Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 3683

Cross-laboratory validation of machine learning models for copper nanocluster synthesis using cloud-based automated platforms

Ricardo Montoya-Gonzalez, [ab] Rosa de Guadalupe González-Huerta,[b] Martha Leticia Hernández-Pichardo [a] and Subha R. Das *[c]

The integration of machine learning (ML) into materials science has the potential to accelerate material discovery and optimize properties. However, the reliability of ML models depends heavily on the consistency and reproducibility of experimental data. In this study, we present a methodology to combine automated, remotely-programmed synthesis protocols with ML to enable data-driven materials discovery. Experiments were programmed and conducted remotely through robotic syntheses at cloud laboratories, using multiple different liquid handlers and spectrometers across two independent facilities (Emerald Cloud Lab, Austin, TX and Carnegie Mellon University Automated Science Lab, Pittsburgh, PA). This multi-instrument approach ensured precise control over reaction parameters, eliminated both operator and instrument-specific variability, and enabled generation of high-quality datasets for ML training. From only 40 training samples, our approach predicts whether specific synthesis parameters will lead to successful formation of copper nanoclusters (CuNCs) with interpretable models providing mechanistic insights through SHAP analysis. Our workflow demonstrates how remotely accessed/cloud laboratory infrastructure coupled with ML can transform traditionally manual processes into autonomous, predictive systems. This multi-instrument validation demonstrates reproducibility critical for reliable ML-driven materials discovery and for advancing automated materials synthesis beyond single-laboratory demonstrations.

Received 30th July 2025
Accepted 28th October 2025

DOI: 10.1039/d5dd00335k

rsc.li/digitaldiscovery

1 Introduction

Artificial intelligence (AI) transforms numerous scientific disciplines, with machine learning (ML) playing a pivotal role as a key variant of AI. Using ML enables systems to discern patterns and make predictions from data, facilitates the analysis of large datasets, identifies complex correlations, and generates predictions more efficiently than traditional methods. An ML enhanced approach is particularly beneficial in materials science, aiding in the exploration of materials with specific properties, design of alloys, prediction of structural stability, and analysis of spectroscopic data, among other applications.[1–3]

Despite the transformative potential of ML in materials science, its application in experimental systems faces significant challenges related to data consistency and quantity.[2,4–6] Materials synthesis involves numerous interdependent variables: selection of reagents, concentrations of reagents, temperature, pH, reaction time, and choice of reaction vessel, where even small changes can dramatically alter outcomes.[7] This sensitivity creates a dual problem – experimental protocols become difficult to reproduce, leading to inconsistent results that generate poor-consistency datasets, while the time-consuming and resource-intensive nature of materials synthesis limits the number of experiments that can be conducted. Even when high-consistency data is obtained, researchers often lack sufficient quantities of data to train robust ML models. Generating comprehensive datasets covering the full parameter space requires hundreds or thousands of experiments—a practical impossibility with traditional manual approaches. Therefore, for ML to be successfully applied to materials science, it remains essential to improve data consistency and data quantity. Data consistency can be improved through enhanced experimental control, while data quantity can be improved through the use of automated, high-throughput experiments. Both limitations have mainly been addressed through the implementation of microfluidic systems and robotic synthesis.[8]

Several approaches have employed ML in experimental materials science, primarily for property optimization[9–11] or to

[a]*Instituto Politécnico Nacional-ESIQIE, Laboratorio de Nanomateriales Sustentables, UPALM, México D.F. 07738, Mexico*

[b]*Instituto Politécnico Nacional, ESIQIE, Laboratorio de Electroquímica y Corrosión, UPALM, Zacatenco, 07738, México City, Mexico*

[c]*Department of Chemistry and Center for Nucleic Acids Science & Technology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. E-mail: srdas@andrew.cmu.edu*

elucidate synthesis parameter effects on specific outcomes,[7,12–14] particularly with complex variable interactions. Tang *et al.*[15] trained a regression ML model to optimize the photoluminescence quantum yield of MoS_2 synthesized *via* chemical vapor deposition. Guda *et al.*[16] explored gold nanoparticle synthesis parameters using an Extra Trees algorithm to predict size and shape. Miyazato *et al.*[17] used ML models to map reaction selectivity in the oxidative coupling of methane, demonstrating ML's value in refining reaction pathway analysis.

Furthermore, several approaches have been made to outline best practices for ML implementation, including data preprocessing, feature engineering, and model evaluation.[2,11,12,18]

Metal nanoclusters (NCs) are materials composed of a few to a hundred atoms with unique properties due to their size comparable to the Fermi wavelength of electrons, resulting in molecular-like properties.[19,20] NCs possess characteristics including fluorescence, low toxicity, large Stokes shift, and good biocompatibility.[21,22] One common synthesis method is traditional wet chemical reduction, typically involving trapping metal ions within pores, channels, a nano-matrix, or a template, followed by reduction using agents like ascorbic acid, sodium borohydride, or hydrazine. While many studies have focused on gold, silver, and copper NCs and their applications, the development of copper NCs (CuNCs) is still comparatively underdeveloped and in its early stages.[23] This is primarily due to challenges, as CuNCs tend to agglomerate and oxidize.[19,24]

In this study, we report a well-documented and reproducible methodology for CuNCs synthesis *via* simple wet chemistry reduction combined with real-time absorbance measurements. The experiments and syntheses were performed in a remotely controlled and automated laboratory. This research proposes a methodology for material design and synthesis aimed at obtaining reliable and reproducible experimental results, validated across different batches and at an independent location, to enable their subsequent integration into ML models. Using a highly controlled protocol with robotic CuNCs synthesis, chosen due to its well-documented nature, the data generated were used to train an ML model capable of predicting synthesis success based on selected parameters. This approach demonstrates the potential of combining automated synthesis with ML to enhance material design and discovery.

2 Experimental

2.1 Synthesis protocol

The procedure begins by adding $CuSO_4$ 1M (Cu) and hexadecyltrimethylammonium bromide 1 M (CTAB) in varying proportions (see SI Table 1) into a 96-well, 2 mL Deep Well Plate, along with 1 mL of H_2O. The mixture is then cooled to 4 °C and stirred at 30 rpm for 1 hour. After this incubation period, ascorbic acid 1 M (AA), sodium hydroxide (NaOH), and 0.8 mL of water are rapidly added, followed by mixing at 300 rpm for 15 minutes. To analyse the reaction, 250 μL aliquots are taken from each well and transferred into a 96-well UV-Star Plate. This plate is then placed into the CLARIOstar absorbance spectrometer, where it is heated to 45 °C. Once this temperature is reached, absorbance spectra are recorded every 43 seconds for 80 minutes.

To measure the reproducibility of the absorbance spectra we calculated the coefficient of variation (CV) of the absorbance intensity at each wavelength, defined as the ratio of the standard deviation to the mean absorbance intensity, expressed as a percentage. This metric quantifies the relative spread of the absorbance values and was used to assess the reproducibility of the measurements.[25,26]

The molar concentrations of the first four samples (SI Table 1) were selected directly from the literature. Samples #5–10 were prepared by incrementing the concentrations of AA and CTAB. Samples #11–20 were prepared by incrementing smaller concentrations of Cu, CTAB, and AA. Samples #21–40 were generated using the Latin Hypercube Sampling method, with absolute concentration bounds set between 0.5 and 5 mM with a sum of 6.25 mM between all reagents.

2.2 ECL command center

Every experiment reported in this work was set up using Command Center Desktop Version: 1.5.134.1. Constellation Version: 2.7.2.2. Mathematica version: 13.3.1 for Microsoft Windows.

2.3 Robotic workcell

Multiple Hamilton Liquid Handler SuperSTAR identical model units were used at both the Austin, TX and CMU Pittsburgh, PA labs. The serial numbers of the different instruments are listed below:

- At Emerald Cloud Lab (Austin, TX): H910, B862, I068
- At CMU Automated Science Lab (Pittsburgh, PA): H924, H688
- Software: Hamilton Microlab STAR Software VENUS two v4.3.0.4686.

Multiple CLARIOstar spectrometers identical model units were used to record absorbance and fluorescence measurements. The serial numbers of the instruments are listed below:

- At Emerald Cloud Lab (Austin, TX): 430-0929, 430-4255
- At CMU Automated Science Lab (Pittsburgh, PA): 430-4371, 430-4373
- Software: CLARIOstar 5.4.

2.4 Machine learning model training

Since the ECL software is based on Mathematica, all data preprocessing, transformation, ML training, and validation were performed using Wolfram Mathematica (version 14.0). All models were trained with the same forty sample information and validated with six never-seen samples. Root mean square error was calculated with eqn (1) where m is the number of values that were us to validate the model, $f(x_i)$ is the predicted output value by the model and y_i is the actual output value

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i) - (y_i)\right)^2} \quad (1)$$

Coefficient of determination (R^2), its expression is shown in eqn (2), where $Stdv_{baseline}$ is the standard deviation of the true values.

$$R^2 = 1 - \frac{RMSE^2}{Stdv_{baseline}{}^2} \quad (2)$$

Each model's hyperparameters were automatically optimized in Mathematica using the Predict function with Quality set as the performance goal, aiming to maximize the overall prediction accuracy.

2.4.1 Linear regression. L2 regularization = 1, max iteration number = 30.

2.4.2 Decision tree. Nodes = 13, leaves = 7, feature fraction = 1.

2.4.3 Random forest. Feature fraction = 1/3, leaf size = 4, trees number = 100, distribution smoothing = 0.5.

2.4.4 Nearest neighbour. Neighbours number = 5, nearest method = KDTree, distribution smoothing = 0.5.

2.4.5 Gradient boosted trees. Max training rounds = 50, leaves number = 25, learning rate = 0.1, max depth = 6, leaf size = 3.

2.4.6 Gaussian process. Estimation method = maximum posterior, nominal covariance type = hamming, numerical covariance type = squared exponential.

2.4.7 Neural network. Fully connected network, depth = 8, number of parameters = 17 700, activation function = SELU, max training rounds = 300.

3 Results and discussion

3.1 Data collection

One of the key aspects of using experimental data for ML training is ensuring its consistency and reproducibility. To address this, we implemented a robotic experimental methodology performed in a robotic workcell (SI Fig. 1), where every synthesis step was executed consistently and in a standardized manner. This approach guarantees high data consistency and ensures that modifications to synthesis parameters can be reliably correlated with their impact on outcomes.

Fig. 1 illustrates the synthesis protocol, adapted from a widely reported method for AuNCs.[27] Briefly, the procedure begins by adding $CuSO_4$ (Cu) and hexadecyltrimethylammonium Bromide (CTAB) solutions in varying proportions (see SI Table 1) along with H_2O. The mixture is cooled and slowly stirred. Subsequently, ascorbic acid (AA) is added, followed by vigorous mixing for 15 minutes. An aliquot is then transferred to a well in a 96-well plate to an absorbance spectrometer, heated to the target reaction temperature, and absorbance spectra are recorded every 43 seconds for no more than 80 minutes to avoid oxidation.[28]

We used the remotely programmed and automated robotic liquid handling enabled by the Emerald Cloud Lab (ECL) at their laboratory in Austin, TX for synthesis. This platform allows precise control of operational conditions, material traceability, and real-time monitoring of experimental progress. Prior to execution, operational parameters such as temperature, stirring

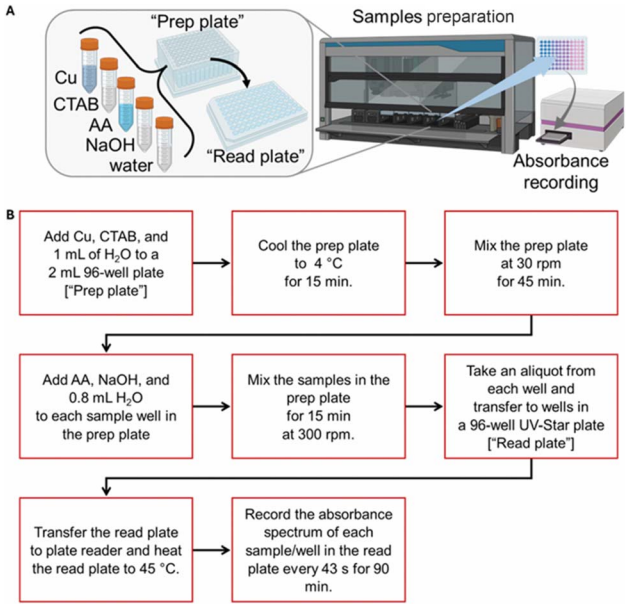

Fig. 1 Experimental robotic synthesis of the copper nanoclusters (A). Experimental setup and (B). Schematic for dispensing the reagents at various molar concentration (see SI Table 1) and subsequent absorbance measurement of CuNCs.

rates, and atmospheric composition can be configured, while ECL also provides the capacity to regulate parameters inherently difficult for human operators during liquid handling, including pipette aspiration and dispensing rates, mixing and equilibration times, and pipette angle and height. Throughout the experiment, critical variables are continuously logged, including these pipetting parameters (SI Fig. 2) and ambient temperature and humidity. After protocol completion, a comprehensive dataset is generated containing full video recordings (SI Video 1), metadata including start/end times and step durations, final volumes, images of stock solutions and resulting samples, recent instrument calibrations, detailed substance handling flowcharts for traceability (SI Fig. 3), and theoretical chemical composition of each sample (SI Table 1) based on actual amounts dispensed and mixed—information later used as features to train the ML model (see features, outcomes and data normalization).

From our synthesis samples, representative absorbance spectra are shown in Fig. 2. In Fig. 2A (sample #17 in SI Table 1), the initial measurement displays a single broad peak around 400 nm, corresponding to the metallic ion complex formed between Cu and CTAB.[22,29] This peak gradually decreases and eventually disappears while absorbance in the 240–300 nm range simultaneously increases until stabilizing in the last 30 minutes. This spectroscopic behavior, combined with the absence of absorbance between 400 and 600 nm where surface plasmon resonance would typically appear, serves as a strong indicator of CuNCs formation, as reported in several studies.[28,30–33] This was further confirmed by the high fluorescence exhibited by successful samples (SI Fig. 4), a characteristic that is restricted to NC composed of fewer than ten copper

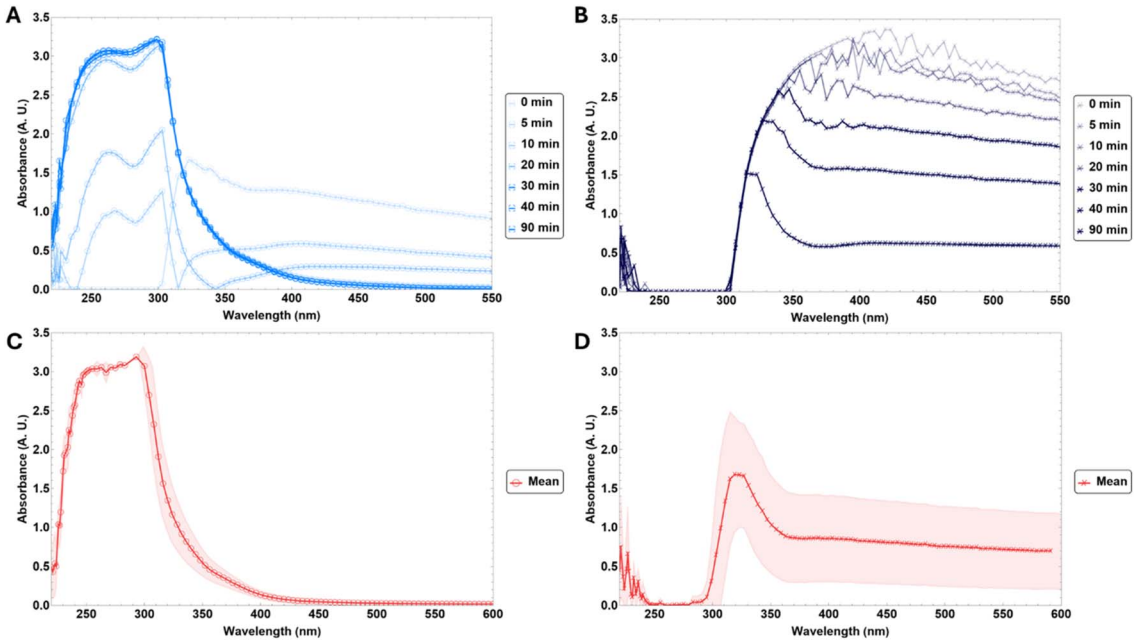

Fig. 2 Representative absorbance spectrum of copper nanoclusters. (A) Successful synthesis and absorbance spectra were recorded every 43 seconds (sample #17). (B) Unsuccessful synthesis, absorbance spectra were recorded every 43 seconds (sample #2). (C) Successful absorbance mean (red) and standard deviation (shaded area) of four batches at 40 minutes (sample #17). (D) Unsuccessful absorbance mean (red) and standard deviation (shaded area) of four batches at 40 minutes (sample #2).

atoms due to discrete electronic transitions enabled by their small size and ligand interactions.[21,34–36] This characteristic was confirmed in only the successful samples, which exhibited high fluorescence (SI Fig. 4). In contrast, Fig. 2B (corresponding to sample #10) shows only the metallic complex absorbance with no characteristic CuNCs signal, indicating an unsuccessful reaction.

To assess reproducibility, samples were synthesized in four different batches. Fig. 2C and 2D show the mean (red) and standard deviation (shaded area) of the absorbance spectra for the successful and unsuccessful samples, respectively, with little detectable difference between them across batches, which resulted in a CV in absorbance intensity of 10–20% for successful samples and 15–25% for unsuccessful samples in the 230–300 nm range (SI Fig. 5). The consistency of these results demonstrates the high reproducibility of this method and the consistency of the collected data. Furthermore, we successfully reproduced the synthesis of sample #17 at a different laboratory location (Carnegie Mellon University (CMU) Automated Lab in Pittsburgh, PA) using the identical robotic protocol and instrumentation and operated remotely through ECL software, with no significant differences observed in the resulting spectra (SI Fig. 6), along with a CV below 50% in the same range of study (SI Fig. 7). We note that the CV was based on the absorbance intensity which may vary due to the age of the stock solutions, especially AA. However, the absorbance peak position that indicated the formation of CuNCs did not vary in any result, demonstrating the high reproducibility of our automated approach.

For samples #3 and #4, the molar concentrations (see SI Table 1) were directly adapted from literature protocols originally developed for gold sphere and rod morphologies, respectively.[16,37] Interestingly, their absorbance spectra (Fig. 3) suggest the possibility of analogous morphological control in CuNCs, though this requires further investigation. Sample #3 exhibits a single peak at 275 nm, reminiscent of the spectroscopic signature typically associated with spherical copper nanostructures.[38,39] In contrast, sample #4 displays two distinct peaks in the same region at 30 minutes, similar to the characteristic dual-peak pattern observed for rod-like nanoparticles. These dual peaks tend to disappear in later reaction stages, possibly due to agglomeration (SI Fig. 8), a behavior consistent with previous copper nanoparticle studies.[40,41]

3.2 Features, outcomes, and data normalization

A key advantage of our standardized robotic synthesis protocol is the ability to modify synthesis parameters with high precision. The molar concentrations between Cu, CTAB, and AA were varied by adding different volumes of stock solutions with identical concentrations. The composition of the sample in each well-determined using ECL software and accounting for actual reagent dispensing and dilution, served as the feature set. To ensure comparability, these compositions were standardized (transformed to have a mean of 0 and standard deviation of 1), which helps improve model performance by making feature scales consistent.

The absorbance signal between 240 and 300 nm is associated with CuNCs formation.[42–44] Therefore, the absorbance ratio at 264 and 294 nm, measured at 40 minutes when the reaction was

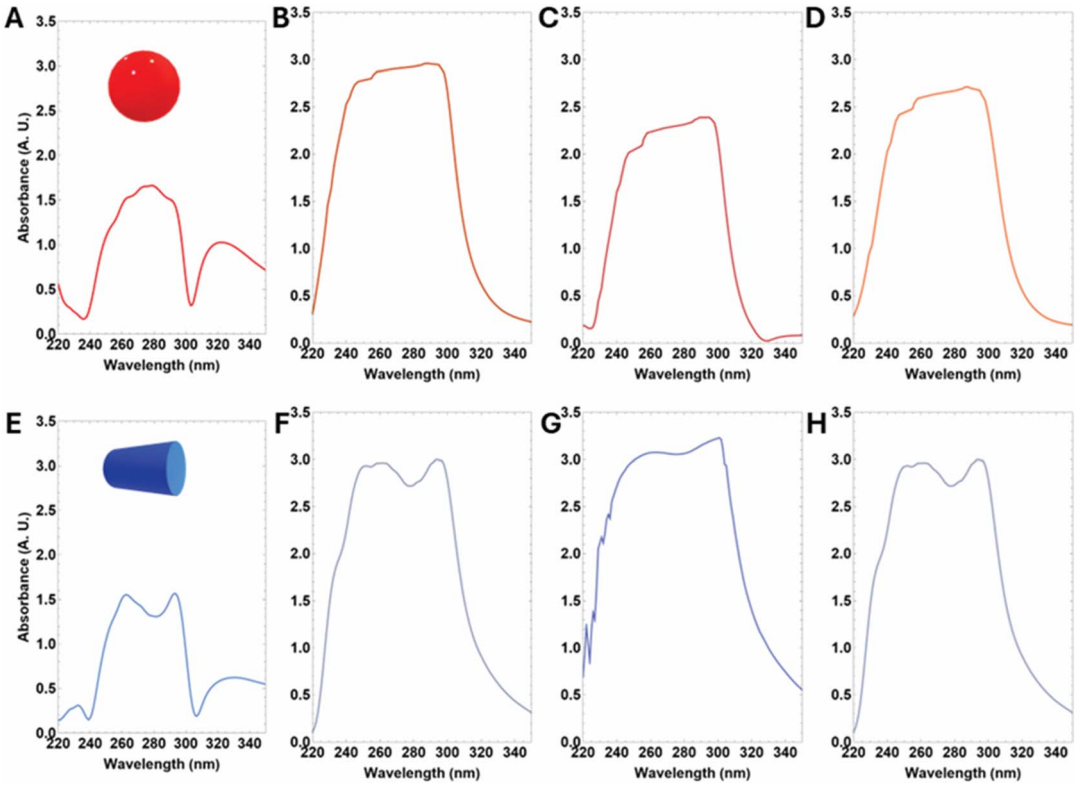

Fig. 3 Absorbance spectra of copper nanoclusters with likely different morphologies, with insets illustrating the suggested shapes. (A–D) Spherical particles (samples #3, #18, #20, and #22, respectively). (E–H) Rod-like particles (samples # 4, #14, #17, and #24, respectively).

nearly complete, was selected as the output metric for training the ML models. Ratios approaching zero corresponded to unsuccessful synthesis (no CuNCs formation), whereas values closer to one indicated successful CuNCs formation.

In the first stage, ten control samples were synthesized to explore possible molar concentrations between Cu, CTAB, and AA. Four ratios were selected from literature sources (samples #1 to #4),[32,45] while the remaining six involved incremental increases of CTAB and AA with all other reagents held constant (samples #5 to #10). This initial approach provided insight into the most relevant reagents concentration for CuNCs formation.

After identifying synthesis parameter boundaries, a second stage was conducted with ten new control samples, selected similarly to the first stage but with fewer variations to refine the sample space understanding (samples 11–20). Additionally, 20 new samples[21–40] were synthesized using Latin Hypercube Sampling (LHS), which ensures even coverage of each dimension in the input space, improving the exploration of molar concentration. Finally, in the third stage, six additional random samples were generated exclusively for model validation, ensuring an independent test set to assess the predictive performance of the machine learning model.

In total, 40 samples were synthesized. Half the samples[1–20] were selected using traditional experimental design, incorporating literature molar concentrations and systematic increments of single reagents. The remaining 20 samples[21–40] were systemically distributed using LHS, ensuring a broader exploration of the parameter space.

Fig. 4 shows a 3D representation of the composition between Cu, CTAB, and AA. The red dots indicate unsuccessful samples, the green dots represent successful samples, and the blue dots correspond to validation samples used for model validation. In Fig. 4A, the composition ratios are displayed on a scale that includes all samples, including First Stage control samples. This visualization highlights the limitations of a classical experimental approach, where systematically increasing reagent concentrations makes it difficult to fully explore the entire sample space. In contrast, Fig. 4B illustrates the distribution of samples on a focused range, displaying only synthesis conditions within the relevant boundaries for NC formation. This method results in a more evenly distributed exploration of the parameter space, ensuring a more comprehensive representation of potential synthesis conditions.

3.3 ML model training and validation

Classical ML approaches such as Linear Regression, Nearest Neighbors, Decision Trees, Random Forest, and Support Vector Machines are preferred for small datasets like ours because these models are interpretable and less prone to overfitting.[2] Table 1 lists the evaluated ML models—that predicts the relative absorbance at 264 and 294 nm—all trained with the same 40 samples and validated with six independent samples, along with key performance metrics: root mean squared error (RMSE),

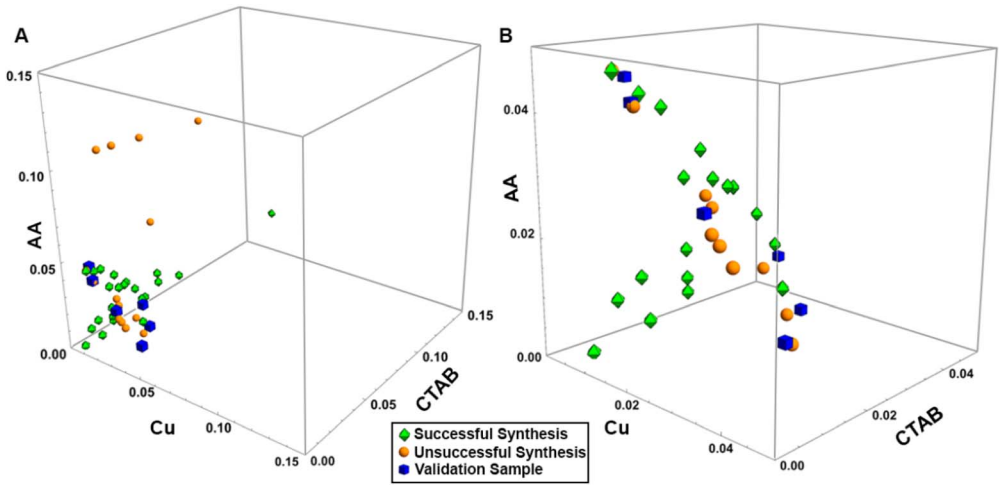

Fig. 4 3D Representation of the same sample composition at different ranges of reagents: Ascorbic Acid (AA), Cetyltrimethylammonium bromide (CTAB), and Copper sulphate (Cu) molar ratios. (A) A broader range that includes all composition ratios, including control samples. (B) A focused range, displaying only synthesis conditions within the relevant boundaries for nanocluster formation.

Table 1 Performance of evaluated models for predicting the relative absorbance at 264 and 294 nm (nanocluster formation). Reported metrics include root mean squared error (RMSE), coefficient of determination (R^2), mean absolute error (MAE). Results are compared against a standard deviation baseline ($Stdv_{baseline}$ = 0.308). The poor performance of the Neural Network model reflects overfitting due to the small dataset size, demonstrating that simpler models are more appropriate for limited training data"

Model	RMSE	MAE	R^2
Decision tree	0.118	0.092	0.853
Linear regression	0.131	0.091	0.820
Random forest	0.151	0.099	0.760
Nearest neighbours	0.163	0.137	0.720
Gradient boosted trees	0.164	0.128	0.716
Gaussian process	0.249	0.191	0.347
Neural network	0.338	0.279	0.200

mean absolute error (MAE), and coefficient of determination (R^2).

The RMSE is compared against the standard deviation baseline ($Stdv_{baseline}$) to assess model utility—if RMSE exceeds $Stdv_{baseline}$, using the mean as a predictor would be preferable, whereas a lower RMSE indicates superior predictive performance. The R^2 value indicates the fraction of variance explained by the model, with a perfect model achieving a value of 1. ML models trained only on LHS and control samples respectively were validated with the same six unseen samples (SI Fig. 9). The best LHS-trained model was LR achieved an RMSE of 0.073 (see SI Fig. 9A), while the best control-trained model had an RMSE of 0.214 (see SI Fig. 9A), highlighting the advantage of using LHS generated molar concentrations.

Two models demonstrated superior performance: Decision Tree (DT) and Linear Regression (LR), selected based on their R^2 values approaching 1 and low RMSE scores. Fig. 5 presents a comparison of DT and LR predictions *versus* actual validation values, with the dashed diagonal line representing the ideal regression where perfect predictions would align. Additionally, SHAP (SHapley Additive exPlanations) value distributions are shown for all features, providing insights into each input variable's contribution to model predictions—values greater than zero indicate a positive impact on CuNCs formation, while a negative value indicates inhibitory effects.

Fig. 5A and B demonstrate that predicted values from both models closely align with the ideal regression line without substantial outliers, confirming strong predictive performance. Given the comparable performance between DT and LR, model selection required additional considerations beyond these metrics. Importantly, when the same validation samples were synthesized at a different location (Pittsburgh) following the identical protocol from the ECL facility at Austin, the LR model maintained reasonable performance (SI Fig. 10) with an RMSE of 0.227. This cross-laboratory validation demonstrates both the reproducibility of our synthesis methodology and the model's capacity to generalize beyond the original training environment.

Feature importance analysis using SHAP values obtained from LR model (Fig. 5C) revealed that AA concentration is the most critical parameter, consistent with first-stage synthesis observations where all samples containing more than 2 mM of AA resulted in unsuccessful synthesis. This behavior may be attributed to competitive interactions between protons and copper ions within the micelle environment.[34,46] Cu concentration emerged as the second most important feature, which is expected since excessive Cu concentration promotes agglomeration, favoring nanoparticle formation over discrete NCs. CTAB concentration showed a smaller but significant impact, reflecting its crucial role in facilitating stable micelle formation during the initial low-temperature mixing step – a prerequisite for successful NCs formation.[47,48]

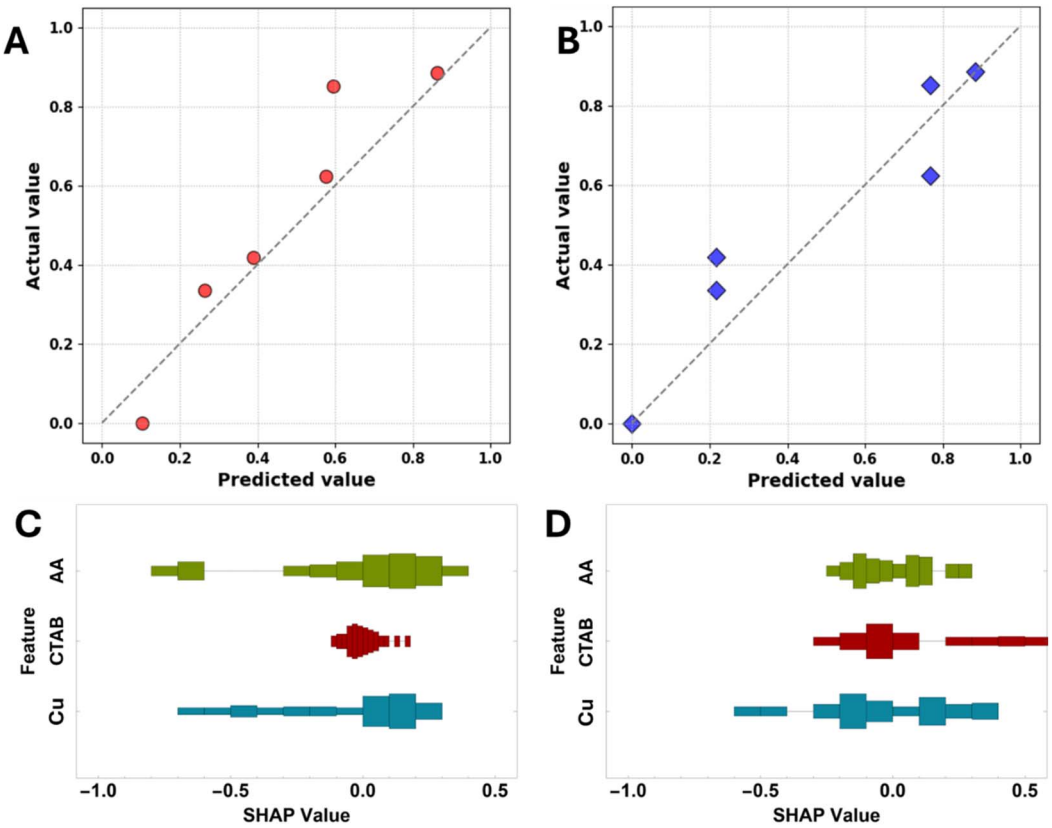

Fig. 5 Comparison plots of the predicted values *versus* validation values and SHAP value distributions for all features, Ascorbic Acid (AA), Cetyltrimethylammonium bromide (CTAB), and Copper sulphate (CuSO4) of different models. (A and C) Linear regression. (B and D) Decision tree.

4 Conclusions

This study establishes a validated framework for integrating machine learning with automated synthesis protocols to predict and control CuNCs formation—a significantly underexplored area compared to extensively studied gold NCs systems. By systematically exploring reagent molar concentrations (Cu, CTAB, and ascorbic acid), we demonstrate that high-quality predictive models can be developed from strategically designed small datasets when coupled with rigorous experimental control.

4.1 Methodological advances and cross-laboratory validation

Our approach transcends traditional trial-and-error synthesis by combining LHS with automated liquid handling and data collection at multiple laboratories. This integration achieved experimental reproducibility, with successful cross-laboratory validation between ECL (Austin, TX) and CMU Automated Science Lab (Pittsburgh, PA) confirming model robustness across different experimental environments. Crucially, our protocols maintained consistency not only across laboratories but also across multiple instruments within each facility—utilizing different Hamilton Liquid Handler units (six total instruments) and CLARIOstar spectrometers (four total instruments)—demonstrating true multi-instrument reproducibility. This multi-instrument validation is critical for ensuring research findings are robust and transferable across different experimental setups. The automated protocols eliminated both operator-dependent and instrument-specific variability, establishing new standards for data consistency and reproducibility in materials discovery.

4.2 Strategic ML implementation for small datasets

Our results demonstrate that sophisticated insights can be extracted from compact, high-quality datasets through intelligent feature engineering. Decision Trees and Linear Regression achieved robust predictive performance ($R^2 > 0.82$, RMSE < 0.15) from just 40 training samples while maintaining interpretability—crucial for understanding underlying chemical mechanisms. SHAP analysis revealed ascorbic acid concentration as the dominant synthesis parameter, providing actionable chemical insights that extend beyond correlation to mechanism-based understanding. Our success with minimal data requirements has significant implications as it suggests that we can develop predictive models with only 40 samples, fewer experiments than traditional approaches, dramatically reducing costs and the experimental burden for developing predictive models. Our approach is particularly valuable for expensive synthesis protocols, rare materials, or hazardous

chemistries where large dataset generation may be impractical or prohibitive.

4.3 Chemical insights and knowledge gap

This work addresses a significant knowledge gap in CuNCs synthesis. While gold NCs have extensive documentation, CuNCs remain underexplored despite advantages in cost, abundance, and catalytic properties. Our systematic exploration revealed critical threshold effects and non-linear relationships between reagent ratios and formation outcomes. The identification of ascorbic acid as the dominant parameter, coupled with CTAB stabilization mechanisms, establishes fundamental design principles for CuNCs synthesis and fills crucial gaps in copper-based NCs chemistry.

4.4 Transformative implications

Our current implementation represents a supervised learning approach, while the validated predictive models establish a necessary foundation for future active learning implementations. Specifically, our models' demonstrated ability to predict synthesis outcomes from just 40 samples could be integrated with Bayesian optimization or other active learning frameworks to guide experimental selection, potentially reducing the required experimental budget by an order of magnitude compared to grid searches. Recent parallel optimization methods developed for cloud laboratories[49] demonstrate the feasibility of such approaches, though their application to materials synthesis remains unexplored. Integration with cloud laboratory infrastructure creates a scalable model for democratizing advanced materials research, enabling researchers without specialized facilities to access automated synthesis capabilities. Our multi-instrument, cross-laboratory validation proves that this approach can maintain reproducibility across different experimental environments while reducing resource consumption. The demonstrated success with CuNCs, validated against parallel gold nanoparticle methodologies, suggests extensibility to other nanomaterial systems. With this validated methodology for combining automated synthesis with ML prediction, we provide a foundation that can be built upon to develop closed-loop optimization systems, ultimately enabling the transition from empirical screening to rational, predictive approaches in materials design.

Author contributions

Ricardo Montoya-Gonzalez and Subha R. Das: these authors designed the project, performed the investigation, and wrote the manuscript. Rosa de G. González-Huerta, Martha L. Hernández-Pichardo and Subha R. Das jointly supervised this work.

Conflicts of interest

The authors declare no competing interests.

Data availability

The supplementary information (SI) included supports the findings presented in the main manuscript. It contains detailed experimental data, additional figures, and visual resources that document the synthesis protocols, validation experiments, and data structure used in the study. Additionally, a video of the liquid handling synthesis system, along with all the absorbance data for the samples, as well as the code used for training and validating the machine learning model are hosted on KiltHub, the Carnegie Mellon University institutional repository and freely accessible DOI: **https://doi.org/10.1184/R1/30062662**. The code also includes the experimental protocol used at both locations. Supplementary information is available. See DOI: **https://doi.org/10.1039/d5dd00335k**.

Acknowledgements

The authors acknowledge Emerald Cloud Lab, Inc. and Deans of the Mellon College of Science and the Office of the VP for Research at Carnegie Mellon University for access to both the automated laboratories ECL (Austin, TX) and the CMU Automated Science Lab (Pittsburgh, PA) including the facilities, instruments, and operators that enabled the remotely programmed experiments presented in this work. SRD and RMG thank Dr Ben Kline of ECL for assistance with the ECL software platform and data parsing. The authors acknowledge Profs. John Kitchin, Newell Washburn and Russell Schwartz for helpful comments and feedback on the manuscript, and Dr Huajin Wang of CMU Libraries for assistance with the KiltHub repository for open access to the data. RMG acknowledges SECIHTI for a doctoral fellowship, as well as the National Polytechnic Institute's International Relations Department for supporting the research visit to CMU. He also thanks the Department of Chemical Engineering at CMU and Professor Ignacio E. Grossmann for hosting the research stay for this project.

References

1 A. E. A. Allen and A. Tkatchenko, Machine learning of material properties: Predictive and interpretable multilinear models, *Sci. Adv.*, 2022, **8**, eabm7185.
2 A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson and T. D. Sparks, Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices, *Chem. Mater.*, 2020, **32**, 4954–4965.
3 C. C. Price, Y. Li, G. Zhou, R. Younas, S. S. Zeng, T. H. Scanlon, J. M. Munro and C. L. Hinkle, Predicting and Accelerating Nanomaterial Synthesis Using Machine Learning Featurization, *Nano Lett.*, 2024, **24**, 14862–14867.
4 F. Q. Imran, D.-H. Kim, S.-J. Bong, S.-Y. Chi and Y.-H. Choi, A Survey of Datasets, Preprocessing, Modeling Mechanisms, and Simulation Tools Based on AI for Material Analysis and Discovery, *Materials*, 2022, **15**, 1428.

5 M. L. Lau, A. Burleigh, J. Terry and M. Long, Materials characterization: Can artificial intelligence be used to address reproducibility challenges?, *J. Vac. Sci. Technol., A*, 2023, **41**, 060801.
6 K. M. Jablonka, D. Ongari, S. M. Moosavi and B. Smit, Big-Data Science in Porous Materials: Materials Genomics and Machine Learning, *Chem. Rev.*, 2020, **120**, 8066–8129.
7 X. Yu, S. Dutta, J. Andreo and S. Wuttke, Identifying synthetic variables influencing the reproducible microfluidic synthesis of ZIF nano- and micro-particles, *Commun. Mater.*, 2024, **5**, 210.
8 S. N. Steinmann, Q. Wang and Z. W. Seh, How machine learning can accelerate electrocatalysis discovery and optimization, *Mater. Horiz.*, 2023, **10**, 393–406.
9 H. Tao, T. Wu, M. Aldeghi, T. C. Wu, A. Aspuru-Guzik and E. Kumacheva, Nanoparticle synthesis assisted by machine learning, *Nat. Rev. Mater.*, 2021, **6**, 701–716.
10 A. S. Nugraha, G. Lambard, J. Na, M. S. A. Hossain, T. Asahi, W. Chaikittisilp and Y. Yamauchi, Mesoporous trimetallic PtPdAu alloy films toward enhanced electrocatalytic activity in methanol oxidation: unexpected chemical compositions discovered by Bayesian optimization, *J. Mater. Chem. A*, 2020, **8**, 13532–13540.
11 T. Kashiwagi, K. Sue, Y. Takebayashi and T. Ono, High-throughput synthesis of silver nanoplates and optimization of optical properties by machine learning, *Chem. Eng. Sci.*, 2022, **262**, 118009.
12 D. Schletz, M. Breidung and A. Fery, Validating and Utilizing Machine Learning Methods to Investigate the Impacts of Synthesis Parameters in Gold Nanoparticle Synthesis, *J. Phys. Chem. C*, 2023, **127**, 1117–1125.
13 H. T. Chiang, K. Vaddi and L. Pozzo, Data-driven exploration of silver nanoplate formation in multidimensional chemical design spaces, *Digital Discovery*, 2024, **3**, 2252–2264.
14 K. Vaddi, H. T. Chiang and L. D. Pozzo, Autonomous retrosynthesis of gold nanoparticles *via* spectral shape matching, *Digital Discovery*, 2022, **1**, 502–510.
15 B. Tang, Y. Lu, J. Zhou, T. Chouhan, H. Wang, P. Golani, M. Xu, Q. Xu, C. Guan and Z. Liu, Machine learning-guided synthesis of advanced inorganic materials, *Mater. Today*, 2020, **41**, 72–80.
16 A. A. Guda, M. V. Kirichkov, V. V. Shapovalov, A. I. Muravlev, D. M. Pashkov, S. A. Guda, A. P. Bagliy, S. A. Soldatov, S. V. Chapek and A. V. Soldatov, Machine Learning Analysis of Reaction Parameters in UV-Mediated Synthesis of Gold Nanoparticles, *J. Phys. Chem. C*, 2023, **127**, 1097–1108.
17 I. Miyazato, S. Nishimura, L. Takahashi, J. Ohyama and K. Takahashi, Data-Driven Identification of the Reaction Network in Oxidative Coupling of the Methane Reaction *via* Experimental Data, *J. Phys. Chem. Lett.*, 2020, **11**, 787–795.
18 P. Karande, B. Gallagher and T. Y.-J. Han, A Strategic Approach to Machine Learning for Material Science: How to Tackle Real-World Challenges and Avoid Pitfalls, *Chem. Mater.*, 2022, **34**, 7650–7665.
19 A. Baghdasaryan and T. Bürgi, Copper nanoclusters: designed synthesis, structural diversity, and multiplatform applications, *Nanoscale*, 2021, **13**, 6283–6340.
20 Y. Lu, W. Wei and W. Chen, Copper nanoclusters: Synthesis, characterization and properties, *Chin. Sci. Bull.*, 2012, **57**, 41–47.
21 M. Lettieri, P. Palladino, S. Scarano and M. Minunni, Copper nanoclusters and their application for innovative fluorescent detection strategies: An overview, *Sens. Actuators Rep.*, 2022, **4**, 100108.
22 S. Maity, D. Bain and A. Patra, Engineering Atomically Precise Copper Nanoclusters with Aggregation Induced Emission, *J. Phys. Chem. C*, 2019, **123**, 2506–2515.
23 S. Shahsavari, S. Hadian-Ghazvini, F. Hooriabad Saboor, I. Menbari Oskouie, M. Hasany, A. Simchi and A. L. Rogach, Ligand functionalized copper nanoclusters for versatile applications in catalysis, sensing, bioimaging, and optoelectronics, *Mater. Chem. Front.*, 2019, **3**, 2326–2356.
24 T.-A. D. Nguyen, Z. R. Jones, B. R. Goldsmith, W. R. Buratto, G. Wu, S. L. Scott and T. W. Hayton, A Cu_{25} Nanocluster with Partial Cu(0) Character, *J. Am. Chem. Soc.*, 2015, **137**, 13319–13324.
25 R. N. Núñez, A. V. Veglia and N. L. Pacioni, Improving reproducibility between batches of silver nanoparticles using an experimental design approach, *Microchem. J.*, 2018, **141**, 110–117.
26 B. Hofko, L. Porot, A. Falchetto Cannone, L. Poulikakos, L. Huber, X. Lu, K. Mollenhauer and H. Grothe, FTIR spectral analysis of bituminous binders: reproducibility and impact of ageing temperature, *Mater. Struct.*, 2018, **51**, 45.
27 M. Zhu, E. Lanni, N. Garg, M. E. Bier and R. Jin, Kinetically Controlled, High-Yield Synthesis of Au_{25} Clusters, *J. Am. Chem. Soc.*, 2008, **130**, 1138–1139.
28 M. Alsawafta, S. Badilescu, M. Packirisamy and V.-V. Truong, Kinetics at the nanoscale: formation and aqueous oxidation of copper nanoparticles, *React. Kinet., Mech. Catal.*, 2011, **104**, 437–450.
29 M. Fernández-Ujados, L. Trapiella-Alfonso, J. M. Costa-Fernández, R. Pereiro and A. Sanz-Medel, One-step aqueous synthesis of fluorescent copper nanoclusters by direct metal reduction, *Nanotechnology*, 2013, **24**, 495601.
30 M. Schmallegger, H. Grützmacher and G. Gescheidt, Bis(acyl)phosphine Oxides as Stoichiometric Photo-Reductants for Copper Nanoparticle Synthesis: Efficiency and Kinetics, *ChemPhotoChem*, 2022, **6**, e202200155.
31 I. Lisiecki, F. Billoudet and M. P. Pileni, Control of the Shape and the Size of Copper Metallic Particles, *J. Phys. Chem.*, 1996, **100**, 4160–4166.
32 M. B. Gawande, A. Goswami, F.-X. Felpin, T. Asefa, X. Huang, R. Silva, X. Zou, R. Zboril and R. S. Varma, Cu and Cu-Based Nanoparticles: Synthesis and Applications in Catalysis, *Chem. Rev.*, 2016, **116**, 3722–3811.
33 S.-H. Wu and D.-H. Chen, Synthesis of high-concentration Cu nanoparticles in aqueous CTAB solutions, *J. Colloid Interface Sci.*, 2004, **273**, 165–169.

34 C. Vázquez-Vázquez, M. Bañobre-López, A. Mitra, M. A. López-Quintela and J. Rivas, Synthesis of Small Atomic Copper Clusters in Microemulsions, *Langmuir*, 2009, **25**, 8208–8216.

35 S. Biswas, A. K. Das and S. Mandal, Surface Engineering of Atomically Precise M(I) Nanoclusters: From Structural Control to Room Temperature Photoluminescence Enhancement, *Acc. Chem. Res.*, 2023, **56**, 1838–1849.

36 J. Benavides, I. Quijada-Garrido and O. García, The synthesis of switch-off fluorescent water-stable copper nanocluster Hg2+ sensors *via* a simple one-pot approach by an *in situ* metal reduction strategy in the presence of a thiolated polymer ligand template, *Nanoscale*, 2020, **12**, 944–955.

37 A. Pakravan, R. Salehi and M. Mahkam, Comparison study on the effect of gold nanoparticles shape in the forms of star, hallow, cage, rods, and Si-Au and Fe-Au core-shell on photothermal cancer treatment, *Photodiagn. Photodyn. Ther.*, 2021, **33**, 102144.

38 C. Wang, Y. Yao and Q. Song, Interfacial synthesis of polyethyleneimine-protected copper nanoclusters: Size-dependent tunable photoluminescence, pH sensor and bioimaging, *Colloids Surf., B*, 2016, **140**, 373–381.

39 P.-C. Chen, Y.-C. Li, J.-Y. Ma, J.-Y. Huang, C.-F. Chen and H.-T. Chang, Size-tunable copper nanocluster aggregates and their application in hydrogen sulfide sensing on paper-based devices, *Sci. Rep.*, 2016, **6**, 24882.

40 Z. Wu, J. Liu, Y. Gao, H. Liu, T. Li, H. Zou, Z. Wang, K. Zhang, Y. Wang, H. Zhang and B. Yang, Assembly-Induced Enhancement of Cu Nanoclusters Luminescence with Mechanochromic Property, *J. Am. Chem. Soc.*, 2015, **137**, 12906–12913.

41 Z. Wu, H. Liu, T. Li, J. Liu, J. Yin, O. F. Mohammed, O. M. Bakr, Y. Liu, B. Yang and H. Zhang, Contribution of Metal Defects in the Assembly Induced Emission of Cu Nanoclusters, *J. Am. Chem. Soc.*, 2017, **139**, 4318–4321.

42 N. T. K. Thanh, N. Maclean and S. Mahiddine, Mechanisms of Nucleation and Growth of Nanoparticles in Solution, *Chem. Rev.*, 2014, **114**, 7610–7630.

43 N. Shahabadi, M. Hakimi, T. Morovati and N. Fatahi, DNA binding affinity of a macrocyclic copper(II) complex: Spectroscopic and molecular docking studies, *Nucleosides, Nucleotides Nucleic Acids*, 2017, **36**, 497–510.

44 D. Li, X. Lei, Q. Liu, Y. Chen, J. Wang, B. Han and G. He, CuNCs assisted by carbon dots to construct Z-type heterostructures: Ultra-high photocatalytic performance and peroxidase-like activity, *Sep. Purif. Technol.*, 2023, **322**, 124133.

45 G. Granata, T. Yamaoka, F. Pagnanelli and A. Fuwa, Study of the synthesis of copper nanoparticles: the role of capping and kinetic towards control of particle size and stability, *J. Nanopart. Res.*, 2016, **18**, 133.

46 G. N. Glavee, K. J. Klabunde, C. M. Sorensen and G. C. Hadjipanayis, Borohydride Reduction of Nickel and Copper Ions in Aqueous and Nonaqueous Media. Controllable Chemistry Leading to Nanoscale Metal and Metal Boride Particles, *Langmuir*, 1994, **10**, 4726–4730.

47 M. S. Saterlie, H. Sahin, B. Kavlicoglu, Y. Liu and O. A. Graeve, Surfactant Effects on Dispersion Characteristics of Copper-Based Nanofluids: A Dynamic Light Scattering Study, *Chem. Mater.*, 2012, **24**, 3299–3306.

48 J. Rozra, I. Saini, A. Sharma, N. Chandak, S. Aggarwal, R. Dhiman and P. K. Sharma, Cu nanoparticles induced structural, optical and electrical modification in PVA, *Mater. Chem. Phys.*, 2012, **134**, 1121–1126.

49 T. S. Frisby, Z. Gong and C. J. Langmead, Asynchronous parallel Bayesian optimization for AI-driven cloud laboratories, *Bioinformatics*, 2021, **37**, i451–i459.