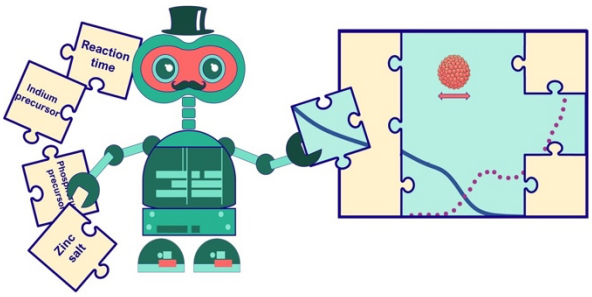
Predicting Indium Phosphide Quantum Dot Properties from Synthetic Procedures Using Machine Learning

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*]

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

Abstract

The prediction 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 the chemical identity or 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. 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. 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 designed an experiment 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 in this study show the ability to design materials with targeted properties and explore underlying reaction mechanisms despite 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 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. 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.

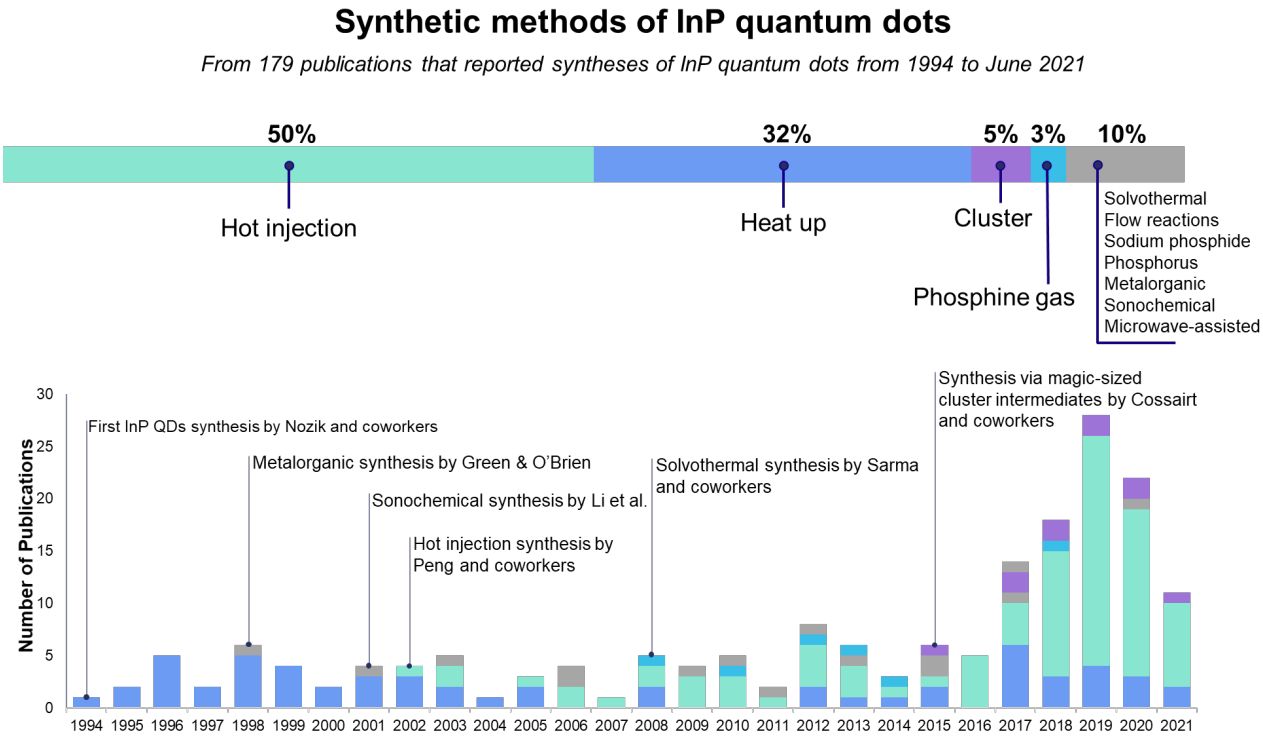


Figure 1. Timeline and number of publications of InP QDs synthesis.

2. Methods

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 the synthetic methods that can affect the accuracy of the predictions, only similar methods, where the reaction is done 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 Feature Selection:

The information extracted from 219 syntheses was split into input features and output targets. With the purpose 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. 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. For example, it has been shown that the final properties and quality of QDs is affected by purification solvents and conditions[32], which are not consistently reported in the literature. Therefore, to evaluate the effect of feature selection on our models, we defined two sets of input features and compiled two datasets: an extended and a condensed dataset (**Figure 2**). In the extended dataset, the additives beyond the indium and phosphorus sources were categorized by their functional groups (e.g., acid, amine, thiol); while the condensed dataset grouped chemicals by their primary assumed role in the synthesis (e.g., ligands). (See the full list of parameters in Table S1)

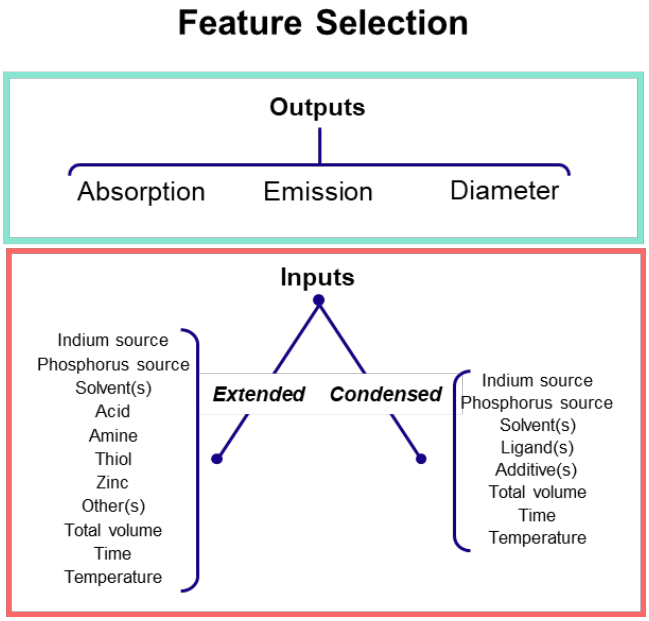


Figure 2. Output and input feature selections.

2.3 Data Imputation

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 used the available data to train predictive machine learning algorithms for each output feature and imputed the missing values. Since absorption was the most frequently reported

output in the initial dataset (205 syntheses), data augmentation 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.

;

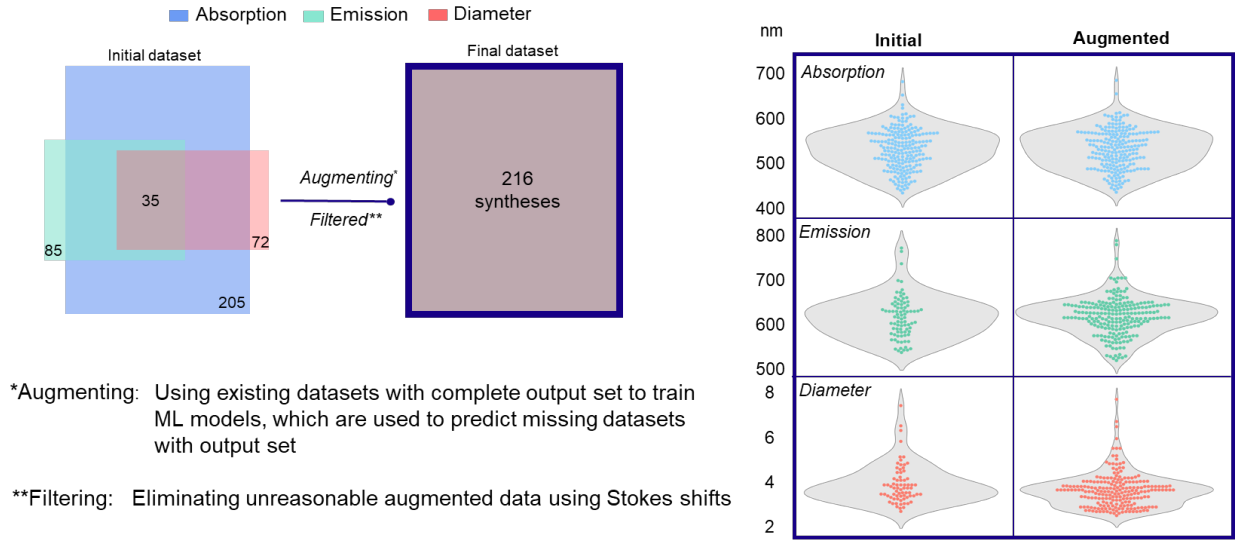


Figure 3. Data imputation process (left) and descriptions of the imputed dataset (right).

2.4 Machine Learning Models and Metrics

Prior to training machine learning models, the numerical parameters in the input set were scaled and the categorical parameters were transformed to numerical features using one-hot encoding and the scikit-learn software package (sklearn)[33]. Next, we compared single- and multi-output regressors for our three output features. Single-output models predict each feature individually, and the features do not depend on each other. Multi-output models predict all output features simultaneously. The output features often depend on each other and on the inputs features[34]. We tested regressors suitable for small datasets such as Extra Trees, Decision Tree, Random Forest, k-NN, Bagging, and Gradient Boosting using sklearn. 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. We optimized parameters using grid search. We used the mean absolute error (MAE) and the coefficient of determination (R^2) as metrics to assess the performance of all models. MAE is sensitive to outliers since it is a linear score, in which all differences are weighted equally. R^2 indicates the proportion of variance for a dependent variable determined by an independent variable. For each model in this study, we report the MAE and R^2 of the predicted set versus the test set.

2.5 Interactive User Interface

To allow external users such as 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[35]. Streamlit is a framework for building interactive web applications with user-friendly components such as buttons, sliders, and plots.

2.6 Predicting Outcomes of New InP QD 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[36–39] 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 S6).

3. Results and Discussion

3.1 Dataset Description

After the data extraction process, the dataset contained 219 syntheses of InP QDs extracted 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[40]. 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. 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 is known to increase the photoluminescence quantum yield and the stability of the InP QDs[41]; 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[42].

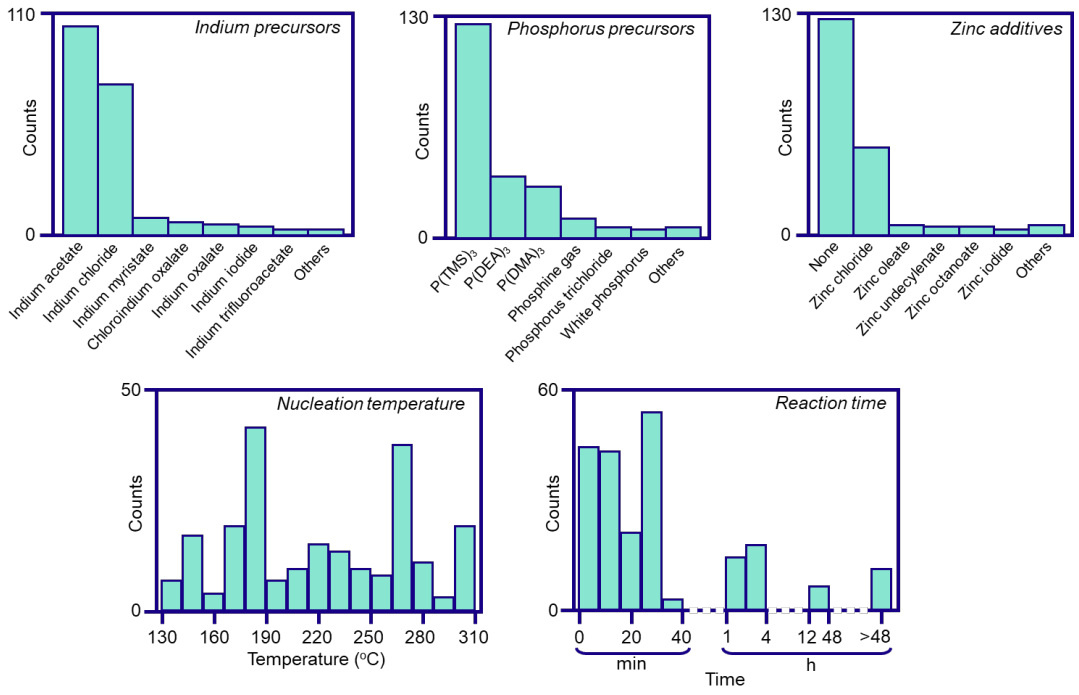


Figure 4. Description of the input set. Histograms of indium, phosphorus precursors, zinc additives, nucleation temperature, and reaction time of the syntheses in the initial dataset.

Complications with extracting data from the literature often arise from inconsistencies in data reporting. To overcome the incompleteness of our initial output set extracted from the literature, 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[43,44]. This method is simple when only a small fraction of the output set is missing and when the missing values can be calculated. However, the problem becomes more challenging when more values are absent. 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. To avoid physically unreasonable data produced from imputation, we calculated the Stokes shift (the difference between peak maxima of absorption and emission spectra) for each synthesis and eliminated syntheses with a negative Stokes shift. The resulting imputed dataset includes 216 syntheses with a complete output set, 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).

3.2 Machine Learning Model Training and Performance

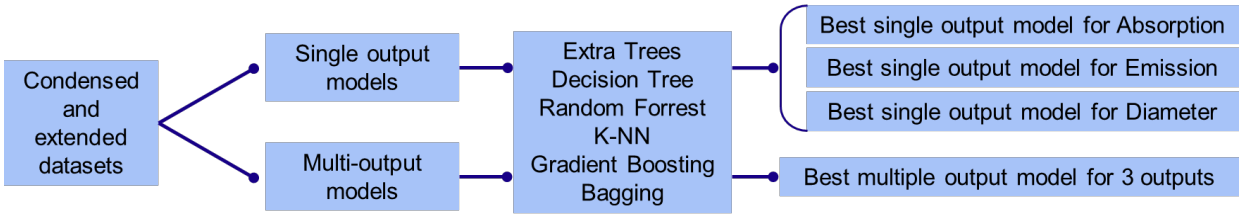


Figure 5. Diagram of the model training process.

The synthesis of InP QDs involves many parameters that play different roles in determining the properties of the final product. Therefore, defining input features for the machine learning models is a crucial step to achieve meaningful results. As mentioned in Section 2.2, a condensed dataset and an extended dataset were defined based on chemical function and chemical identity, respectively. Both datasets were used in a model training process (**Figure 5**), where two classes of machine learning models, single-output and multi-output, were employed. For each class, 6 different algorithms that are well-suited for small datasets were adapted. **Figure 6** shows the performance of the best model for each output in each study case. Extra Trees and Decision Tree algorithms gave the lowest MAE in all cases. The single-output Extra Trees and Decision Tree algorithms for the condensed dataset achieved the lowest MAEs for the prediction of absorption and diameter, respectively, while the single-output Extra Trees model that used the extended dataset gave the lowest MAE for emission prediction.

Although multi-output models were expected to give better predictions due to the strong correlation between the three output features, single-output models showed lower MAEs for both datasets (See Supporting Information Section S3 for Pearson correlations). This observation indicates that the algorithms were unable to capture the correlation among the QD properties with the given datasets, or that there are synthetic variables that affect one output more significantly than others. Another possibility is that this was a result of our method of data imputation, where we imputed each output feature separately, and had to fill in emission wavelength and diameter for many syntheses. This observation may also be affected by the lack of important synthetic and post-synthetic parameters being reported such as injection rate[45], purification solvents[32], and specification as to whether the data are for purified or in-situ samples.

While the models using the condensed dataset gave better predictions for absorption wavelength and diameter, the models using the extended dataset showed modestly better performance for emission wavelength. R^2 values for diameter for all cases were relatively low because this outcome is the most absent in the initial dataset and the way that particle sizes are measured from TEM can be inaccurate and prone to user error[46]. We also observed an underfitting behavior in the Decision Tree model for diameter in the single-output model that used the condensed dataset (**Figure 6A**), which can occur when the influence of a few parameters is significantly higher than others. Using the best model for our four study cases, we were able to identify temperature and time as the most influential synthetic parameters, an expected result (**Figure 7**). More interestingly, the models also recognized the importance of zinc additives, which aligns with the reported observations of spectral shifts and size changes when a zinc salt is present in the synthesis[41,47].

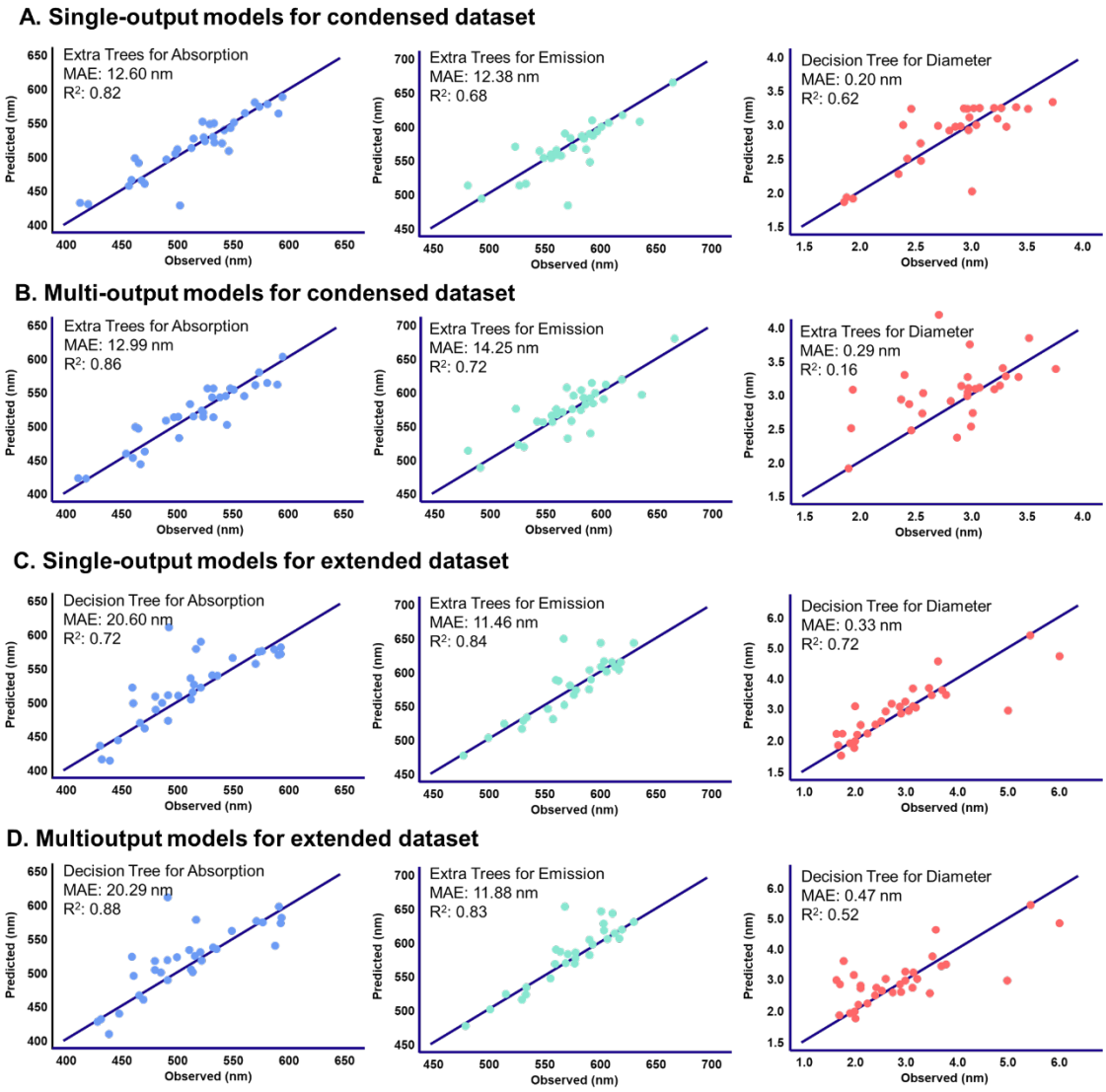


Figure 6. The observed vs predicted plots and the performance of single-output and multi-ouput models for the three outputs using the condensed and extended datasets.

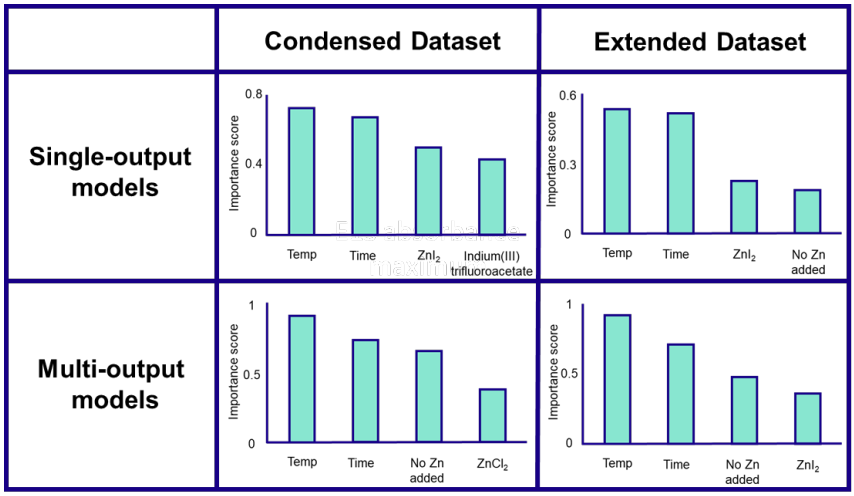


Figure 7. Feature importance charts of the best model for each study case.

3.3 Models Using Datasets with Only Hot Injection Synthesis for InP

Our next step was to improve the accuracy of the machine learning prediction by using datasets that contained only hot injection syntheses. Hot injection syntheses were filtered from both condensed and extended datasets and resulted in 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 diameters and absorption wavelengths. 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

3.4 Models for the Hot Injection Synthesis of CdSe

To further evaluate the validity and show the utility of the imputing method for small datasets, we revised and extended the CdSe 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 S9 of the Supporting Information). 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 (GBM) algorithm[25] (**Figure 8**). 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 to the GBM 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 9**).

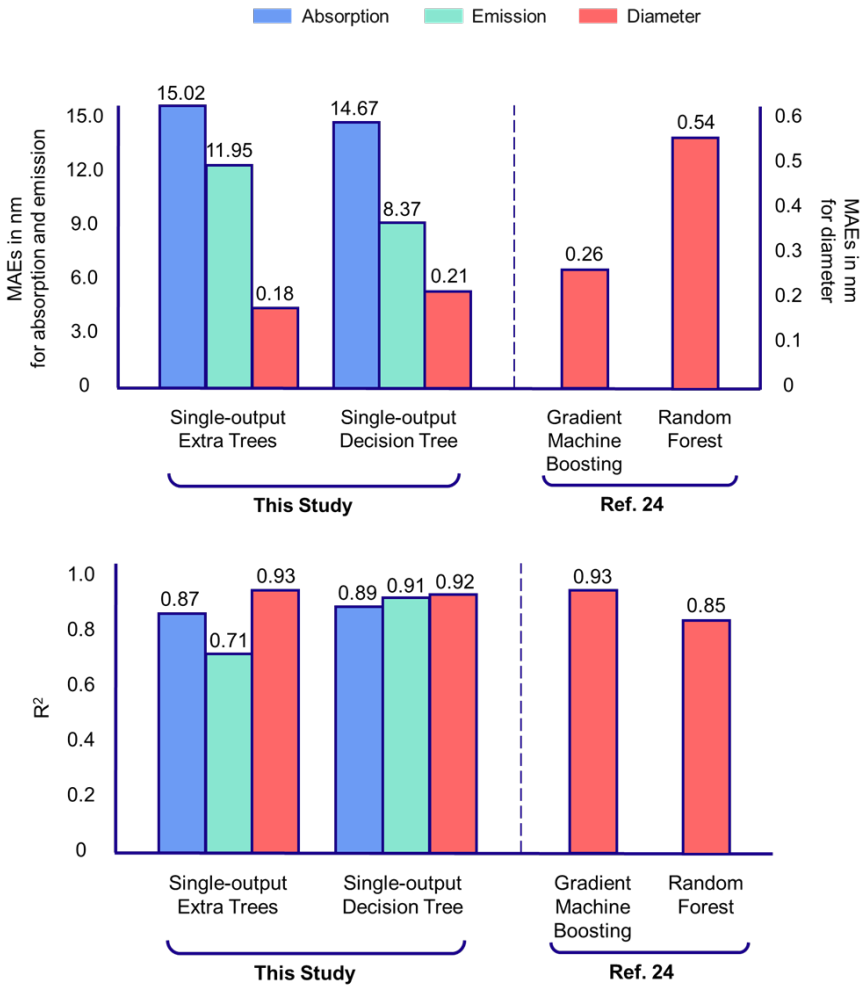


Figure 8. MAEs and R^2 values comparison of the two models between this study and ref 24.

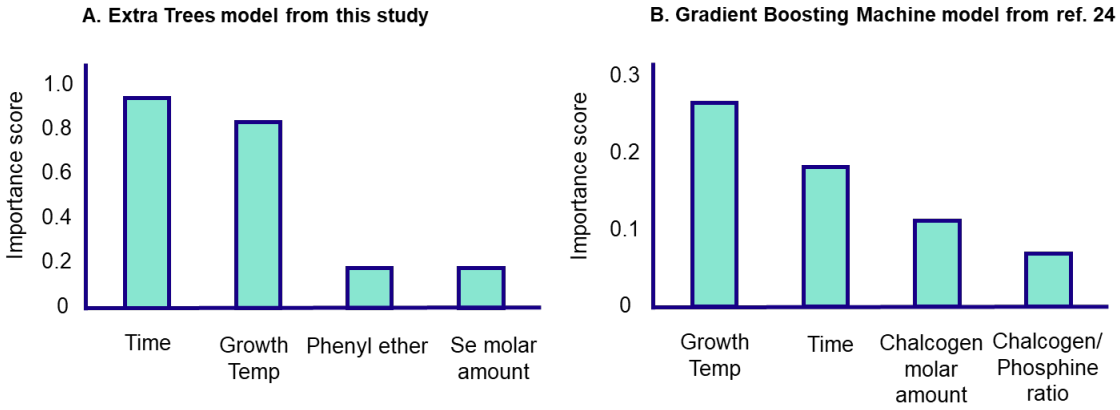


Figure 9. Feature importance charts of A. Extra Trees model from this study and B. Gradient Boosting Machine model from ref 24.

3.5 Interactive Web Applications and Comparison with Experimental Results

From the best machine learning models that used datasets with all appropriate synthetic methods described in section 3.2, we deployed four Streamlit web apps that enabled real time reaction analysis and prediction. Each web app included sections where synthetic conditions can be imported to a chosen machine learning algorithm for predictions of InP QD optical properties (**Figure 10**). This allowed 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 apps 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 11**). These observations are consistent with the reported literature[41,48].

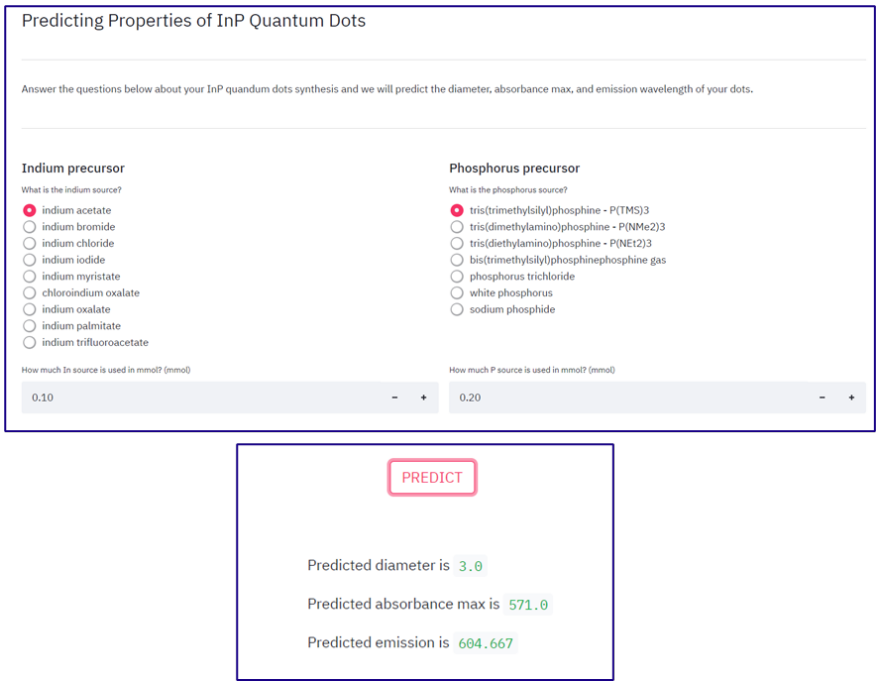


Figure 10. Interactive web applications by Streamlit for predicting InP QD properties from synthetic conditions.

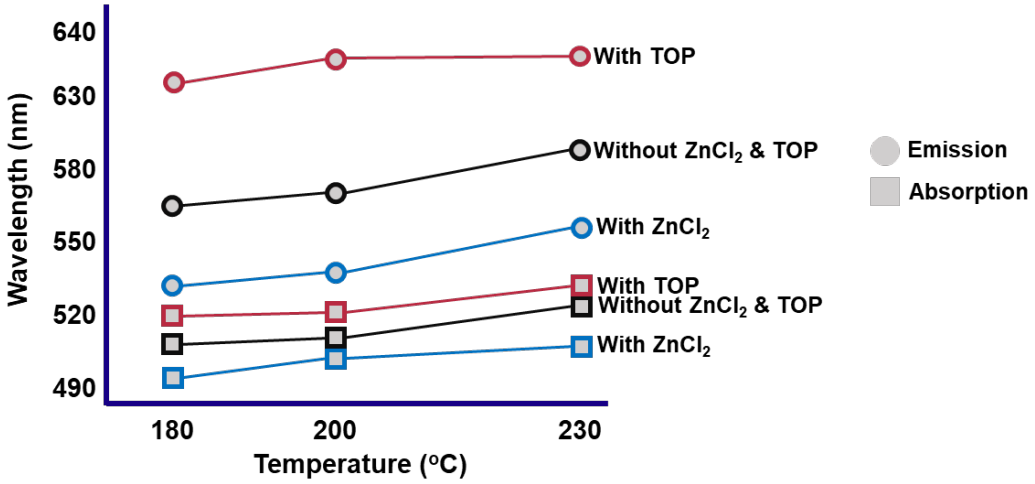


Figure 11. Predicted emission (circles) and absorption (squares) wavelengths from the Streamlit webapp using single-output algorithms and the condensed dataset with all methods. Reaction conditions include 0.1 mmol of $InCl_3$, 1 mL of oleylamine, 0.15 mmol of $P(DEA)_3$, nucleation temperature at 180 °C, 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).

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 S6). We entered these synthetic conditions into our four web apps to obtain predicted values computed by our chosen machine learning models. These values were compared with our experimentally measured values of absorption and emission wavelengths, and particle sizes determined from TEM. MAEs on the experimental values are higher than the ones from the test sets, except for diameter predictions using the extended dataset (**Figure 12**). The large MAEs arise from the small size of the experimental set, which makes the impact of outliers on MAE more pronounced. However, some models were able to predict the synthetic outcomes with absolute differences as small as 1 nm for absorption, 3 nm for emission, and 0.07 nm for diameter (See Table S7-10 for more details). Since preliminary analysis from section 3.3 suggested that models using datasets with only hot injection reactions have improved accuracy, we tested their performance with our experimental results. Despite the improved training metrics originally observed, the performance of these models when compared to our experimental set showed only a modest improvement in prediction accuracy (**Figure 13**).

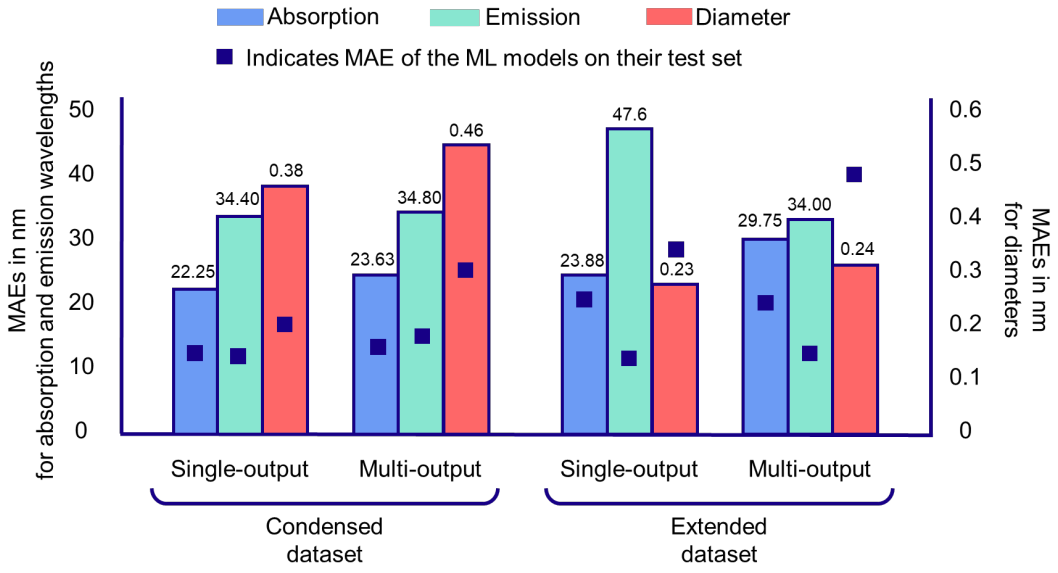


Figure 12. Mean absolute errors (MAE) in nm of model predictions vs experimental results. Algorithms used datasets with all synthetic methods. Dark blue squares indicate MAEs of the models on their test set.

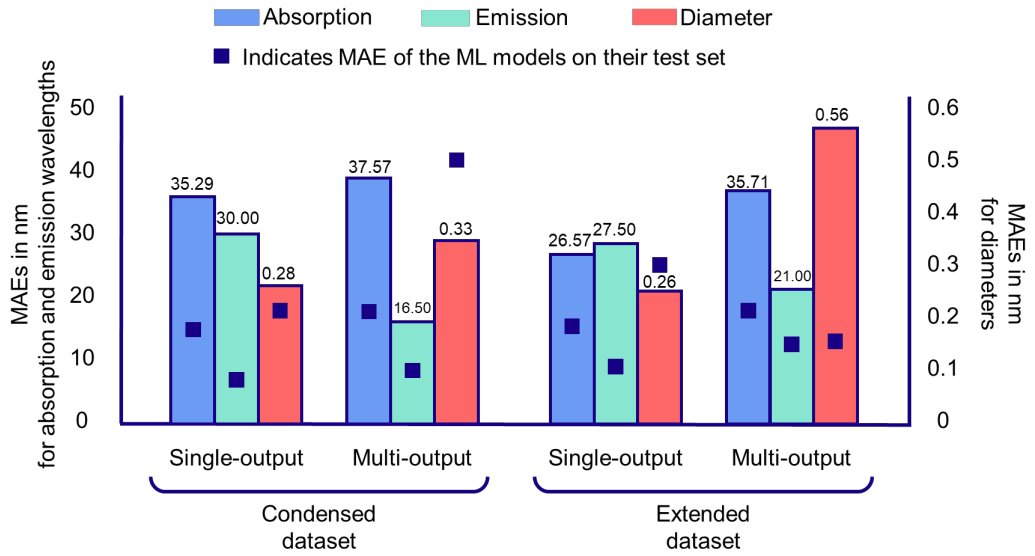


Figure 13. Mean absolute errors (MAE) in nm of model predictions vs experimental results. Algorithms used the datasets with only hot injection method. Dark blue squares indicate MAEs of the models on their test set.

Among all the web apps we have deployed, the one from single-output algorithms for the hot injection extended dataset showed the best performance and consistency across test sets and the experimental set, therefore, we launched this app for public users via https://share.streamlit.io/cossairt-lab/indium-phosphide/Hot_injection/hot_injection_prediction.py. In addition, this webapp also includes the prediction of CdSe QD optical properties from our study in section 3.4. We hope that by sharing this webapp with other researchers, more chemical insights of InP synthesis from the machine learning will be discovered. 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 a low synthesis variety in the dataset. We expect the performance of the webapp to improve when a larger dataset becomes available. A disadvantage of the webapp is that new synthetic conditions such as precursors cannot be entered, so one can only examine procedures from the existing dataset.

Finally, we targeted ~600 nm–absorbing QDs using synthetic conditions and precursors from an existing procedure[49]. Using the suggested reaction conditions obtained from the webapp, we were able to synthesize InP QDs with desired optical properties with high accuracy (**Figure 14**). For absorption and emission wavelengths, we also found that there was a noticeable difference between samples before and after purification. This observation justified 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.

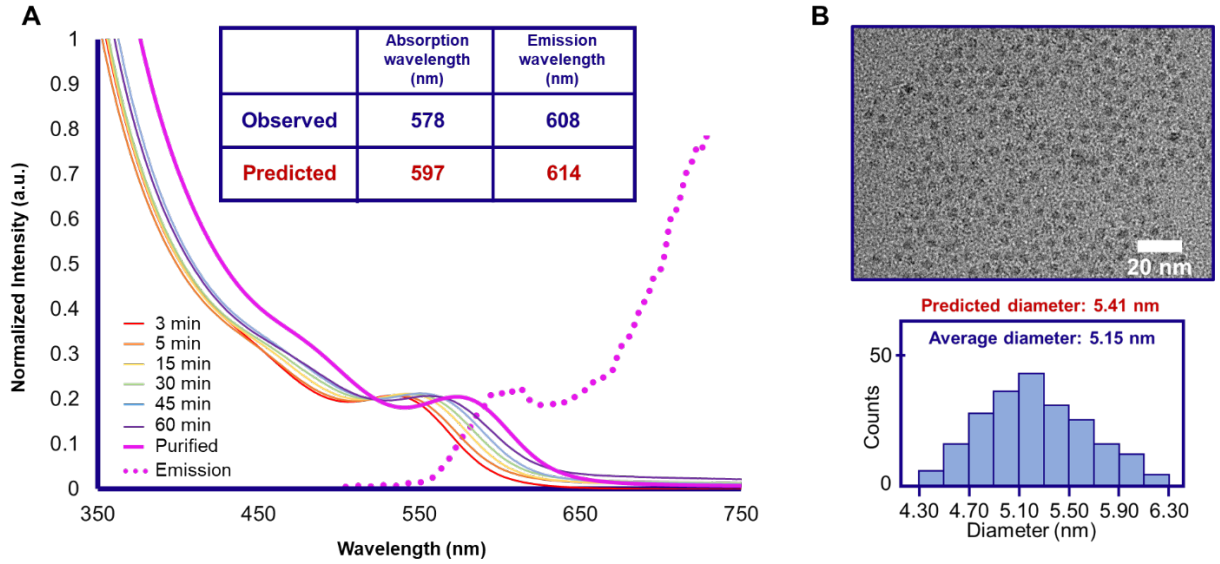


Figure 14. A: UV-Vis spectra of timed aliquots and emission spectrum of the purified product from the reaction using 0.40 mmol indium acetate, 1.45 mmol myristic acid, and 0.20 mmol $P(SiMe_3)_3$ injected at 315 °C. The nucleation temperature was 310 °C. B: A TEM image of the purified particles with an average diameter of 5.15 nm.

4. Conclusions

We have trained and used machine learning models to predict the properties of InP QDs based on synthetic input features. Using an imputed dataset, the descriptor space was defined in two ways (condensed and extended) to study the best approach for input feature selection. While models using the condensed dataset had better performance in predicting absorption, models using the extended dataset gave improved predictions for emission wavelength and diameter. Single-output and multi-output machine learning algorithms were applied in this study. The single-output models showed enhanced performance over the multi-output models. 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 and machine learning training were applied to both InP and CdSe hot injection datasets. Furthermore, we deployed a web app that external users can access to predict InP and CdSe synthetic outcomes using our best algorithms. Using this web app, we were able to test our models with newly adapted InP syntheses and synthesize InP QDs with 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 in order to obtain accurate predictions. Overall, this work provides a procedure to preprocess datasets, train machine learning 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.

Corresponding Author

*cossairt@uw.edu

Funding Sources

National Science Foundation OMA-1936100 and DMR-2019444.

Acknowledgements

We thank Dr. Stephanie Valleau and Dr. David Beck for their input and scientific guidance through the Data Intensive Research Enabling Clean Technologies (DIRECT) program at the University of Washington. 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) 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.
(33) Pedregosa, F. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
(34) 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.
(35) *Https://Docs.Streamlit.Io/Library/Get-Started*.
(36) 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.
(37) 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.
(38) 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.
(39) 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.

(40) 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.
(41) 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.
(42) 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.
(43) 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*.
(44) 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.
(45) 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.
(46) 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.
(47) 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.
(48) 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.
(49) 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.