Predicting Indium Phosphide Quantum Dot Properties from Synthetic Procedures Using Machine Learning

Hao A. Nguyen[a], Florence Y. Dou[a], Nayon Park[a], Shenwei Wu[a], Harrison Sarsito[b], Benedicte Diakubama[b], Helen Larson[a], Emily Nishiwaki[a], Micaela Homer[a], Melanie Cash[a], Brandi M. Cossairt[a]

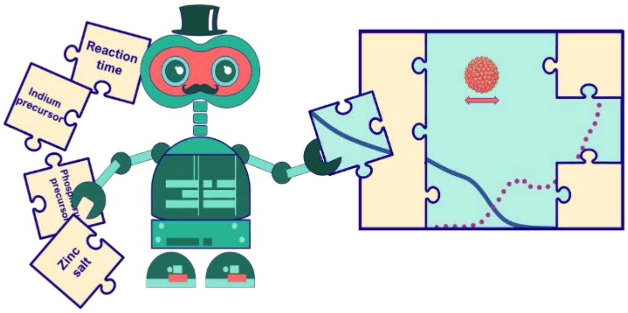
Affiliations

[a] Department of Chemistry, University of Washington, Seattle, WA 98195, USA

[b] Department of Chemical Engineering, University of Washington, Seattle, WA 98195, USA

Abstract

Predictions of chemical reaction outcomes using machine learning (ML) has emerged as a powerful tool for advancing materials synthesis. However, this approach requires large and diverse datasets, which are extremely limited in the field of nanomaterials synthesis due to inconsistent and non-standardized reporting in the literature and a lack of understanding of synthetic mechanisms. In this study, we extracted parameters of InP quantum dot (QD) syntheses as our inputs, and resultant properties (absorption, emission, diameter) as our outputs from 72 publications. We “filled in” missing outputs using a data imputation method to prepare a complete dataset containing 216 entries for training and testing predictive ML models. We defined the descriptor space in two ways (condensed and extended) based on either chemical identity or the role of reagents to explore the best approach for categorizing input features. We achieved mean absolute errors (MAEs) as low as 20.29, 11.46, and 0.33 nm for absorption, emission, and diameter respectively with our best ML model across diverse synthetic methods. We used these models to deploy an accessible and interactive webapp for designing syntheses of InP (https://share.streamlit.io/cossairt-lab/indium-phosphide/Hot_injection/hot_injection_prediction.py). Using this webapp, we investigated the power of ML to uncover chemical trends in InP syntheses, such as the effects of common additives, like zinc salts and trioctylphosphine. We also designed and conducted new experiments based on extensions of literature procedures and compared our experimentally measured properties to predictions, thus evaluating the “real-life” accuracy of our models. Conversely, we used inverse-design to obtain InP QDs with specific properties. Finally, we applied the same approach to train, test, and launch predictive models for CdSe QDs by expanding a previously published dataset. Altogether, our data pre-processing method and ML implementations demonstrate the ability to design materials with targeted properties and explore underlying reaction mechanisms even when faced with limited data resources.



1. INTRODUCTION

Indium phosphide quantum dots (QDs) are a promising alternative to traditional Cd- and Pb- based materials for lighting, displays, and optoelectronic technologies[1–3]. However, due to its increased covalency, limitations in easily accessible precursors, and inherent distinctions in precursor reactivity and valency, the synthesis of InP has been met by more challenges when compared to their II-VI and IV-VI counterparts in terms of extracting generalizable design principles and targeted properties[4]. Since the first InP QD synthesis in 1994 that reported the use of chloroindium oxalate combined with tris(trimethylsilyl)phosphine ($P(TMS)_3$) in a mixture of trioctylphosphine (TOP) and trioctylphosphine oxide (TOPO) using a heat-up method[5], intense effort has been devoted to exploring new synthetic methodologies and new precursors (**Figure 1**). The most important synthetic developments include the hot-injection method that typically produces ensembles with a high degree of monodispersity[6], the magic-sized cluster-mediated method that exploits our understanding of the non-classical growth mechanisms observed under certain reaction conditions[7,8], and the microwave-assisted method that uses inductive heating and in situ fluoride generation to develop a scalable InP synthetic platform that results in luminescent InP cores directly out of the synthesis[9]. Efforts to replace the highly reactive and challenging to handle tris(trimethylsilyl)phosphine ($P(TMS)_3$) precursor to better separate nucleation and growth have resulted in a variety of new phosphorus precursors such as aminophosphines[10], tris(trimethylgermyl)phosphine[11], phosphine gas[12], and white phosphorus[13]. In general, synthetic development has focused on narrowing size distributions, increasing quantum yields, and exploring more environmentally benign reagents. Other important considerations in this regard are tunability and reproducibility in particle size and emission wavelength, which are governed by different synthetic factors including but not limited to nucleation temperature, reaction time, precursor conversion kinetics, additives, and post-synthetic manipulations. Often, QDs with distinct sizes and excitonic emission wavelengths are isolated by taking aliquots from the reaction mixture at different reaction times. However, maximizing material yield and achieving precise synthetic control and reproducibility over particle size and emission wavelengths of InP QDs still remain a challenge.

In recent years, machine learning (ML) has emerged as a powerful tool to accelerate chemical reaction design and materials discovery. ML techniques are effective at inferring patterns and uncovering trends from complex chemical processes or mechanisms when a database of a reasonable size is available. In the field of nanomaterials, ML has been used to extract data[14–16], discover novel materials[17–19], optimize chemical reactions[20–22], reveal underlying mechanisms[23,24], and predict synthetic outcomes[25]. For example, support vector machine classification and regression models were used to synthetically control layer thickness of perovskite halide nanoplatelets[26]. In another application, Bayesian optimization was applied to improve monodispersity of PbS QDs, leading to the narrowest reported half-width at half-maximum of absorbance of this material[27]. In 2020, Santos and coworkers published a study wherein different ML algorithms were applied to identify influential synthetic parameters and to predict the final size of a variety of metal chalcogenide QDs, including CdSe, CdS, PbS, PbSe, and ZnSe[25]. The Gradient Boosting Machine algorithm used in that study resulted in a high R^2 value and revealed that growth temperature and time are the most influential synthetic parameters. In addition, several groups have used automated technology with feedback learning mechanisms to generate their own synthesis parameter space to create nanocrystals, including InP[28], with desired characteristics[20,29]. The accuracy of predictions is typically limited by the size of the dataset, and the completeness and quality (i.e., cover a wide distribution of parameter space). While there are many valuable materials databases such as the Inorganic Crystal Structure Database, NREL Materials Database, Materials Project, Stanford Catalysis-Hub, and PubChem, in the field of nanomaterials, there are a limited number of adequate datasets largely due to inconsistencies in reporting and the lack of an organized, centralized data repository.

In this work, we employ different predictive ML algorithms to gain insights into reaction condition control over particle diameter, absorption, and emission wavelength of InP QDs from reported data. ML methods are appropriate to help us gain deeper understanding of InP QD synthesis because of the complexity of factors that affect the physical and electronic structure of the QDs. In principle, particle diameter, excitonic absorption, and band-edge emission should be connected, but from experimental observations, nuances related to surface chemistry, stoichiometry, and size and morphological heterogeneity make direct correlations less obvious. We demonstrate a dataset pre-processing technique to overcome the challenge of having limited data from the literature. Different approaches to define input descriptors and machine learning model types are explored to find the best strategy for reaction prediction. Finally, we deploy an accessible user interface for external users and apply this interface to compare the results of new experiments with predicted results obtained from the ML models.

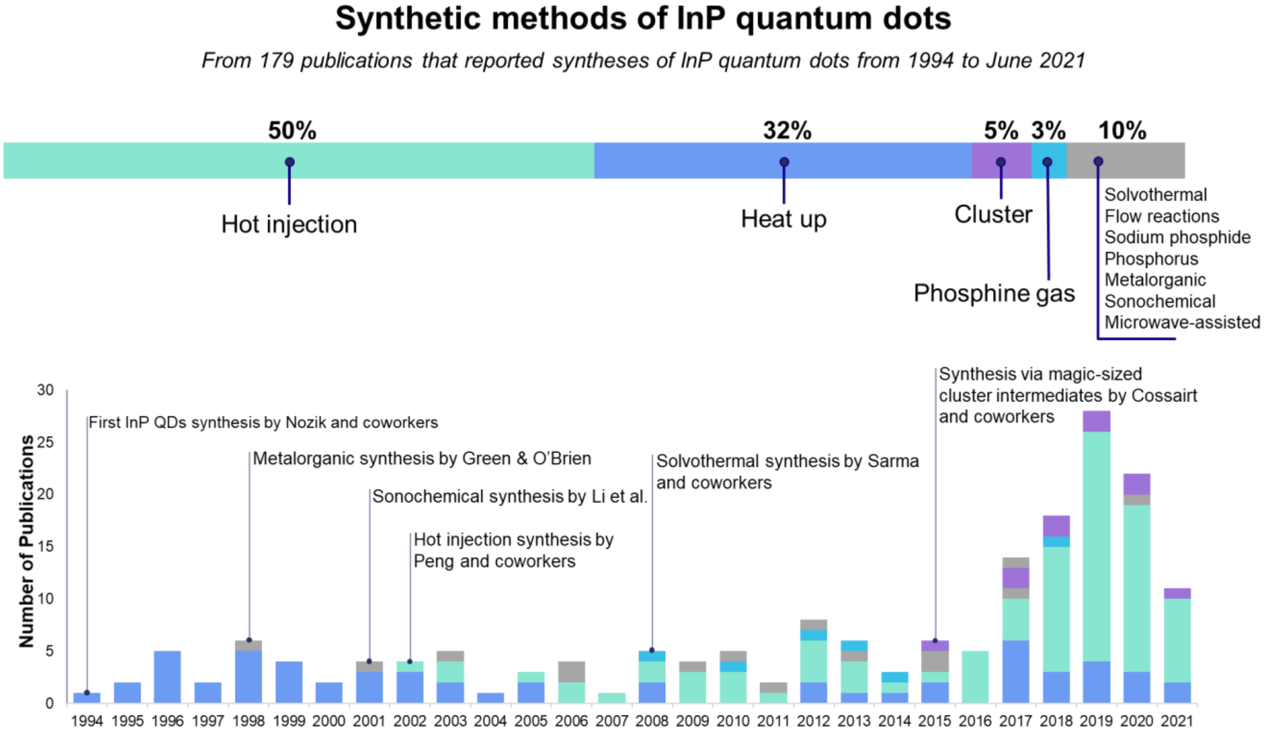


Figure 1. Timeline and number of publications of InP QDs synthesis.

2. METHODOLOGY

2.1 Data Acquisition

The dataset was created by manually extracting reaction conditions and resultant size and optical properties reported in the literature using Web of Science and Scifinder with search terms: “indium phosphide”, “indium phosphide quantum dots”, “InP”, and “III-V quantum dots”. We identified 179 articles from 1994 to June 2021 that reported syntheses of InP QDs. We then classified the articles by synthetic methods (e.g., heat-up, hot injection, magic-sized cluster-mediated, etc.). Since there are significant practical differences among these synthetic methods that can affect the accuracy of the predictions, only similar methods, where the reaction is performed using batch-type techniques with molecular indium and phosphorus precursors, were used for further data extraction. We also excluded syntheses that did not include any size, absorption, or emission data. This process resulted in an initial dataset that included 219 syntheses from 72 different articles, in which the hot injection method, heat-up method, reactions using phosphine gas, reactions using white phosphorus, and reactions using sodium phosphide make up 73%, 19%, 5%, 2%, and 1% of the syntheses respectively. An illustration of how the data extraction was done can be found in *Figure S1*.

2.2 Datasets

The data extracted from the 219 syntheses were split into input features and output targets. With the goal of predicting properties of QDs, the output targets contained particle diameter in nm measured directly from transmission electron microscopy (TEM), absorption wavelength in nm, and photoluminescence (PL) emission wavelength in nm. Although the three chosen outputs are physically related, e.g., QD size can be theoretically determined by the excitonic peak from absorption spectra, we wanted to investigate the ability of the ML models to recognize these relationships.

While defining the output set was straightforward, determining the input features required more consideration. In general, the performance of a predictive model depends on finding representative input features[30,31]. Furthermore, using too many input features may lead to overfitting. This becomes challenging, especially for predictive chemical synthesis models, where the outcomes of syntheses are non-trivially affected by unknown, unreported, and/or seemingly trivial parameters. Therefore, to evaluate the effect of feature selection on our models, we defined two sets of input features and compiled two datasets: an extended dataset with 22 features and a condensed dataset with 18 features (**Figure 2B**). In the extended dataset, the additives beyond the indium and phosphorus sources were categorized by their functional groups (e.g., carboxylic acid, amine, thiol); while the condensed dataset grouped chemicals by their primary assumed role in the synthesis (e.g., ligands, solvents). (See the full list of input features in *Table S1*). Using the extended dataset, we hoped to uncover trends of additives based on chemical identity such as fatty amines, zinc salts, and thiol-containing ligands that may play more than one role in the synthesis[32–35]. For example, since thiols and fatty amines are sometimes used as both coordinating solvents and capping ligands that directly affect the optical properties of QDs, it is more reasonable to separate these features from each other and from other features (solvents and acids) in the dataset with the risk of having a high dimensionality. On the other hand, features in the condensed dataset were chosen to reduce the number of input variables for better ML performance with the tradeoff of not observing unique behaviors of some additives. Prior to training machine learning models, the continuous values (In amount, P amount, reaction time, etc.) in the input set were scaled and the categorical features (In source, P source, etc.) were transformed to numerical features using one-hot encoding and the scikit-learn software package (sklearn)[36].

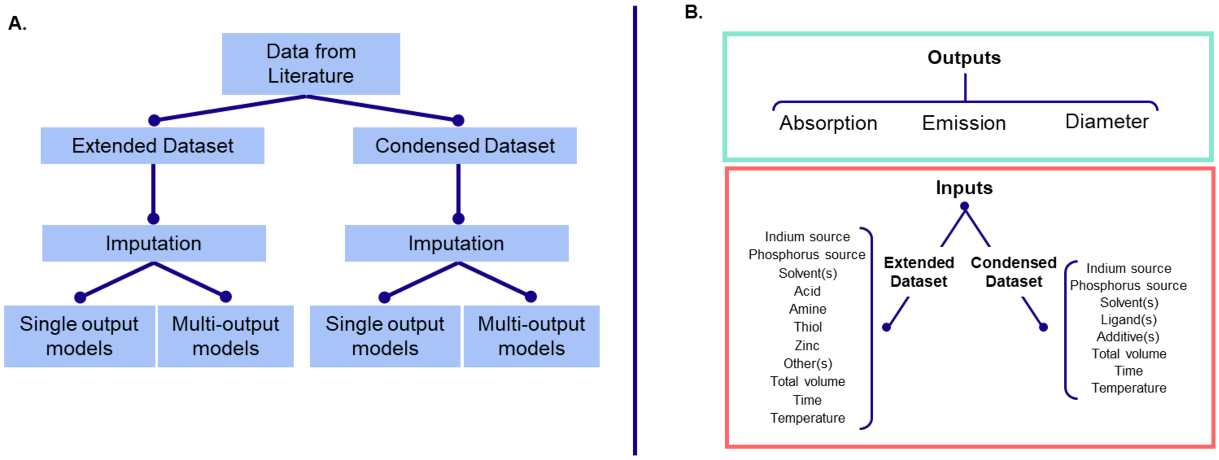
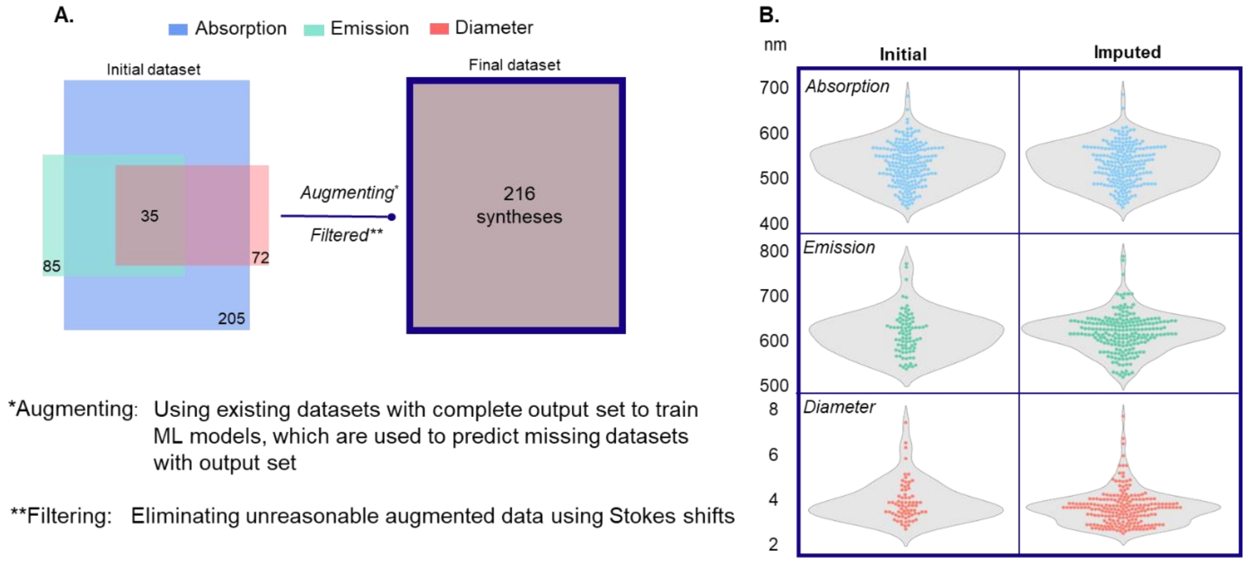


Figure 2. A. Workflow. **B.** Output and input feature selection.

2.3 Data Imputation

After defining the two datasets, we applied a data imputation process to both datasets. One of the biggest challenges when applying machine learning to materials chemistry is the lack of sufficient data. In our initial dataset, only 35 out of 219 syntheses had a complete set of output target values, because only a few articles reported all three targeted properties of InP QDs (**Figure 3** – left). To "fill in" the output target values, we performed a data imputation process. Data imputation, or imputing, is a technique used for filling in missing entries in the dataset, when values are not measured or reported[37,38]. This method is simple when only a small fraction of the output set is missing and when the missing values can be calculated or easily predicted.

In our study, since the three outputs are physically related to each other, i.e., optical properties in QDs are influenced by the size of the particles, which are in turn governed by synthetic parameters, we imputed the missing values by training a predictive model for each output feature, using the initial input set and the available output entries as training data. Since absorption was the most frequently reported output in the initial dataset (205 syntheses), data imputation was performed on absorption first, followed by emission, and finally diameter. Each imputative model was tuned by an exhaustive grid search to find the best parameters. (See details in *Supporting Information S2*). We then eliminated any syntheses that gave negative Stokes shift values, resulting in a final dataset of 216 syntheses, where excitonic absorption maxima ranged between 397 and 729 nm, band edge PL emission ranged between 470 and 775 nm, and diameters ranged between 1.5 and 8.3 nm (**Figure 3** – right).



*Augmenting: Using existing datasets with complete output set to train ML models, which are used to predict missing datasets with output set

**Filtering: Eliminating unreasonable augmented data using Stokes shifts

Figure 3. Data imputation process (left) and descriptions of the imputed dataset (right).

2.4 Machine Learning Models and Metrics

After filling in the missing output targets, we trained our datasets by both single- and multi-output regressors. Single-output models predict each target individually, and the features do not depend on each other. Multi-output models predict all output targets simultaneously, and the output targets depend on each other and on the inputs features[39]. We tested six regression algorithms suitable for small datasets: Extra Trees, Decision Tree, Random Forest, k-NN, Bagging, and Gradient Boosting using sklearn. To create representative samples for testing and training, we performed random sampling and stratified sampling methods for our datasets; and used Extra Trees and Decision Tree models to evaluate which train/test partitions give better pre-training performance (See *Supporting Information S4*). For the stratified sampling method, we sorted our data based on the values in the "emission_nm" column and put them into 6 'bins': [450, 500), [500, 550), [550, 600), [600, 650), [650, 700), and [700, 800). Then we sampled uniformly from each bin. Dividing this dataset by these bins avoids clustering of data since some specific QD sizes are more common synthetically than others. For all models, the datasets were split into 85% for training and 15% for testing. Results for a 70/30 train/test partition are also shown in the *Supporting Information S5*. We optimized the hyperparameters for each model using grid search. The final hyperparameters used for each model are listed in the *Supporting Information S11*. We used the mean absolute error (MAE), the coefficient of determination (R^2), and relative absolute errors (RAE) as metrics to assess the performance of all models. MAEs are sensitive to outliers since it is a linear score, in which all differences are weighted equally. Using MAEs also helps compare performances across datasets and models for three different output targets in a direct and intuitive manner. R^2 indicates the proportion of variance for a dependent variable determined by an independent variable. RAEs consider all errors equally important and provide informative metrics to non-experts in the field of QDs. For each model in this study, we report the MAE, R^2, and RAE of the predicted set versus the test set.

2.5 Syntheses

We conducted 8 new syntheses of InP QDs to test the prediction accuracy of our models. The experiments were designed based on four procedures found in the literature[40–43] with minor adjustments such that all reaction parameters were not already included as entries in the dataset used to train the machine learning models. The reaction parameters were also selected such that they were not easily extrapolated from the parent procedures. (See synthesis details in *Supplemental Information S7*).

3. RESULTS AND DISCUSSION

3.1 Data Description

After the data extraction process, the dataset contained 219 syntheses of InP QDs from 72 papers, with an average of 3 syntheses per paper. However, the dataset is biased towards hot-injection syntheses, with 71% of entries from this method. This bias reflects the most used technique to synthesize InP QDs found in the literature, since the hot-injection method has been proposed to assist the formation of monodisperse InP QDs due to rapid nucleation at elevated temperature[44]. Despite this bias, we also included comparable methods in the dataset to maximize the size and diversity of inputs in our dataset, even though every synthetic parameter (e.g., temperature ramp rate) could not be captured due to limited and inconsistent reporting. As can be seen, the most common In and P precursors were indium acetate, indium chloride, and $P(TMS)_3$ (**Figure 4**). The addition of zinc salts is known to increase the photoluminescence quantum yield and the stability of the InP QDs[45]; around 41% of the syntheses in the dataset include a Zn additive, with $ZnCl_2$ being the most common. The reaction temperatures ranged from 130 to 310 °C, in which the lowest temperatures correspond to reactions using chloroindium oxalate, and the highest temperatures correspond to reactions using indium tris(N,N'-diisopropylacetamidinato), indium trifluoroacetate, indium oxalate, indium palmitate, and indium myristate. Across the dataset, the reaction times were concentrated below 1 hour, which is related to the widespread use of the hot-injection method. In contrast, the heat-up procedure requires much longer reaction times, due to progressive heating and typically lower precursor reactivity, resulting in long supersaturation times[46].

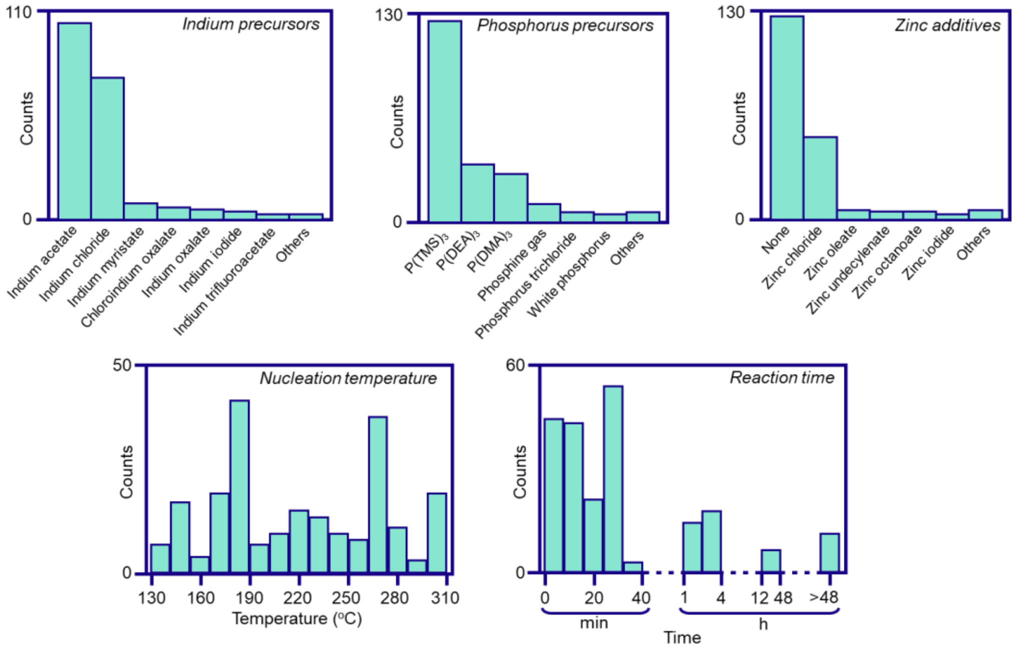
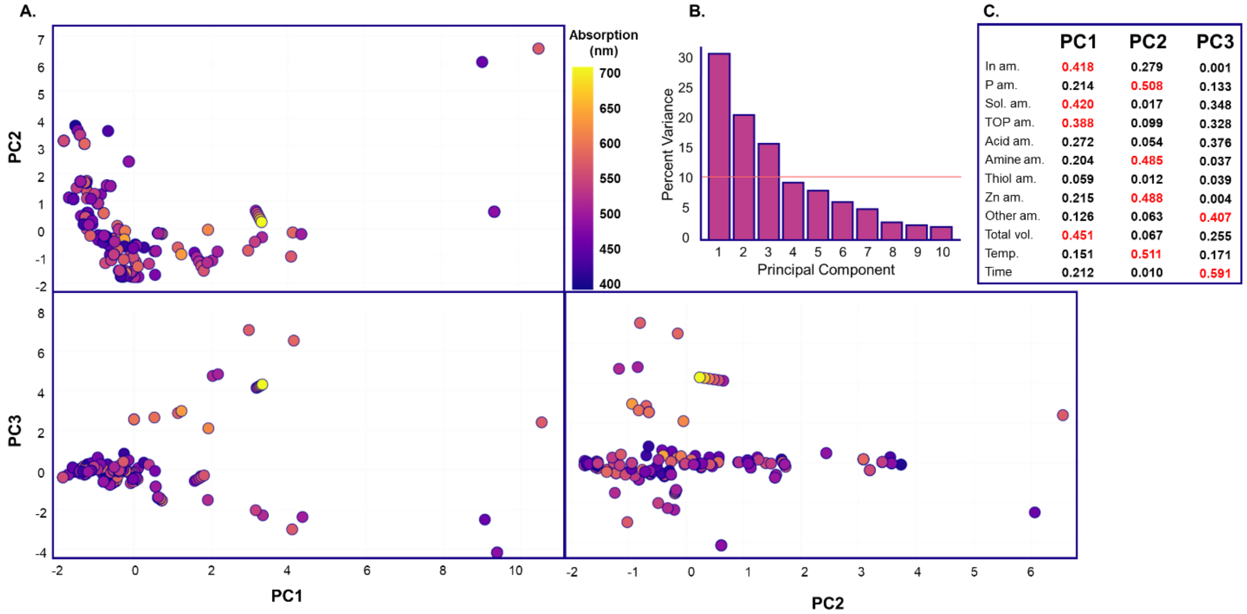


Figure 4. Description of the input set. Histograms of indium, phosphorus precursors, zinc additives, nucleation temperature, and reaction time of the syntheses in the initial dataset.

Principal component analysis (PCA, **Figure 5A**) was performed using continuous features in the extended dataset (In amount, P amount, etc.). The scree plot (**Figure 5B**) identifies the three directions (PC1, PC2, and PC3 capturing 29.68%, 10.01%, and 15.95% of the dataset respectively) along which the data have the largest spread. The features contributing the most to PC1 are total volume of reaction, solvent amount, and Indium precursor amount; while reaction temperature and phosphorus precursor amount contribute the most to PC2, and reaction time contributes the most to PC3. When shown as coefficients of PC1 vs PC2 and PC1 vs PC3, no clear relationship between these PCs and absorption wavelengths is observed. The spanning of syntheses along PC1 indicates that most InP QD syntheses in the dataset were conducted on a similar scale, while there are three syntheses that have a significantly larger scale than the rest of the dataset. The spanning of syntheses along PC3 indicates that most of the syntheses in the dataset were run for a similar duration, owing to their use of the hot injection method (typically less than 1h), while a portion of the syntheses were run for a much longer time.



	PC1	PC2	PC3
In am.	0.418	0.279	0.001
P am.	0.214	0.508	0.133
Sol. am.	0.420	0.017	0.348
TOP am.	0.388	0.099	0.328
Acid am.	0.272	0.054	0.376
Amine am.	0.204	0.485	0.037
Thiol am.	0.059	0.012	0.039
Zn am.	0.215	0.488	0.004
Other am.	0.126	0.063	0.407
Total vol.	0.451	0.067	0.255
Temp.	0.151	0.511	0.171
Time	0.212	0.010	0.591

Figure 5. Principal component analysis (PCA) of the extended dataset before imputation, where continuous features are considered. **A.** PCA biplots with syntheses plotted in two dimensions using their projections onto the first three principal components (PC). The syntheses are colored according to the absorption wavelength (nm) of the synthetic outcomes. **B.** Scree plot indicating the variance of the PCs when PCA is applied to the dataset. **C.** Table of loadings shows which features contribute the most to the PCs.

The plot of absorption peak versus the emission peak from the datasets before and after imputation (**Figure 6A**) suggests a linear relationship between these two output targets. **Figure 6B** displays the dependence of the Stokes shift on the first excitonic absorption peak. Our observation from the dataset before imputation (e.g., only reported values) agrees with the size-dependent behavior of Stokes shifts in InP QDs in that Stokes shift increases as QD size decreases or as absorption peak energy increases[47].

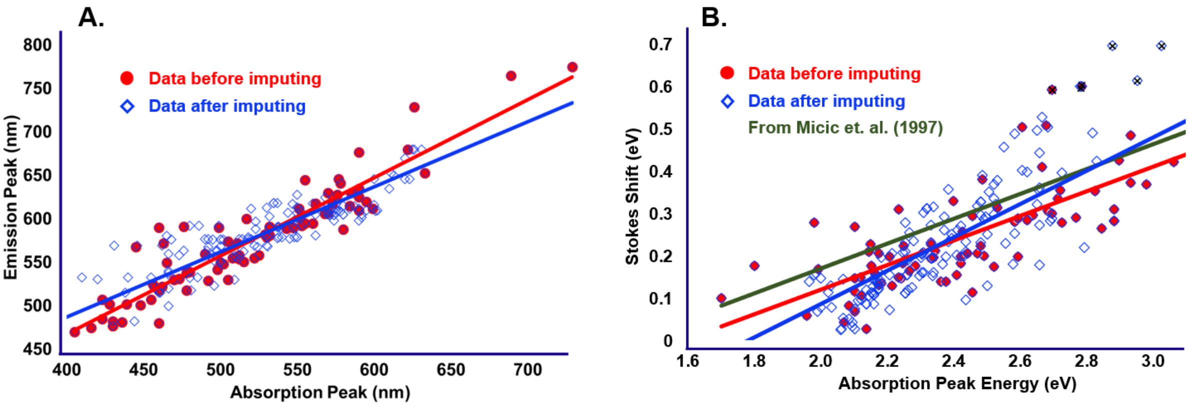


Figure 6. Plots and best fit lines after excluding outliers (indicated by black crosses) for datasets before and after imputation. **A.** Emission peak versus the first excitonic absorption peak in nm and **B.** First excitonic absorption energy peak versus Stokes shift (energy difference between emission peak and absorption peak). Green line is from Micic et. al.

In **Figure 7**, we plotted the band gap energy versus InP QD size and generated sizing curves from both the initial dataset and the imputed dataset. In the small particle size range (2 – 4 nm), both curves fit reasonably well with the empirical sizing curve developed by Micic et. al.[5], the calculated sizing curves from Cho et. al.[48] and Baskoutas & Terzis[49], but the curve after imputing deviates in the larger size range The inverse square fitted sizing curve for the data before imputing is

$$\boldsymbol{E_0 = 1.69 + \frac{1}{0.206\, d^{\,2}}} \qquad (1)$$

and the curve for the data after imputing is

$$\boldsymbol{E_0 = 1.94 + \frac{1}{0.343\, d^{\,2}}} \qquad (2)$$

where E_0 is the band gap in eV and d is the QD diameter in nm. The most likely reason for the deviation in the curves at larger sizes is the limited available data for high quality InP QDs larger than 4 nm. However, challenges associated with sample size polydispersity and surface oxidation may also be convoluting the reported data. It is also interesting to see the difference in the sizing curves before and after imputing in comparison to the empirically derived sizing curve from Micic. The imputed dataset seems to fit better than the raw dataset in the smallest size regime, but imputation appears to overestimate the band gap in the 2.5-3.5 nm regime in some cases leading to a higher degree of curvature in the line of best fit.

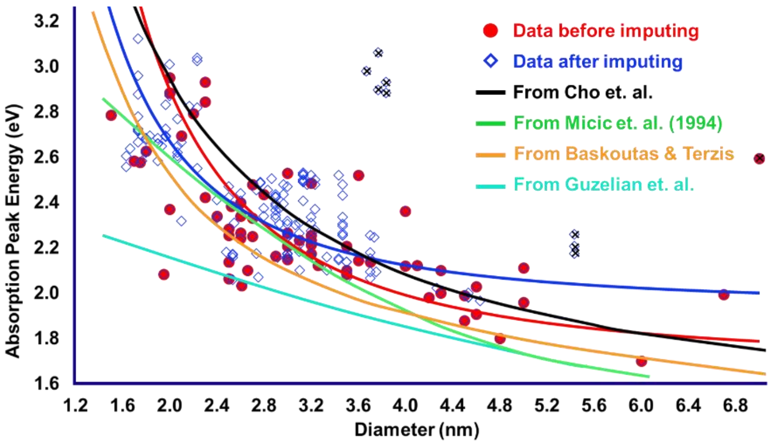


Figure 7. Band gap energy (eV) versus InP QD diameter determined by TEM. Red circles represent data points from the initial dataset, blue diamonds represent data points after the imputation step, the red and blue lines represent the sizing curves for InP QDs using dataset before and after imputation respectively after excluding outliers (indicated by black crosses). The cyan and green lines represent empirical sizing curves developed by Guzelian et. al.[50] and Micic et. al.[5], respectively. The black line represents the sizing curve calculated using density functional theory by Cho et. al.[48] The orange line represents the sizing curved calculated by the potential-morphing method on effective mass approximation by Basloutas & Terzis[49].

3.2 Model Performance

Figure 8 shows the performance of the best model for each output in each study case. Performance data for all other models are listed in *Supporting Information S5*. For most cases, the Extra Trees algorithm outperformed other algorithms. Considering that the datasets in this study are small, unbalanced, and contain noise, randomized tree-based algorithms such as Extra Trees would be expected to perform better than other methods, such as single decision tree or boosting algorithms. Extra Trees algorithm uses the entire set of learning entries to develop the tree and the decision rule is selected randomly, therefore bias in the datasets is minimized[51].

Among the three output targets, predictions of emission were the best, followed by absorption, and finally diameter. The differences in predictions among different synthetic outcomes might be attributed to the correlation between the reported outcome values and reported synthetic conditions. Emission and absorption peaks are often used to monitor QD reactions, while particle diameter, determined by TEM, must be done many hours after the synthesis finishes and most often following purification. Further, the size measurements are usually done manually without established best practices in the community. Therefore, data on particle size is not consistent and hence, more prone to poor correlations with synthetic conditions, leading to poor predictions when synthetic conditions are the descriptors.

Although multi-output models were expected to give better predictions due to the strong correlation between the three output targets (See *Supporting Information Section S3* for Pearson correlations), single-output models showed better performance for both the condensed and the extended datasets, indicating that assuming a relationship among the output targets did not improve, but worsened the predictions. While there seems to be a linear relationship between emission and absorption wavelength (**Figure 6A**), the datasets failed to reflect the expected relationship between particle size and absorption. Thus, for these datasets, using different model selections for each output, would give a better prediction performance.

When comparing the performance of the two descriptor sets using the models that gave the lowest MAEs, models that used the condensed dataset were expected to show better performance since they have lower dimensionality from manually combining some input features while retaining the same information. **Figure 8** shows that models using the condensed dataset gave better predictions for absorption wavelength and diameter than models using the extended dataset, while emission prediction accuracy seemed to be similar in both cases. However, for the single-output Decision Tree model for diameter that used the condensed dataset, we saw that many predictions were centered around 3.3 nm for the observed range of 3 – 4 nm (**Figure 8A** – right). This might be caused by the low complexity of the model and/or oversimplification of the descriptor set, leading to inaccurate predictions when a few input features have significantly higher influence on the model than others. When the Decision Tree model is applied to the extended dataset with similar complexity, this behavior seemed to be eliminated (**Figure 7C** – right). Thus, we note that to improve ML model performance for QD synthesis, oversimplification of feature engineering may directly affect the prediction accuracy.

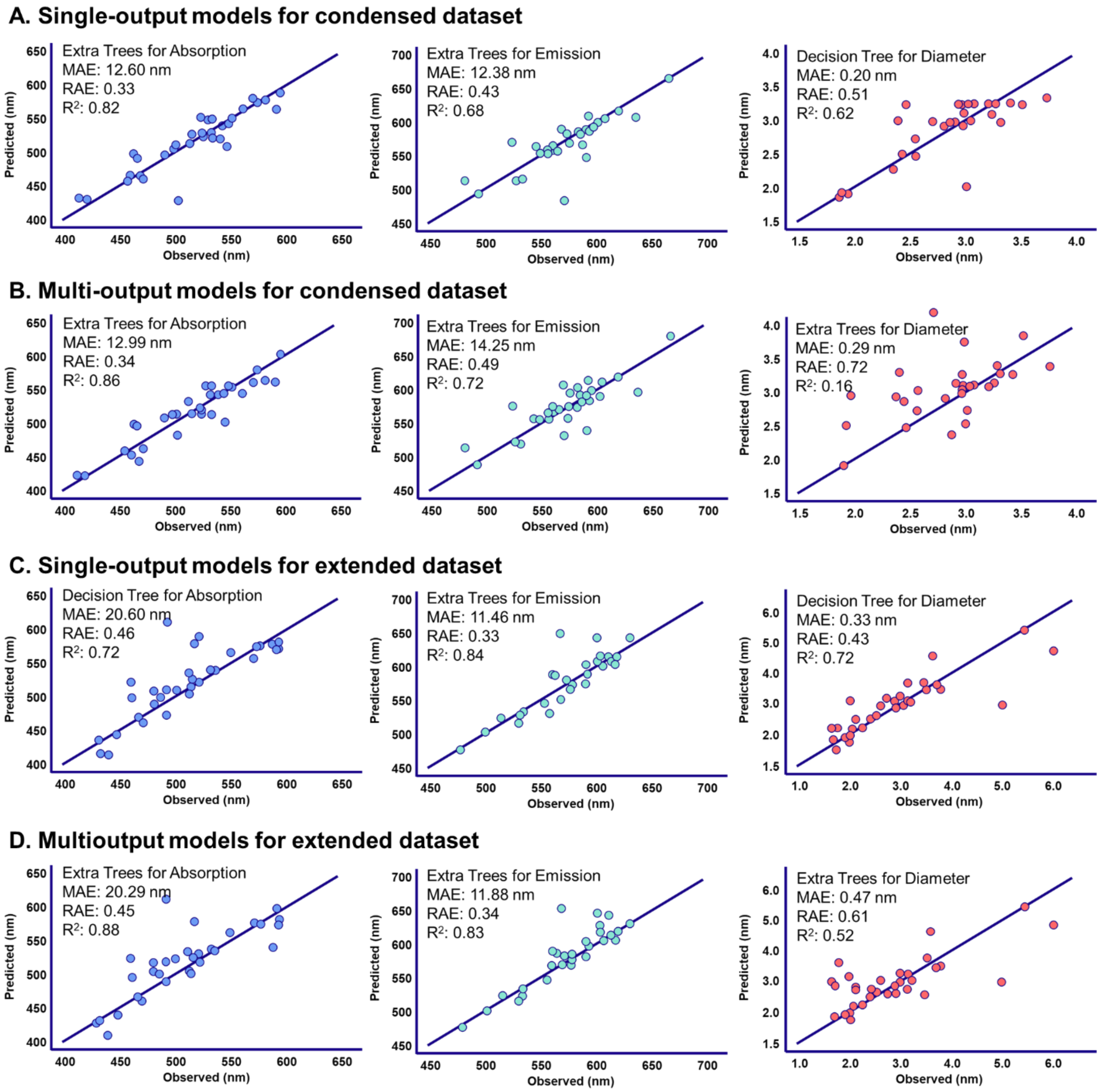


Figure 8. Parity plots of observed vs predicted values and the performance of single-output and multi-ouput models for the three outputs using the condensed and extended datasets.

3.3 Validation

Applying a complex algorithm to a small dataset can result in significant overfitting that leads to misleading predictions. Here, we used different methods, statistically and experimentally, to detect overfitting and test the accuracy of our ML models.

3.3.1 Stratified k-Fold Validation

We first used the stratified k-fold validation method on both the condensed dataset and the extended dataset to justify the accuracy of our ML models. The data points were divided into 5 groups based on their emission wavelength output to ensure that test sets are uniformly sampled across the dataset (**Figure 9**).

Then, a stratified test/train split of the dataset was performed to achieve the ratio of 15/85, consistent with the ratio used in this study. For the 4 cases, we applied the same ML algorithms as shown in **Figure 8** and evaluated their performance by MAEs. **Figure 10** indicates that the accuracy of all models was consistent over 5 iterations, and no considerable overfitting was observed. It should be noted that the hyperparameters used in the models for this validation step were adopted from the models in **Figure 8**, which led to a slightly higher MAEs when compared to the values in **Figure 8**.

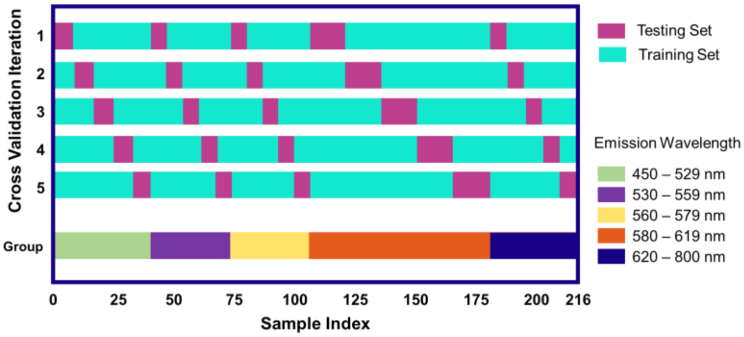


Figure 9. Visualization of the stratified k-fold validation in this study.

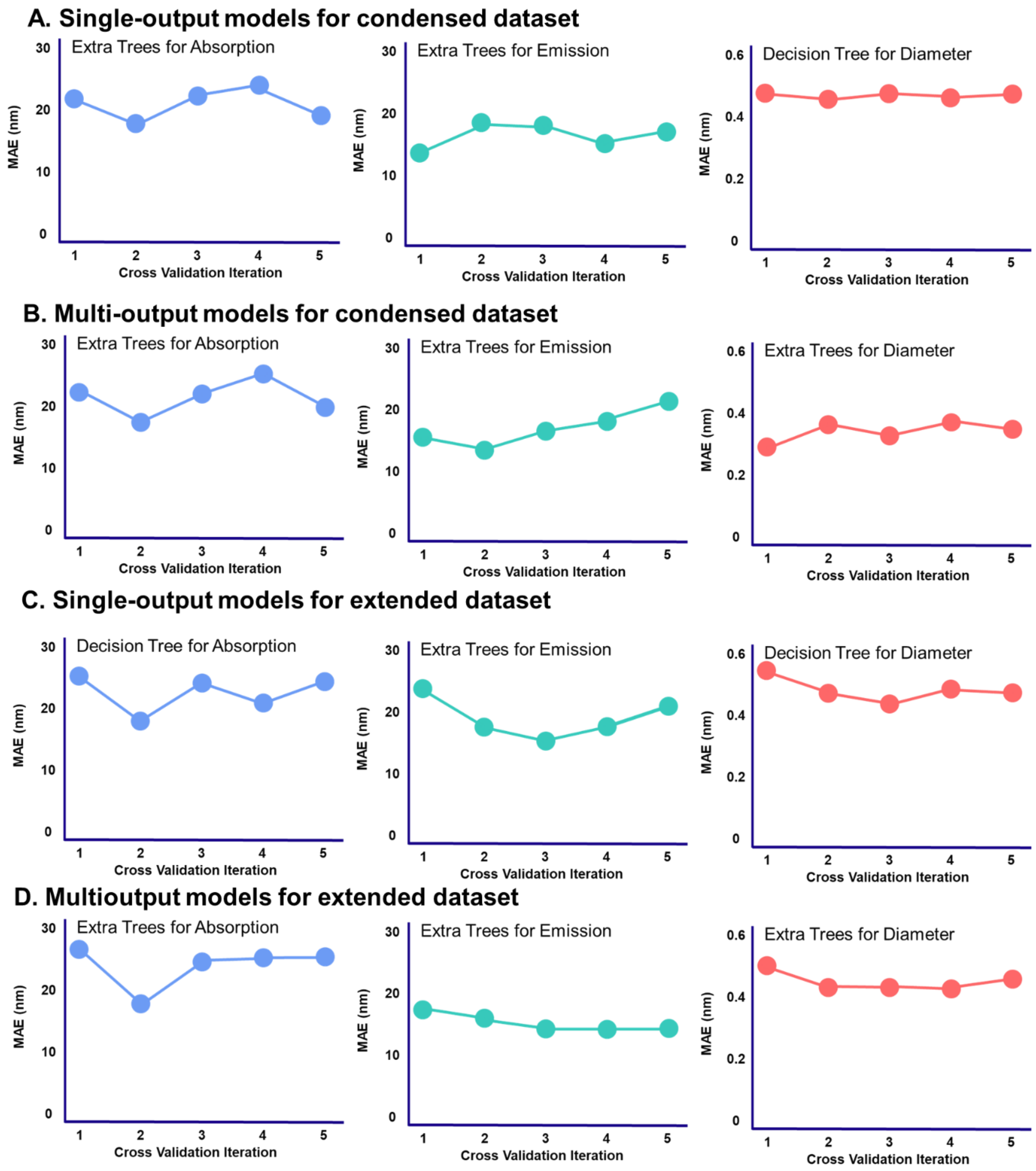


Figure 10. Mean absolute errors over 5 iterations of stratified k-fold validation.

3.3.2 Comparison with Non-imputed Models

Next, we trained and tested models with the initial or non-imputed datasets. Details on data processing and ML training on these datasets are shown in the *Supplemental Information Section S6*. Due to the small dataset size (205 datapoints for absorption, 85 datapoints for emission, and 72 datapoints for diameter), predictions using the non-imputed datasets gave higher errors especially for emission and diameter targets (**Figure 11**). This result indicates that it is necessary and reliable to effectively impute the missing data to improve the performance of predictive models for QD synthesis when the available datasets are limited.

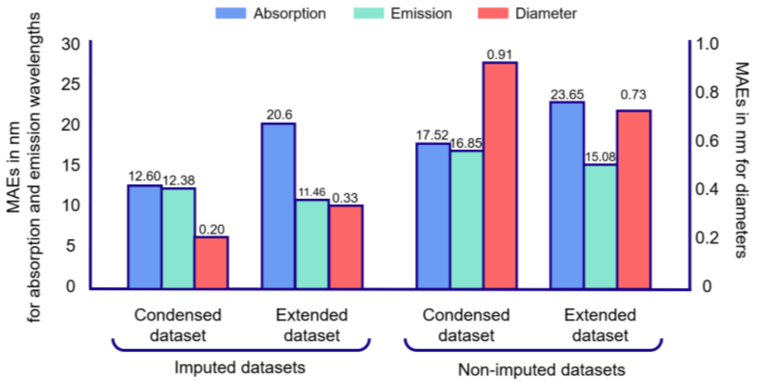


Figure 11. Performance comparison of models using the imputed datasets and the non-imputed datasets.

3.3.3 Comparison with Experimental Data

To further test the practical accuracy of the models, we conducted a series of 8 InP QD syntheses. The synthetic procedures were designed by varying the reaction conditions of existing syntheses of InP QDs found in the literature, such that they would not be entries in the initial dataset, and not easily extrapolated from the original reports (*Section S7*). The QDs from each synthesis were characterized by UV-Vis and photoluminescence spectroscopy, and the particle sizes were determined by TEM analysis. Only 5 out of 8 batches of InP QDs showed strong luminescence because as-prepared InP NCs generally exhibit poor luminescence due to non-radiative channels originating from surface states. The parity plots in **Figure 12** showed that the models correctly predicted the actual synthetic outcomes in many cases. While predictions of experimental absorption and particle size had similar accuracy as the test sets, MAEs for emission predictions were high because there were only 5 datapoints for emission and MAEs are sensitive to large errors. It should be noted that models using the extended dataset had a much better performance than the models using the condensed dataset for prediction of particle size.

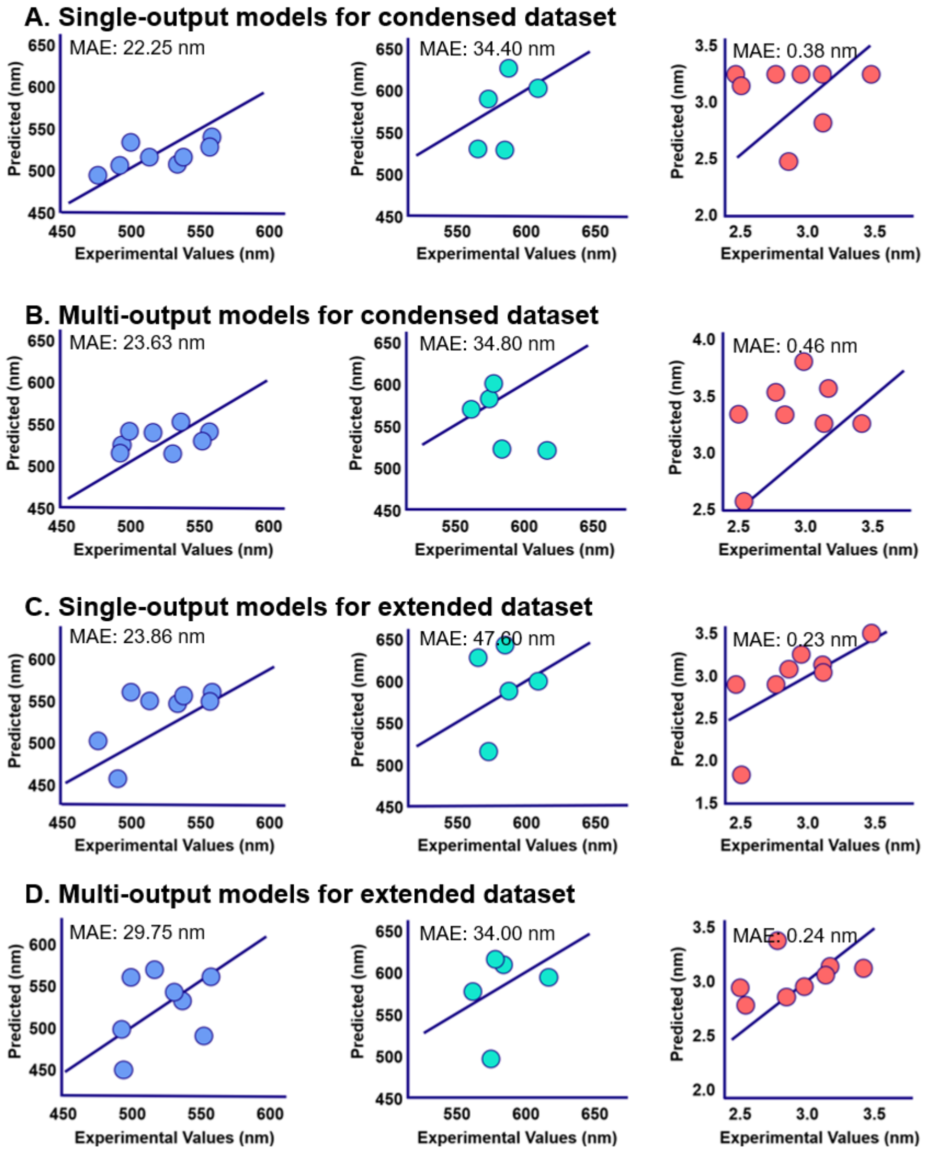


Figure 12. Parity plots of experimental values vs. predicted values from the ML models.

3.4 Interactive User Interface

To allow external users, including researchers with no background in machine learning, to use our model to predict InP QDs synthesis outcomes and explore new synthetic methods, we deployed a user interface using an open-source Python library provided by Streamlit[52]. Streamlit is a framework for building interactive web applications with user-friendly components such as buttons, sliders, and plots. From the best ML models in this study, we deployed a Streamlit web app that enabled real time reaction analysis and prediction https://share.streamlit.io/cossairt-lab/indium-phosphide/Hot_injection/hot_injection_prediction.py. The web app includes sections where users answer questions about QD synthetic conditions to get a prediction of diameter, emission, and the first excitonic absorption peak with a prediction interval as uncertainty. We anticipate that this webapp will enable more chemical insights into InP synthesis from machine learning. Although the best models from this study were used, inaccurate predictions i.e., absorption wavelength higher than emission wavelength, can sometimes be seen from the webapp due to inconsistency and low synthesis variety in the dataset. We expect the performance of the webapp to improve when a larger dataset becomes available.

3.5 Synthetic Insights

Using the best model for our four study cases, we calculated the feature importance from each model. Feature importance reflects the extent to which a variable is used for accurate predictions (i.e., the more a model uses a variable, the more important it is). Specifically in the case of Extra Trees and Decision Tree algorithms, feature importance is computed as the normalized total reduction of the criterion brought by that feature, which is also known as the Gini importance. As expected, temperature and time were found to be most important in all cases as they directly nucleation and growth kinetics. Interestingly, the presence of zinc additives also plays an important role (**Figure 13**) consistent with the reported observations of spectral shifts and size changes when a zinc salt is present in the synthesis[45,53].

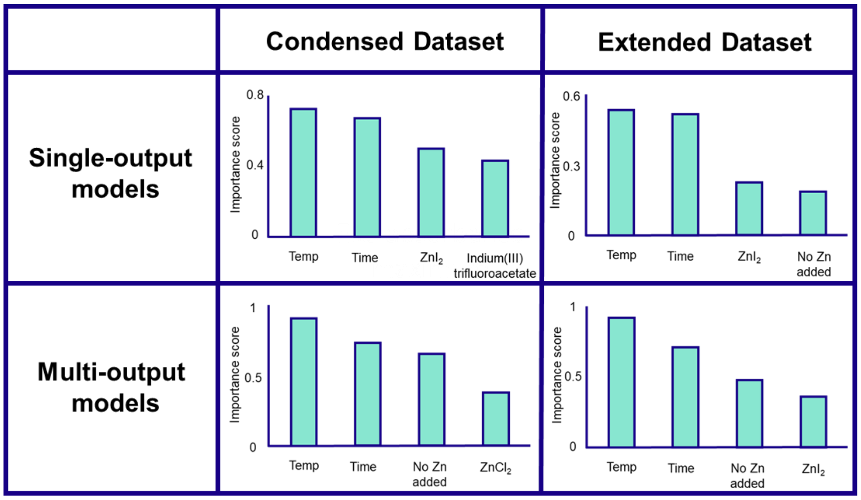


Figure 13. Feature importance charts for the best model in each study case.

As discussed in the above section, the webapp allows us to explore the chemical intuition of our algorithms beyond basic statistical metrics and discover synthetic trends without conducting actual experiments. For example, predicted outcomes from the web-app suggested that for a typical hot-injection synthesis where $InCl_3$ reacts with tris(diethylamino)phosphine, the presence of TOP redshifts the emission and absorption maxima, while the presence of a zinc halide salt results in spectral blueshifts (**Figure 14**). These observations are consistent with the reported literature[45,54].

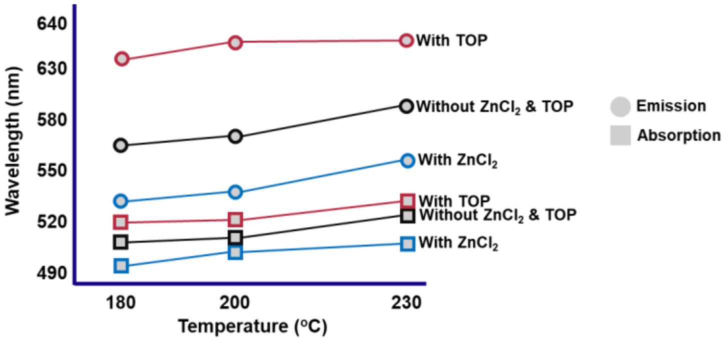


Figure 14. Predicted emission (circles) and absorption (squares) wavelengths from the Streamlit webapp using single-output algorithms and the condensed dataset with all methods. Reaction conditions include 0.1 mmol of $InCl_3$, 1 mL of oleylamine, 0.15 mmol of $P(DEA)_3$, nucleation temperature at 180 ^oC, reaction time of 2 min, with 0.3 mmol $ZnCl_2$ (blue outlines), or with 0.2 mL TOP (pink outlines), or without both $ZnCl_2$ and TOP (black outlines).

3.6 Limitations

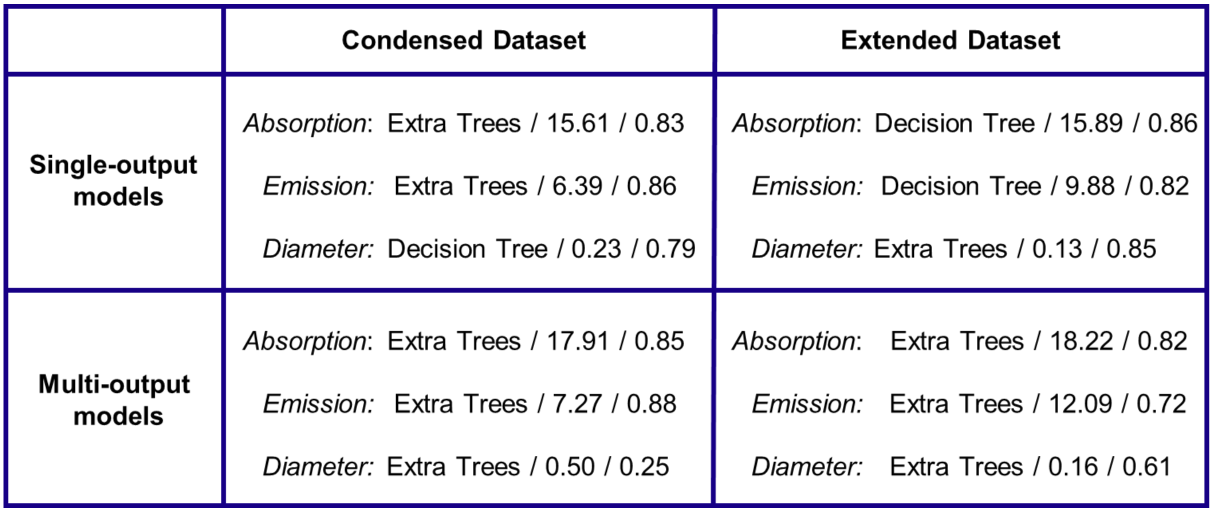
Despite their accuracy, there are several inevitable limitations of the ML models that arise from the available data and the nature of QD synthesis. The novelty of this study is based on the collection, imputation, and ML training of reported data in the literature, however the presence of unknown or unreported contaminants, and the related lack of standardization in reporting data in the literature heavily influences these ML results. As shown in *Figure S1* and discussed in Section 2.1, many publications did not report details in the synthesis, which significantly decreases the size of the dataset despite the large number of published reports of InP QD synthesis in the past decades. Other synthetic parameters that also affect the synthetic outcomes but are often not mentioned include injection rate[55], solvents and precursors purity[56,57], and QD purification status and methods[58–60]. Moreover, heterogeneity of experimental condition in different labs also impacts the synthetic outcomes. For example, it has been shown that presence of trace water can affect the size of InP QDs[61,62]. Another inconsistency in reporting synthetic results comes from uncertainties associated with using TEM to determine particle size and size distribution. As discussed in detail by Pyrz and Buttrey[63], many decisions during image acquisition and size determination can lead to over- and under-estimation of particle size, especially for smaller particles. Those decisions include optimization of measurement resolution, limiting electron beam damage, proper determination of particle boundaries, and reliable quantification of particle size distributions. Data diversification plays an important role in the performance of ML models. In this study, a diverse dataset that consists of many syntheses of a variety of particle size or emission peaks would help improve ML model performance. However, since synthesis of blue-emitting (<480 nm) or small InP QDs is still challenging, and current applications of InP often make use of QDs in the 2 – 4 nm size in range, the data inevitably concentrated around a small range of particle sizes, leading to less accurate predictions for synthesis of QDs outside of that range.

4. APPLICATIONS

4.1 Predicting InP QD Hot Injection Synthesis Outcomes

We applied the process of data preparation, data imputation, and ML training from this study to other datasets with similar size. First, we prepared new condensed and extended datasets that have only hot injection syntheses by filtering our initial datasets. These new datasets contained 157 syntheses. The results (**Table 1**) showed improvement in R^2 values for all outputs and lower MAEs for emission predictions but demonstrated modest differences in MAEs for diameter and absorption wavelength. Similar to the previous observation, models using the condensed dataset and single-output algorithms have better performance than models using the extended dataset and multi-output algorithms, respectively. It should be noted that single-output algorithms using the hot injection dataset could achieve MAEs as low as 0.13 nm for diameter and 6.39 nm for emission wavelength predictions. The algorithms were also able to identify temperature and time as the most influential parameters that affect the synthetic outcomes (*Figure S13*).

Table 1. Performance of the best algorithms using the hot injection dataset (Output: Model / MAE in nm / R^2)



	Condensed Dataset	Extended Dataset
Single-output models	*Absorption*: Extra Trees / 15.61 / 0.83 *Emission:* Extra Trees / 6.39 / 0.86 *Diameter:* Decision Tree / 0.23 / 0.79	*Absorption*: Decision Tree / 15.89 / 0.86 *Emission:* Decision Tree / 9.88 / 0.82 *Diameter:* Extra Trees / 0.13 / 0.85
Multi-output models	*Absorption*: Extra Trees / 17.91 / 0.85 *Emission:* Extra Trees / 7.27 / 0.88 *Diameter:* Extra Trees / 0.50 / 0.25	*Absorption*: Extra Trees / 18.22 / 0.82 *Emission:* Extra Trees / 12.09 / 0.72 *Diameter:* Extra Trees / 0.16 / 0.61

4.2 Predicting CdSe QD Hot Injection Synthesis Outcomes

To further evaluate the reliability and show the utility of the imputing method for small datasets, we revised and extended the CdSe QD dataset from Baum et. al.[25] to include absorption and emission wavelengths in the output set. The revised dataset contained 233 hot injection syntheses of CdSe QDs, in which absorption wavelength is absent in 38 syntheses (16%) and emission wavelength is absent in 77 syntheses (33%). The dataset preprocessing, data imputation, model tuning, model training, and user interface creation were done in the same manner of the InP study. For feature selection, we reduced the number of input features from 27 to 15 since models with fewer input variables typically give better performance[30] (details on feature selection can be found in *Section S10*). Compared to the InP models for the hot injection dataset, CdSe models showed better performance for all three output features, especially for diameter. This is likely a result of the original study's focus on diameter, whose values were not limited to TEM measurements, but were also calculated from absorption spectra. Further, a much smaller portion of the dataset was missing absorption and emission entries, perhaps reflecting the inherent poor emissivity of InP QDs, thus reducing prediction bias. Results from the hot injection models also showed that single-output models outperformed multi-output models with MAEs as low as 14.67, 8.37, and 0.18 nm for absorption wavelength, emission wavelength, and particle diameter, respectively. R^2 values for diameter from the Extra Trees and Decision Tree algorithms are comparable to the value from the reported Gradient Boosting Machine algorithm[25] (**Figure 15**). Examining feature importance in our study showed that reaction time and growth temperature are the most influential factors in the synthesis of CdSe. This is consistent with the Gradient Boosting Machine model from Baum et. al., however, in this study the two most important variables have a significantly higher influence on the synthesis than other variables (**Figure 16**).

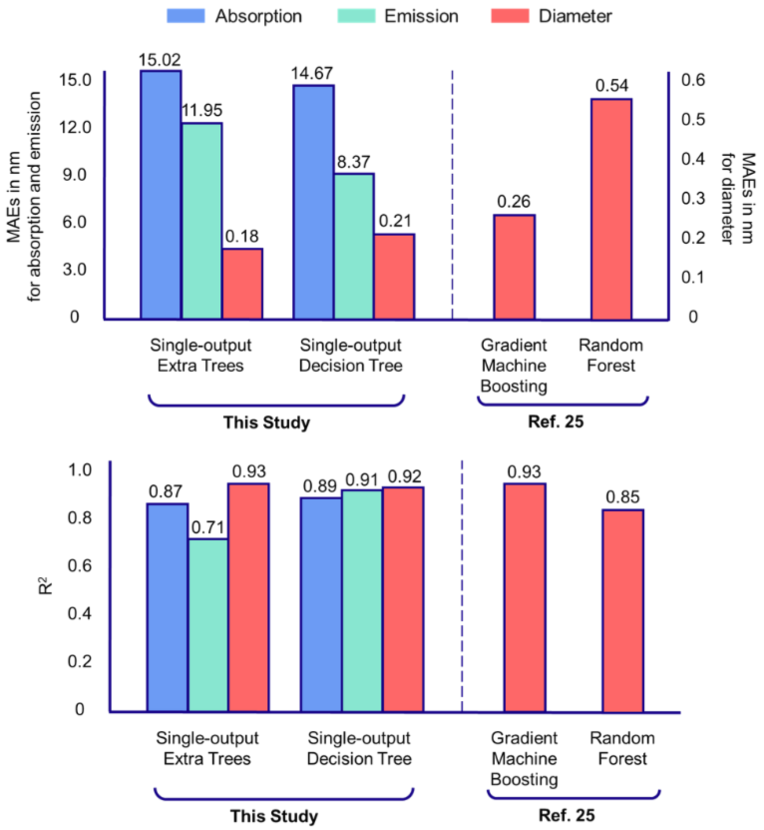


Figure 15. MAEs and R^2 values comparison of the two models between this study and ref 25.

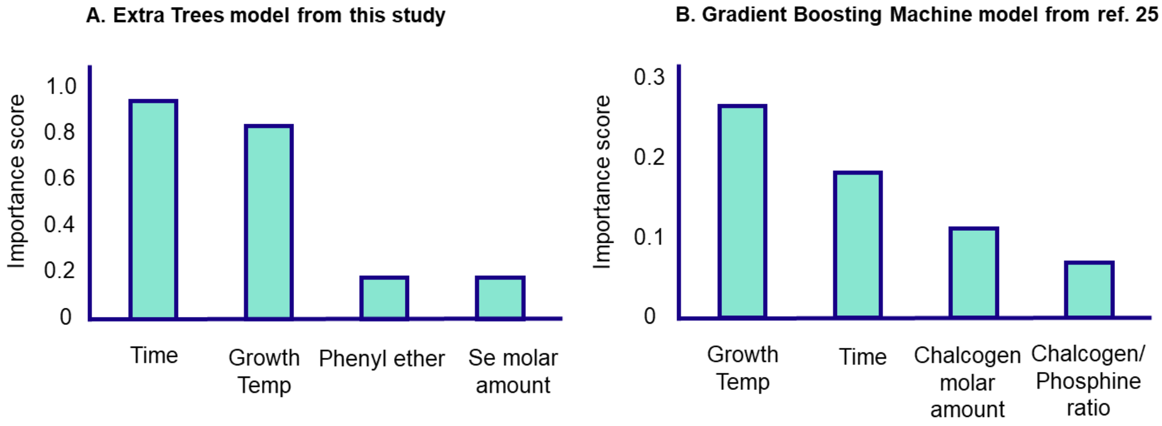
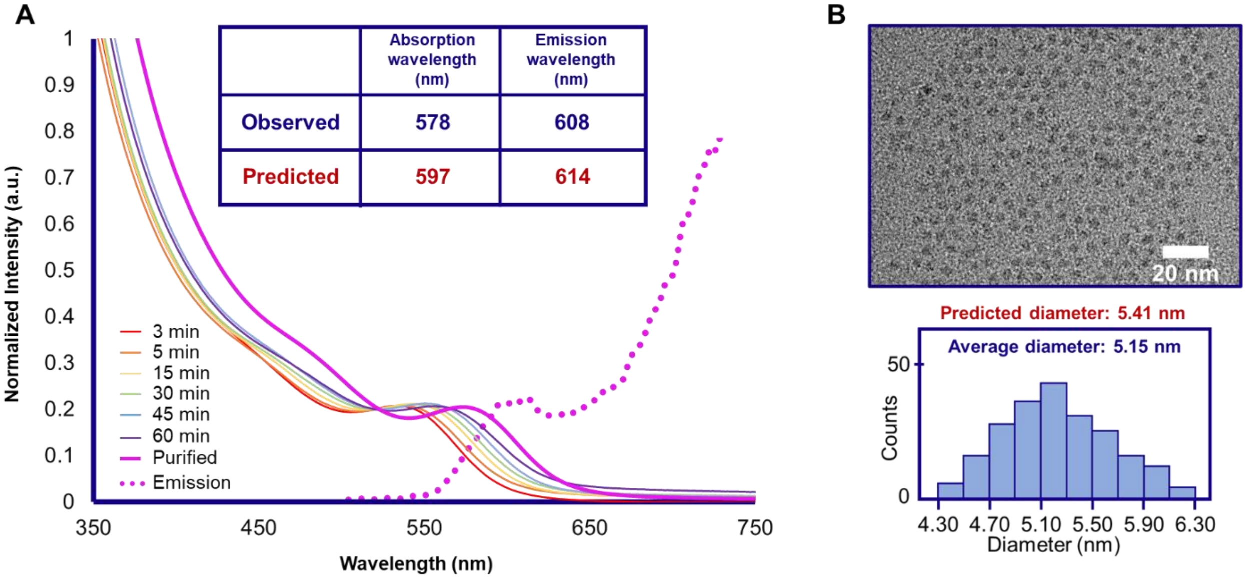


Figure 16. Feature importance charts of **A.** Extra Trees model from this study and **B.** Gradient Boosting Machine model from ref 25.

4.3 Inverse-Design Using the Streamlit User Interface

Finally, we targeted 600 nm–absorbing QDs using synthetic conditions and precursors from an existing procedure[64]. Using the Streamlit webapp, we entered the synthetic conditions from the procedure, modified chemicals to what were available to us, and adjusted the reaction temperature and time to achieve the desired synthetic outcome. We conducted the experiment and were able to synthesize InP QDs with desired optical properties with high accuracy (**Figure 17**). For absorption and emission wavelengths, we also found that there was a noticeable difference between samples before and after purification. This observation justifies our previous hypotheses that the inconsistency from reported values from the literature can strongly affect the accuracy of prediction, that our syntheses were a mix of purified and in situ data entries, and that there are many unreported factors that can also play a role in achieving precise optical properties.



	Absorption wavelength (nm)	Emission wavelength (nm)
Observed	578	608
Predicted	597	614

Figure 17. A. UV-Vis spectra of timed aliquots and emission spectrum of the purified product from the reaction using 0.40 mmol indium acetate, 1.45 mmol myristic acid, and 0.20 mmol $P(SiMe_3)_3$ injected at 315 °C. The nucleation temperature was 310 °C. **B.** A TEM image of the purified particles with an average diameter of 5.15 nm.

5. CONCLUSION

We have trained and used ML models to predict the properties of InP QDs based on synthetic conditions. The descriptor space was defined in two ways (condensed and extended) to study the best approach for predicting QD synthesis outcomes where the available data is limited. We tested single-output and multi-output ML algorithms and found that single-output models showed enhanced performance over the multi-output models despite the physical relationships among the output targets (diameter, absorption and emission wavelengths). The performance of the models was validated in different ways, including stratified k-fold validation, comparison with non-imputed datasets, and comparison with newly collected experimental data. From the model estimation errors, we found that reaction temperature, time, and the addition of zinc salts were the most influential synthetic parameters. The same dataset pre-processing, imputation, and ML training were applied to both InP and CdSe hot injection datasets, resulting in accurate predictions for these two cases. Furthermore, we deployed a web-app that employs our best algorithms so that external users can use them to predict InP and CdSe synthetic outcomes. Using this web app, we were able to test our models with newly adapted InP syntheses that targeted and achieved desired optical properties. The webapps also allowed us to investigate the limitations of the ML approach in this study. Because the algorithms cannot recognize new precursors, reaction conditions need to be closely based on existing procedures to obtain accurate predictions. Overall, this work provides a procedure to preprocess datasets, train ML models, and implement models for public users in the field of nanocrystal synthesis, especially where available datasets are small and incomplete.

Supporting Information

Electronic supplementary information (ESI) available: Additional details for data acquisition, data imputation, Pearson correlation, datasets, code files, machine learning modeling and experimental methods. See DOI: XXXXXX.

Codes and datasets are also available at https://github.com/Cossairt-Lab/Indium-Phosphide

This article is available as a preprint: Nguyen, H.; Dou, F.; Park, N.; Wu, S.; Sarsito, H.; Diakubama, B.; Larson, H.; Nishiwaki, E.; Homer, M.; Cash, M.; Cossairt, B. Predicting Indium Phosphide Quantum Dot Properties from Synthetic Procedures Using Machine Learning. ChemRxiv, 2022, https://chemrxiv.org/engage/chemrxiv/article-details/621d2ec8011b5834cbb8ad0d.

Corresponding Author

*cossairt@uw.edu

Funding Sources

National Science Foundation OMA-1936100 and DMR-2019444.

Acknowledgements

This material is based upon work supported by the National Science Foundation under award number OMA-1936100 and DMR-2019444. Part of this work was conducted at the Molecular Analysis Facility, a National Nanotechnology Coordinated Infrastructure site at the University of Washington which is supported in part by the National Science Foundation (grant NNCI-1542101), the University of Washington, the Molecular Engineering & Sciences Institute, and the Clean Energy Institute.

References

(1) Eren, G. O.; Sadeghi, S.; Bahmani Jalali, H.; Ritter, M.; Han, M.; Baylam, I.; Melikov, R.; Onal, A.; Oz, F.; Sahin, M.; Ow-Yang, C. W.; Sennaroglu, A.; Lechner, R. T.; Nizamoglu, S. Cadmium-Free and Efficient Type-II InP/ZnO/ZnS Quantum Dots and Their Application for LEDs. *ACS Appl. Mater. Interfaces* **2021**, *13* (27), 32022–32030. https://doi.org/10.1021/acsami.1c08118.

(2) Sadeghi, S.; Bahmani Jalali, H.; Melikov, R.; Ganesh Kumar, B.; Mohammadi Aria, M.; Ow-Yang, C. W.; Nizamoglu, S. Stokes-Shift-Engineered Indium Phosphide Quantum Dots for Efficient Luminescent Solar Concentrators. *ACS Appl. Mater. Interfaces* **2018**, *10* (15), 12975–12982. https://doi.org/10.1021/acsami.7b19144.

(3) Saeboe, A. M.; Nikiforov, A. Yu.; Toufanian, R.; Kays, J. C.; Chern, M.; Casas, J. P.; Han, K.; Piryatinski, A.; Jones, D.; Dennis, A. M. Extending the Near-Infrared Emission Range of Indium Phosphide Quantum Dots for Multiplexed In Vivo Imaging. *Nano Lett.* **2021**, *21* (7), 3271–3279. https://doi.org/10.1021/acs.nanolett.1c00600.

(4) Kim, Y.; Chang, J. H.; Choi, H.; Kim, Y.-H.; Bae, W. K.; Jeong, S. III–V Colloidal Nanocrystals: Control of Covalent Surfaces. *Chem. Sci.* **2020**, *11* (4), 913–922. https://doi.org/10.1039/C9SC04290C.

(5) Micic, O. I.; Curtis, C. J.; Jones, K. M.; Sprague, J. R.; Nozik, A. J. Synthesis and Characterization of InP Quantum Dots. *J. Phys. Chem.* **1994**, *98* (19), 4966–4969. https://doi.org/10.1021/j100070a004.

(6) Battaglia, D.; Peng, X. Formation of High Quality InP and InAs Nanocrystals in a Noncoordinating Solvent. *Nano Lett.* **2002**, *2* (9), 1027–1030. https://doi.org/10.1021/nl025687v.

(7) Gary, D. C.; Flowers, S. E.; Kaminsky, W.; Petrone, A.; Li, X.; Cossairt, B. M. Single-Crystal and Electronic Structure of a 1.3 Nm Indium Phosphide Nanocluster. *J. Am. Chem. Soc.* **2016**, *138* (5), 1510–1513. https://doi.org/10.1021/jacs.5b13214.

(8) 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. https://doi.org/10.1021/acs.chemmater.6b03408.

(9) Gerbec, J. A.; Magana, D.; Washington, A.; Strouse, G. F. Microwave-Enhanced Reaction Rates for Nanoparticle Synthesis. *J. Am. Chem. Soc.* **2005**, *127* (45), 15791–15800. https://doi.org/10.1021/ja052463g.

(10) Tessier, M. D.; Dupont, D.; De Nolf, K.; De Roo, J.; Hens, Z. Economic and Size-Tunable Synthesis of InP/ZnE (E = S, Se) Colloidal Quantum Dots. *Chem. Mater.* **2015**, *27* (13), 4893–4898. https://doi.org/10.1021/acs.chemmater.5b02138.

(11) Harris, D. K.; Bawendi, M. G. Improved Precursor Chemistry for the Synthesis of III–V Quantum Dots. *J. Am. Chem. Soc.* **2012**, *134* (50), 20211–20213. https://doi.org/10.1021/ja309863n.

(12) Vinokurov, A. A.; Dorofeev, S. G.; Znamenkov, K. O.; Panfilova, A. V.; Kuznetsova, T. A. Synthesis of InP Quantum Dots in Dodecylamine from Phosphine and Indium(III) Chloride. *Mendeleev Commun.* **2010**, *20* (1), 31–32. https://doi.org/10.1016/j.mencom.2010.01.012.

(13) Bang, E.; Choi, Y.; Cho, J.; Suh, Y.-H.; Ban, H. W.; Son, J. S.; Park, J. Large-Scale Synthesis of Highly Luminescent InP@ZnS Quantum Dots Using Elemental Phosphorus Precursor. *Chem. Mater.* **2017**, *29* (10), 4236–4243. https://doi.org/10.1021/acs.chemmater.7b00254.

(14) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *Npj Comput. Mater.* **2019**, *5* (1), 83. https://doi.org/10.1038/s41524-019-0221-0.

(15) Jensen, Z.; Kim, E.; Kwon, S.; Gani, T. Z. H.; Román-Leshkov, Y.; Moliner, M.; Corma, A.; Olivetti, E. A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. *ACS Cent. Sci.* **2019**, *5* (5), 892–899. https://doi.org/10.1021/acscentsci.9b00193.

(16) Mukaddem, K. T.; Beard, E. J.; Yildirim, B.; Cole, J. M. ImageDataExtractor: A Tool To Extract and Quantify Data from Microscopy Images. *J. Chem. Inf. Model.* **2020**, *60* (5), 2492–2509. https://doi.org/10.1021/acs.jcim.9b00734.

(17) Li, Z.; Najeeb, M. A.; Alves, L.; Sherman, A. Z.; Shekar, V.; Cruz Parrilla, P.; Pendleton, I. M.; Wang, W.; Nega, P. W.; Zeller, M.; Schrier, J.; Norquist, A. J.; Chan, E. M. Robot-Accelerated Perovskite Investigation and Discovery. *Chem. Mater.* **2020**, *32* (13), 5650–5663. https://doi.org/10.1021/acs.chemmater.0c01153.

(18) Masood, H.; Toe, C. Y.; Teoh, W. Y.; Sethu, V.; Amal, R. Machine Learning for Accelerated Discovery of Solar Photocatalysts. *ACS Catal.* **2019**, *9* (12), 11774–11787. https://doi.org/10.1021/acscatal.9b02531.

(19) Vasylenko, A.; Gamon, J.; Duff, B. B.; Gusev, V. V.; Daniels, L. M.; Zanella, M.; Shin, J. F.; Sharp, P. M.; Morscher, A.; Chen, R.; Neale, A. R.; Hardwick, L. J.; Claridge, J. B.; Blanc, F.; Gaultois, M. W.; Dyer, M. S.; Rosseinsky, M. J. Element Selection for Crystalline Inorganic Solid Discovery Guided by Unsupervised Machine Learning of Experimentally Explored Chemistry. *Nat. Commun.* **2021**, *12* (1), 5561. https://doi.org/10.1038/s41467-021-25343-7.

(20) 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), 2001626. https://doi.org/10.1002/adma.202001626.

(21) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3* (12), 1337–1344. https://doi.org/10.1021/acscentsci.7b00492.

(22) Kim, J. Y.; Steeves, A. H.; Kulik, H. J. Harnessing Organic Ligand Libraries for First-Principles Inorganic Discovery: Indium Phosphide Quantum Dot Precursor Design Strategies. *Chem. Mater.* **2017**, *29* (8), 3632–3643. https://doi.org/10.1021/acs.chemmater.7b00472.

(23) Meng, F.; Li, Y.; Wang, D. Predicting Atomic-Level Reaction Mechanisms for SN2 Reactions via Machine Learning. *J. Chem. Phys.* **2021**, *155* (22), 224111. https://doi.org/10.1063/5.0074422.

(24) Komp, E.; Valleau, S. Machine Learning Quantum Reaction Rate Constants. *J. Phys. Chem. A* **2020**, *124* (41), 8607–8613. https://doi.org/10.1021/acs.jpca.0c05992.

(25) 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. https://doi.org/10.1021/acs.jpcc.0c05993.

(26) Braham, E. J.; Cho, J.; Forlano, K. M.; Watson, D. F.; Arròyave, R.; Banerjee, S. Machine Learning-Directed Navigation of Synthetic Design Space: A Statistical Learning Approach to Controlling the Synthesis of Perovskite Halide Nanoplatelets in the Quantum-Confined Regime. *Chem. Mater.* **2019**, *31* (9), 3281–3292. https://doi.org/10.1021/acs.chemmater.9b00212.

(27) 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. https://doi.org/10.1021/acsnano.9b03864.

(28) 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. https://doi.org/10.1039/D1NR05497J.

(29) Bezinge, L.; Maceiczyk, R. M.; Lignos, I.; Kovalenko, M. V.; deMello, A. J. Pick a Color MARIA: Adaptive Sampling Enables the Rapid Identification of Complex Perovskite Nanocrystal Compositions with Defined Emission Characteristics. *ACS Appl. Mater. Interfaces* **2018**, *10* (22), 18869–18878. https://doi.org/10.1021/acsami.8b03381.

(30) Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; New York: Springer, 2013.

(31) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3* (7–8), 1157–1182. https://doi.org/10.1162/153244303322753616.

(32) Clarke, M. T.; Viscomi, F. N.; Chamberlain, T. W.; Hondow, N.; Adawi, A. M.; Sturge, J.; Erwin, S. C.; Bouillard, J.-S. G.; Tamang, S.; Stasiuk, G. J. Synthesis of Super Bright Indium Phosphide Colloidal Quantum Dots through Thermal Diffusion. *Commun. Chem.* **2019**, *2* (1), 36. https://doi.org/10.1038/s42004-019-0138-z.

(33) Suh, Y.-H.; Lee, S.; Jung, S.-M.; Bang, S. Y.; Yang, J.; Fan, X.-B.; Zhan, S.; Samarakoon, C.; Jo, J.-W.; Kim, Y.; Choi, H. W.; Occhipinti, L. G.; Lee, T. H.; Shin, D.-W.; Kim, J. M. Engineering Core Size of InP Quantum Dot with Incipient ZnS for Blue Emission. *Adv. Opt. Mater.* **2022**, *10* (7), 2102372. https://doi.org/10.1002/adom.202102372.

(34) Jiang, W.; Choi, Y.; Chae, H. Efficient Green Indium Phosphide Quantum Dots with Tris(Dimethylamino)-Phosphine Phosphorus Precursor for Electroluminescent Devices. *J. Mater. Sci. Mater. Electron.* **2021**, *32* (4), 4686–4694. https://doi.org/10.1007/s10854-020-05206-5.

(35) Riehle, F. S.; Yu, K. Role of Alcohol in the Synthesis of CdS Quantum Dots. *Chem. Mater.* **2020**, *32* (4), 1430–1438. https://doi.org/10.1021/acs.chemmater.9b04009.

(36) Pedregosa, F. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(37) Irwin, B. W. J.; Mahmoud, S.; Whitehead, T. M.; Conduit, G. J.; Segall, M. D. Imputation versus Prediction: Applications in Machine Learning for Drug Discovery. *Future Drug Discov.* **2020**, *2* (2), FDD38. https://doi.org/10.4155/fdd-2020-0008.

(38) Guo, C.-Y.; Yang, Y.-C.; Chen, Y.-H. The Optimal Machine Learning-Based Missing Data Imputation for the Cox Proportional Hazard Model. *Front. Public Health* **2021**, *9*.

(39) Borchani, H.; Varando, G.; Bielza, C.; Larrañaga, P. A Survey on Multi-Output Regression. *WIREs Data Min. Knowl. Discov.* **2015**, *5* (5), 216–233. https://doi.org/10.1002/widm.1157.

(40) Lee, S. H.; Kim, Y.; Jang, H.; Min, J. H.; Oh, J.; Jang, E.; Kim, D. The Effects of Discrete and Gradient Mid-Shell Structures on the Photoluminescence of Single InP Quantum Dots. *Nanoscale* **2019**, *11* (48), 23251–23258. https://doi.org/10.1039/C9NR06847C.

(41) Kim, H.-J.; Jo, J.-H.; Yoon, S.-Y.; Jo, D.-Y.; Kim, H.-S.; Park, B.; Yang, H. Emission Enhancement of Cu-Doped InP Quantum Dots through Double Shelling Scheme. *Materials* **2019**, *12* (14). https://doi.org/10.3390/ma12142267.

(42) Stein, J. L.; Holden, W. M.; Venkatesh, A.; Mundy, M. E.; Rossini, A. J.; Seidler, G. T.; Cossairt, B. M. Probing Surface Defects of InP Quantum Dots Using Phosphorus Kα and Kβ X-Ray Emission Spectroscopy. *Chem. Mater.* **2018**, *30* (18), 6377–6388. https://doi.org/10.1021/acs.chemmater.8b02590.

(43) Min, C.-H.; Joo, J. Studies on the Effect of Acetate Ions on the Optical Properties of InP/ZnSeS Core/Shell Quantum Dots. *J. Ind. Eng. Chem.* **2020**, *82*, 254–260. https://doi.org/10.1016/j.jiec.2019.10.021.

(44) 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. https://doi.org/10.1021/acs.chemmater.5b00286.

(45) Kirkwood, N.; De Backer, A.; Altantzis, T.; Winckelmans, N.; Longo, A.; Antolinez, F. V.; Rabouw, F. T.; De Trizio, L.; Geuchies, J. J.; Mulder, J. T.; Renaud, N.; Bals, S.; Manna, L.; Houtepen, A. J. Locating and Controlling the Zn Content in In(Zn)P Quantum Dots. *Chem. Mater.* **2020**, *32* (1), 557–565. https://doi.org/10.1021/acs.chemmater.9b04407.

(46) van Embden, J.; Chesman, A. S. R.; Jasieniak, J. J. The Heat-Up Synthesis of Colloidal Nanocrystals. *Chem. Mater.* **2015**, *27* (7), 2246–2285. https://doi.org/10.1021/cm5028964.

(47) Mićić, O. I.; Cheong, H. M.; Fu, H.; Zunger, A.; Sprague, J. R.; Mascarenhas, A.; Nozik, A. J. Size-Dependent Spectroscopy of InP Quantum Dots. *J. Phys. Chem. B* **1997**, *101* (25), 4904–4912. https://doi.org/10.1021/jp9704731.

(48) Cho, E.; Jang, H.; Lee, J.; Jang, E. Modeling on the Size Dependent Properties of InP Quantum Dots: A Hybrid Functional Study. *Nanotechnology* **2013**, *24* (21), 215201. https://doi.org/10.1088/0957-4484/24/21/215201.

(49) Baskoutas, S.; Terzis, A. F. Size-Dependent Band Gap of Colloidal Quantum Dots. *J. Appl. Phys.* **2006**, *99* (1), 013708. https://doi.org/10.1063/1.2158502.

(50) Guzelian, A. A.; Katari, J. E. B.; Kadavanich, A. V.; Banin, U.; Hamad, K.; Juban, E.; Alivisatos, A. P.; Wolters, R. H.; Arnold, C. C.; Heath, J. R. Synthesis of Size-Selected, Surface-Passivated InP Nanocrystals. *J. Phys. Chem.* **1996**, *100* (17), 7212–7219. https://doi.org/10.1021/jp953719f.

(51) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006**, *63* (1), 3–42. https://doi.org/10.1007/s10994-006-6226-1.

(52) *Https://Docs.Streamlit.Io/Library/Get-Started.*

(53) Stein, J. L.; Mader, E. A.; Cossairt, B. M. Luminescent InP Quantum Dots with Tunable Emission by Post-Synthetic Modification with Lewis Acids. *J. Phys. Chem. Lett.* **2016**, *7* (7), 1315–1320. https://doi.org/10.1021/acs.jpclett.6b00177.

(54) Zhang, X.; Lv, H.; Xing, W.; Li, Y.; Geng, C.; Xu, S. Trioctylphosphine Accelerated Growth of InP Quantum Dots at Low Temperature. *Nanotechnology* **2021**, *33* (5), 055602. https://doi.org/10.1088/1361-6528/ac3180.

(55) Achorn, O. B.; Franke, D.; Bawendi, M. G. Seedless Continuous Injection Synthesis of Indium Phosphide Quantum Dots as a Route to Large Size and Low Size Dispersity. *Chem. Mater.* **2020**, *32* (15), 6532–6539. https://doi.org/10.1021/acs.chemmater.0c01906.

(56) Shallcross, R. C.; Graham, A. L.; Karayilan, M.; Pavlopoulous, N. G.; Meise, J.; Pyun, J.; Armstrong, N. R. Influence of the Processing Environment on the Surface Composition and Electronic Structure of Size-Quantized CdSe Quantum Dots. *J. Phys. Chem. C* **2020**, *124* (39), 21305–21318. https://doi.org/10.1021/acs.jpcc.0c05622.

(57) 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. https://doi.org/10.1021/nl801692g.

(58) Kowalczyk, B.; Lagzi, I.; Grzybowski, B. A. Nanoseparations: Strategies for Size and/or Shape-Selective Purification of Nanoparticles. *Curr. Opin. Colloid Interface Sci.* **2011**, *16* (2), 135–148. https://doi.org/10.1016/j.cocis.2011.01.004.

(59) Morris-Cohen, A. J.; Donakowski, M. D.; Knowles, K. E.; Weiss, E. A. The Effect of a Common Purification Procedure on the Chemical Composition of the Surfaces of CdSe Quantum Dots Synthesized with Trioctylphosphine Oxide. *J. Phys. Chem. C* **2010**, *114* (2), 897–906. https://doi.org/10.1021/jp909492w.

(60) Taylor, D. A.; Teku, J. A.; Cho, S.; Chae, W.-S.; Jeong, S.-J.; Lee, J.-S. Importance of Surface Functionalization and Purification for Narrow FWHM and Bright Green-Emitting InP Core–Multishell Quantum Dots via a Two-Step Growth Process. *Chem. Mater.* **2021**, *33* (12), 4399–4407. https://doi.org/10.1021/acs.chemmater.1c00348.

(61) Vikram, A.; Zahid, A.; Bhargava, S. S.; Keating, L. P.; Sutrisno, A.; Khare, A.; Trefonas, P.; Shim, M.; Kenis, P. J. A. Mechanistic Insights into Size-Focused Growth of Indium Phosphide Nanocrystals in the Presence of Trace Water. *Chem. Mater.* **2020**, *32* (8), 3577–3584. https://doi.org/10.1021/acs.chemmater.0c00781.

(62) Xie, L.; Harris, D. K.; Bawendi, M. G.; Jensen, K. F. Effect of Trace Water on the Growth of Indium Phosphide Quantum Dots. *Chem. Mater.* **2015**, *27* (14), 5058–5063. https://doi.org/10.1021/acs.chemmater.5b01626.

(63) Pyrz, W. D.; Buttrey, D. J. Particle Size Determination Using TEM: A Discussion of Image Acquisition and Analysis for the Novice Microscopist. *Langmuir* **2008**, *24* (20), 11350–11360. https://doi.org/10.1021/la801367j.

(64) Gary, D. C.; Cossairt, B. M. Role of Acid in Precursor Conversion During InP Quantum Dot Synthesis. *Chem. Mater.* **2013**, *25* (12), 2463–2469. https://doi.org/10.1021/cm401289j.