CHEMISTRY OF MATERIALS

pubs.acs.org/cm

Article

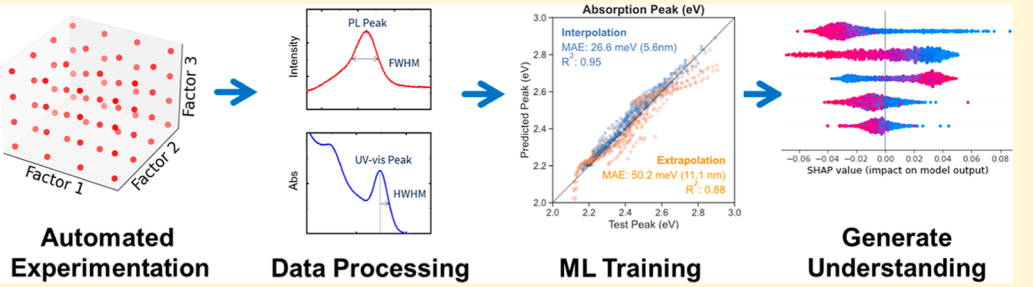
Understanding Hot Injection Quantum Dot Synthesis Outcomes Using Automated High-Throughput Experiment Platforms and Machine Learning

Rui Hua Jeff Xu, Logan P. Keating, Ajit Vikram, Moonsub Shim, and Paul J. A. Kenis*

Cite This: https://doi.org/10.1021/acs.chemmater.3c02751

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Machine learning (ML) has demonstrated potential toward accelerating synthesis planning for various material systems. However, ML has remained out of reach for many materials scientists due to the lack of systematic approaches or heuristics for developing ML workflows for material synthesis. In this work, we report an approach for selecting ML algorithms to train models for predicting nanomaterial synthesis outcomes. Specifically, we developed and used an automated batch microreactor platform to collect a large experimental data set for hot-injection synthesis outcomes of CdSe quantum dots. Thereafter, this data set was used to train models for predicting synthesis outcomes using various ML algorithms. The relative performances of these algorithms were compared for experimental data sets of different sizes and with different amounts of noise added. Neural-network-based models show the most accurate predictions for absorption and emission peak, while a cascade approach for predicting full width at half-maximum was shown to be superior to the direct approach. The SHapley Additive exPlanations (SHAP) approach was used to determine the relative importance of different synthesis parameters. Our analyses indicate that SHAP importance scores are highly dependent on feature selection and highlight the importance of developing inherently interpretable models for gaining insights from ML workflows for material synthesis.

1. INTRODUCTION

Semiconductor quantum dots (QDs) offer the potential to design nanomaterials with widely tunable optoelectronic properties by control over size, shape, phase, and composition.[1,2] They demonstrate promise in a wide range of applications such as catalysis, coatings, imaging, sensing, displays, photovoltaics, and more.[1,3−17] In many of these applications, tight control over QD properties (shape, size distribution, composition, and heterostructures) is critical as these parameters determine key optical and electronic properties including absorption and emission. Several studies have sought to understand QD synthesis by uncovering mechanisms of their formation (reaction, nucleation, and growth steps) and by demonstrating methods to tune synthesis parameters to achieve specific QD properties.[18−32]

While these studies have led to significant advances in understanding the nucleation and growth of colloidal QDs, gaps in understanding QD formation still remain due to a number of challenges: (1) huge synthesis parameter space involved in QD synthesis coupled with slow throughput of manual experiments, complicating systematic experimental exploration of the parameter space to uncover general trends, and (2) inconsistent reproducibility in traditional batch synthesis methods leading to high noise in literature data.[33−36] In addition, publication bias (researchers typically only reporting methods that work) further limits broad availability of critical information across synthesis parameter space (which would include successful and unsuccessful experimental data), which in turn hampers researchers in uncovering the complex relationships between synthesis parameters on nanoparticle properties.[37,38]

A number of studies have demonstrated that machine learning (ML) can be a powerful tool to accurately model complex

Received: October 25, 2023
Revised: January 18, 2024
Accepted: January 19, 2024

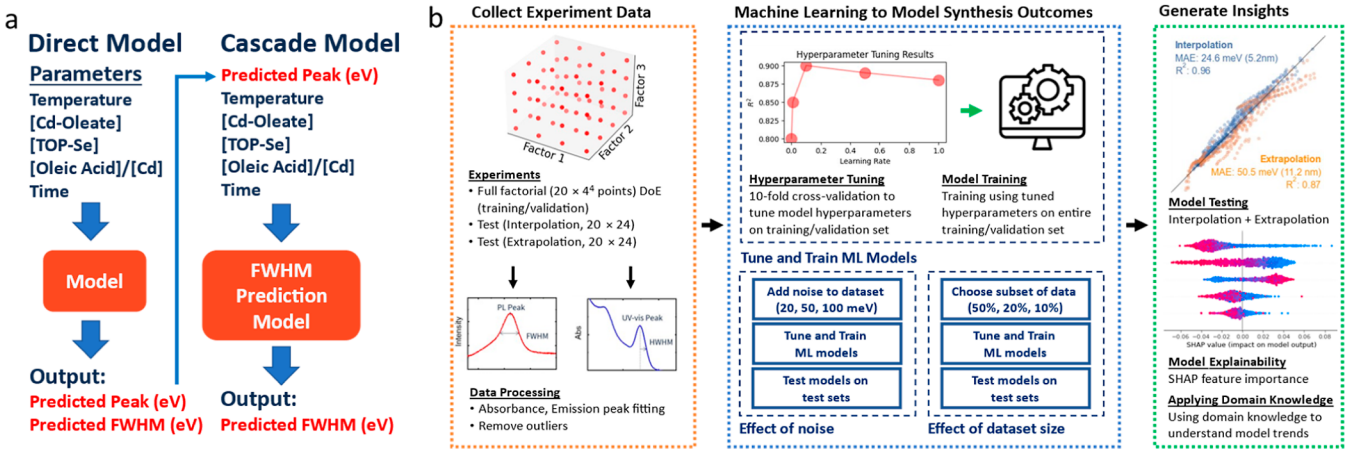


Figure 1. (a) Direct vs cascade single-output models used for synthesis outcome prediction. (b) Experiment and ML workflow for assessing performance of ML algorithms for QD synthesis prediction.

parameter spaces of different classes of materials.[39] ML tools have been used on well-curated computational data sets and text-mined data sets from the literature and have demonstrated potential in material discovery.[40−46] They have also been used on large data sets generated from automated synthesis and characterization platforms to accurately predict material synthesis outcomes.[47−51] Despite these successes, ML application toward synthesis prediction still faces the challenges of accessibility and interpretability. Useful heuristics for choosing ML models and features are not available, hampering materials scientists who are looking to incorporate ML into their workflows. Most ML workflows applied to materials science use *uninterpretable* "black-box" models to map the output to input variables. Deducing actionable insights from such models is difficult at best. These black-box models thus have limited scientific utility.[52] A more systematic approach for the selection of ML algorithms and features is needed to enable nonexperts to build ML workflows into the materials research pipeline. At the same time, model explainability tools should be employed systematically to obtain physical insights from the ML models.

In the area of QD synthesis planning, ML models trained on literature data have been used to predict QD synthesis outcomes based on synthesis parameters for II−VI materials and for InP.[45,46] Prior work has also demonstrated approaches to use automated flow synthesis platforms for generating large combinatorial QD synthesis data sets able to train models for predicting QD synthesis outcomes.[53] Unfortunately, literature data are generally inadequate for benchmarking different ML algorithms due to the combination of noise in QD synthesis, inadequate exploration of synthesis parameter spaces, and bias in literature reporting (lack of reporting of negative synthesis outcomes). To develop generalizable heuristics to apply ML algorithms for mapping synthesis parameter spaces of QD materials, high-quality, reproducible experimental data sets are needed to train and test different ML algorithms.

In this work, we report an automated batch reactor platform to investigate the synthesis parameter space of the CdSe hot-injection synthesis. Synthesis data generated from this automated platform will be used to train different ML algorithms for predicting QD properties to evaluate common ML regression models. We demonstrate a workflow to test the accuracy of ML models as well as their efficiency and robustness to noise. Model explainability methods will then be applied to the trained models to determine the effects of different synthesis parameters on CdSe hot injection outcomes, and these findings will be discussed in the context of prior work. Furthermore, we highlight best practices for applying ML to experimental nanomaterial synthesis data sets and to generating experimental data sets for ML modeling. We also suggest future opportunities for applying ML tools to nanomaterial systems.

2. METHODS

2.1. Automated Experiment Platform. To build the automated batch reactor platform, microfluidic flow switches (Fluigent M-SWITCH) and solenoid valves (Clippard NR1-2-24-P) were used for flow control, while temperature control for the reactor and precursors was achieved using Omega CS8DPT PID temperature controllers. Stainless steel (50 mL, KD Scientific) and glass (5 mL, Hamilton Gastight) syringes connected to syringe pumps (Harvard 33DDS, Chemyx Fusion 4000) were used for transferring precursors into the batch reactor, and reactor pressure was controlled using a Fluigent Flow-EZ pressure controller (PC). The reactor is stirred at a fixed stir rate using a glass stir rod (6 mm × 5 mm, Big Science Inc.) and a Fisherbrand Isotemp stirrer. Absorption spectroscopy was measured using a 2 mm path length UV−vis flow cell connected to an Ocean Optics Flame-S-UV-vis-ES spectrometer and an Ocean Optics DH-MINI UV−vis light source. Photoluminescence spectroscopy was performed by using another spectrometer with a 385 nm excitation source (Ocean Optic LLS). All components are automated on LabVIEW.

2.2. CdSe Hot-Injection Synthesis and Data Analysis. CdSe was synthesized using hot injection in octadecene (ODE) solvent and oleic acid (OA) ligand as described previously.[54] This method was adapted for our automated platform by preparing a 300 mM Cadmium-oleate solution and a 300 mM TOP-Se solution as Cd and Se precursors, respectively. CdO (99.5%), ODE (90%), selenium (99.95%), trioctylphosphine (TOP, 97%), and OA (90%) were all obtained from Sigma-Aldrich for precursor preparation. Detailed preparation steps are listed in Section S1. Hot injection reactions are initiated by injecting the TOP-Se precursor into the heated batch reactor loaded with ODE, OA, and Cd-oleate, and aliquots are sampled at specified time intervals for optical characterization.

Peaks for absorption are analyzed using a half-Gaussian fit on the lowest energy peak or shoulder to determine the peak position, while PL emission peaks are analyzed using a Gaussian fit on the highest energy emission peak. Broad peaks corresponding to surface trap states for PL are not recorded. Peak analysis codes are available on our GitHub repository (https://github.com/AutonomousWorkflows/CdSe_ML).

2.3. ML Benchmarking Workflow. *2.3.1. ML Models.* Five different ML models will be investigated in this work: ridge regression, decision tree, random forest (RF), artificial neural networks (NN), and

https://doi.org/10.1021/acs.chemmater.3c02751
Chem. Mater. XXXX, XXX, XXX−XXX

Monte Carlo neural networks (MCNN).[55−58] Tree-based models and NN models have been widely applied toward materials synthesis and this work seeks to systematically compare these models for their accuracy and robustness in QD synthesis prediction.[45−48,53] Linear and tree-based models are all implemented on scikit while NN-based models are implemented on keras.[59,60] A more thorough discussion of these models is available in Section S2. All data generated and codes used for training and testing models are collated in our GitHub depository (https://github.com/AutonomousWorkflows/CdSe_ML).

In the prediction of full width at half-maximum (fwhm), both direct and cascade (stacked) models will be tested (Figure 1a). Direct models predict single outputs directly from model inputs, while cascade models will predict fwhm using the predicted peak as an input. This approach has been shown in prior work to provide more accurate fwhm predictions than direct models as fwhm is closely correlated with peak.[47]

2.3.2. Model Prediction Performance. The overall ML workflow for this work is shown in Figure 1b. Data are split by experiment number for 10-fold cross-validation (group *K*-fold splits) to simulate the reality where several time points are generated for each batch experiment. Model hyperparameters are then tuned using 10-fold cross-validation using the bayesopt package before the models are then trained on the entire training set with tuned hyperparameters.[61] Details on data normalization and hyperparameter tuning are described in Sections S3 and S4, respectively. The trained models are then tested on the test set, where mean absolute errors (MAE) and coefficient of determination (R^2) values are then collected. The training and validation data sets are then modified by adding random noise or by removing training data points to investigate the effect of noise and data set size on model performance.

2.3.3. Model Explainability Using SHAP. After building and testing the models, feature importance scores from the model are obtained using SHapley Additive exPlanations (SHAP), a model-agnostic approach to explain model predictions locally by computing the contribution of each feature to a specific predicted data point.[62] Combining these local Shapley values allows us to generate global feature importance scores that demonstrate the global impact of individual features on the model's predictions and allows us to apply domain knowledge to generate new insights or propose new hypotheses. Trends gathered from SHAP analysis will be evaluated in the context of prior proposed mechanisms for QD formation: precursor reaction, nucleation, and growth.[63]

3. RESULTS AND DISCUSSION

3.1. Automated Synthesis Platform. We designed and built an automated batch microreactor to investigate the CdSe hot-injection synthesis. We selected a batch instead of a flow reactor due to its suitability for the commonly reported hot-injection QD synthesis method in the literature. Table 1 lists some of the advantages and disadvantages of batch and flow reactors. Flow reactors are easier to automate and have superior heat transfer properties but are also prone to clogging, wide residence time distribution, and inefficient experiment throughput at long residence times: they are particularly suitable for fast and relatively clean reactions at small time scales due to the rapid heating and cooling rates and the small reactor volumes. Conversely, batch reactors are generally more robust, less prone to clogging, and are able to handle a wider variety of precursors at the cost of being harder to automate and clean.

Table 1. Summary of the Advantages and Disadvantages of Batch and Flow Reactors with Regards to Automation

	batch		flow
characteristics	hot-injection	heat-up	microfluidic flow
mass transfer	slow	moderate	**fast**
heat transfer	scale-dependent	scale-dependent	**excellent**
scalability	poor	good	**excellent**
viscous solutions	**yes**	**yes**	no
solid precursors	**yes**	**yes**	no
precursor additions	**yes**	**yes**	yes (3-phase flow)
long reaction times	**efficient**	**efficient**	**inefficient**

We selected hot-injection synthesis of CdSe QDs as it is well established, allowing for comparison of trends observed in data from our work with the existing literature. We decided that a batch reactor is more suitable as the time scale of most CdSe synthesis reactions (minutes) makes it more time-efficient to sample multiple aliquots from a single batch reaction, and clogging due to solidification of the Cd precursor can be avoided.[45,64,65]

The automated batch reactor (Figure 2) comprises a temperature-controlled stirred stainless-steel microreactor (1−3 mL working volume) and was tested to be capable of sustained operation at 400 °C. Aliquots from the reactor can be withdrawn at specified time intervals by pressurizing the reactor with a PC and using a solenoid valve. A 1.2 mL reaction volume was chosen as it is small enough to enable rapid changes in reactant temperature while containing sufficient volume for the withdrawal of 20 × 40 μL aliquots at different time points. Precursors for the reaction are stored in Pyrex glass bottles and loaded into the batch reactor using syringe pumps and a microfluidic flow switch, while a heating tape controlled by a temperature controller was used to keep a stainless steel syringe of Cd-oleate above the solidification temperature of approximately 50 °C. Aliquots from the reactor are cooled using 1/16″ coiled water-cooled copper tubing and diluted before flowing into flow cells for spectroscopic characterization. A detailed stepwise procedure for each reaction is listed in Section S5, while an image of the assembled automated reactor platform is shown in Figure S1.

3.2. Validation of the Automated Synthesis Platform. To validate the reproducibility of the automated platform, we ran repeated hot injection syntheses using the same reaction conditions. Results are shown in Figure 3. Each data point shows the mean peak position (Figure 3a) or fwhm (Figure 3b) for each reaction condition over 5 runs, with the error bars representing the standard deviations of the five runs. Our automated batch reactor platform is highly reproducible, producing QDs with an average standard deviation of under 0.5 nm (2 meV) and coefficient of variation (CV) < 0.1%, even for data points at short reaction times. This demonstrates high reproducibility in the precursor injection step that matches prior work on automated hot injection of CdSe showing a 0.2% CV for the absorption peak. These values are over an order of magnitude lower than the CV range obtained for manual batch synthesis demonstrated in the same work (2.5% CV).[50] The reproducibility experiments (Figure 3) also reveal that the emission peak fwhm is generally lower than the absorption peak fwhm, an observation that will be further explained in Section 3.4.

3.3. Experimental Data Set Collection. The training/validation experimental synthesis data set is collected for ML using grid sampling of the synthesis parameter space, by varying Cd concentration ([Cd]), TOP-Se concentration ([Se]), [OA]/[Cd] ratio, and temperature to generate 256 hot injection experiments (Table 2). Each hot-injection experiment is run for 10 min and sampled at predetermined time intervals. These experimental parameters are chosen based on ranges used in similar CdSe hot-injection synthesis described in the literature, referencing the hot-injection synthesis database collated in prior

https://doi.org/10.1021/acs.chemmater.3c02751
Chem. Mater. XXXX, XXX, XXX−XXX

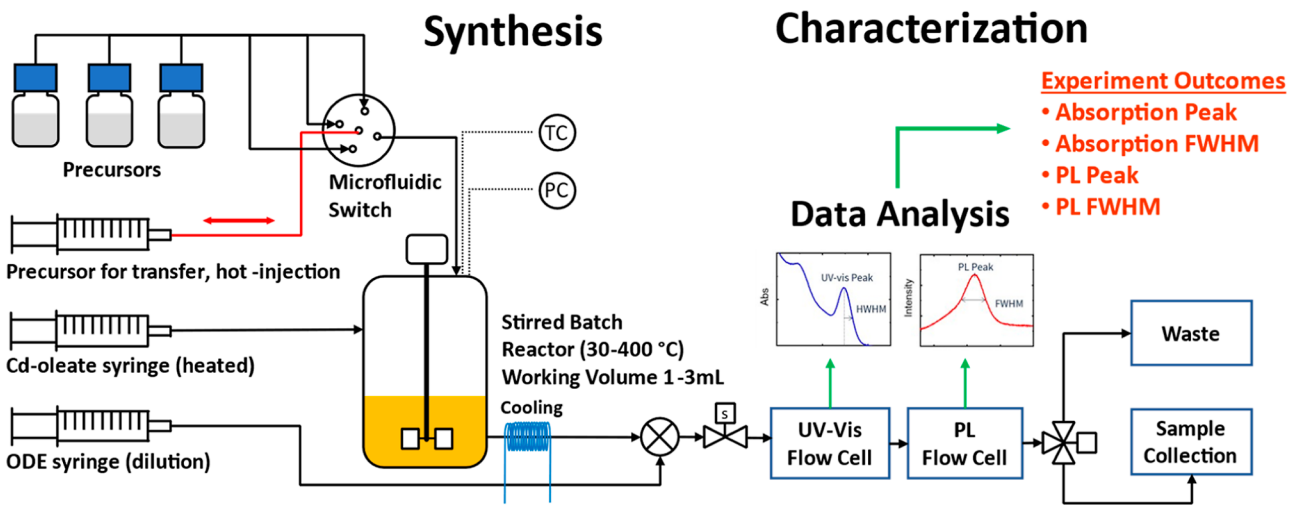


Figure 2. Schematic of a modular automated batch reactor setup for hot-injection synthesis of CdSe, comprising modules for synthesis (left) and characterization (right). The reactor is sampled periodically to characterize each batch reaction.

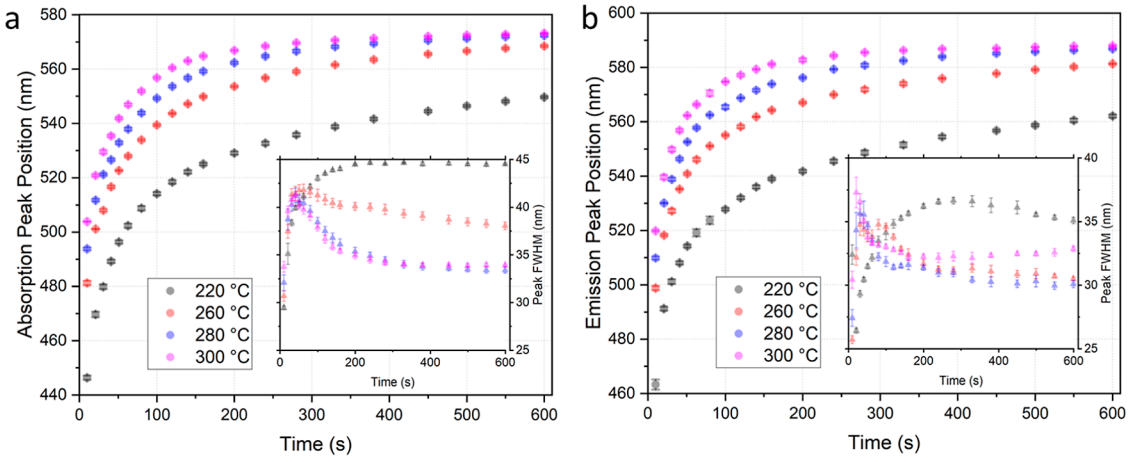


Figure 3. UV−vis (a) and PL (b) peaks and fwhm (inset) of repeated runs on an automated reactor platform demonstrating the reproducibility of synthesis platform and data analysis workflow. Each data point shows 5 runs with 10 mM Cd-oleate, 10 mM TOP-Se, 40:1 OA/Cd ratio.

work.[45] The stir rate was set to 800 rpm to ensure sufficient micro-mixing and thus uniform size distributions while reducing splatter in our setup.[66]

Two experimental test data sets are collected to test the ability of each of the tested models for interpolation and extrapolation. Twenty-four experiments are systematically chosen from points inside the training distribution away from points in the training/validation data set, while another 24 experiments were chosen from points outside the training distribution. The specific experimental conditions for the test data sets are listed in Table 2.

3.4. Experimental Data Set Visualization. 256 experiments (Table 2) were conducted over the course of 80 h to generate a total of 4946 (absorption) and 5088 (emission) data points, which corresponds to 97 and 99% of the total measured points, respectively. Some of the acquired raw spectra exhibited insufficient absorption or emission signal to allow for automated analysis by our peak-fitting algorithms, especially for specific reaction conditions such as those at low concentrations, temperatures, and times. Scatterplots for absorption and emission peaks versus fwhm are shown in Figure 4a. These data reveal that at the reaction conditions covered, this chemistry appears to be suitable for synthesizing monodisperse (<200 meV fwhm) nanoparticles with absorption energies from 2.1 to 3.0 eV and emission energies from 2.1 to 2.8 eV. A higher temperature appears to have the potential to yield larger particles with lower absorption energy. Similar plots plotted in units of nanometers are shown in Figure S2. Of note in this data set is that while observing the fitted peaks and fwhm of individual experiments, both peak and fwhm have relatively low absolute fitting errors (5 meV or 1 nm). However, because of the difference in ranges of peak and fwhm (~1 vs ~0.1 eV as seen in Figure 4b), a greater *relative* noise is present in the fwhm data that will affect the final coefficient of determination (R^2) values of the trained models (Figure S3).

Violin plots of absorption and emission peaks and fwhm are shown in Figure 4b. Emission peaks appear to have fwhm values slightly lower than those of absorption peaks. This difference in peak width stems from particles of different sizes emitting at different brightnesses: specific particles might be emitting more than others, which leads to an apparent narrowing of emission fwhm.[67] Another explanation for this observation is the overestimation of the absorption fwhm from the peak fitting algorithm due to the overlap of the absorption peak with the bulk onset absorption.

https://doi.org/10.1021/acs.chemmater.3c02751
Chem. Mater. XXXX, XXX, XXX−XXX

Table 2. Experiment Parameters for All-Automated Hot-Injection Synthesis

	training/validation data			
temperature (°C)	210	240	270	300
Cd concentration (mM)	5	10	20	40
TOP-Se concentration (mM)	5	10	20	40
OA/Cd ratio	3	10	20	40
	test data (interpolation)			
temperature (°C)	225	255	285	
Cd concentration (mM)	7	30		
TOP-Se concentration (mM)	7	30		
OA/Cd ratio	7.5	35		
	test data (extrapolation)			
temperature (°C)	205	210	305	310
Cd concentration (mM)	5	50		
TOP-Se concentration (mM)	5	50		
OA/Cd ratio	20	100	150	
	common parameters			
time (s)	process parameters			
10, 20, 30, 40, 50, 60,80	injection volume: 300 μL			
100, 120, 140, 160, 200	total reaction volume: 1200 μL			
240, 280, 330, 380, 450	800 rpm stir rate			
500, 550, 600	600 mbar reactor pressure (gauge)			

Correlation heatmaps for input and output variables are shown in Figure 4c. Cd-concentration ([Cd]), OA to Cd ratio ([OA]/[Cd]), temperature, and time exhibit a negative correlation to particle absorption energy; in other words, increasing these variables increases particle size. Conversely, Se-concentration [Se] is positively correlated to particle absorption energy, and increasing [Se] appears to decrease particle size. Such findings are consistent with literature observations that will be further discussed in Section 3.5. fwhm also appears to be strongly correlated with peak position, consistent with literature observations, and will also be further discussed in Section 3.5.

3.5. ML Model Performance for Synthesis Outcome Prediction. In this section, we will systematically investigate the capacity of various ML models to capture the trends in the experimental data set through their cross-validation performance. Next, the interpolation and extrapolation test sets will be used to test the best performing ML models to evaluate their performance on previously unseen data. Thereafter, the impact of noise and data set size on the prediction performance of various ML models will be evaluated.

3.5.1. Cross-Validation Performance. We performed 10-fold cross-validation for all ML models, the results of which are shown in units of energy (meV) in Figure 5. Equivalent results in units of wavelength (nm) can be found in Figure S4. The errors and R^2 values of a cross-validation analysis provide a measure for the ability of a model to capture trends in the data set. This approach is often used for ML model hyperparameter tuning and ML model selection. Cross-validation performance for peak prediction (Figure 5a) reveals a MAE of 20 meV (5 nm) for most models, which is relatively good performance for particles with peak energies between 2.1 and 3.0 eV, but it is much greater than the average experimental noise of about 2 meV (0.5 nm) as seen in Figure 3. Figure 5a also shows the inability of linear models such as ridge regression to predict QD synthesis outcomes due to the highly nonlinear effect of particle size on

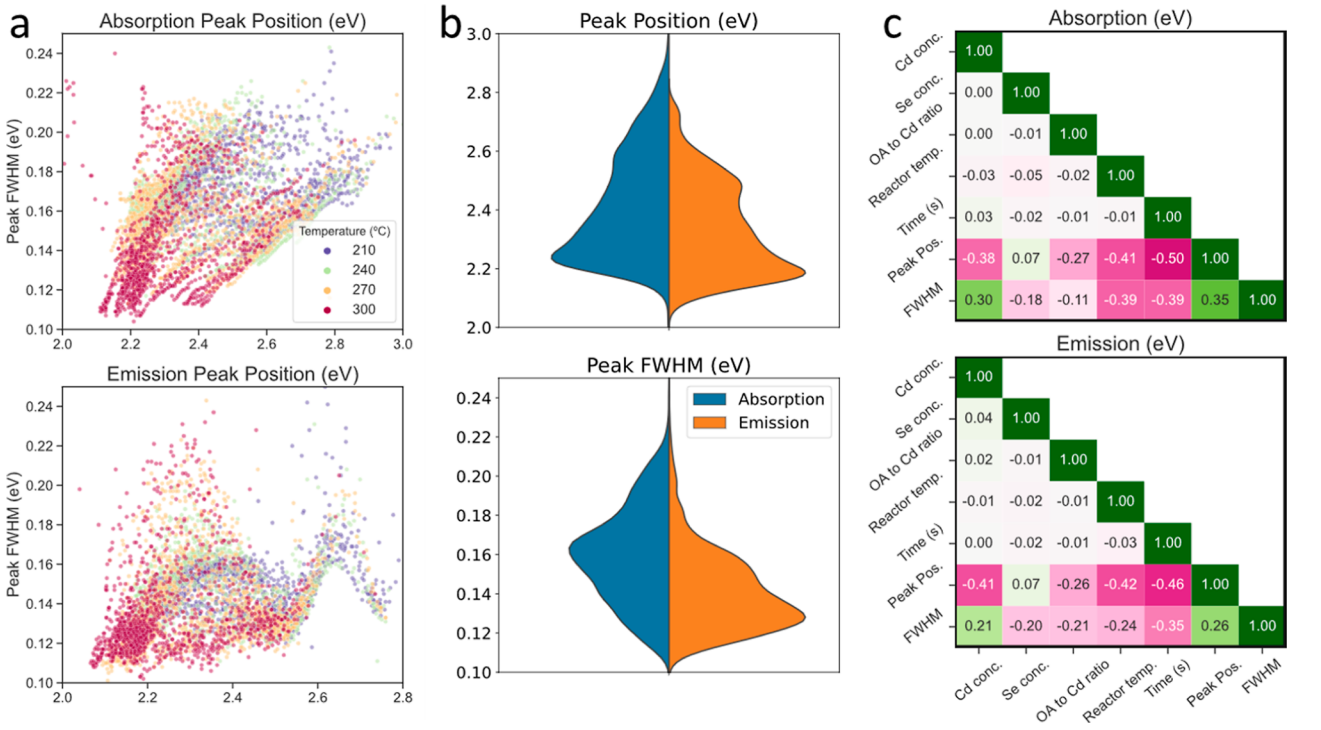


Figure 4. (a) Plot of fwhm vs Peak on a training/validation data set for absorption and emission; color of points indicates reaction temperature. (b) Histograms of peak and fwhm for absorption and emission. (c) Correlation heatmap of the training/validation data set demonstrating the influence of different synthesis parameters on reaction outcomes (peak, fwhm).

https://doi.org/10.1021/acs.chemmater.3c02751
Chem. Mater. XXXX, XXX, XXX−XXX

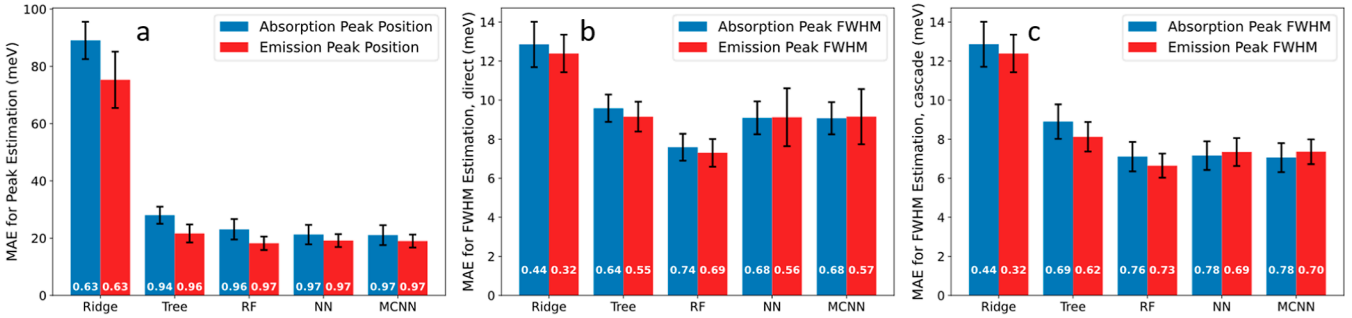


Figure 5. Cross-validation performance for each model type for the (a) peak position, (b) fwhm (direct), and (c) fwhm (cascade). R^2 values are indicated in each bar in white. Models with greater R^2 values and lower MAEs are more suitable to fit the data set.

process parameters. Tree-based models and NNs show reasonable prediction accuracy with high R^2 values and a low MAE, with RF slightly outperforming the single decision tree. RF models are bagging ensembles of decision trees and use the average of multiple different trees, each trained on a different subset of the data set to make predictions, which leads to more robust predictions. Both NN and MCNN demonstrate a similar performance. Prediction of absorption peak position by the models is generally worse compared to prediction of emission peak position, probably due to the peak fitting algorithm's ability to curve fit emission peaks more accurately.

Cross-validation performance for fwhm prediction using a direct model is shown in Figure 5b, while the cascade model is shown in Figure 5c. These figures demonstrate models for peak fwhm prediction showing lower R^2 values than those for the peak position, despite the low MAE values of 7 to 13 meV (1.5−2.5 nm). This is because fwhm is distributed between 100 and 220 meV (Figure 4b) and the peak fitting algorithms generate greater relative noise in the data set, as visualized in examples shown in Supporting Information Figure 3. The cascade model also demonstrates a slight but noticeable reduction in prediction errors compared to the direct model due to the correlation between fwhm and peak energy (Figure 4c): particles display lowered homogeneous and inhomogeneous line width as they increase in size. Homogeneous peak broadening stems from the intrinsic properties of individual semiconductor nanoparticles while inhomogeneous peak broadening stems from the heterogeneity of particle properties in an ensemble (e.g., size dispersion).[23] Generally, the homogeneous line width of smaller particles is higher due to the increased coupling between the exciton and surface phonon modes.[68] In this case, however, the majority of contributions to fwhm stem from size dispersion since the values of fwhm (>100 meV) in this data set are much higher than the minimum experimentally recorded ensemble line widths of 30−40 meV reported previously.[69] This demonstrates that larger particles tend to have more narrow size distributions in our data set due to size-focusing caused by factors such as diffusion-limited growth.[70]

3.5.2. Performance of ML Models for Synthesis Outcome Prediction on Unseen Data. After model hyperparameters were tuned for the RF and MCNN models, the entire training and validation set was used for model training and tested using the experimental test sets (Figure 6). Parity plots for other models are shown in Figure S5. Figure 6a,b and show the parity plots for absorption peak position and fwhm for RF regression and MCNN, respectively, while Figure 6c,d shows the parity plots for emission peak position and fwhm. MCNN shows good performance in predicting absorption and emission peaks, with the interpolation test set giving $R^2 > 0.96$ and a MAE of 24.6 meV (5.2 nm) and 19.3 meV (4.3 nm), respectively. This test accuracy is similar to prior work reported on flow synthesis of CdSe with similar precursors and additional dodecylamine ligands (Emission MAE = 4.2 nm).[53] This test performance is also significantly better than models trained on (mostly hot-injection) InP synthesis outcomes from the literature (absorption MAE = 14.7 nm, emission MAE = 8.4 nm), which demonstrate the limitations of using relatively sparse and noisy literature data for training ML models.[46]

In comparison of RF and MCNN models, it appears that the tree-based RF model interpolates relatively poorly. This is due to the way trees make splits based on the input feature's statistics. Because grid sampling was used to generate the input points, the individual splits are limited in resolution (only 4 points per varied experiment parameter), which can affect interpolation accuracy, especially if the test points are far from the training points.[56] On the other hand, NN-based models perform better for interpolation, as they work by creating a best-fit function that correlates output to input variables, which gives more accurate predictions than trees that split based on feature statistics.

This highlights the issue of sampling and its effect on the model performance. It appears that grid sampling with only 4 points per parameter in the parameter space explored for this synthesis system is unable to generate tree-based models with high prediction accuracy. Alternative sampling methods including random sampling or quasi-random sampling such as Latin-hypercube sampling can generate more accurate regression trees than our experiment data set since they give a wider resolution for tree-based models to make splits on compared to grid sampling.[71] Choosing the right sampling method for data set collection is essential for any experimental ML workflow. While grid sampling can allow us to systematically explore every variable, it is inefficient for building high-quality regressors for predicting experiment outcomes.[77−79] We advise experimentalists to use random or quasi-random sampling for collecting data sets in larger parameter spaces.

For fwhm prediction, cascade models appear to perform slightly better than direct models: the MCNN cascade model results exhibit relatively low MAEs of 7.4 meV (1.6 nm) and 6.2 meV (1.5 nm) for absorption and emission, respectively, which is close to the expected noise from peak fitting (~4 meV). This is a noticeable improvement in prediction performance over prior work using flow reactors for CdSe QD synthesis (MAE = 2.5 nm).[53] This improvement in fwhm prediction performance from cascade model stems from the previously mentioned correlation between absorption peak position and fwhm.[70] Direct models have access to the exact same information as cascade models but

https://doi.org/10.1021/acs.chemmater.3c02751
Chem. Mater. XXXX, XXX, XXX−XXX

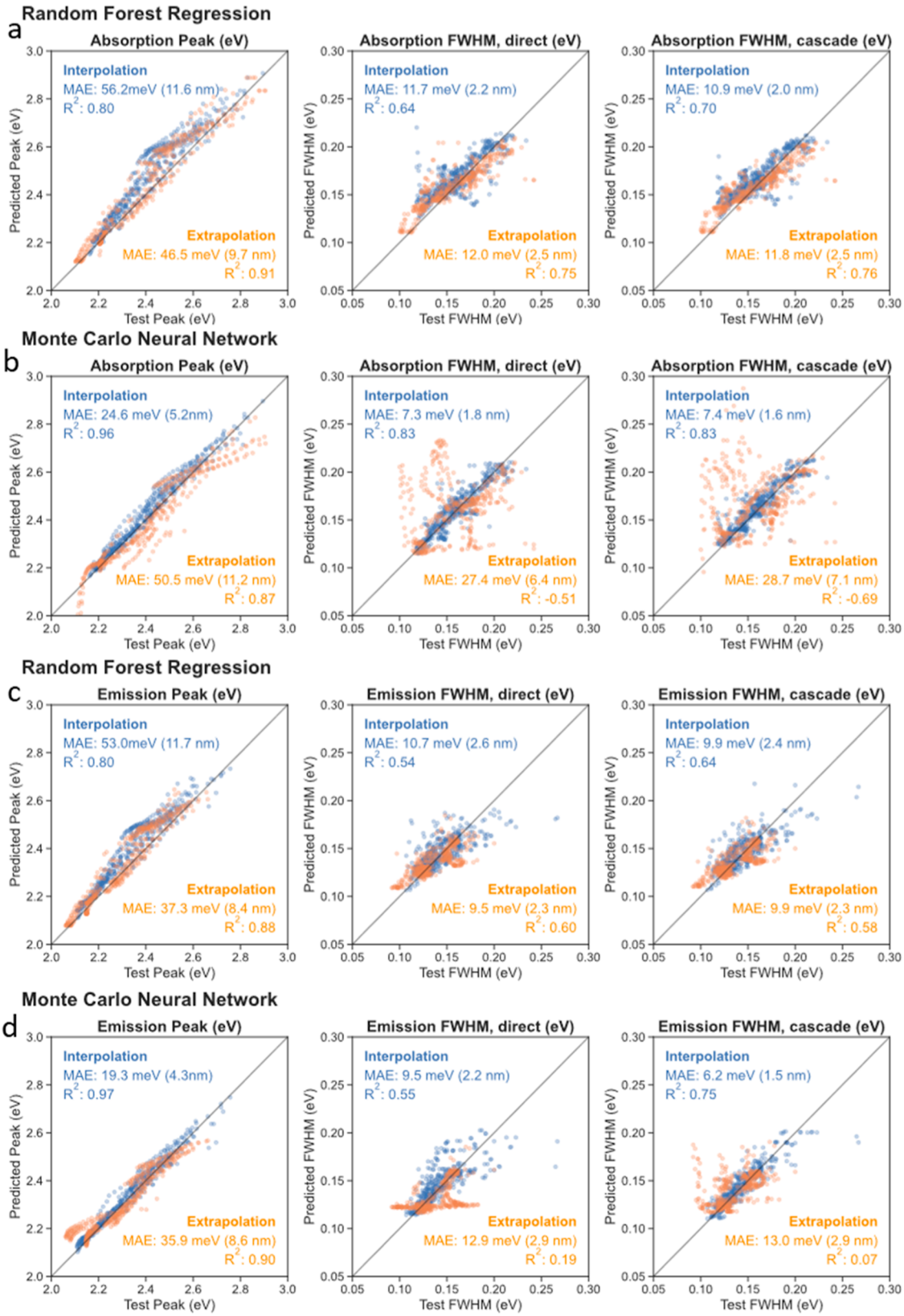


Figure 6. Parity plots for UV absorption peak position and fwhm for (a) random forest regression and (b) MCNN. Parity plots for PL emission peak position and fwhm for (c) random forest regression and (d) MCNN.

yield less accurate predictions since all learning systems are somewhat naïve, and simply adding an additional feature "peak position", even if it includes prediction errors, will help a model significantly by giving it an anchor point, aiding the prediction of fwhm. A cascade approach explicitly tells a model that the peak is an important feature for predicting fwhm and as such provides more accurate predictions. The improvement in prediction accuracy using a cascade approach demonstrates the use of

https://doi.org/10.1021/acs.chemmater.3c02751
Chem. Mater. XXXX, XXX, XXX−XXX

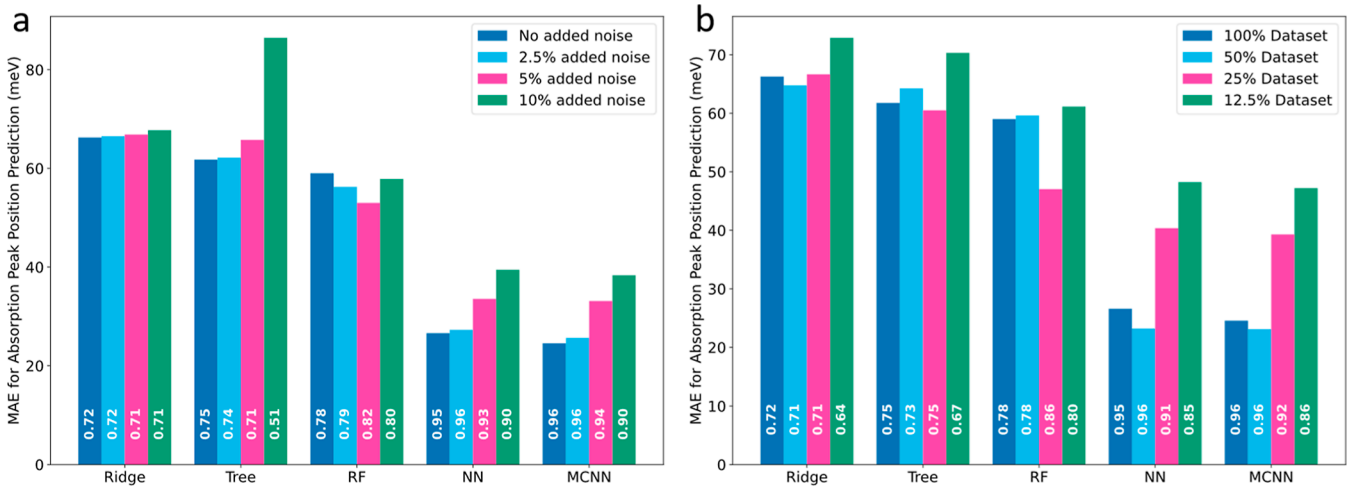


Figure 7. Plot of MAE and R^2 values (in white) for predictions (interpolation test set) of absorption peak position as a function of (a) noise added to the training set and (b) size of the data set.

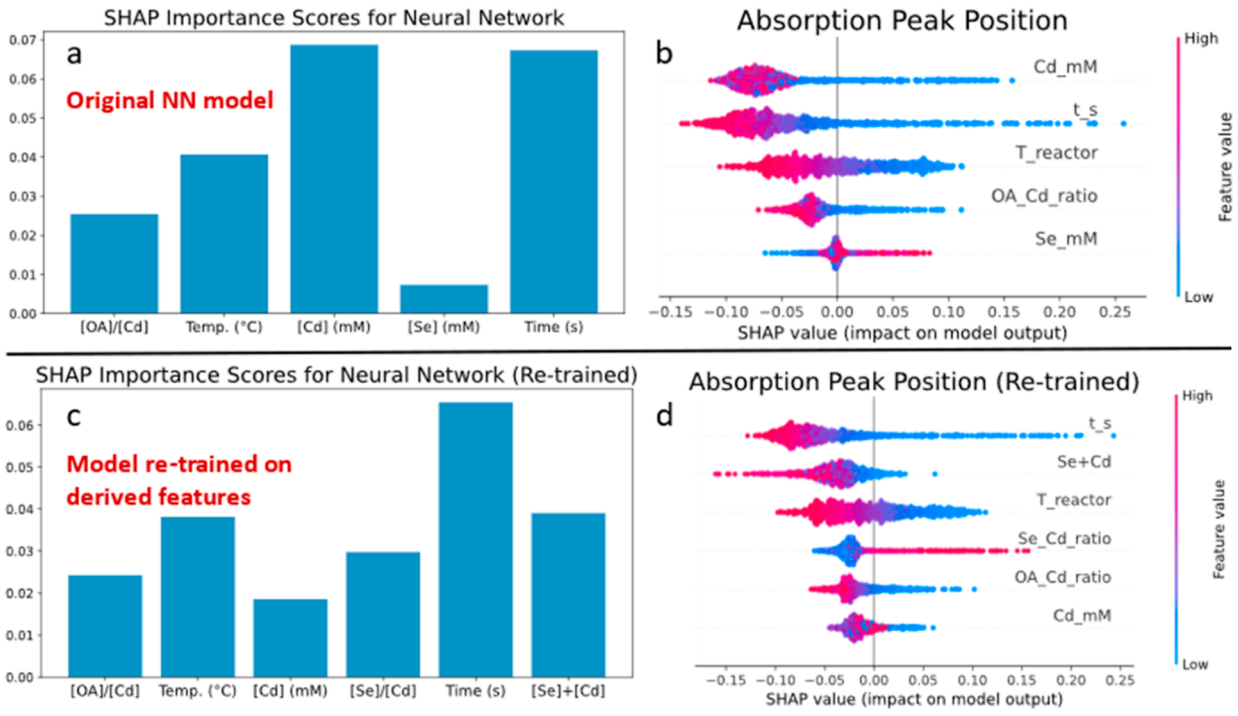


Figure 8. (a) Feature importance of the NN model on absorption and emission peaks. (b) SHAP summary plots for absorption peak, using TOP-Se concentrations. (c,d) Feature importance and summary plots of the NN model trained using derived features of [Se]/[Cd] ratio and total precursor concentrations [Se] + [Cd]. Higher values of absorption energy on the summary plots denote higher absorption energy and smaller particle diameter.

domain knowledge to design model architectures and is highly recommended for the design of experimental ML workflows.

Extrapolation data sets are included in our analysis because often points of interest in any material discovery or optimization campaign may lie outside the initial predetermined search space.[72] Unfortunately, in this study, no model performs well for points outside the training distribution due to the way the models work. Tree-based models appear to perform relatively well for our extrapolation test set because they extrapolate based on the splits closest to the unknown extrapolated point, and the closeness of the extrapolation data points with our training set allows these predictions to be relatively accurate.[73] On the other hand, NN-based models only seek to create a best-fit function that fits the points inside the training set, and this functional form cannot be guaranteed to satisfy experimental trends without explicit constraints set by a priori assumptions, which makes extrapolation unreliable.

3.5.3. Effect of Noise on Model Performance. To investigate the effect of noise on model performance, Gaussian noise with standard deviations of 2.5, 5, and 10% was added to the training/validation data set. Then, the models studied here were again tuned and trained for absorption peak prediction before testing on the interpolation data set. The performances of all models in terms of MAE and R^2 values for the cases with 2.5, 5, and 10% standard deviations as well as for the case without added noise are compared in Figure 7a. Ridge regression, as a linear model, appears to be the least sensitive to noise as the data set is relatively large compared to the number of variables. Among

https://doi.org/10.1021/acs.chemmater.3c02751
Chem. Mater. XXXX, XXX, XXX−XXX

tree-based models, single decision trees are especially sensitive to noise as regression trees work by splitting along feature statistics and assigning values to individual leaves. Adding noise to the data set directly affects the assigned values and leads to greatly diminished accuracy. On the other hand, RF models are a bagging ensemble of trees and are relatively insensitive to noise because these RF models are made up of numerous trees, each trained on a randomly selected subset of the data and the mean is taken over all the RF model's constituent trees to make predictions. This suggests the potential of using bagging ensembles of simpler models to improve model performance when the data set is noisy, albeit at the cost of data interpretability. Among the NN-based models (NN, MCNN), noise leads to poorer model performance, but both still show relatively accurate peak energy prediction with MAE <40 meV (<10 nm) even at 10% noise.

3.5.4. Effect of Data Set Size on Model Performance. Different subsets of the training/validation set (50, 25, and 12.5%) were randomly selected for model tuning, training, and testing. The experimental data set is split based on experiment number instead of randomly across all variables to simulate the effect of executing fewer batch experiments. Ridge regression and tree-based models show similar performance for peak prediction as data set size decreases. Even down to 12.5% of the experimental data set (only 614 data points), they maintain their test performance for this specific data set. However, NN models show that model accuracy is highly correlated to data availability, with a doubling of the MAE as data set size is reduced from 4946 to 614 data points. This demonstrates the inherent limitations of using artificial NNs for synthesis prediction: while NNs can model any arbitrary function accurately, in practice, their performance drops rapidly for smaller data sets.

In practice, however, the experimenter must carefully consider both the capabilities of the experimental platform (e.g., throughput, noise), the capacity of the model to capture nonlinear trends in the data set, and the interpretability of the model when selecting models. In specific cases where data sets are small and noisy, simpler interpretable linear models are far more practical to model synthesis space.

3.6. Investigating Model Explainability Tools to Generate Knowledge. The NN model for absorption peak prediction was used to investigate the relationship between the synthesis parameters and particle absorption peak position. SHAP values were calculated from 1000 points randomly selected from the synthesis parameter space to generate importance scores (Figure 8a) and SHAP summary plots (Figure 8b) as well as to visualize the effects of individual inputs on model predictions (Figure 8a,b).

3.6.1. SHAP Feature Importance vs Linear Correlation. Figure 8a shows the relative importance of synthesis parameters on the absorption energy. Time is the most important variable, as particles start out small and grow rapidly during the initial stages, leading to larger particles and lower absorption energy with time (Figure 3). Temperature is also a relatively important factor, with higher temperatures correlated with larger particle sizes and lower absorption energies. Surprisingly, Cd-oleate concentration is the second most important variable after time, much more important than the [OA]/[Cd] ratio and TOP-Se concentration. As seen in Figure 8b, both concentrations of Cd-oleate and the [OA]/[Cd] ratio are negatively correlated with particle absorption energy while the TOP-Se concentration is slightly positively correlated. Overall, SHAP feature importance scores agree with the correlation heatmap (Figure 4c), indicating that time is the most important factor for peak position prediction. However, the correlation heatmap that quantifies linear correlation might under-report the importance of factors that are nonlinearly correlated.

3.6.2. Effect of Feature Selection on SHAP Feature Importance Scores. Based on the results from Figure 8a,b, the Cd-oleate concentration appears to be an extremely important variable, while the TOP-Se concentration is relatively unimportant. This is contrary to past work that reported a significant effect of Cd/Se ratios on particle size.[74] To investigate whether using different features for peak energy prediction will generate different feature importance, we repeated the model tuning and training workflow using the derived feature of the [Se]/[Cd] ratio in lieu of [Se] and added an additional derived feature [Cd] + [Se]. This model performed similarly well to the original model using the raw concentration features (MAE = 21.0 meV (4.4 nm), R^2 = 0.97). SHAP feature importance scores and summary plots for this model (Figure 8c,d) indicate that time and temperature are still major factors with feature importance similar to that for the previous model, but the order of feature importance for the concentrations and ratios changes. In this case, both [Se]/[Cd] ratio and [Se] + [Cd] total concentration are more important features, and the importance of Cd-oleate concentration drops. This means that the model relies on these two derived features far more than Cd-oleate concentrations.

This observation calls into question the generalizability of model explainability tools such as SHAP. In both models, the inputs contain the same information, yet using [Se] alone as a feature suggests that the TOP-Se concentration is a minor factor in determining QD size, while using a more knowledge informed [Se]/[Cd] ratio shows TOP-Se to be a relatively important factor in determining particle size. Our experiments and prior work agree that a larger [Se]/[Cd] ratio leads to smaller particle sizes.[75] Explaining black box models with model explainability tools like SHAP runs the risk of underestimating the importance of specific factors since the models can make similarly accurate predictions very differently based on the choice of input features. This highlights the need for applying domain knowledge toward proper feature selection when training and explaining black-box models such as NNs or alternatively the need to develop inherently interpretable models in scientific ML, models that allow us to unambiguously interpret the correlations between output and input variables.

3.6.3. Explaining SHAP Trends with Domain Knowledge. To gain insights into the feature importance scores and trends from SHAP, we need to apply domain knowledge to put these trends into context. Generally, the formation of nanoparticles such as QDs can be split into three steps: reaction between precursors to form monomers, formation of nuclei, and growth of nuclei. Every step in this process is influenced by process parameters in a nontrivial way to affect the final ensemble property of particle size, which determines the measured absorption energy. Most particles exhibit a positive correlation between reaction time and particle size due to the growth of particles over time as the precursors react. In the case of our data set, particle size also appears to increase with temperature. This is attributed to fewer nuclei being formed at high temperatures due to the increasing solubility of the CdSe monomer at higher temperatures. Fewer nuclei formed lead to the monomer pool getting deposited onto fewer particles and thus increasing the average size of QDs formed.

https://doi.org/10.1021/acs.chemmater.3c02751
Chem. Mater. XXXX, XXX, XXX−XXX

As seen in Figure 8a,b, the absolute Se concentration appears to be an unimportant feature that is only slightly positively correlated to the absorption energy. A higher TOP-Se concentration in the reaction leads to smaller particles, and vice versa. However, Figure 8c,d demonstrates that the Se concentration is relatively important, with both the total precursor concentration ([Se] + [Cd]) and [Se]/[Cd] ratios having relatively high feature importance scores. A higher Se/Cd ratio leads to a greater number of nuclei formed and thus a smaller particle as the monomer pool gets deposited onto a greater number of particles. This is because a high Se/Cd ratio favors Se-terminated particles that are more difficult to dissolve compared to Cd-terminated particles.[75] Figure 8d also shows that increasing the total precursor concentration ([Se] + [Cd]) leads to larger particle sizes. Increasing the precursor concentration leads to an increased pool of available monomers that can deposit onto nuclei to form larger particles.

Besides precursor concentration, the ratio of ligands to precursors (OA/Cd ratio) is also a significant factor in determining particle size, with a higher OA/Cd ratio yielding larger particles. Because OA stabilizes Cd species in solution, higher OA/Cd ratios can dissolve smaller particles, leading to fewer nuclei and thus larger average nanocrystal size.[76] Overall, the trends predicted by our model for particle size prediction agree with the prior trends reported in the literature.

Ultimately, the combination of SHAP feature importance scores and domain knowledge can generate only hypotheses to be tested or validated with prior data. Because the models themselves are inherently uninterpretable, it is difficult to determine which hypotheses are more likely. To address this limitation, we can collect more data regarding the reaction, such as particle concentration, reaction yield, or reaction byproducts. Alternatively, inherently interpretable ML models able to suggest reaction rate laws can narrow the pool of available hypotheses for testing.

4. DISCUSSION OF LIMITATIONS AND FUTURE OPPORTUNITIES

Despite the relatively high prediction accuracy demonstrated for ML algorithms trained with experimental data sets obtained with our automated batch synthesis platform, this approach still has limitations. First, the use of a flow cell for UV−vis and PL spectroscopy coupled with periodic sampling of reactor contents means that we cannot precisely control the dilution of the sampled aliquot (needed to reduce absorption maxima with reaction progression) and are thus unable to collect spectroscopic data with identical peak intensity, leading to some noisiness in fwhm prediction, which in turn affects the training data set quality. Second, Peak-fitting for UV−vis absorption has no baseline subtraction and leads to a slightly overestimated fwhm and peak energy due to the convolution of the bulk onset absorption with the primary absorption peak. The combined contribution of these two factors leads to an overestimated, noisier data set for absorption peak position and absorption peak fwhm, which explains why the models for PL peak position and PL peak fwhm appear to perform better (Figure 6).

Despite these sources of noise from experimental and signal processing limitations, this work still generated a high-quality experimental data set for the hot-injection synthesis of CdSe quantum dots. This experimental data set includes the raw absorption, fitted peak, and fwhm data; it is openly available and can serve as a benchmark data set for testing of future ML algorithms. Furthermore, this work has demonstrated a high-quality model for predicting the synthesis outcome of CdSe hot-injection reactions—predictions from this model can be used to generate testable hypotheses that can provide new insights. This model can also be used as a digital twin of a real experiment system for use as a simulated experiment for developing active learning or adaptive sampling approaches to actively map the synthesis parameter space for QD synthesis. These active learning platforms can then be integrated into automated QD synthesis platforms to generate autonomous workflows that can be used for efficiently optimizing or mapping synthesis parameter space.[80] Prior work has demonstrated this integration for flow reactors, but the use of an automated batch reactor in our case opens the door to a wider range of precursors and possible heterostructure syntheses via multistep batch reactions.[47,48] Combining the automated platform with active learning algorithms will enable autonomous exploration of synthesis space beyond what was explored here, including different ligands and additives (e.g., phosphines) known to influence reaction outcomes. Such efforts can generate high-quality models to enhance the understanding of QD formation mechanisms.

The modular nature of our automated platform could enable the integration of characterization tools beyond optical spectroscopy. For example, small-angle X-ray scattering from synchrotron radiation sources can be used to directly determine the phases and quantify the concentrations and particle size distributions for synthesized QDs to generate more data that can be used to build more complex models for QD synthesis. Similarly, characterization tools, such as dynamic light scattering, can be used to quantify particle concentrations and sizes from batch synthesis platforms. The generated data can be used for more in-depth kinetic modeling that will improve the understanding of the mechanisms behind QD synthesis.

5. CONCLUSIONS

In this work, we used an automated batch synthesis platform to generate a large experimental data set for the hot-injection synthesis of CdSe QDs. We demonstrated a platform for performing automated experiments as well as the systematic testing and selection of ML algorithms to train models for accurately predicting synthesis outcomes. A comparison of experimental data and ML model outcomes revealed the strengths and weaknesses of various ML algorithms toward predicting synthesis outcomes for nanomaterial synthesis systems of various complexity, sizes, and noisiness and several best practices toward experimental data set collection, model selection, and domain knowledge application for model design highlighted.

SHAP analysis was used to reveal trends from the best trained models. Applying domain knowledge toward sensible feature selection turns out to be essential for maximizing the utility of model explainability tools. The dependence of SHAP importance scores on feature selection highlights the need for knowledge-informed feature selection for ML workflows and alternative ML models that are inherently interpretable.

This work enables the development of intelligent systems for understanding nanoparticle synthesis by both providing a high-quality, publicly available experimental data set that can be used for testing experimental ML workflows and by demonstrating examples of using model explanations to generate understanding about nanomaterial synthesis systems. The modular nature of the automated platform and the ML implementations also allow for further integration with other more informative character-

https://doi.org/10.1021/acs.chemmater.3c02751
Chem. Mater. XXXX, XXX, XXX−XXX

ization methods or more interpretable ML workflows and enable the development of autonomous workflows that can efficiently explore synthesis parameter space beyond what was explored here, to allow for further acceleration of developing an understanding of nanomaterial synthesis.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemmater.3c02751.

Detailed information for precursor preparation steps, ML models chosen, and specifics of model training and tuning and equivalents of Figures 4−7 plotted in units of wavelength (nm) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Paul J. A. Kenis − *Department of Chemical and Biomolecular Engineering, Urbana, Illinois 61801, United States;* orcid.org/0000-0001-7348-0381; Email: kenis@illinois.edu

Authors

Rui Hua Jeff Xu − *Department of Chemical and Biomolecular Engineering, Urbana, Illinois 61801, United States*

Logan P. Keating − *Department of Materials Science & Engineering, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-1106-9658

Ajit Vikram − *Department of Chemical and Biomolecular Engineering, Urbana, Illinois 61801, United States*

Moonsub Shim − *Department of Materials Science & Engineering, Urbana, Illinois 61801, United States;* orcid.org/0000-0001-7781-1029

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemmater.3c02751

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge financial support from the National Science Foundation under research grant no. 1825356 and support from the Grainger College of Engineering under the Strategic Research Initiative for conducting the research reported in this work. We thank Dr. Gryphon Drake (now University of South Carolina) for demonstrating manual batch synthesis of CdSe nanoparticles. We also thank Dr. Jiali Li and Dr. Flore Mekki-Berrada (National University of Singapore) for invaluable discussions on implementing ML algorithms and model explainability tools for experimental datasets.

■ REFERENCES

(1) García de Arquer, F. P.; Talapin, D. V.; Klimov, V. I.; Arakawa, Y.; Bayer, M.; Sargent, E. H. Semiconductor Quantum Dots: Technological Progress and Future Challenges. *Science* **2021**, *373* (6555), No. eaaz8541.

(2) Owen, J.; Brus, L. Chemical Synthesis and Luminescence Applications of Colloidal Semiconductor Quantum Dots. *J. Am. Chem. Soc.* **2017**, *139* (32), 10939−10943.

(3) Choi, M. K.; Yang, J.; Hyeon, T.; Kim, D.-H. Flexible Quantum Dot Light-Emitting Diodes for next-Generation Displays. *npj Flexible Electron.* **2018**, *2* (1), 10.

(4) Martynenko, I. V.; Litvin, A. P.; Purcell-Milton, F.; Baranov, A. V.; Fedorov, A. V.; Gun'ko, Y. K. Application of Semiconductor Quantum Dots in Bioimaging and Biosensing. *J. Mater. Chem. B* **2017**, *5* (33), 6701−6727.

(5) Byers, R. J.; Hitchman, E. R. Quantum Dots Brighten Biological Imaging. *Prog. Histochem. Cytochem.* **2011**, *45* (4), 201−237.

(6) van der Bok, J. C.; Dekker, D. M.; Peerlings, M. L. J.; Salzmann, B. B. V.; Meijerink, A. Luminescence Line Broadening of CdSe Nanoplatelets and Quantum Dots for Application in W-LEDs. *J. Phys. Chem. C* **2020**, *124* (22), 12153−12160.

(7) Nan, W.; Niu, Y.; Qin, H.; Cui, F.; Yang, Y.; Lai, R.; Lin, W.; Peng, X. Crystal Structure Control of Zinc-Blende CdSe/CdS Core/Shell Nanocrystals: Synthesis and Structure-Dependent Optical Properties. *J. Am. Chem. Soc.* **2012**, *134* (48), 19685−19693.

(8) Huang, L.; Ye, Z.; Yang, L.; Li, J.; Qin, H.; Peng, X. Synthesis of Colloidal Quantum Dots with an Ultranarrow Photoluminescence Peak. *Chem. Mater.* **2021**, *33* (5), 1799−1810.

(9) Hadar, I.; Hitin, G. B.; Sitt, A.; Faust, A.; Banin, U. Polarization Properties of Semiconductor Nanorod Heterostructures: From Single Particles to the Ensemble. *J. Phys. Chem. Lett.* **2013**, *4* (3), 502−507.

(10) Hu, L.; Zhao, Q.; Huang, S.; Zheng, J.; Guan, X.; Patterson, R.; Kim, J.; Shi, L.; Lin, C.-H.; Lei, Q.; Chu, D.; Tao, W.; Cheong, S.; Tilley, R. D.; Ho-Baillie, A. W. Y.; Luther, J. M.; Yuan, J.; Wu, T. Flexible and Efficient Perovskite Quantum Dot Solar Cells via Hybrid Interfacial Architecture. *Nat. Commun.* **2021**, *12* (1), 466.

(11) Kagan, C. R.; Lifshitz, E.; Sargent, E. H.; Talapin, D. V. Building Devices from Colloidal Quantum Dots. *Science* **2016**, *353* (6302), aac5523.

(12) Ko, Y.-H.; Prabhakaran, P.; Choi, S.; Kim, G.-J.; Lee, C.; Lee, K.-S. Environmentally Friendly Quantum-Dot Color Filters for Ultra-High-Definition Liquid Crystal Displays. *Sci. Rep.* **2020**, *10* (1), 15817.

(13) Bao, J.; Bawendi, M. G. A Colloidal Quantum Dot Spectrometer. *Nature* **2015**, *523* (7558), 67−70.

(14) Siddiqi, I. Engineering High-Coherence Superconducting Qubits. *Nat. Rev. Mater.* **2021**, *6* (10), 875−891.

(15) Yang, Z.; Gao, M.; Wu, W.; Yang, X.; Sun, X. W.; Zhang, J.; Wang, H.-C.; Liu, R.-S.; Han, C.-Y.; Yang, H.; Li, W. Recent Advances in Quantum Dot-Based Light-Emitting Devices: Challenges and Possible Solutions. *Mater. Today* **2019**, *24*, 69−93.

(16) Kirmani, A. R.; Luther, J. M.; Abolhasani, M.; Amassian, A. Colloidal Quantum Dot Photovoltaics: Current Progress and Path to Gigawatt Scale Enabled by Smart Manufacturing. *ACS Energy Lett.* **2020**, *5* (9), 3069−3100.

(17) Carey, G. H.; Abdelhady, A. L.; Ning, Z.; Thon, S. M.; Bakr, O. M.; Sargent, E. H. Colloidal Quantum Dot Solar Cells. *Chem. Rev.* **2015**, *115* (23), 12732−12763.

(18) Gary, D. C.; Glassy, B. A.; Cossairt, B. M. Investigation of Indium Phosphide Quantum Dot Nucleation and Growth Utilizing Triarylsilylphosphine Precursors. *Chem. Mater.* **2014**, *26* (4), 1734−1744.

(19) Gary, D. C.; Terban, M. W.; Billinge, S. J. L.; Cossairt, B. M. Two-Step Nucleation and Growth of InP Quantum Dots via Magic-Sized Cluster Intermediates. *Chem. Mater.* **2015**, *27* (4), 1432−1441.

(20) Wang, F.; Tang, R.; Buhro, W. E. The Trouble with TOPO; Identification of Adventitious Impurities Beneficial to the Growth of Cadmium Selenide Quantum Dots, Rods, and Wires. *Nano Lett.* **2008**, *8* (10), 3521−3524.

(21) Wang, F.; Tang, R.; Kao, J. L.-F.; Dingman, S. D.; Buhro, W. E. Spectroscopic Identification of Tri-n-Octylphosphine Oxide (TOPO) Impurities and Elucidation of Their Roles in Cadmium Selenide Quantum-Wire Growth. *J. Am. Chem. Soc.* **2009**, *131* (13), 4983−4994.

(22) Kopping, J. T.; Patten, T. E. Identification of Acidic Phosphorus-Containing Ligands Involved in the Surface Chemistry of CdSe Nanoparticles Prepared in Tri-N-Octylphosphine Oxide Solvents. *J. Am. Chem. Soc.* **2008**, *130* (17), 5689−5698.

(23) Nguyen, H. A.; Dixon, G.; Dou, F.; Gallagher, S.; Gibbs, S.; Ladd, D.; Marino, E.; Ondry, J.; Shanahan, J.; Vasileiadou, E.; et al. Design Rules for Obtaining Narrow Luminescence from Semiconductors Made in Solution. *Chem. Rev.* **2023**, *123* (12), 7890−7952.

(24) Wang, F.; Richards, V. N.; Shields, S. P.; Buhro, W. E. Kinetics and Mechanisms of Aggregative Nanocrystal Growth. *Chem. Mater.* **2014**, *26* (1), 5−21.
(25) Cossairt, B. M. Shining Light on Indium Phosphide Quantum Dots: Understanding the Interplay among Precursor Conversion, Nucleation, and Growth. *Chem. Mater.* **2016**, *28* (20), 7181−7189.
(26) Frenette, L. C.; Krauss, T. D. Uncovering Active Precursors in Colloidal Quantum Dot Synthesis. *Nat. Commun.* **2017**, *8* (1), 2082.
(27) García-Rodríguez, R.; Hendricks, M. P.; Cossairt, B. M.; Liu, H.; Owen, J. S. Conversion Reactions of Cadmium Chalcogenide Nanocrystal Precursors. *Chem. Mater.* **2013**, *25* (8), 1233−1249.
(28) Owen, J. S.; Chan, E. M.; Liu, H.; Alivisatos, A. P. Precursor Conversion Kinetics and the Nucleation of Cadmium Selenide Nanocrystals. *J. Am. Chem. Soc.* **2010**, *132* (51), 18206−18213.
(29) Joo, J.; Pietryga, J. M.; McGuire, J. A.; Jeon, S.-H.; Williams, D. J.; Wang, H.-L.; Klimov, V. I. A Reduction Pathway in the Synthesis of PbSe Nanocrystal Quantum Dots. *J. Am. Chem. Soc.* **2009**, *131* (30), 10620−10628.
(30) Steckel, J. S.; Yen, B. K. H.; Oertel, D. C.; Bawendi, M. G. On the Mechanism of Lead Chalcogenide Nanocrystal Formation. *J. Am. Chem. Soc.* **2006**, *128* (40), 13032−13033.
(31) Evans, C. M.; Evans, M. E.; Krauss, T. D. Mysteries of TOPSe Revealed: Insights into Quantum Dot Nucleation. *J. Am. Chem. Soc.* **2010**, *132* (32), 10973−10975.
(32) He, X.-H.; Yin, Y.-L.; Tang, G.; Liu, Y.; Jiang, F.-L. Role of Trioctylphosphine in the Synthesis of Quantum Dots: A Modulator of Nucleation, Growth, and Solubility. *J. Phys. Chem. C* **2023**, *127* (10), 5021−5028.
(33) Abdel-Latif, K.; Bateni, F.; Crouse, S.; Abolhasani, M. Flow Synthesis of Metal Halide Perovskite Quantum Dots: From Rapid Parameter Space Mapping to AI-Guided Modular Manufacturing. *Matter* **2020**, *3* (4), 1053−1086.
(34) Voznyy, O.; Levina, L.; Fan, J. Z.; Askerka, M.; Jain, A.; Choi, M. J.; Ouellette, O.; Todorović, P.; Sagar, L. K.; Sargent, E. H. Machine Learning Accelerates Discovery of Optimal Colloidal Quantum Dot Synthesis. *ACS Nano* **2019**, *13* (10), 11122−11128.
(35) Nightingale, A. M.; De Mello, J. C. Microscale Synthesis of Quantum Dots. *J. Mater. Chem.* **2010**, *20* (39), 8454−8463.
(36) Vikram, A.; Zahid, A.; Bhargava, S. S.; Jang, H.; Sutrisno, A.; Khare, A.; Trefonas, P.; Shim, M.; Kenis, P. J. A. Unraveling the Origin of Interfacial Oxidation of InP-Based Quantum Dots: Implications for Bioimaging and Optoelectronics. *ACS Appl. Nano Mater.* **2020**, *3* (12), 12325−12333.
(37) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki-Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819−4827.
(38) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angew. Chem. Int. Ed.* **2022**, *61* (29), No. e202204647.
(39) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547−555.
(40) Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials Discovery and Design Using Machine Learning. *J. Mater.* **2017**, *3* (3), 159−177.
(41) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56* (10), 1894−1904.
(42) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29* (21), 9436−9444.
(43) Kim, E.; Jensen, Z.; Van Grootel, A.; Huang, K.; Staib, M.; Mysore, S.; Chang, H. S.; Strubell, E.; McCallum, A.; Jegelka, S.; Olivetti, E. Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (3), 1194−1201.
(44) Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E. A.; Ceder, G. Opportunities and Challenges of Text Mining in Materials Research. *iScience* **2021**, *24* (3), 102155.
(45) Baum, F.; Pretto, T.; Köche, A.; Santos, M. J. L. Machine Learning Tools to Predict Hot Injection Syntheses Outcomes for II-VI and IV-VI Quantum Dots. *J. Phys. Chem. C* **2020**, *124* (44), 24298−24305.
(46) Nguyen, H. A.; Dou, F. Y.; Park, N.; Wu, S.; Sarsito, H.; Diakubama, B.; Larson, H.; Nishiwaki, E.; Homer, M.; Cash, M.; Cossairt, B. M. Predicting Indium Phosphide Quantum Dot Properties from Synthetic Procedures Using Machine Learning. *Chem. Mater.* **2022**, *34* (14), 6296−6311.
(47) Vikram, A.; Brudnak, K.; Zahid, A.; Shim, M.; Kenis, P. J. A. Accelerated Screening of Colloidal Nanocrystals Using Artificial Neural Network-Assisted Autonomous Flow Reactor Technology. *Nanoscale* **2021**, *13* (40), 17028−17039.
(48) Epps, R. W.; Bowen, M. S.; Volk, A. A.; Abdel-Latif, K.; Han, S.; Reyes, K. G.; Amassian, A.; Abolhasani, M. Artificial Chemist: An Autonomous Quantum Dot Synthesis Bot. *Adv. Mater.* **2020**, *32* (30), 1−9.
(49) Gongora, A. E.; Xu, B.; Perry, W.; Okoye, C.; Riley, P.; Reyes, K. G.; Morgan, E. F.; Brown, K. A. A Bayesian Experimental Autonomous Researcher for Mechanical Design. *Sci. Adv.* **2020**, *6* (15), No. eaaz1708.
(50) Chan, E. M.; Xu, C.; Mao, A. W.; Han, G.; Owen, J. S.; Cohen, B. E.; Milliron, D. J. Reproducible, High-Throughput Synthesis of Colloidal Nanocrystals for Optimization in Multidimensional Parameter Space. *Nano Lett.* **2010**, *10* (5), 1874−1885.
(51) Li, J.; Chen, T.; Lim, K.; Chen, L.; Khan, S. A.; Xie, J.; Wang, X. Deep Learning Accelerated Gold Nanocluster Synthesis. *Adv. Intell. Syst.* **2019**, *1* (3), 1900029.
(52) Esterhuizen, J. A.; Goldsmith, B. R.; Linic, S. Interpretable Machine Learning for Knowledge Generation in Heterogeneous Catalysis. *Nat. Catal.* **2022**, *5* (3), 175−184.
(53) Orimoto, Y.; Watanabe, K.; Yamashita, K.; Uehara, M.; Nakamura, H.; Furuya, T.; Maeda, H. Application of Artificial Neural Networks to Rapid Data Analysis in Combinatorial Nanoparticle Syntheses. *J. Phys. Chem. C* **2012**, *116* (33), 17885−17896.
(54) Yu, W. W.; Peng, X. Formation of High-Quality CdS and Other II-VI Semiconductor Nanocrystals in Noncoordinating Solvents: Tunable Reactivity of Monomers. *Angew. Chem. Int. Ed.* **2002**, *41* (13), 2368−2371.
(55) McDonald, G. C. Ridge Regression. *Wiley Interdiscip. Rev. Comput. Stat.* **2009**, *1* (1), 93−100.
(56) Breiman, L. *Classification and Regression Trees*; Routledge, 2017.
(57) Anderson, J. A. *An Introduction to Neural Networks*; MIT press, 1995.
(58) Gal, Y.; Ghahramani, Z. *33rd Int. Conf. Mach. Learn. ICML 2016*; PMLR, 2016; Vol. 3, pp 1651−1660.Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning
(59) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, C. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825−2830.
(60) Chollet, F.; et al. Keras. GitHub. 2015, https://github.com/keras-team/keras (accessed Oct 03, 2023).
(61) Gabry, J.; Simpson, D.; Vehtari, A.; Betancourt, M.; Gelman, A. Visualization in Bayesian Workflow. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2019**, *182* (2), 389−402.
(62) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; Vol. *30.Advances in Neural Information Processing Systems*
(63) Chu, D. B. K.; Owen, J. S.; Peters, B. Nucleation and Growth Kinetics from LaMer Burst Data. *J. Phys. Chem. A* **2017**, *121* (40), 7511−7517.
(64) Trachsel, F.; Günther, A.; Khan, S.; Jensen, K. F. Measurement of Residence Time Distribution in Microfluidic Systems. *Chem. Eng. Sci.* **2005**, *60* (21), 5729−5737.

https://doi.org/10.1021/acs.chemmater.3c02751
Chem. Mater. XXXX, XXX, XXX−XXX

(65) Chan, E. M.; Alivisatos, A. P.; Mathies, R. A. High-Temperature Microfluidic Synthesis of CdSe Nanocrystals in Nanoliter Droplets. *J. Am. Chem. Soc.* **2005**, *127* (40), 13854−13861.
(66) Salaheldin, A. M.; Walter, J.; Herre, P.; Levchuk, I.; Jabbari, Y.; Kolle, J. M.; Brabec, C. J.; Peukert, W.; Segets, D. Automated Synthesis of Quantum Dot Nanocrystals by Hot Injection: Mixing Induced Self-Focusing. *Chem. Eng. J.* **2017**, *320*, 232−243.
(67) Efros, A. L.; Nesbitt, D. J. Origin and Control of Blinking in Quantum Dots. *Nat. Nanotechnol.* **2016**, *11* (8), 661−671.
(68) Kang, S.; Kim, Y.; Jang, E.; Kang, Y.; Han, S. Fundamental Limit of the Emission Linewidths of Quantum Dots: An Ab Initio Study of CdSe Nanocrystals. *ACS Appl. Mater. Interfaces* **2020**, *12* (19), 22012−22018.
(69) Cho, W.; Kim, S.; Coropceanu, I.; Srivastava, V.; Diroll, B. T.; Hazarika, A.; Fedin, I.; Galli, G.; Schaller, R. D.; Talapin, D. V. Direct Synthesis of Six-Monolayer (1.9 Nm) Thick Zinc-Blende CdSe Nanoplatelets Emitting at 585 Nm. *Chem. Mater.* **2018**, *30* (20), 6957−6960.
(70) Reiss, H. The Growth of Uniform Colloidal Dispersions. *J. Chem. Phys.* **1951**, *19* (4), 482−487.
(71) Shirley, P. Discrepancy as a Quality Measure for Sample Distributions. *Eurographics* **1991**, *91*, 183−194.
(72) Mekki-Berrada, F.; Ren, Z.; Huang, T.; Wong, W. K.; Zheng, F.; Xie, J.; Tian, I. P. S.; Jayavelu, S.; Mahfoud, Z.; Bash, D.; Hippalgaonkar, K.; Khan, S.; Buonassisi, T.; Li, Q.; Wang, X. Two-Step Machine Learning Enables Optimized Nanoparticle Synthesis. *npj Comput. Mater.* **2021**, *7* (1), 55.
(73) Xu, K., Li, J., Zhang, M., Du, S. S., Kawarabayashi, K., Jegelka, S. How Neural Networks Extrapolate: From Feedforward to Graph Neural Networks. **2020**, arXiv Prepr. arXiv:2009.11848 (accessed Oct 03, 2023).
(74) Krishnadasan, S.; Brown, R. J. C.; DeMello, A. J.; DeMello, J. C. Intelligent Routes to the Controlled Synthesis of Nanoparticles. *Lab Chip* **2007**, *7* (11), 1434−1441.
(75) Tamukong, P. K.; Peiris, W. D. N.; Kilina, S. Computational Insights into CdSe Quantum Dots' Interactions with Acetate Ligands. *Phys. Chem. Chem. Phys.* **2016**, *18* (30), 20499−20510.
(76) Oh, N.; Shim, M. Metal Oleate Induced Etching and Growth of Semiconductor Nanocrystals, Nanorods, and Their Heterostructures. *J. Am. Chem. Soc.* **2016**, *138* (33), 10444−10451.
(77) Liashchynskyi, P., Liashchynskyi, P. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. **2019**, arXiv Prepr. arXiv:1912.06059 (accessed Oct 03, 2023).
(78) Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13* (2), 281−305.
(79) Mishra, S.; Rusch, T. K. Enhancing Accuracy of Deep Learning Algorithms by Training with Low-Discrepancy Sequences. *SIAM J. Numer. Anal.* **2021**, *59* (3), 1811−1834.
(80) Ren, Z.; Ren, Z.; Zhang, Z.; Buonassisi, T.; Li, J. Autonomous Experiments Using Active Learning and AI. *Nat. Rev. Mater.* **2023**, *8* (9), 563−564.

https://doi.org/10.1021/acs.chemmater.3c02751
Chem. Mater. XXXX, XXX, XXX−XXX