflected in calculated RMSEs (Figure S7a,c). Overall

RMSE revealed trends similar to MAEs except for the predictions of the M1 machine when the temperature and time are included in the feature set.

## CONCLUSION

A database was created from a meta-analysis of 407 literature examples of CDs synthesized with citric acid and urea or ethylenediamine as precursors. Initial data analysis showed that a few key features, such as reaction method, solvent, purification method, and precursors, correlate more with the emission wavelength of CDs than reaction temperature and time. Three ML machines were built to capture the nonlinear correlations in the database and make predictions. Two separate sets of experiments, with 16 reactions each, were carried out to generate new CDs, and their emission wavelengths were recorded as examples within these additional test sets. The M1 machine is based on a regression model, and it predicts emission characteristics of blue and green CDs reasonably well. The M2 and M3 (hybrid) machines employ two ANNs in series: a classification model for color prediction and a regression model for wavelength prediction using the predicted color from the first model as a feature. M2 and M3 achieved significantly higher accuracy than M1 as they have lower MAEs. M2 and M3 are better than M1 at predicting the emission of red CDs, with a slight trade-off in accuracy for the predictions of blue CDs. Additionally, the impact of temperature and time in tuning CD emission were explored. It seems that adding these two parameters can improve the overall prediction because it improves the predictions for blue CDs in the classification model and improves the prediction of all CDs from the hybrid machines. The implementation of an ensemble method for the classification model could further stabilize the overall predictability in the hybrid machines. Overall, our results show that the emission of CDs can be adjusted more effectively by changing reaction method, purification method, and solvent than other experimental factors. Future studies comparing a series of these features in parallel are needed to better determine their impact on CD properties, as most present studies have only varied reaction temperature and time. The tools that have been developed in this work, such as the M3 hybrid machine with an average MAE of 27 nm, should be useful in the prediction of emission of novel CDs. In this way, a few promising reaction examples can be selected from the model for the synthesis of CDs of specific colors, thereby saving significant effort in the synthesis optimization process.

## ASSOCIATED CONTENT

### Data Availability Statement

The data sets used for training and validation are available in the GitHub repository, <https://github.com/rxhernandez/MLCD>. In addition, that repository contains all the scripts used to train, test, and run the three machines, M1, M2, and M3.

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.2c01007>.

Database construction methods, experimental methods, and hyperparameter optimization methods and several supporting tables and figures related to optimization of the methods (PDF)

## AUTHOR INFORMATION

### Corresponding Authors

Christy L. Haynes – Department of Chemistry, University of Minnesota–Twin Cities, Minneapolis, Minnesota 55455, United States; [orcid.org/0000-0002-5420-5867](https://orcid.org/0000-0002-5420-5867); Email: [chaynes@umn.edu](mailto:chaynes@umn.edu)

Rigoberto Hernandez – Department of Chemistry and Departments of Chemical and Biomolecular Engineering and Materials Science and Engineering, Johns Hopkins University, Baltimore, Maryland 21218, United States; [orcid.org/0000-0001-8526-7414](https://orcid.org/0000-0001-8526-7414); Email: [hernandez@jhu.edu](mailto:hernandez@jhu.edu)

### Authors

Ravithree D. Senanayake – Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21218, United States; [orcid.org/0000-0002-4727-0521](https://orcid.org/0000-0002-4727-0521)

Xiaoxiao Yao – Department of Chemistry, University of Minnesota–Twin Cities, Minneapolis, Minnesota 55455, United States; [orcid.org/0000-0001-7999-2898](https://orcid.org/0000-0001-7999-2898)

Clarice E. Froehlich – Department of Chemistry, University of Minnesota–Twin Cities, Minneapolis, Minnesota 55455, United States; [orcid.org/0000-0001-8862-785X](https://orcid.org/0000-0001-8862-785X)

Meghan S. Cahill – Department of Chemistry, University of Minnesota–Twin Cities, Minneapolis, Minnesota 55455, United States; [orcid.org/0000-0002-1514-7625](https://orcid.org/0000-0002-1514-7625)

Trevor R. Sheldon – Department of Chemistry, University of Minnesota–Twin Cities, Minneapolis, Minnesota 55455, United States; [orcid.org/0000-0002-4851-6204](https://orcid.org/0000-0002-4851-6204)

Mary McIntire – Department of Chemistry, University of Minnesota–Twin Cities, Minneapolis, Minnesota 55455, United States; [orcid.org/0000-0003-1124-5970](https://orcid.org/0000-0003-1124-5970)

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.jcim.2c01007>

### Author Contributions

<sup>1</sup>R.D.S. and X.Y. contributed equally to this work.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grant No. CHE-2001611, the NSF Center for Sustainable Nanotechnology (CSN). The CSN is part of the Centers for Chemical Innovation Program. The computing resources necessary for this work were provided in part by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (NSF) Grant Number ACI-1548562 through allocation CTS090079, and the Advanced Research Computing at Hopkins (ARCH) high-performance computing (HPC) facilities supported by NSF Grant Number OAC-1920103. We thank Dr. Baoyue Fan from Andreas Stein's group for assistance in the synthesis of CDs by hydrothermal/solvothermal treatment and pyrolysis. Figure 1d was created with Lucid Visual Collaboration Suite. The table of contents image was created with Biorender.com.

## REFERENCES

- (1) Zhi, B.; Yao, X.; Cui, Y.; Orr, G.; Haynes, C. L. Synthesis, Applications and Potential Photoluminescence Mechanism of Spectrally Tunable Carbon Dots. *Nanoscale* **2019**, *11*, 20411–20428.

- (2) Pandit, S.; Banerjee, T.; Srivastava, I.; Nie, S.; Pan, D. Machine Learning-Assisted Array-Based Biomolecular Sensing Using Surface-Functionalized Carbon Dots. *ACS Sens.* **2019**, *4*, 2730–2737.
- (3) Dhenadhayalan, N.; Lin, K.-C.; Saleh, T. A. Recent Advances in Functionalized Carbon Dots toward the Design of Efficient Materials for Sensing and Catalysis Applications. *Small* **2020**, *16*, 1905767.
- (4) Meng, W.; Bai, X.; Wang, B.; Liu, Z.; Lu, S.; Yang, B. Biomass-Derived Carbon Dots and Their Applications. *Energy Environ. Mater.* **2019**, *2*, 172–192.
- (5) Tao, H.; Yang, K.; Ma, Z.; Wan, J.; Zhang, Y.; Kang, Z.; Liu, Z. In Vivo NIR Fluorescence Imaging, Biodistribution, and Toxicology of Photoluminescent Carbon Dots Produced from Carbon Nanotubes and Graphite. *Small* **2012**, *8*, 281–290.
- (6) Song, Y.; Zhu, S.; Zhang, S.; Fu, Y.; Wang, L.; Zhao, X.; Yang, B. Investigation From Chemical Structure to Photoluminescent Mechanism: A Type of Carbon Dots From the Pyrolysis of Citric Acid and an Amine. *J. Mater. Chem. C* **2015**, *3*, 5976–5984.
- (7) Li, L.; Li, Y.; Ye, Y.; Guo, R.; Wang, A.; Zou, G.; Hou, H.; Ji, X. Kilogram-Scale Synthesis and Functionalization of Carbon Dots for Superior Electrochemical Potassium Storage. *ACS Nano* **2021**, *15*, 6872–6885.
- (8) Hong, Q.; Wang, X.-Y.; Gao, Y.-T.; Lv, J.; Chen, B.-B.; Li, D.-W.; Qian, R.-C. Customized Carbon Dots with Predictable Optical Properties Synthesized at Room Temperature Guided by Machine Learning. *Chem. Mater.* **2022**, *34*, 998–1009.
- (9) Ding, H.; Zhou, X.-X.; Wei, J.-S.; Li, X.-B.; Qin, B.-T.; Chen, X.-B.; Xiong, H.-M. Carbon Dots with Red/Near-Infrared Emissions and Their Intrinsic Merits for Biomedical Applications. *Carbon* **2020**, *167*, 322–344.
- (10) Yuan, F.; Yuan, T.; Sui, L.; Wang, Z.; Xi, Z.; Li, Y.; Li, X.; Fan, L.; Tan, Z. a.; Chen, A.; Jin, M.; Yang, S. Engineering Triangular Carbon Quantum Dots with Unprecedented Narrow Bandwidth Emission for Multicolored LEDs. *Nat. Commun.* **2018**, *9*, 2249.
- (11) Zhu, Z.; Zhai, Y.; Li, Z.; Zhu, P.; Mao, S.; Zhu, C.; Du, D.; Belfiore, L. A.; Tang, J.; Lin, Y. Red Carbon Dots: Optical Property Regulations and Applications. *Mater. Today Commun.* **2019**, *30*, 52–79.
- (12) Sun, S.; Chen, J.; Jiang, K.; Tang, Z.; Wang, Y.; Li, Z.; Liu, C.; Wu, A.; Lin, H. Ce6-Modified Carbon Dots for Multimodal-Imaging-Guided and Single-NIR-Laser-Triggered Photothermal/Photodynamic Synergistic Cancer Therapy by Reduced Irradiation Power. *ACS Appl. Mater. Interfaces* **2019**, *11*, 5791–5803.
- (13) Tian, Z.; Zhang, X.; Li, D.; Zhou, D.; Jing, P.; Shen, D.; Qu, S.; Zboril, R.; Rogach, A. L. Full-Color Inorganic Carbon Dot Phosphors for White-Light-Emitting Diodes. *Adv. Opt. Mater.* **2017**, *5*, 1700416.
- (14) Jiang, K.; Sun, S.; Zhang, L.; Lu, Y.; Wu, A.; Cai, C.; Lin, H. Red, Green, and Blue Luminescence by Carbon Dots: Full-Color Emission Tuning and Multicolor Cellular Imaging. *Angew. Chem., Int. Ed.* **2015**, *54*, 5360–5363.
- (15) Zhang, M.; Hu, L.; Wang, H.; Song, Y.; Liu, Y.; Li, H.; Shao, M.; Huang, H.; Kang, Z. One-Step Hydrothermal Synthesis of Chiral Carbon Dots and Their Effects on Mung Bean Plant Growth. *Nanoscale* **2018**, *10*, 12734.
- (16) Liu, M. L.; Yang, L.; Li, R. S.; Chen, B. B.; Liu, H.; Huang, C. Z. Large-Scale Simultaneous Synthesis of Highly Photoluminescent Green Amorphous Carbon Nanodots and Yellow Crystalline Graphene Quantum Dots at Room Temperature. *Green Chem.* **2017**, *19*, 3611–3617.
- (17) Holá, K.; Sudolská, M.; Kalytchuk, S.; Nachtigallová, D.; Rogach, A. L.; Otyepka, M.; Zboril, R. Graphitic Nitrogen Triggers Red Fluorescence in Carbon Dots. *ACS Nano* **2017**, *11*, 12402–12410.
- (18) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555.
- (19) Han, Y.; Tang, B.; Wang, L.; Bao, H.; Lu, Y.; Guan, C.; Zhang, L.; Le, M.; Liu, Z.; Wu, M. Machine-Learning-Driven Synthesis of Carbon Dots with Enhanced Quantum Yields. *ACS Nano* **2020**, *14*, 14761–14768.
- (20) Chen, T.; Guestrin, C. In Xgboost: A Scalable Tree Boosting System In *Proceedings of The 22nd ACM sigkdd International Conference on Knowledge discovery and Data Mining*, 2016; pp 785–794.
- (21) Wang, X.-Y.; Chen, B.-B.; Zhang, J.; Zhou, Z.-R.; Lv, J.; Geng, X.-P.; Qian, R.-C. Exploiting Deep Learning for Predictable Carbon Dot Design. *Chem. Commun.* **2021**, *57*, 532–535.
- (22) Chollet, F. *Deep Learning with Python*, 1st ed.; Manning Publications Co.: Greenwich, CT, 2017.
- (23) Alaya, M. Z.; Bussy, S.; Gaïffas, S.; Guilloux, A. Binarisity: A Penalization for One-Hot Encoded Features in Linear Supervised Learning. *J. Mach. Learn. Res.* **2019**, *20*, 1–34.
- (24) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140.
- (25) Bauer, E.; Kohavi, R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Mach. Learn.* **1999**, *36*, 105–139.
- (26) Cunningham, P.; Carney, J.; Jacob, S. Stability Problems with Artificial Neural Networks and the Ensemble Solution. *Artif. Intell. Med.* **2000**, *20*, 217–225.
- (27) Dietterich, T. G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Mach. Learn.* **2000**, *40*, 139–157.
- (28) Barrow, D. K.; Crone, S. F. In Cropping (Cross-Validation Aggregation) for Forecasting—A Novel Algorithm of Neural Network Ensembles on Time Series Subsamples, *The 2013 International Joint Conference on Neural Networks, Dallas, TX, Aug 4–9, 2013*; Dallas, TX.
- (29) Khwaja, A. S.; Naeem, M.; Anpalagan, A.; Venetsanopoulos, A.; Venkatesh, B. Improved Short-Term Load Forecasting Using Bagged Neural Networks. *Electr. Power Syst. Res.* **2015**, *125*, 109–115.
- (30) Daly, C. A., Jr; Hernandez, R. Optimizing Bags of Artificial Neural Networks for the Prediction of Viability from Sparse Data. *J. Chem. Phys.* **2020**, *153*, 054112.
- (31) Nikolaidis, A.; Solon Heinsfeld, A. S.; Xu, T.; Bellec, P.; Vogelstein, J.; Milham, M. Bagging Improves Reproducibility of Functional Parcellation of the Human Brain. *Neuroimage* **2020**, *214*, 116678.
- (32) Fix, E.; Hodges, J. L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *Int. Stat. Rev.* **1989**, *57*, 238–247.
- (33) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175–185.
- (34) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
- (35) Daly, C. A., Jr; Hernandez, R. Learning From The Machine: Uncovering Sustainable Nanoparticle Design Rules. *J. Phys. Chem. C* **2020**, *124*, 13409–13420.
- (36) Géron, A. *Hands-On Machine Learning With Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media, Inc: 2017.
- (37) Chollet, F. *Keras, GitHub*; <https://github.com/keras-team/keras> 2015.
- (38) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint* 2016; arXiv:1603.04467.
- (39) Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **2007**, *9*, 10–20.
- (40) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **2007**, *9*, 90–95.
- (41) McKinney, W. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Austin, TX, 2010; pp 51–56.
- (42) Van Der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.
- (43) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.

Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(44) Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B. E.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J. B.; Grout, J.; Corlay, S. *Jupyter Notebooks—Publishing Format for Reproducible Computational Workflows*; 2016; Vol. 2016.

(45) Harper, M.; Weinstein, B.; tgwoodcock; Simon, C.; chebee7i; Morgan, W.; Knight, V.; Swanson-Hysell, N.; Evans, M.; Zgainsforth; Badger, T. G.; SaxonAnglo; Greco, M.; Zuidhof, G. python-ternary: Ternary Plots in Python, Version 1.0.6. *Zenodo* 2019.

(46) Xu, Y.; Wu, M.; Liu, Y.; Feng, X.-Z.; Yin, X.-B.; He, X.-W.; Zhang, Y.-K. Nitrogen-Doped Carbon Dots: A Facile and General Preparation Method, Photoluminescence Investigation, and Imaging Applications. *Eur. J. Chem.* **2013**, *19*, 2276–2283.

(47) Park, Y.; Kim, Y.; Chang, H.; Won, S.; Kim, H.; Kwon, W. Biocompatible Nitrogen-Doped Carbon Dots: Synthesis, Characterization, and Application. *J. Mater. Chem. B* **2020**, *8*, 8935–8951.

(48) Zhi, B.; Yao, X.; Wu, M.; Mensch, A.; Cui, Y.; Deng, J.; Duchimaza-Heredia, J. J.; Trerayapiwat, K. J.; Niehaus, T.; Nishimoto, Y.; Frank, B. P.; Zhang, Y.; Lewis, R. E.; Kappel, E. A.; Hamers, R. J.; Fairbrother, H. D.; Orr, G.; Murphy, C. J.; Cui, Q.; Haynes, C. L. Multicolor Polymeric Carbon Dots: Synthesis, Separation and Polyamide-Supported Molecular Fluorescence. *Chem. Sci.* **2021**, *12*, 2441–2455.

(49) Yin, B.; Deng, J.; Peng, X.; Long, Q.; Zhao, J.; Lu, Q.; Chen, Q.; Li, H.; Tang, H.; Zhang, Y.; Yao, S. Green Synthesis of Carbon Dots with Down- and Up-Conversion Fluorescent Properties for Sensitive Detection of Hypochlorite With a Dual-Readout Assay. *Analyst* **2013**, *138*, 6551–6557.

(50) Barati, A.; Shamsipur, M.; Arkan, E.; Hosseinzadeh, L.; Abdollahi, H. Synthesis of Biocompatible and Highly Photoluminescent Nitrogen Doped Carbon Dots From Time: Analytical Applications and Optimization Using Response Surface Methodology. *Mater. Sci. Eng., C* **2015**, *47*, 325–332.