UHASSELT

KNOWLEDGE IN ACTION

Faculteit Wetenschappen

master in materiomics

Masterthesis

Machine Learning Small Datasets for Cu Nanoparticles: enhancing Experimental and Computational lab-scale data

Brent Motmans

Scriptie ingediend tot het behalen van de graad van master in materiomics

PROMOTOR :
Prof. dr. dr. Danny VANPOUCKE
Prof. dr. An HARDY

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

2024
2025

UHASSELT

KNOWLEDGE IN ACTION

Faculteit Wetenschappen

master in materiomics

Masterthesis

Machine Learning Small Datasets for Cu Nanoparticles: enhancing Experimental and Computational lab-scale data

Brent Motmans

Scriptie ingediend tot het behalen van de graad van master in materiomics

PROMOTOR :

Prof. dr. dr. Danny VANPOUCKE

Prof. dr. An HARDY

Machine Learning Small Datasets for Cu Nanoparticles: enhancing Experimental and Computational lab-scale data

Brent Motmans[1,2,], Digvijay Ghogare[1,3,4], Thijs van Wijk[2,3], An Hardy[1,3] and Danny E.P. Vanpoucke[2,3]*

[1] Hasselt University, Institute for Materials Research (imo-imomec), DESINe group, Martelarenlaan 42, B-3500 Hasselt, Belgium

[2] Hasselt University, Institute for Materials Research (imo-imomec), Quantum & Artificial inTelligence design Of Materials (QuATOMs), Martelarenlaan 42, B-3500 Hasselt, Belgium

[3] imo-imomec, Wetenschapspark 1, B-3590 Diepenbeek, Belgium

[4] Electrochemistry Excellence Centre (ELEC), Materials & Chemistry Unit, Flemish Institute for Technological Research (VITO), Boeretang 200, Mol, 2400 Belgium

E-mail: brent.motmans@student.uhasselt.be

Keywords: machine learning, microwave-assisted polyol synthesis, small datasets, copper nanoparticles

Abstract: Copper nanoparticles (Cu NPs) have a broad applicability, yet their synthesis is sensitive to subtle changes in reaction parameters. This sensitivity, combined with the time- and resource-intensive nature of experimental optimisation, poses a major challenge in achieving reproducible and size-controlled synthesis. Additionally, while Machine Learning (ML) shows promise in materials research, its application is often limited by scarcity of high-quality experimental datasets. This study explores ML to predict the size of Cu NPs from microwave-assisted polyol synthesis using small datasets generated from 25 in-house syntheses. Latin Hypercube Sampling is applied to efficiently cover the parameter space of precursor concentration, temperature, and reaction time. Ensemble regression models, built with the AMADEUS framework, successfully predict particle sizes with high accuracy, outperforming classical statistical approaches. Feature selection reduces model complexity and improves generalisability. Additionally, classification models using both random forests and Large Language Models (LLMs) are evaluated to distinguish between large and small particles.

While random forests show moderate performance, LLMs offer no significant advantages under data-scarce conditions. Overall, this study demonstrates that carefully curated small datasets, paired with robust classical ML, can effectively predict the synthesis of Cu NPs and highlights that for lab-scale studies, complex models like LLMs offer limited benefit over simpler techniques.

Dutch Abstract: Koper-nanodeeltjes (Cu NP's) hebben een brede toepasbaarheid, maar hun synthese is gevoelig voor kleine veranderingen in reactieparameters. Deze gevoeligheid, in combinatie met het tijdrovende en arbeidsintensieve karakter van experimentele optimalisatie, vormt een grote uitdaging voor reproduceerbare synthese met gecontroleerde deeltjesgrootte. Bovendien is *Machine Learning* (ML) weliswaar veelbelovend gebleken voor materiaalonderzoek, maar wordt toepassing ervan vaak beperkt door gebrek aan hoogwaardige experimentele datasets. Deze studie onderzoekt ML om de grootte van Cu NP's gevormd met microgolf geassisteerde polyolsynthese te voorspellen met kleine datasets, gegenereerd uit 25 intern uitgevoerde syntheses. *Latin Hypercube Sampling* wordt gebruikt om de parameterruimte van precursorconcentratie, temperatuur en reactietijd efficiënt te samplen. Ensemble-regressiemodellen, gebouwd met het *AMADEUS-framework*, voorspellen met hoge nauwkeurigheid de deeltjesgrootte en presteren daarmee beter dan klassieke statistische benaderingen. Featureselectie vermindert complexiteit van het model en verbetert generaliseerbaarheid. Daarnaast worden classificatiemodellen, gebaseerd op zowel traditionele *random forests* als *Large Language Modellen* (LLM's), geëvalueerd om onderscheid te maken tussen grote en kleine deeltjes. Terwijl *random forests* matig presteren, bieden LLM's geen duidelijke verbeteringen in omstandigheden met weinig gegevens. Over het algemeen toont dit onderzoek dat zorgvuldig samengestelde kleine datasets, in combinatie met robuuste klassieke ML, synthese van Cu NP's effectief kan voorspellen en dat voor laboratoriumonderzoek complexe LLM's geen voordelen bieden.

1. Introduction

In recent years, metal nanoparticles (NPs) have become key elements in a wide range of technological applications due to their size- and shape-dependent properties, which can differ significantly from the bulk properties.[1-4] These typical properties, ranging from confinement and surface plasmon resonance (SPR) to increased catalytic activity, make metal NPs widely applicable in electronics[2], catalysis[3] and medicine[4]. Controlling these properties, however, requires very precise control over various characteristics of the NPs, including their morphology (*i.e.,* size and shape), composition, and surface chemistry down to the nanometre scale. Achieving this level of control remains one of the biggest challenges in the synthesis of NPs, as small variations in the synthesis can lead to final particles with entirely different properties. In classical approaches, such synthesis procedures are developed based on an iterative process and empirically guided decisions, rather than relying on data-driven methodologies, resulting in a time-consuming and labour-intensive process. As the complexity of nanomaterials continues to increase, there is a need for more systematic and efficient methods to optimise and develop synthesis processes.[5]

To tackle the challenges in NP synthesis and accelerate the development of synthesis procedures, Artificial intelligence (AI) and machine learning (ML) are playing an increasingly important role (**Figure 1a**).[6] This increasing trend in the use of AI and ML for scientific research was recently confirmed by the 2024 Nobel Prizes in Physics[7] and Chemistry[8] for AI-related research. The growing impact of these technologies is mainly driven by advances in computing power, which have enabled a rapid expansion of AI and ML applications in various disciplines. These technologies are causing significant changes in research methodologies, especially in the field of materials science, where they offer tools for accelerating discoveries, improving optimisation and guiding experimental design.[9-12] One of their most important advantages in materials science is enabling predictive modelling and uncovering structure-property relationships, which significantly accelerates the discovery, optimisation, and development of synthesis procedures from scratch for new materials.[9-12] Furthermore, it opens up new possibilities for understanding complex physicochemical phenomena that were previously difficult to analyse.[9-12] Scientists are increasingly trying to take advantage of these powerful tools and are realising the potential to significantly improve synthesis development and the possibilities to adapt material properties faster to specific needs.[13]

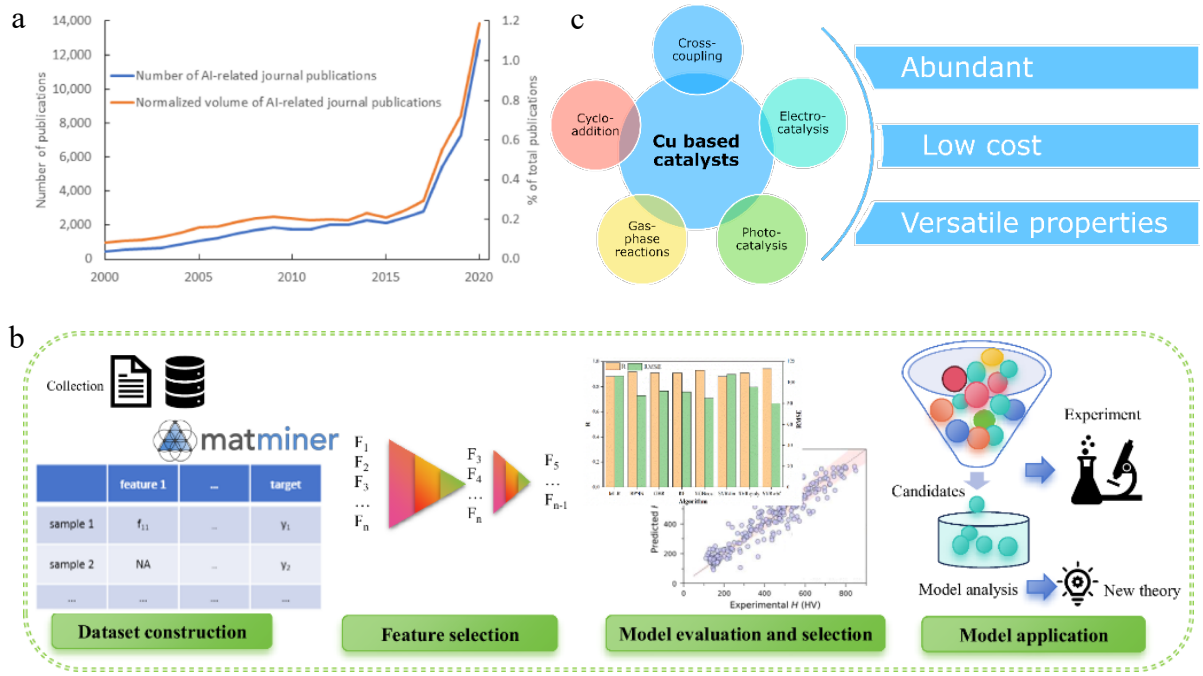

Figure 1 *- a) Evolution of AI related publications in chemistry research[6]; b) Workflow for ML use in chemistry research[18]; c) Cu NP applications.*

However, the growing interest in data-driven synthesis development is not just the result of increasing technological capabilities. It is also a direct response to a fundamental bottleneck in inorganic synthetic development. Inorganic synthesis method development is often done empirically and lacks a deep mechanistic insight.[14] Traditional development of synthesis methods relies on a one-variable-at-a-time (OVAT) method. This is a slow and inefficient approach that often struggles to identify the complex and high-dimensional interactions in the parameter space.[14,15] Thus, in this context, ML can play an important role in better identifying the complex interactions between parameters via a data-driven approach to accelerate synthesis development.[15]

In this study, the use of ML techniques for predictive inorganic NP synthesis, using small lab-scale datasets, is investigated. This includes a comparison of the performance of different ML techniques using small lab-scale datasets.

1.1. AI and ML for experimental Copper NP synthesis

AI encompasses all systems that imitate human intelligence for tasks like pattern recognition and decision-making, while ML is a specific subset of AI that focuses on algorithms that can learn from data and improve their performance over time (**Figure 2**). Zooming in further on ML, a distinction can be made between classic ML algorithms and Deep Learning algorithms. Classic ML includes regression-based models and Random Forest algorithms. Deep

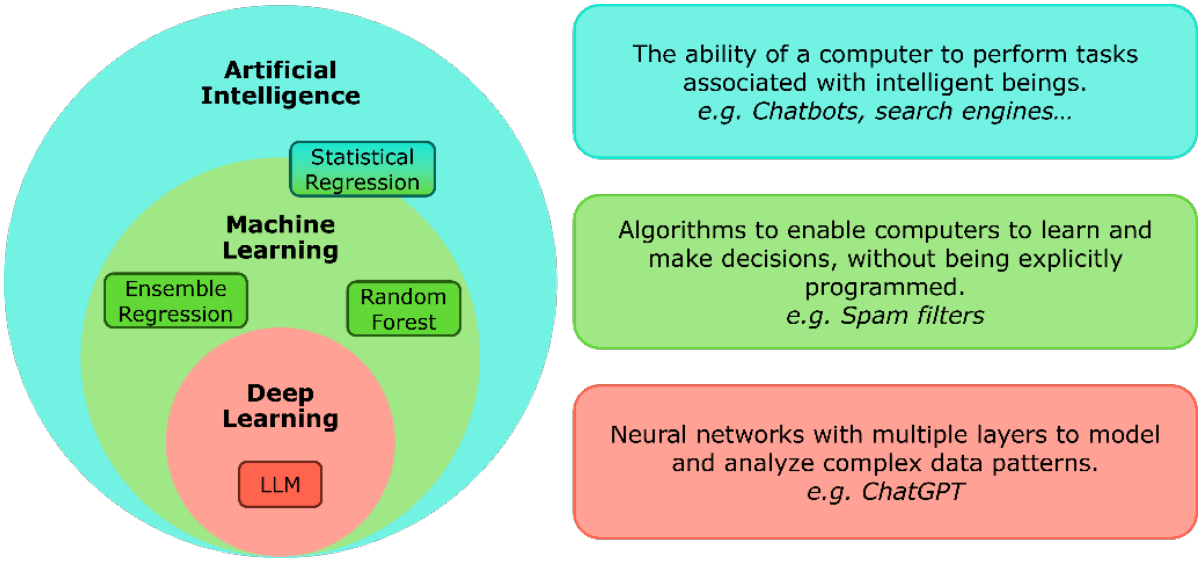

Figure 2 *– Definitions and relationships of AI, ML and Deep Learning.*

Learning models are based on so-called Neural Networks (NN).[16] Well-known examples of deep learning include Large Language Models (LLMs) such as ChatGPT. Besides the recent developments in AI, classical Design-of-Experiments (DoE) has evolved as a statistical methodology for systematically planning and analysing experiments.[17]

Among the various areas being transformed by AI and ML, inorganic synthesis is one in which ML can offer a great deal of assistance (**Figure 1b**).[18] Due to the high sensitivity to small variations in the reaction conditions and the wide variety of possible outcomes, developing these synthesis methods is often complex and time-consuming.[5] By using ML, it becomes easier to delve deeper into the complexity of an inorganic synthesis process to find out which specific parameters have a significant influence on the synthesis outcome. Guda *et al.* demonstrated the use of ML in gold NP synthesis to classify synthesis outcomes (success, shape, and formation kinetics) based on the used concentrations of gold precursor, reducing agent, and surfactant.[19] In addition, ML also allows the use of a closed-loop system, where an ML model automatically launches new experiments based on the results of previous experiments to speed up the development of a synthesis procedure and enable high-throughput experimentation. Since different parameters are often connected in a complex way, it can be very time and resource-consuming to find the optimal combination of parameters via a OVAT method. Here, ML models can be used to make suggestions for the next set of experiments based on data from a previous one, and so arrive more quickly at the most optimal parameters (*i.e.,* to obtain a desired outcome).[20,21]

Specifically, in NP synthesis, precise control over size, shape, and composition is essential for tailoring properties to specific applications. Because of their numerous applications in catalysts, there is currently a strong focus on investigating the synthesis of nanoparticles with specific sizes and shapes.[3] The morphology of the formed nanoparticles is often sensitive to small changes in the synthesis parameters. As a result, the synthesis of a specific morphology requires time-intensive development of the synthesis process. Currently, a OVAT method is often used for this purpose. Using ML for synthesis development can provide a more efficient alternative to the OVAT approach.[22] Pellegrino *et al.* showed how a ML model can be used to predict the synthesis of TiO_2 NPs. They used NNs to predict the size, dispersion, and aspect ratio of the TiO_2 NPs based on the used synthesis parameters. Through reverse engineering, they obtained the most optimal synthesis parameters. Therefore, the NNs were coupled to a genetic algorithm.[23] Another example of ML applied to NP synthesis is the work by Williams et al., who used a combination of DoE and ML to predict the yield of Ag nanowires in a polyol-based flow reactor. Their models, based on random forests, achieved high accuracy in predicting reaction yield based on the used reaction parameters.[24]

As a practical case study, the prediction of Copper (Cu) NP synthesis is used in this study. Cu NPs hold significant promise in catalysis due to their abundance, cost-effectiveness, and versatile properties (**Figure 1c**). An application of Cu NPs as photocatalysts can be found in the coupling of nitroaromatics, where it gives an alternative to classical reactions with diazonium salts that generate a lot of organic waste.[25] Another example is the use of Cu NPs in the catalytic electroreduction of CO_2. Reske *et al.* showed that decreasing the size of Cu NPs significantly improves their catalytic activity and shifts the product distribution toward H_2 and CO, while suppressing hydrocarbon formation.[26] Through various synthesis methods, it is possible to make different types of Cu NPs. For instance, the oxidation state of Cu can differ, and particles of different shapes and sizes can be synthesised.[3] Therefore, the wide variety of possible synthesis outcomes, as well as the large parameter space, complicates developing a synthesis procedure to obtain specific properties. Due to their promising functional properties and complex synthesis, Cu NPs are used in this study as a model system to investigate the use of ML for synthesis prediction.

1.2. Overcoming critical challenges for ML-enhanced experimentation

While the potential of ML models to streamline the synthesis of Cu NPs is clear, their successful application depends on overcoming a critical challenge: the need for robust and

qualitative datasets. For inorganic synthesis, the available data is often limited and frequently has a strong bias, as only positive synthesis methods are usually published.[27,28] Therefore, collecting a dataset via data and text mining from scientific publications would result in an inconsistent dataset that does not properly describe the entire parameter space. Also, there are many different synthesis methods for making Cu NPs[3], so data mining would produce a dataset that is far too complex. While ideally, the full complexity of the system is represented, this is not realistic within the scope of a lab-scale study using small datasets. Including all synthesis variations would drastically increase the dimensionality of the parameter space, requiring far more data to train robust models. Because the performance and limits of ML models are mainly determined by the quality of the dataset used, it is crucial to work with high-quality and reliable data.[18] For this reason, the dataset used in this study was created and synthesised entirely in-house. This allows for maximum control over both the quality and relevance of the data, which is essential for robust and reproducible model development.

1.2.1. Microwave-assisted polyol route for Cu NPs synthesis

Despite there being many different possible synthesis methods for Cu nanoparticles, only Microwave (MW)-assisted polyol synthesis is considered in this study.[29] Using the polyol method has several important advantages. It uses a high-boiling solvent, making high-temperature syntheses relatively easy to perform. In addition, the polyol can also act as a reducing agent to convert Cu^{2+} or Cu^{+} to metallic Cu without the need for additional reagents.[30] This way, the parameter space can be significantly constrained, which also reduces the number of data points required to achieve a good ML model. The use of MW heating allows syntheses to be carried out in a relatively short period. It also ensures more uniform heating of the reaction mixture because the MW radiation is selectively absorbed by the polar polyol molecules.[29] This uniform heating results in a more consistent and better control over synthesis parameters such as temperature and heating rate.[29,30]

The first step in the polyol reduction (**Scheme 1**) is the dissolution of the copper acetate ($Cu(OAc)_2$) precursor in the polyol (ethylene glycol). During the dissolution, the Cu^{2+} ions are stabilised due to an initial complexation by ethylene glycol molecules, resulting in a turquoise solution. A green precipitate will first form upon heating, followed by conversion to an alkoxide intermediate. Finally, the Cu^{2+} will be reduced to Cu^{0} NPs upon further heating.[29,31]

a) $Cu(CH_3COO)_2 \xrightarrow{dissolve} Cu^{2+} + 2\ CH_3COO^-$

b) $Cu^{2+} + 3\ C_2H_6O_2 \xrightarrow{complexation} Cu[(C_2H_6O_2)_3]^{2+}$

c) $3\ Cu[(C_2H_6O_2)_3]^{2+} + 2\ CH_3COO^- \xrightarrow{heating} Cu_3(CO_2CH_3)_2(C_2H_4O_2)_2 + 7\ C_2H_6O_2 + 4\ H^+$

d) $Cu_3(CH_3COO)_2(C_2H_4O_2)_2 + C_2H_6O_2 \xrightarrow{heating} 3\ Cu(C_2H_4O_2) + 2\ CH_3COO^- + 2\ H^+$

e) $Cu(C_2H_4O_2) \xrightarrow{reduction} Cu^0 + C_2H_6O_2\ oxidation\ products$

Scheme 1 *– a) The polyol reduction of Cu starts with dissolving of the $Cu(OAc)_2$ precursor salt in ethylene glycol; b) Next, the Cu^{2+} is complexed by ethylene glycol molecules to stabilise the ions in solution; c) Upon heating, first a green precipitate will form; d) Upon further heating, an alkoxide intermediate is formed; e) Finally, the alkoxide intermediate is reduced to metal Cu and ethylene glycol is oxidised to e.g., glycolaldehyde, oxalate, glyoxal, and CO_2.*

However, the above method, which only uses polyol for the reduction, has the disadvantage that polyols are only mild reducing agents, which means that the reaction is relatively slow and still requires relatively high temperatures (>200 °C), while the Cu NPs obtained are relatively large.[29,31] The reduction can be accelerated by adding a base to the solution. The added base plays an important role in two steps of the reaction mechanism. First, the base helps in dissolving the $Cu(OAc)_2$ precursor by deprotonating the hydroxyl groups of the polyol, and therefore facilitating the complexation.[29] Besides, the base also activates the polyol molecules by partial deprotonation of the α-CH_2, generating a reactive alkoxide electron donor.[29,31,32] When adding a base to the reaction mixture, therefore, the reduction will occur faster and at lower temperatures (160-180 °C), but also the obtained NPs will be smaller and more monodisperse.[31]

1.2.2. ML using small datasets

Although constructing an in-house dataset addresses the issues of bias and inconsistency, it introduces the challenge of working with a limited number of data points, requiring careful design to ensure the dataset's effectiveness. Since ML usually needs large datasets to generalise as much as possible, it is necessary to configure the limited dataset as optimally as possible. The resulting parameter space for the MW-assisted polyol synthesis (Cu precursor concentration, time and temperature) should subsequently be described as best as possible by the constructed dataset. This can be obtained by using Latin hypercube sampling (LHS).[33] Classic OVAT methods (**Figure 3a**) keep all variables constant except for one, meaning no interactions between variables can be detected. Classic DoE (**Figure 3b**) varies multiple variables simultaneously in a structured manner, allowing interactions between variables to be detected.[33,34] LHS works similarly to DoE but provides random sampling

(**Figure 3c-d**). LHS divides the entire parameter space into equal intervals, after which it is filled via random sampling. This ensures good coverage of the entire parameter space, even with a limited number of samples. As a result, LHS helps to avoid introducing artificial correlations between variables, which can occur in structured sampling methods, such as DoE, where parameters vary in a fixed pattern. By randomly selecting values from each interval independently for each variable, LHS ensures that parameter values are sampled in an uncorrelated way, increasing the robustness of the experimental design. [33,34]

Despite good sampling and consistent dataset construction, the number of data points is still very small for training ML models.[35,36] To be able to work with this limited number of data points in ML models, it is necessary to use ensemble regression models (**Figure 4**).[35,36] This involves training different regression models on a randomly selected subset of the dataset, after which the average is taken to arrive at the best possible overall model.[35,36] An additional advantage of these regression models is that they allow solving inverse problems. Then, the trained model can help to identify the optimal synthesis parameters to achieve a desired synthesis outcome.

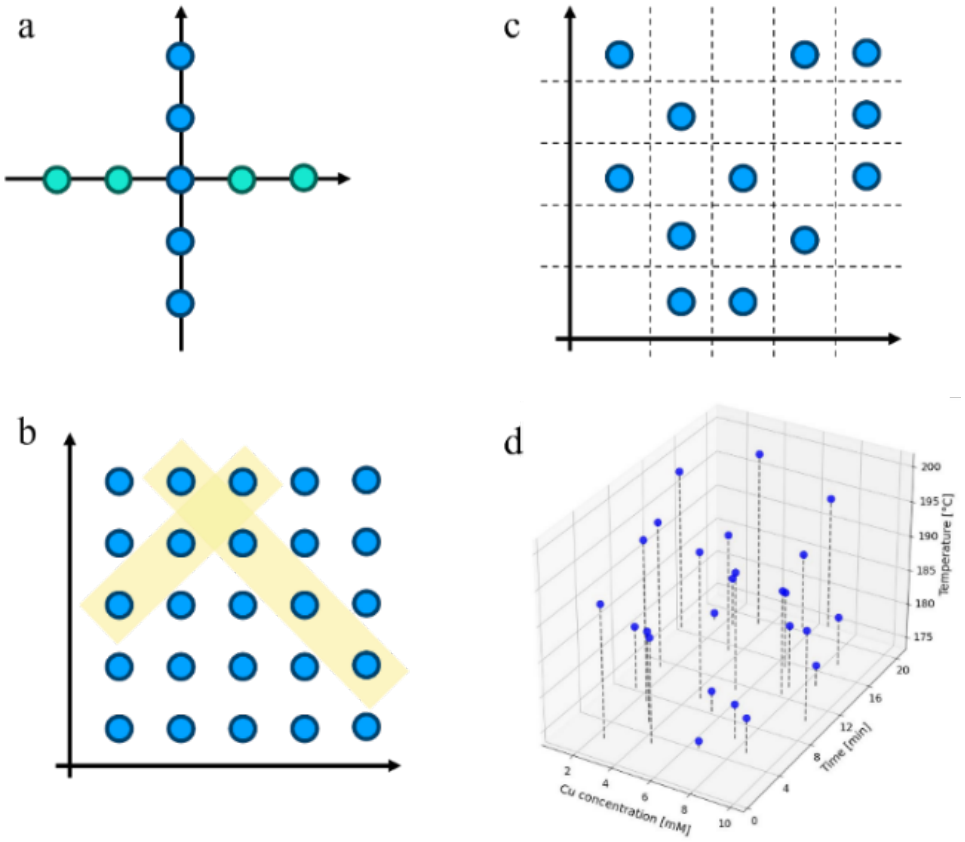

Figure 3 *- Comparison of different sampling methods. a) Classical one-variable-at-a-time experimentation; b) Classical Design-of-Experiment approach; c) Latin Hypercube Sampling(LHS); d) Distribution of experiments sampled using LHS.*

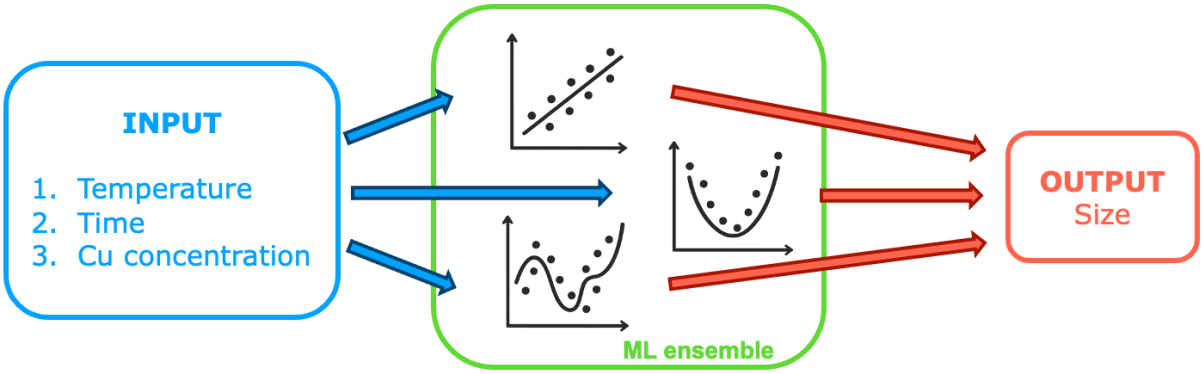

Figure 4 *– Representation of the use of ML ensembles to handle small datasets.*

1.3. Comparing ML performance in inorganic synthesis prediction

This study investigates how ML can predict and help develop inorganic synthesis procedures. As a model system, the MW-assisted polyol synthesis of Cu NPs is modelled based on the in-house constructed dataset. With this dataset, different ML models and statistical approaches are used and compared to model the synthesis process. Firstly, ensemble regression models are used, as these have been shown to work effectively for small datasets in previous research.[35,36] These ensemble regression models are built using repeated fitting of the same model architecture (*e.g.*, 3rd degree polynomial regression) on random subsets of the dataset, and aim to predict particle size as a numerical value using the synthesis parameters as features. Finally, these different regression models are averaged to give the final ensemble model. This method is compared with a classic statistical method, such as used in DoE, where one regression model is fitted to the entire dataset.

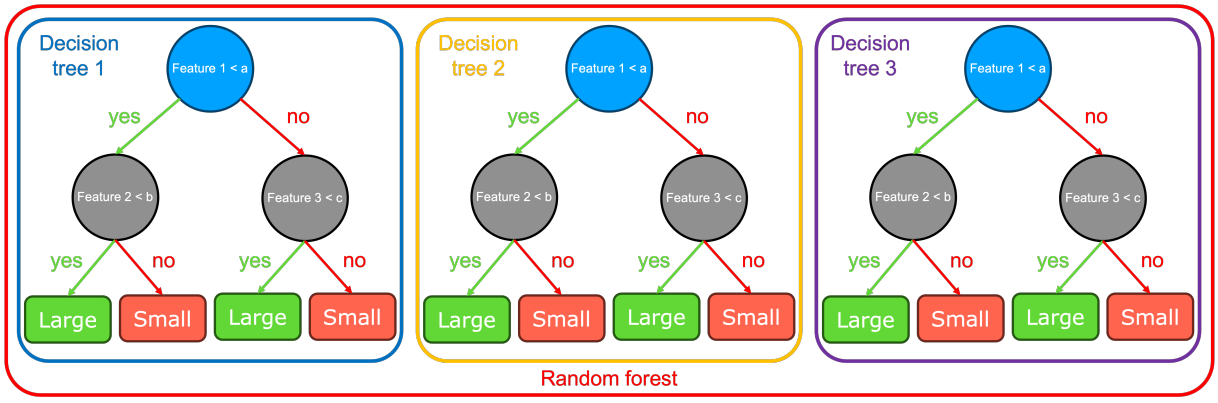

Figure 5 *– Illustrative and simplified example of a random forest classification model. Each decision tree has a hierarchical structure of layers consisting of decision nodes. In each decision node, the tree splits further, and a specific path is followed through the tree based on trained selection criteria. At the end, a leaf node is then reached that predicts whether the entered criteria will lead to small or large particles. A random forest consists of several independent decision trees, each of which individually predicts whether it will be small or large particles. The final prediction of the forest is then the class (small or large) most predicted by the different individual trees.*

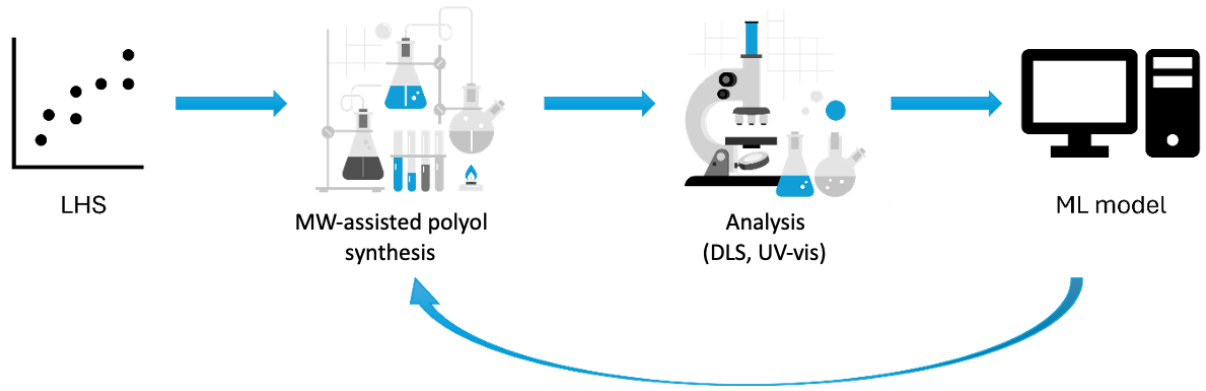

Figure 6 *- Visualisation of the workflow followed to create the dataset and model the syntheses.*

In addition, modelling the synthesis is approached as a classification problem. Here, the data is divided into a group of large and a group of small particles relative to the median size. The models then no longer predict a numerical value for the particle size, but predict which group the particles will belong to based on the synthesis parameters used. To this end, the performance of a complex LLM[37,38] is compared with that of a classic random forest (**Figure 5**). A random forest is an ensemble method that builds multiple decision trees on different data subsets and aggregates their predictions through majority voting.[16]

2. Methods

The study follows a specific workflow to model the synthesis (**Figure 6**). In the first step, the investigated parameter space is sampled via LHS to achieve optimal space coverage. Subsequently, the sampled syntheses are carried out in the lab via MW-assisted polyol synthesis. The results are then analysed via dynamic light scattering (DLS) to determine the particle size of the Cu NPs. Also, the SPR peak is determined using UV-vis measurements. The synthesis parameters and corresponding particle sizes form the dataset used to model the synthesis process. Finally, the results of the ML models can be used to assist in further experimental syntheses.

2.1. Experimental synthesis

2.1.1. Materials

For all experimental syntheses, absolute ethanol (≥ 99.8%, analytical grade) obtained from Fisher Chemical and copper(II) acetate monohydrate ($Cu(OAc)_2 \bullet H_2O$, ACS reagent), tetramethylammonium hydroxide (TMAH, 25 wt. % in H_2O) and ethylene glycol (≥ 99%) obtained from Sigma Aldrich were used. All reagents were used as received without further purification.

2.1.2. Sampling synthesis parameters

LHS is used to obtain an optimised dataset with minimal correlation between the various experiments and optimal coverage of the parameter space under investigation.[31] The parameter space under investigation comprises three dimensions:

a) Cu precursor concentration (1 mM to 10 mM)
b) Reaction time (1 minute to 20 minutes)
c) Reaction temperature (175 °C to 200 °C)

Using LHS, 25 experiments are selected that fill this parameter space as optimally as possible. The selected experiments are summarised in **Table S1**.

2.1.3. Cu NPs synthesis

The Cu NPs were synthesised via the MW-assisted polyol route[39] in a CEM Discover SP MW reactor (**Figure 7**). The synthesis parameters used can be found in **Table S1**. First, the amount of $Cu(OAc)_2$ is weighed into the reaction vial. It is then dissolved in 5 ml of ethylene glycol under sonication. After the Cu precursor has completely dissolved, the tetramethylammonium hydroxide (TMAH) is added to the solution and further mixed and degassed under sonication. TMAH is consistently used at a molar ratio of 4:1 relative to the Cu precursor. The reaction vessel is then placed in the microwave reactor, with the headspace filled with inert Argon up to a pressure of 5 bar, and is first stirred for 2 minutes at 80 °C to ensure complete dissolution of the precursor. Afterwards, the temperature is raised to the desired reaction temperature with continuous stirring and maintained for the desired reaction time (**Table S1**). After the reaction time has passed, the reaction is quenched by actively cooling the vessel with compressed air. Now, a brown-red suspension is obtained, which is typical for Cu NPs.[30] To collect the Cu NPs, the reaction mixture is diluted with 15 ml of ethanol and centrifuged at 10000 rpm for 10 minutes. Subsequently, the supernatant is removed, and the resulting NPs are again dispersed in 20 ml of ethanol.

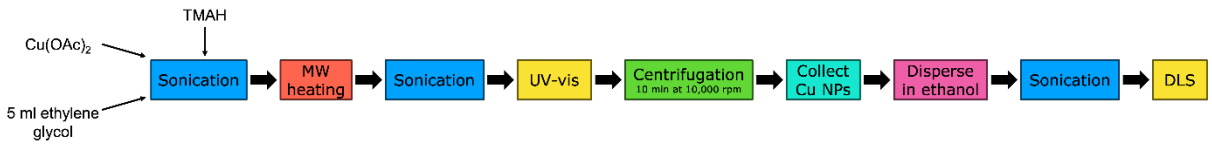

***Figure 7** -The synthesis starts with dissolving $Cu(OAc)_2$ in 5 ml of ethylene glycol and sonication. Next, TMAH is added and sonicated again, after which the solution is placed in the MW reactor. After the reaction, a brown-red dispersion of Cu NPs is obtained, which is sonicated and analysed using UV-vis. Next, the suspension is centrifuged, and the obtained particles are redispersed in ethanol, after which the particle size is determined using DLS measurements.*

2.1.4. Characterization

The SPR of the synthesised Cu NPs in the ethylene glycol dispersion is determined via UV-vis measurements with the Agilent Cary 5000 spectrophotometer equipped with a dual-beam configuration. Measurements were performed in quartz cuvettes (10 mm path length) over a wavelength range of 300–800 nm. Baseline correction was applied using the ethylene glycol solvent as reference.

The size of the synthesised Cu NPs is determined via DLS measurements (**Figure 7**) with a Zetasizer Nano ZS, equipped with a 633 nm He–Ne laser. For these measurements, the ethanolic Cu dispersions are sonicated for at least 10 minutes. Subsequently, five drops of the NP dispersion are diluted with ethanol and measured.

2.2. Computational modelling

To model the synthesis, various individual models are constructed based on different methods and their performance is compared. In the first part, regression models are used to predict the exact particle size based on the synthesis parameters used (temperature, time and Cu concentration). An initial method used for this are ensemble regression models. These are compared with a model resulting from a classical statistical regression. In a second part, a classification problem is examined in which the synthesis parameters are used to predict whether the particles obtained will be large or small. For this classification, advanced LLMs are compared with the classical random forest.

2.2.1. Ensemble regression models

The first generation of ensemble models based on scikit-learn[16,40] builds the ensemble through repeated iterations of model training on random subsets of the dataset, using the temperature, time, and Cu concentration as features and the particle size as target.[40] For each ensemble model, 2,500 random 80/20 train-test splits are made, on which a model is trained and evaluated each time. The predictions for the test data of each model are kept track of, and the ensemble prediction for a data point is taken as the average of the corresponding predictions of the individual models for that data point. Due to instability in the hyperparameter tuning as a result of the small dataset, among other things, it was decided to switch to the AMADEUS framework.

The following generation ensemble regression models are built using the scikit-learn-based AMADEUS framework (**Figure 8**).[35,40,41] AMADEUS creates pasting-type ensemble models consisting of 100 base instances. Each of these instances is trained on a different 80/20

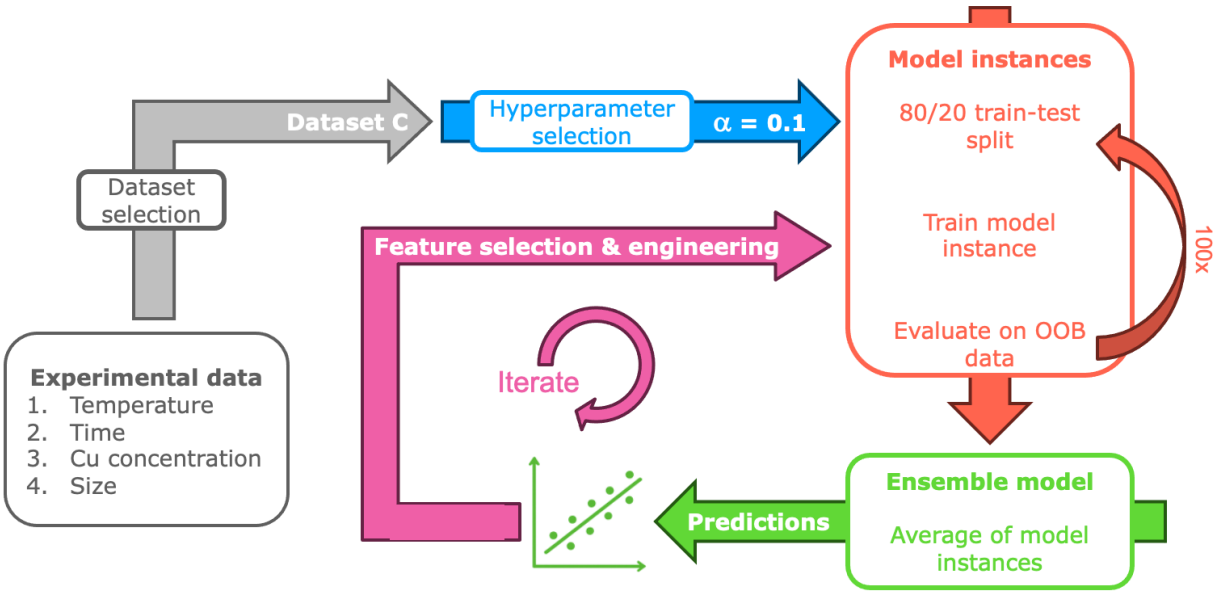

Figure 8 *– The computational modelling starts with selecting the best performing dataset. Next, the optimal value for the regularisation strength α is selected. Then, 100 model instances are trained on random subsets of the dataset. Once all model instances are trained, the ensemble model is the average of all instances. Finally, for next generation models, feature selection and engineering are performed.*

split of the in-bag and out-of-bag (OOB) parts of the complete dataset. The final ensemble model is a single base instance equal to the average of the 100 base model instances. Both linear and polynomial models are considered, and LASSO regularisation is used.[42] In a first step, it is examined how best to deal with the multimodal data points by comparing three different datasets. Next, a hyperparameter selection for the LASSO regularisation is performed. Finally, a final model is constructed through feature engineering and reduction.

Performance metrics:

Three performance metrics are defined to determine and compare the accuracy and generalisability of the various models.

a) Root-mean-squared error (RMSE) measures the average quadratic deviation between the predicted value ($\hat{y}_i$) and the experimental value (y_i). Because RMSE penalises large errors more heavily than small errors, it is highly sensitive to outliers.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=n}^{1}(\hat{y}_i - y_i)^2} \quad (1)$$

b) Mean-average error (MAE) calculates the average absolute error in the predictions.

$$MAE = \frac{1}{n}\sum_{i=n}^{1}|\hat{y}_i - y_i| \quad (2)$$

c) R^2 is a measure of the correlation between the predicted value ($\hat{y}_i$) and the experimental value (y_i). An R^2=1 means that the predicted values perfectly match the experimental

values, while an R^2=0 means that the model performs no better than predicting the average.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

Within the AMADEUS framework, a further distinction is made between three different types of error. The error in the training data (MAE and RMSE) is the error that the model makes on the data on which it was effectively trained. This error is an important indicator to check that the model is not underfit. However, a low training error is no guarantee of a good model, as it can also be the result of overfitting, especially with small datasets. The OOB error (MAE and RMSE) is the error the trained model makes on its test data. This is data the model has not yet seen during training. This error is a good indicator of the generalisability of the model. A low training error and high OOB error are very strong indicators of a model that is overfitting. Finally, there is the error of the average model (MAE, RMSE and R^2). This is calculated by having the final ensemble model make predictions for all data points in the entire dataset.

Dataset selection:

In the first series of models generated using AMADEUS, the optimal dataset is determined to train the regression models. To this end, the performance of three different datasets is compared using the RMSE, MAE and R^2 of the best-performing ensemble per dataset.

a) Dataset A: All 25 data points are included in the dataset, and the particle size predicted is the smallest particle size present in the DLS spectrum (**Table S2**).
b) Dataset B: All 25 data points are included in the dataset, and the particle size predicted corresponds to the peak with the highest percentage contribution in the DLS spectrum (**Table S3**).
c) Dataset C: Only the 18 data points with a monomodal DLS spectrum are included in the dataset (**Table S4**).

The various performance metrics show that dataset C performs significantly better than the other two datasets (**Table 1**). Therefore, only dataset C will be considered further. The trained models in this study thus focus on predicting the particle size for a monomodal distribution and therefore cannot predict situations where a multimodal distribution is obtained. Further research can investigate the cause of these multimodal distributions and train models that take them into account, but this is beyond the scope of this study.

Table 1 *– Performance metrics of the best-performing ensemble average for the different datasets.*

	RMSE [nm]	MAE [nm]	R^2
Dataset A	63.46	50.50	0.42
Dataset B	59.44	46.06	0.35
Dataset C	34.09	24.28	0.73

Hyperparameter tuning:

Due to the limited size of the dataset, the regularisation strength parameter α cannot be determined via automated optimisation. Instead, it is selected based on visual comparison of four values (α = 1.0, 0.1, 0.01, and 0.001). For this purpose, polynomial models up to order eight are optimised. Analysing the MAE of the average ensemble models shows an overall optimal performance for regularisation strength α = 0.1 (**Figure 9**). Also, the hyperparameter tuning shows no further improvements above polynomial degree three. Therefore, no features above degree three will be further considered to reduce the risk of overfitting.

Feature selection:

When using a small dataset, the growing number of features in polynomial models quickly leads to overfitting. Regularisation reduces the number of relevant features, reducing the risk of overfitting. The AMADEUS framework tracks how often a polynomial feature is retained within the ensemble, providing a measure of importance for the features. A feature with an importance of 100% will have a non-zero coefficient in all model instances in the ensemble. If a feature has an importance of 75%, this means that in 25% of the model instances in the ensemble, the coefficient is zero for this feature. **Figure 10** summarises the feature importances for all features up to degree three.

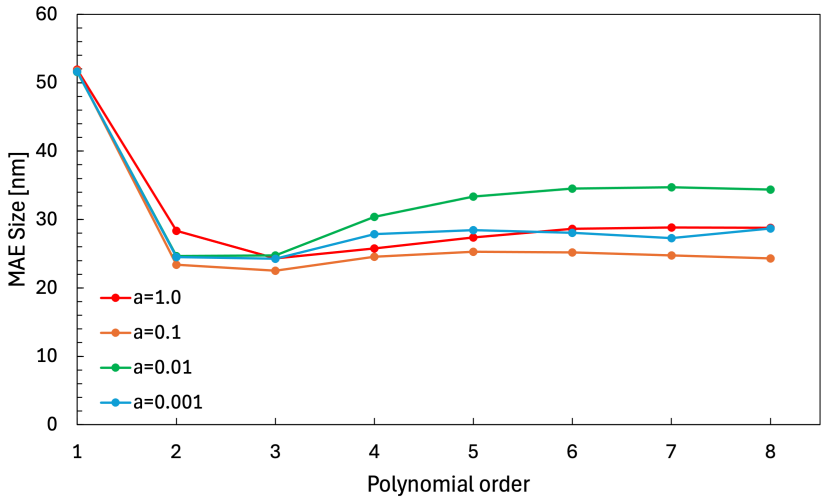

Figure 9 *- Hyperparameter optimisation for size prediction with dataset C. LASSO regularised polynomials up to order eight are compared, and MAE of the average ensemble models is shown for α=1.0 (red), α=0.1 (orange), α=0.01 (green), and α=0.001 (blue).*

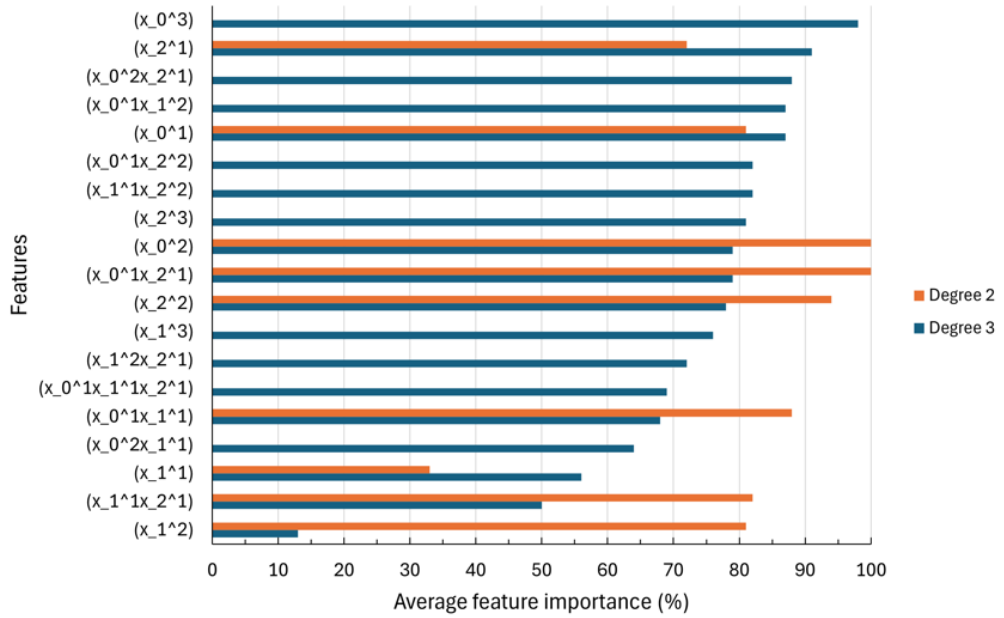

Figure 10 *- Overview of feature importance, sorted by increasing importance in the third degree polynomial, in the polynomial models with degree two (orange) and degree three (blue), and trained on dataset C with a=0.1. x_0 = Cu precursor concentration, x_1 = Reaction temperature and x_2 = Reaction time.*

Here, it is clear that the Cu precursor concentration plays an important role, followed by the reaction time to a somewhat lower extent. Temperature, on the other hand, mostly plays a role in interaction parameters. Based on the trends in feature importance, seven features are selected:

a) [Cu]
b) Temperature
c) Time
d) $[Cu]^2$
e) $[Cu]^3$
f) [Cu] x Temperature
g) [Cu] x Time

[Cu], temperature and time are all three retained as features, as these are the base parameters used in synthesis. $[Cu]^2$ and $[Cu]^3$ are included as they play relatively important roles in both the second and third degree models. The interaction terms [Cu] x Temperature and [Cu] x Time are used, given their high importance in the second-degree polynomial; higher-order interaction terms are not included to limit the model complexity and hence the probability of overfitting. Using these features, a linear regression model is trained using the AMADEUS framework, and iteratively, features are further engineered to obtain the optimal combination.

2.2.2. Classical statistical methods (DoE)

Using DoE software JMP Pro (SAS Institute Inc.), a linear regression model is constructed for dataset C, including terms up to the second degree. Next, insignificant terms are

removed until only significant terms remain. The available dataset is not split into training and test data; all data points are used to fit the regression model. The performance of the model is determined by calculating the expected size for all data points using the equation obtained.

2.2.3. Classification analysis

Here, a binary classification is performed on the particle size using dataset A (**Table S2**). The aim is to predict whether the resulting particles will be large or small based on the synthesis parameters. Particles are defined as large if they are larger than the median size in the dataset; if they are smaller than the median size, they are defined as small. For this classification task, the performance of complex LLMs is compared with that of a classic random forest classifier.

Performance metrics:

Accuracy and kappa (κ) are used as performance metrics to assess the classification models' performance.

a) Accuracy is the most commonly used error metric for classification problems and indicates the percentage of correct predictions. An accuracy of 0.9, therefore, means that 90% of the predictions were made correctly.

$$accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (4)$$

b) κ also indicates the agreement between predictions and actual labels, but corrects for coincidental agreements. It therefore indicates how well the model performs compared to simple random guessing. If κ is 1, there is a perfect match between predictions and actual labels. If κ is 0, the model's performance is the same as if random guesses had been made, and if κ is negative, the model performs even worse than random guessing.

$$\kappa = \frac{accuracy - p_e}{1 - p_e} \quad (5)$$

Here, p_e is the outcome expected by chance, calculated from the proportions of the real ($P_{real,i}$) and predicted classes ($P_{predicted,i}$). The proportion $P_{real,i}$ refers to the fraction of samples that truly belong to class *i*, and $P_{predicted,i}$ refers to the fraction of samples predicted to be in class *i*.

$$p_e = \sum_i (P_{real,i} \times P_{predicted,i}) \quad (6)$$

Large Language Models:

Two different LLMs were trained and compared with each other for the binary classification task: GPT-J and LLAMA 3.1. The training of the LLMs was carried out by dr.

Joren Van Herck at prof. dr. Berend Smit's group at the École Polytechnique Fédérale de Lausanne (EPFL). The pretrained LLMs are further trained using transfer learning, similar to previous work[37,38], to predict the outcome as large particles (1) or small particles (0) based on the experimental input parameters. The LLMs are trained to use the following prompt as input:

'What is the size of nanoparticles synthesized with a <Cu conc> mol/litre Copper Acetate precursor concentration, a <TMAH conc> mol/liter tetramethylammoniumhydroxide concentration, heated for <time> minutes at <temperature> degrees celcius?'

Here, <parameter> serves as a placeholder for the reaction parameter's value. The LLM will then answer the prompt with either the value 0 for experimental small NPs or 1 for experimental big particles.

The maximum number of training data was set to 15. The number of test data remained constant over all runs, *e.g.*, 5. First, the number of epochs was screened: 10, 25, 50, and 100. Five unique runs were performed for every pair of training size-epoch experiments to get the average metrics. The GPT-J model was used for this screening. Since it is a binary classification of a balanced dataset, all accuracies above 50% are an improvement over random guessing. The first experiments with 25 epochs were not successful (**Figure S4**). The models could not capture the prompt/completion structure and returned invalid completions for the test data, for example, '-9223372036854775808'. For a higher number of epochs, the fine-tuned model returns valid completions ('1' or '0').

Random Forest classifier:

An ensemble approach of random forest models via scikit-learn is used for binary classification. Hyperparameter selection was used to determine the optimal value for n_estimators, the number of decision trees within a forest, and max_depth, the maximum depth of a tree. The n_estimators was set to 200 and max_depth to 3. With these hyperparameters, 2,500 random forests are trained on a new random subset of the data. 20% of the total dataset serves as OOB data. Each generated model is tested on its OOB data, and these predictions are logged. After training and evaluating all 2,500 models, the OOB predictions of all models are used to obtain a final classification through hard voting. For the OOB predictions, the number of times they were classified as small (0) and the number of times as big (1) are counted for each data point. The class with the majority of votes will be the final predicted class for that datapoint.

3. Results and discussion

3.1. Cu NPs synthesis

A crucial step in modelling the MW-assisted polyol synthesis of Cu NPs is the construction of a high-quality dataset. This dataset was created by performing 25 experimental syntheses according to the protocol described earlier (Section 2.1.3). The experiments were selected using LHS and are summarised in **Table S1**. The DLS measurements show a peak above 5000 nm for five samples. However, these peaks are not included in the analysis. Due to their very low intensity (<1.5%), it can be assumed that these results are due to contamination, such as dust particles, during the measurements. The synthesis yielded Cu NPs with a particle size between 40 nm and 350 nm (**Table S5**). The DLS shows a bimodal spectrum for five samples and a trimodal spectrum for two samples. The remaining samples show a monomodal spectrum (**Figure 11a**). Analysis of the peaks in the DLS spectra corresponding to the smallest particle size gives a median particle size of 157.9 nm (**Figure 11b**).

Figure 11a clearly shows that the samples with a multimodal DLS spectrum are the samples with the smallest detected particle size. It is also noticeable that the slope of the smallest sizes at these multimodal data points is less steep than that of the monomodal data points. More in-depth analysis of the reaction parameters shows that the multimodal results mainly occur at a 6 mM or higher Cu concentration. A possible explanation could be that in these situations, the NPs were still in their growth phase and the reaction had not yet been completed at the time of quenching. However, this would then be expected to occur at lower temperatures and/or reaction times, which is not particularly the case. At this point, it is

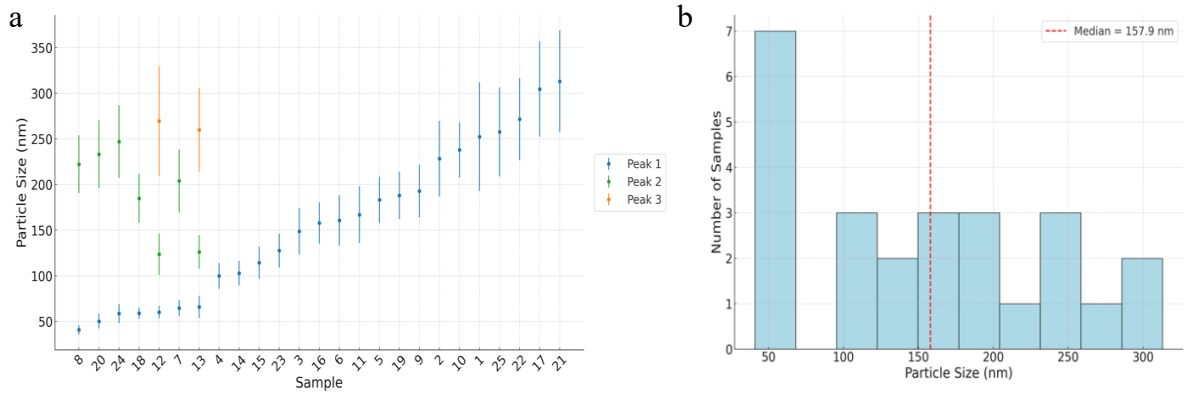

Figure 11 *– a) DLS results showing particle sizes and standard deviations for all detected peaks per sample. Each point represents a population of a distinct size (Peak 1 in blue, Peak 2 in green, and Peak 3 in orange), with error bars indicating the corresponding standard deviation. Samples are sorted in ascending order based on the size of Peak 1. Peaks above 5000 nm were excluded due to their negligible intensity (<1.5%); b) Distribution of the smallest detected particle size per sample as measured by DLS. The red line indicates the median particle size.*

therefore not yet possible to provide a conclusive explanation for the formation of the multimodal particle size distributions, and further research is needed.

Based on the UV-vis spectra (**Figure S5-Figure S6**) and the SPR location (**Table S5**), it can be verified for the different samples that Cu metal (Cu^0) NPs were effectively formed. Cu^0 NPs exhibit a sharp SPR peak around 600 nm, which can shift depending on the size of the NPs.[30] The different UV-vis spectra show a clear SPR peak between 590 nm and 630 nm for most samples, which is a strong indication of the presence of Cu^0 NPs. Samples 4, 6, and 19 do not show a clear SPR feature, which may be due to low particle concentration, oxidation[43], or the presence of ultrasmall particles (<5 nm), which do not exhibit a well-defined SPR; for the other samples, at least some Cu^0 is present. Given the limited yield of the syntheses, no X-ray diffraction measurements can be performed to provide a conclusive answer, but the combination of the red-brown colour of the solution and the UV-vis spectra strongly indicates that Cu^0 NPs have indeed been formed. To further support this conclusion, Selected Area Electron Diffraction analysis was conducted on sample 22 using a transmission electron microscope operated at an accelerating voltage of 120 kV. The resulting diffraction rings (**Figure 12**) were indexed to the face-centred cubic structure of copper, with prominent rings corresponding to the (111), (200), and (220) planes. The primary diffraction rings were assigned the following d-spacings: 2.08 Å for the (111) plane, 1.83 Å for the (200) plane, and 1.29 Å for the (220) plane. These values agree with the standard values for FCC copper (JCPDS card no. 04-0836)[44], confirming the formation of Cu^0 NPs.

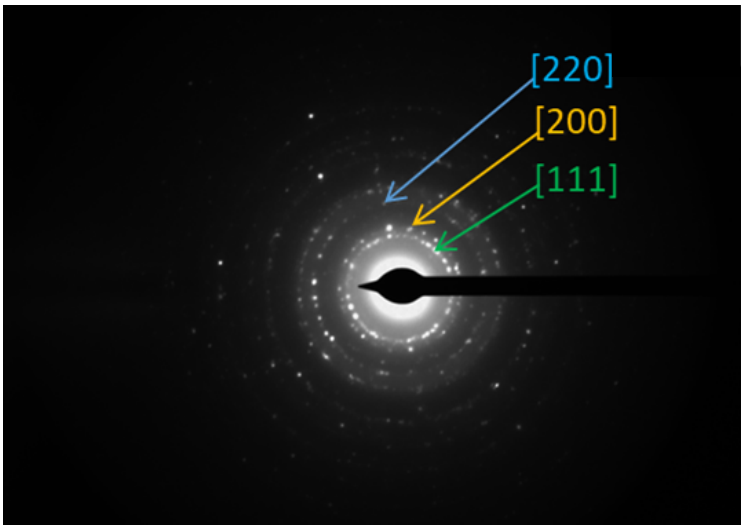

Figure 12 *- Selected area electron diffraction pattern of Cu^0 nanoparticles from sample 22, recorded at an acceleration voltage of 120 kV. The presence of concentric diffraction rings indicates a polycrystalline material. The rings are indexed to the face-centred cubic (FCC) crystal lattice of metallic copper, with corresponding (111), (200) and (220) planes. The measured d-spacings for these planes are 2.08 Å, 1.83 Å, and 1.29 Å, respectively, in good agreement with standard reference values.[44]*

Table 2 *- Performance metrics of different regression model generations.*

	Number of features	Training RMSE [nm]	Training MAE [nm]	OOB RMSE [nm]	OOB MAE [nm]	Average RMSE [nm]	Average MAE [nm]	R^2
Linear	3					81.10	68.80	-0.55
Polynomial	19					118.00	79.40	-2.29
2nd generation	19	22.17	16.05	86.52	68.51	32.64	22.53	0.75
3rd generation	7	29.40	20.47	51.85	41.36	32.61	22.28	0.75
4th generation	11	27.78	19.79	63.66	52.18	33.07	23.35	0.74
5th generation	6	30.57	22.39	49.90	40.92	33.03	23.81	0.74
DoE	2					40.91	33.54	0.60

3.2. Ensemble regression models and size prediction

In the first phase of computational modelling, ensemble regression models and classic statistical regression models are developed to predict a numerical value for the obtained Cu NP size, based on the synthesis parameters. Different generations of models are trained to investigate the influence of different features on model performance.

3.2.1. First generation

The first generation models are used as baseline models. These models do not yet have any predictive value. The linear ensemble has an R^2 of -0.55 (**Table 2**), which indicates that the model's predictions are slightly worse than the average prediction. The graphical representations of the predictions (**Figure 13**) also show that the predictions broadly fluctuate around the average. The third-order polynomial ensemble has an even lower R^2 of -2.29 (**Table 2**) and indicates graphically (**Figure 13**) that the predictions are virtually random. More complex models are therefore needed. Because of the extremely poor performance of the polynomial model, the linear model will be used as the baseline model.

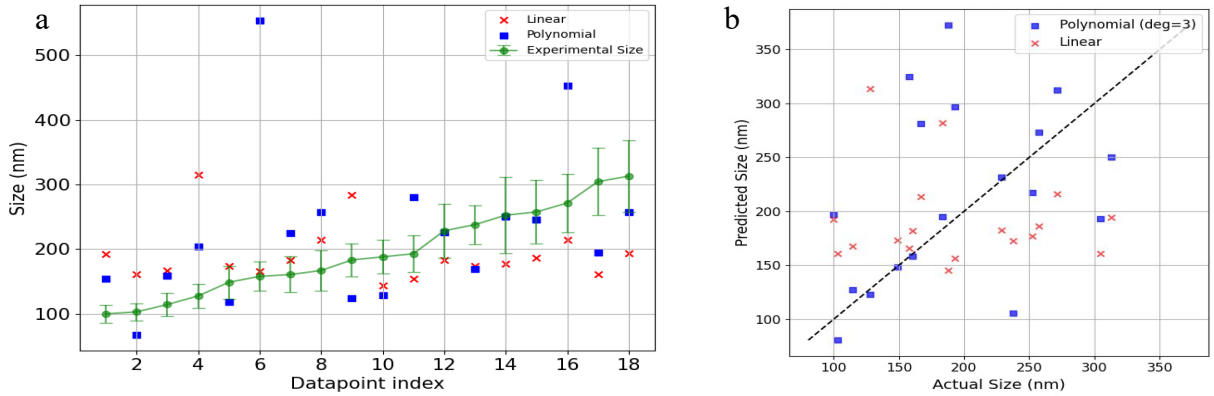

Figure 13 *- a) Plot of the experimental measured particle size and their standard deviation in ascending order (green) with the corresponding predicted particle size of the linear ensemble model (red) and the polynomial ensemble model of degree three (blue); b) Parity plot of the actual, experimental size against the corresponding predicted size using the linear ensemble model (red) and the polynomial ensemble model of degree 3 (blue).*

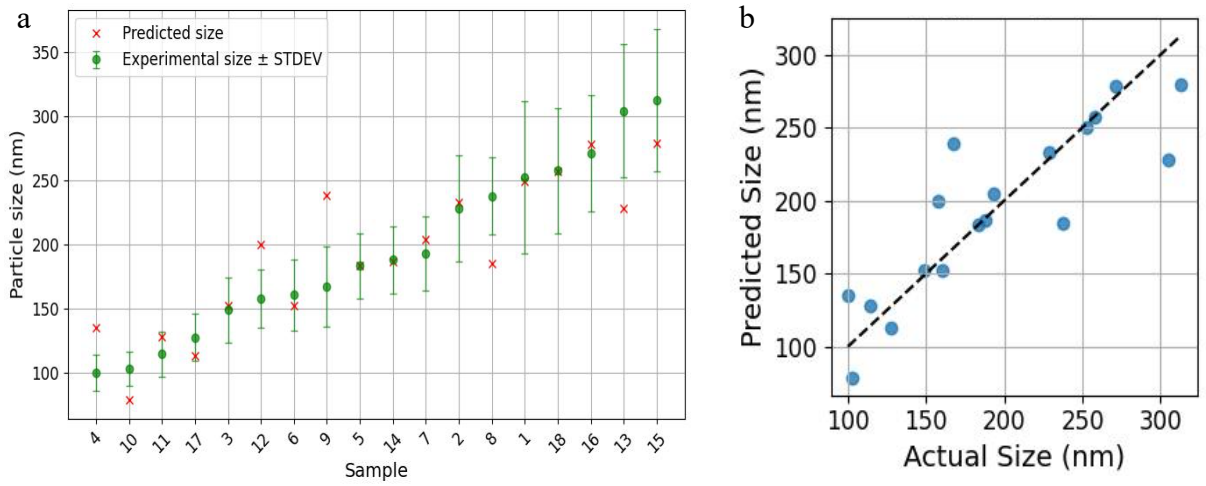

Figure 14 *- a) Plot of the experimentally measured particle size and their standard deviation in ascending order (green) with the corresponding predicted particle size of the second generation ensemble model (red); b) Parity plot of the actual, experimental size against the corresponding predicted size using the second generation ensemble model.*

3.2.2. Second generation

As second generation (**Figure 14**), we consider the complete third-degree polynomial model with regularisation strength $\alpha = 0.1$ generated via the AMADEUS framework during hyperparameter tuning. Here, there is a significant improvement in performance metrics (**Table 2**), with an R^2 rising to 0.75 and RMSE and MAE of the ensemble average falling from 81.10 nm to 32.64 nm and from 68.80 nm to 22.53 nm, respectively. The more complex error handling and model building of the AMADEUS framework have a strong positive influence on the quality of the model, highly outperforming the baseline model.

3.2.3. Third generation

Despite the strong performance of the second generation model, its main disadvantage is the use of an extensive set of 19 features. This increases the risk of overfitting, especially when working with small datasets. In addition, a larger feature set results in a more complex model and, therefore, higher computational costs. The aim of the next generations is thus to reduce model complexity drastically, on the one hand, improve generalisability and, on the other hand, reduce computational costs. In the third generation the model (**Figure 15**) complexity is strongly reduced by using only the seven features ([Cu], Time, Temperature, $[Cu]^2$, $[Cu]^3$, [Cu]xTime, [Cu]xTemperature) selected based on the feature importances (**Figure 10**). A linear regression model is built using AMADEUS with regularisation strength $\alpha = 0.1$ using these features.

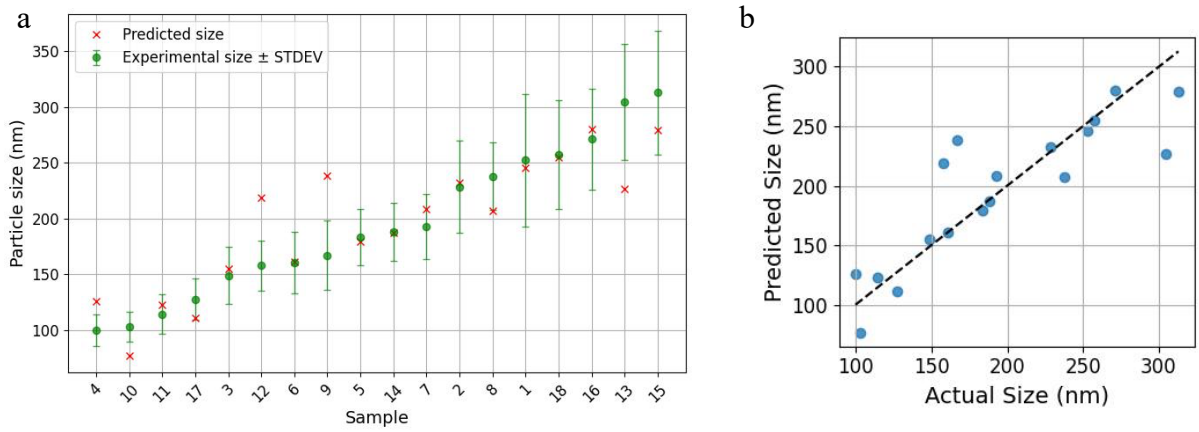

Figure 15 *- a) Plot of the experimentally measured particle size and their standard deviation in ascending order (green) with the corresponding predicted particle size of the third generation ensemble model (red); b) Parity plot of the actual, experimental size against the corresponding predicted size using the third generation ensemble model.*

The ensemble average now shows hardly any difference in performance compared to the second generation, with almost identical RMSE, MAE and R^2 (**Table 2**), despite the significant reduction in the number of features used. In addition, the training and OOB errors show that the third generation is more generalisable than the second generation (**Table 2**). The training error of the third generation is only slightly higher (MAE of 20.47 nm compared to 16.05 nm), but the OOB error is drastically reduced from 68.51 to 41.36 for the MAE. The smaller difference between training and OOB error indicates that the model is less overfitted and generalises better.

3.2.4. Fourth generation

Here, the feature set used in the previous generation is expanded again with four additional features. This expansion aims to investigate the influence of non-polynomial features (based on domain knowledge) on the performance of the model. To this end, four additional features are added to the set.

a) $e^{[Cu]}$ is added because [Cu] often has a high and consistent importance even in higher powers.

b) [Cu]/Time can be interpreted as a measure of the amount of cations available per unit of time during the growth of the particles.

c) Log([Cu]xTime) provides a measure of the total exposure of the reduction medium to Cu ions.

d) Time/Temp as a measure of the thermal load during the reaction. A higher value indicates milder reaction conditions.

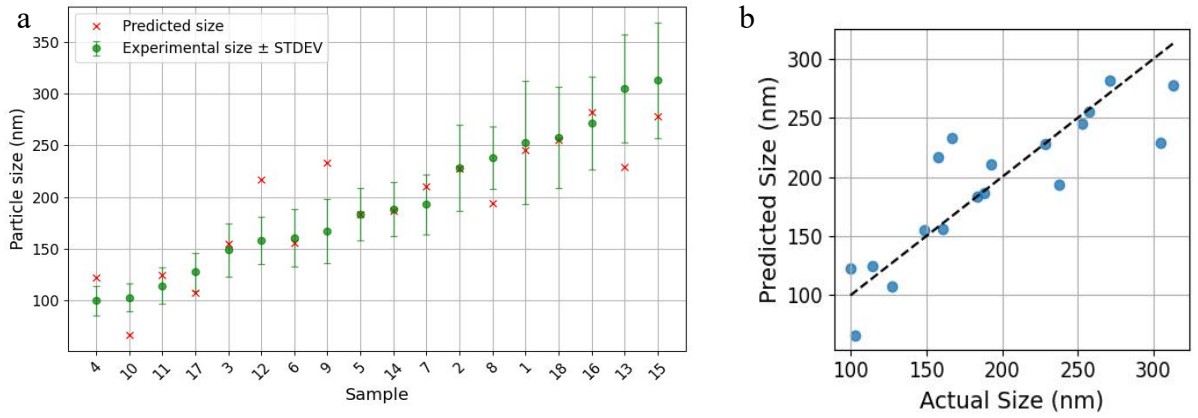

Figure 16 *- a) Plot of the experimentally measured particle size and their standard deviation in ascending order (green) with the corresponding predicted particle size of the fourth generation ensemble model (red); b) Parity plot of the actual, experimental size against the corresponding predicted size using the fourth generation ensemble model.*

Using this more complex feature set, a linear regression model is built again using AMADEUS with regularisation strength $\alpha = 0.1$ (**Figure 16**). As expected, there is again a slight decline in the performance of the model (**Table 2**) due to the increased complexity, which leads to stronger overfitting and therefore poorer generalisability. However, given the limited difference in performance between this and the previous generation, a comprehensive analysis of the feature importances in this model will be carried out in the final generation. The aim is to arrive at a final reduced model with a good balance between sufficient complexity to maintain predictive value and not too much complexity to ensure sufficient generalisability.

3.2.5. Fifth generation

Finally, the fourth generation model is again reduced in complexity. Therefore, an extensive analysis of feature importance is conducted. Various models were trained based on the extensive feature list of the fourth generation. One, two or three of the features were systematically disabled in the model to determine the influence of each feature on the model's performance. Here, four features used in the fourth generation have a rather low importance of less than 85% on average over all models. Therefore, only the six features with an average of more than 90% feature importance are kept in the final generation.

a) Temperature (T) b) $e^{[Cu]}$

c) [Cu] x Time (t) d) [Cu] / Time

e) $[Cu]^3$ f) log([Cu] x Time)

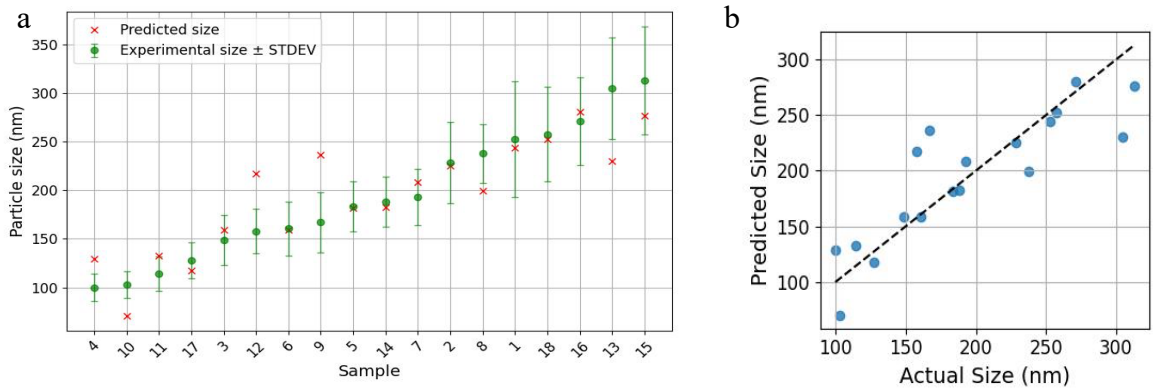

Figure 17 *- a) Plot of the experimentally measured particle size and their standard deviation in ascending order (green) with the corresponding predicted particle size of the fifth generation ensemble model (red); b) Parity plot of the actual, experimental size against the corresponding predicted size using the fourth generation ensemble model.*

Using these six features, the final fifth generation model (**Figure 17**) is obtained via a linear regression using AMADEUS with regularisation strength $\alpha = 0.1$ (**equation 7**). Now, all used features have an average importance of 98% or higher. The ensemble average now has an RMSE, MAE and R^2 of 33.03 nm, 23.81 nm and 0.74 (**Table 2**), respectively, which is hardly any difference compared to the more extensive fourth generation. Because the complexity of the model has been greatly reduced in this generation, the MAE for the OOB predictions decreases from 52.18 nm for the fourth generation to 40.92 nm for the fifth generation (**Table 2**). Therefore, this fifth generation generalises significantly better compared to the fourth generation.

$$Cu\ NP\ size = 194.66 - 12.59T - 36.16[Cu]t - 123.47[Cu]^3 + 108.12e^{[Cu]} + 32.30\frac{[Cu]}{t} + 63.00\log([Cu]t) \quad (7)$$

Despite the very limited dataset, it is possible to predict the particle size of the Cu NPs quantitatively based on the synthesis parameters used, with an R^2 of 0.74 and an MAE of 23.81 nm. For most of the data points, the predicted particle size by the ensemble model is also within the experimental STDEV. Throughout the different generations of ensemble regression models, reducing the number of features used has little effect on the performance of the ensemble predictions, while the OOB error decreases slightly when the number of features is reduced, showing that the model is less overfitted and therefore generalises better. Reducing the number of features to six in the fifth generation has two important advantages over the full third-degree polynomial model from the second generation. On the one hand, it generalises better, and on the other hand, the model is much less complex, which also reduces the computational cost.

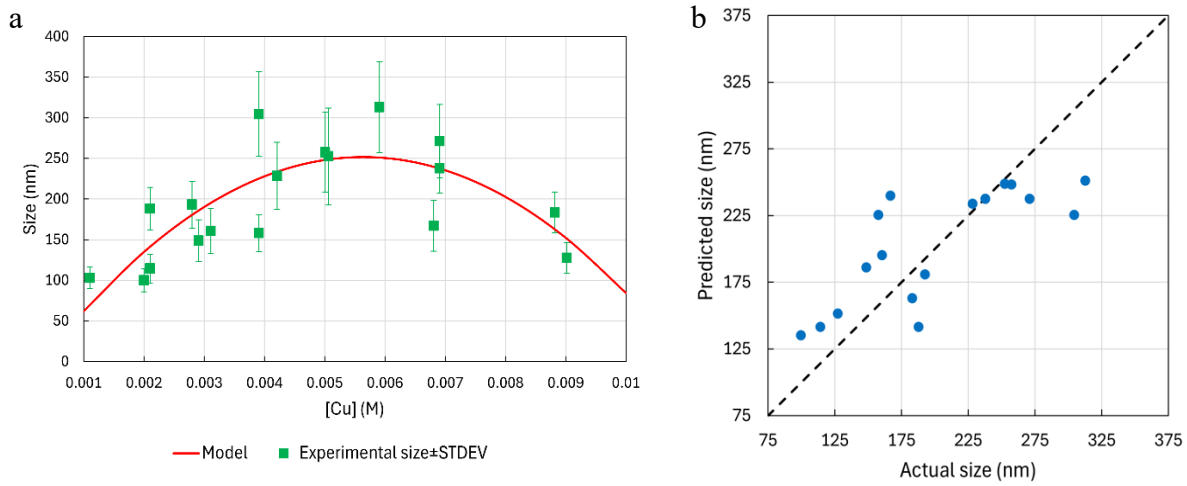

Figure 18 *- a) Plot of the experimental measured particle sizes and their standard deviation (green) and the statistical model (red); b) Parity plot of the actual, experimental size against the corresponding predicted size using the statistical model.*

3.2.6. Statistical approach (DoE)

In the classical statistical approach, only the concentration of the Cu precursor appears to significantly influence the particle size, resulting in a quadratic model (**Figure 18**) given by **equation 8**. This model has an R^2 of 0.60 and an MAE of 33.54 nm (**Table 2**), which is significantly lower compared to the ML regression models. In addition, it is unclear whether the model is sufficiently generalised, as it is based on the entire dataset. An OOB error is therefore unavailable, which can provide an indication of the generalisability.

$$Cu\ NP\ size = 156.79 + 18\,530.65[Cu] + ([Cu] - 0.005) \times \left(-8\,801\,229.71([Cu] - 0.005)\right) \quad (8)$$
$$= -63.24 + 106\,542.95[Cu] - 8\,801\,229.71[Cu]^2$$

Compared to the fifth generation ensemble regression model, the classical regression model shows a significant decrease in performance, with R^2 decreasing from 0.74 to 0.60 and MAE rising from 23.81 nm to 33.54 nm. Since classic regression is performed on the entire dataset, no information is available about the OOB error, making it difficult to conclude how the generalisability of both models relates to each other. Since the ensemble regression models are built on random subsets of the entire dataset, it can be expected that they will be better able to generalise than the classical regression model that was trained on the entire dataset. On the other hand, the classical regression model consists of only two terms containing the precursor concentration, compared to the six features in the ensemble regression model, which results in a simpler model and therefore lower computational costs.

3.3. Classification

In the second part of the modelling, the synthesis process is treated as a binary classification problem in which the synthesis parameters are used to predict whether the Cu NPs obtained will be large or small. For this approach as a classification problem, the performance of complex LLMs, trained using transfer learning, is compared with a classic random forest model, trained from scratch.

For LLM-based classification, there is a clear difference between the performance of GPT-J and LLAMA. Where GPT-J performs slightly worse than random guessing, with an accuracy of 0.48 and a κ of -0.04 (**Table 3**), the use of LLAMA 3.1 results in an increase in accuracy to 0.64 and an increase in κ to 0.28 (**Table 3**), although this is a significant performance improvement, the overall performance is relatively low. One of the possible reasons for this limited performance is that the size of the dataset is very limited. Small datasets already poses major challenges for classic ML, such as regression and decision trees, but the use of such small datasets is particularly challenging for complex models as LLMs. The classic random forest classifier has an overall performance between the two LLMs, with an accuracy of 0.56 and a κ of 0.11 (**Table 3**). In this case, the performance is therefore slightly better than simply guessing at random.

The relatively limited difference in performance between LLMs and classic random forests raises questions about the usefulness of complex neural networks in the context of limited lab-scale studies. It is clear that neural networks require a much larger dataset to achieve proper performance. When the dataset is limited, as it is here, there is hardly any difference in performance between neural networks and simpler classical models such as decision trees. Despite the lack of difference in performance, LLMs are much more complex and require much higher computing costs compared to classical models, which can make LLMs less efficient if the computational costs and energy requirements are considered.

Table 3 *- Performance metrics of binary classification models.*

	Accuracy	κ
GPT-J	0.48	-0.04
LLAMA 3.1	0.64	0.28
Random forest	0.56	0.11

4. Conclusions

It can be concluded that even with a small lab-scale dataset, classical ensemble regression ML can make quantitative predictions about the Cu NP synthesis results. More complex models, such as LLMs, on the other hand, clearly have much more difficulty learning from these small lab-scale datasets, as even a simple task such as binary classification is difficult to accomplish. Further research may reveal whether expanding the experimental dataset can lead to further improvements in the performance of the various ML models. However, improving performance by further expanding the dataset raises the new question of what the added value of using ML for synthesis procedure development is. If the models were trained using a nearly exhaustive sampling of the parameter space under consideration, the result would be a computationally expensive database, which would also be time-consuming to construct, rather than a predictive model that helps to accelerate the development of a synthesis procedure. Based on this study, it can be concluded that it makes more sense to invest in creating a high-quality lab-scale dataset in combination with fine-tuned classical ML models, such as ensemble regression, than to invest in complex and expensive LLMs.

Acknowledgements

Firstly, I would like to thank my supervisors, Prof. dr. An Hardy and Prof. dr. dr. Danny Vanpoucke, for giving me the opportunity to carry out the research for this thesis within the DESINe and QuATOMs groups. I would also like to thank them for sharing their expert knowledge and experience in the completion of this thesis. In addition, I would like to thank my mentors, Digvijay Ghogare and Thijs Van Wijk, for their excellent guidance and detailed feedback during the various stages of the research and the writing of the thesis. Furthermore, I would like to thank Pieter de Meyer for developing and sharing the results of the classical regression model, and Prof. dr. Berend Smit and dr. Joren van Herck of EPFL for training the LLMs and sharing their results. Also, I would like to thank Sander Stulens for performing the SAED measurements. Finally, I would also like to thank my friends, family and fellow students for their motivation and support over the past five years.

References

[1] Daniel, M.-C., & Astruc, D. (2004). Gold nanoparticles: Assembly, supramolecular chemistry, quantum-size-related properties, and applications toward biology, catalysis, and nanotechnology. Chemical Reviews, 104(1), 293–346.

[2] Zhai, T., Li, L., Wang, X., Fang, X., Bando, Y., & Golberg, D. (2010). Recent developments in one-dimensional inorganic nanostructures for photodetectors. Advanced Functional Materials, 20(24), 4233–4248.

[3] Gawande, M. B., Goswami, A., Felpin, F.-X., Asefa, T., Huang, X., Silva, R., Zou, X., Zboril, R., & Varma, R. S. (2016). Cu and Cu-based nanoparticles: Synthesis and applications in catalysis. *Chemical Reviews, 116*(6), 3722–3811.

[4] Pankhurst, Q. A., Connolly, J., Jones, S. K., & Dobson, J. (2003). Applications of magnetic nanoparticles in biomedicine. *Journal of Physics D: Applied Physics, 36*(13), R167–R181.

[5] Baig, N., Kammakakam, I., & Falath, W. (2021). Nanomaterials: A review of synthesis methods, properties, recent progress, and challenges. Materials Advances, 2(6), 1821–1871.

[6] Baum, Z. J., Yu, X., Ayala, P. Y., Zhao, Y., Watkins, S. P., & Zhou, Q. (2021). Artificial intelligence in chemistry: Current trends and future directions. *Journal of Chemical Information and Modeling, 61*(7), 3197–3212.

[7] Nobel Prize Outreach. (2024). *The Nobel Prize in Physics 2024 – Press release*. NobelPrize.org. https://www.nobelprize.org/prizes/physics/2024/press-release/

[8] Nobel Prize Outreach. (2024). *The Nobel Prize in Chemistry 2024 – Press release*. NobelPrize.org. https://www.nobelprize.org/prizes/chemistry/2024/press-release/

[9] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. Nature, 559(7715), 547–555.

[10] Schmidt, J., Marques, M. R. G., Botti, S., & Marques, M. A. L. (2019). Recent advances and applications of machine learning in solid-state materials science. NPJ Computational Materials, 5(1), 83.

[11] Sanchez-Lengeling, B., & Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. Science, 361(6400), 360–365.

[12] Huang, G., Guo, Y., Chen, Y., & Nie, Z. (2023). Application of machine learning in material synthesis and property prediction. Materials, 16(17), 5977.

[13] Agrawal, A., & Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials, 4*(5), 053208.

[14] Williamson, A. J., & Brutchey, R. L. (2023). *Using data-driven learning to predict and control the outcomes of inorganic materials synthesis*. Inorganic Chemistry, 62(49), 19284–19295.

[15] Che Lah, N. A. (2025). *Properties optimisation of nanostructures via machine learning: Progress and perspective*. Engineering Analysis with Boundary Elements, 171, 106063.

[16] Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

[17] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE, 104*(1), 148–175.

[18] Ma, Y., Xu, P., Li, M., Ji, X., Zhao, W., & Lu, W. (2024). The mastery of details in the workflow of materials machine learning. *NPJ Computational Materials, 10*, Article 141.

[19] Guda, A. A., Kirichkov, M. V., Shapovalov, V. V., Muravlev, A. I., Pashkov, D. M., Guda, S. A., Bagliy, A. P., Soldatov, S. A., Chapek, S. V., & Soldatov, A. V. (2023). Machine learning analysis of reaction parameters in UV-mediated synthesis of gold nanoparticles. *The Journal of Physical Chemistry C, 127*(2), 1097-1108.

[20] Tao, H., Wu, T., Kheiri, S., Aldeghi, M., Aspuru-Guzik, A., & Kumacheva, E. (2021). Self-driving platform for metal nanoparticle synthesis: Combining microfluidics and machine learning. *Advanced Functional Materials, 31*(2106725).

[21] Epps, R. W., Bowen, M. S., Volk, A. A., Abdel-Latif, K., Han, S., Reyes, K. G., Amassian, A., & Abolhasani, M. (2020). Artificial chemist: An autonomous quantum dot synthesis bot. *Advanced Materials, 32*(2001626).

[22] Che Lah, N. A. (2025). Properties optimisation of nanostructures via machine learning: Progress and perspective. Engineering Analysis with Boundary Elements, 171, 106063.

[23] Pellegrino, F., Isopescu, R., Pellutiè, L., & others. (2020). Machine learning approach for elucidating and predicting the role of synthesis parameters on the shape and size of TiO_2 nanoparticles. *Scientific Reports, 10*, 18910.

[24] Williams, M. K., Batchelor, R., Curry, R. J., & Lermusiaux, P. (2023). Optimizing silver nanowire synthesis: Machine learning improves and predicts yield for a polyol, millifluidic flow reactor. Applied Nanoscience, 13, 3157–3168.

[25] Guo, X., Hao, C., Jin, G., Zhu, H.-Y., & Guo, X.-Y. (2014). Copper nanoparticles on graphene support: An efficient photocatalyst for coupling of nitroaromatics in visible light. *Angewandte Chemie International Edition, 53*(7), 1973–1977.

[26] Reske, R., Mistry, H., Behafarid, F., Roldan Cuenya, B., & Strasser, P. (2014). Particle size effects in the catalytic electroreduction of CO2 on Cu nanoparticles. *Journal of the American Chemical Society, 136*(17), 6978–6986.

[27] Chen, C., Zuo, Y., Ye, W., Li, X., Deng, Z., & Ong, S. P. (2020). A critical review of machine learning of energy materials. *Advanced Energy Materials, 10*(19), 1903242.

[28] Raccuglia, P., Elbert, K. C., Adler, P. D. F., Falk, C., Wenny, M. B., Mollo, A., Zeller, M., Friedler, S. A., Schrier, J., & Norquist, A. J. (2016). Machine-learning-assisted materials discovery using failed experiments. Nature, 533(7601), 73–76.

[29] Fiévet, F., Ammar-Merah, S., Brayner, R., Chau, F., Giraud, M., Mammeri, F., Peron, J., Piquemal, J.-Y., Sicard, L., & Viau, G. (2018). The polyol process: A unique method for easy access to metal nanoparticles with tailored sizes, shapes and compositions. *Chemical Society Reviews, 47*(14), 5187–5233.

[30] Blosi, M., Albonetti, S., Dondi, M., Ferretti, M., & Martelli, C. (2011). Microwave-assisted polyol synthesis of Cu nanoparticles. *Journal of Nanoparticle Research, 13*(1), 127138.

[31] Teichert, J., Doert, T., & Ruck, M. (2018). Mechanisms of the polyol reduction of copper(II) salts depending on the anion type and diol chain length. *Dalton Transactions, 47*(44), 14085–14093.

[32] Carpenter, E. E., Carroll, K. J., Reveles, J. U., Shultz, M. D., & Khanna, S. N. (2011). Preparation of elemental Cu and Ni nanoparticles by the polyol method: An experimental and theoretical approach. *The Journal of Physical Chemistry C, 115*(6), 2656–2664.

[33] Borisut, P., & Nuchitprasittichai, A. (2023). Adaptive Latin hypercube sampling for a surrogate-based optimization with artificial neural network. *Processes, 11*(11), 3232.

[34] Braham, E. J., Davidson, R. D., Al-Hashimi, M., Arróyave, R., & Banerjee, S. (2020). Navigating the design space of inorganic materials synthesis using statistical methods and machine learning. *Dalton Transactions, 49*(33), 11480–11488.

[35] Vanpoucke, D. E. P., van Knippenberg, O. S. J., Hermans, K., Bernaerts, K. V., & Mehrkanoon, S. (2020). Small data materials design with machine learning: When the average model knows best. *Journal of Applied Physics, 128*(5), 054901.

[36] Vanpoucke, D. E. P., Delgove, M. A. F., Stouten, J., Noordijk, J., De Vos, N., Matthysen, K., Deroover, G. G. P., Mehrkanoon, S., & Bernaerts, K. V. (2022). A machine learning approach for the design of hyperbranched polymeric dispersing agents based on aliphatic polyesters for radiation-curable inks. *Polymer International, 71*(8), 966–975.

[37] Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A., & Smit, B. (2024). Leveraging large language models for predictive chemistry. *Nature Machine Intelligence, 6*(2), 161–169.

[38] Van Herck, J., Gil, M. V., Jablonka, K. M., Abrudan, A., Anker, A. S., Asgari, M., Blaiszik, B., Buffo, A., Choudhury, L., Corminboeuf, C., ... & Smit, B. (2025). Assessment of fine-tuned large language models for real-world chemistry and material science applications. *Chemical Science, 16*(3), 670–684.

[39] Torras, M., & Roig, A. (2021). Copper oxide nanocubes wrapping metals by microwave synthesis. *Crystal Growth & Design, 21*(9), 5027–5035.

[40] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[41] Danny E. P. Vanpoucke, AMADEUS v0.2. Available: https://github.com/DannyVanpoucke/Amadeus (2020).

[42] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

[43] Vargas-Urbano, M. A., Marín, L., Castillo, W. M., Rodríguez, L. A., Magén, C., Manotas-Albor, M., Diosa, J. E., & Gross, K. (2022). Effect of ethylene glycol: Citric acid molar ratio and pH on the morphology, vibrational, optical and electronic properties of TiO_2 and CuO powders synthesized by Pechini method. *Materials*, *15*(15), 5266.

[44] International Centre for Diffraction Data. (2001). *PDF-2: Database card No. 04-0836, Copper (Cu), face-centered cubic*. Newtown Square, PA: ICDD.

Supporting Information

Table S1 *- Sampled synthesis parameters used in the first series of 25 experimental syntheses.*

Sample	Theoretical Cu concentration[a] [mM]	Effective Cu precursor[b] [mM]	TMAH[c] [µl]	Time[d] [min]	Temperature[e] [°C]
MOTB241202_1	5.00	5.06	36.0	8	196
MOTB241202_2	4.00	4.21	29.0	4	188
MOTB241202_3	3.00	2.91	21.8	6	198
MOTB241202_4	2.00	2.00	14.6	7	184
MOTB241202_5	9.00	8.82	65.4	4	180
MOTB241202_6	3.00	3.11	21.8	1	194
MOTB241202_7	10.00	10.22	72.8	9	188
MOTB241206_8	9.00	8.82	65.4	16	182
MOTB241206_9	3.00	2.80	21.8	17	182
MOTB241206_10	7.00	6.91	51.0	20	194
MOTB241206_11	7.00	6.81	51.0	14	186
MOTB241206_12	6.00	6.01	43.6	10	192
MOTB241206_13	8.00	8.31	58.2	12	184
MOTB241206_14	1.00	1.10	7.2	15	198
MOTB241206_15	2.00	2.10	14.6	10	196
MOTB241210_16	4.00	3.91	29.0	18	200
MOTB241210_17	4.00	3.91	29.0	14	192
MOTB241210_18	8.00	8.31	58.2	5	180
MOTB241210_19	2.00	2.10	14.6	17	176
MOTB241212_20	8.00	7.91	58.2	11	190
MOTB241212_21	6.00	5.91	43.6	7	178
MOTB241212_22	7.00	6.91	51.0	3	176
MOTB241212_23	9.00	9.02	65.4	13	178
MOTB241212_24	6.00	6.11	43.6	19	186
MOTB241212_25	5.00	5.01	36.0	2	190

[a)]Theoretical concentration of $Cu(OAc)_2$ precursor in 5 ml ethylene glycol in mM; [b)]Effective Cu concentration of $Cu(OAc)_2$ used; [c)]Volume of 2.75 M TMAH in H_2O added to 5 ml ethylene glycol in µl; [d)]Time in minutes reaction mixture is kept on defined temperature; [e)]Temperature in °C of reaction mixture for defined reaction time.

Table S2 *- Dataset A containing 25 datapoints with particle size, the smallest size detected by DLS.*

Sample	Effective Cu concentration[a] [mM]	TMAH[b] [µl]	Time[c] [min]	Temperatur[d] [°C]	Size ± STDEV[e] [nm]
MOTB241202_1	5.06	36.0	8	196	252.50 ± 59.37
MOTB241202_2	4.21	29.0	4	188	228.40 ± 41.47
MOTB241202_3	2.91	21.8	6	198	148.80 ± 25.56
MOTB241202_4	2.00	14.6	7	184	99.90 ± 14.07
MOTB241202_5	8.82	65.4	4	180	183.30 ± 25.42
MOTB241202_6	3.11	21.8	1	194	160.70 ± 27.62
MOTB241202_7	10.22	72.8	9	188	64.65 ± 8.80
MOTB241206_8	8.82	65.4	16	182	40.85 ± 5.20
MOTB241206_9	2.80	21.8	17	182	192.90 ± 28.88
MOTB241206_10	6.91	51.0	20	194	237.90 ± 30.29
MOTB241206_11	6.81	51.0	14	186	167.00 ± 31.15
MOTB241206_12	6.01	43.6	10	192	60.10 ± 6.88
MOTB241206_13	8.31	58.2	12	184	66.02 ± 11.94
MOTB241206_14	1.10	7.2	15	198	102.90 ± 13.51
MOTB241206_15	2.10	14.6	10	196	114.40 ± 17.64
MOTB241210_16	3.91	29.0	18	200	157.90 ± 22.70
MOTB241210_17	3.91	29.0	14	192	304.50 ± 52.21
MOTB241210_18	8.31	58.2	5	180	59.09 ± 6.13
MOTB241210_19	2.10	14.6	17	176	188.10 ± 26.04
MOTB241212_20	7.91	58.2	11	190	50.36 ± 8.16
MOTB241212_21	5.91	43.6	7	178	312.90 ± 55.82
MOTB241212_22	6.91	51.0	3	176	271.60 ± 45.13
MOTB241212_23	9.02	65.4	13	178	127.60 ± 18.55
MOTB241212_24	6.11	43.6	19	186	58.69 ± 10.61
MOTB241212_25	5.01	36.0	2	190	257.60 ± 48.85

[a)]Effective Cu concentration of $Cu(OAc)_2$ used in 5 ml ethylene glycol; [b)]Volume of 2.75 M TMAH in H_2O added to 5 ml ethylene glycol in µl; [c)]Time in minutes reaction mixture is kept on defined temperature; [d)]Temperature in °C of reaction mixture for defined reaction time; [e)] Size of the NPs and the standard deviation (STDEV) in nm as measured using DLS.

Table S3 *- Dataset B containing 25 datapoints with particle size, the size with the highest percentage contribution in DLS.*

Sample	Effective Cu concentration[a] [mM]	TMAH[b] [µl]	Time[c] [min]	Temperatur[d] [°C]	Size ± STDEV[e] [nm]
MOTB241202_1	5.06	36.0	8	196	252.50 ± 59.37
MOTB241202_2	4.21	29.0	4	188	228.40 ± 41.47
MOTB241202_3	2.91	21.8	6	198	148.80 ± 25.56
MOTB241202_4	2.00	14.6	7	184	99.90 ± 14.07
MOTB241202_5	8.82	65.4	4	180	183.30 ± 25.42
MOTB241202_6	3.11	21.8	1	194	160.70 ± 27.62
MOTB241202_7	10.22	72.8	9	188	204.00 ± 34.54
MOTB241206_8	8.82	65.4	16	182	222.30 ± 31.62
MOTB241206_9	2.80	21.8	17	182	192.90 ± 28.88
MOTB241206_10	6.91	51.0	20	194	237.90 ± 30.29
MOTB241206_11	6.81	51.0	14	186	167.00 ± 31.15
MOTB241206_12	6.01	43.6	10	192	269.50 ± 60.33
MOTB241206_13	8.31	58.2	12	184	66.02 ± 11.94
MOTB241206_14	1.10	7.2	15	198	102.90 ± 13.51
MOTB241206_15	2.10	14.6	10	196	114.40 ± 17.64
MOTB241210_16	3.91	29.0	18	200	157.90 ± 22.70
MOTB241210_17	3.91	29.0	14	192	304.50 ± 52.21
MOTB241210_18	8.31	58.2	5	180	184.90 ± 26.97
MOTB241210_19	2.10	14.6	17	176	188.10 ± 26.04
MOTB241212_20	7.91	58.2	11	190	50.36 ± 8.16
MOTB241212_21	5.91	43.6	7	178	312.90 ± 55.82
MOTB241212_22	6.91	51.0	3	176	271.60 ± 45.13
MOTB241212_23	9.02	65.4	13	178	127.60 ± 18.55
MOTB241212_24	6.11	43.6	19	186	58.69 ± 10.61
MOTB241212_25	5.01	36.0	2	190	257.60 ± 48.85

[a)]Effective Cu concentration of $Cu(OAc)_2$ used in 5 ml ethylene glycol; [b)]Volume of 2.75 M TMAH in H_2O added to 5 ml ethylene glycol in µl; [c)]Time in minutes reaction mixture is kept on defined temperature; [d)]Temperature in °C of reaction mixture for defined reaction time; [e)] Size of the NPs and the standard deviation (STDEV) in nm as measured using DLS.

Table S4 *- Dataset C containing only 18 datapoints with monomodal DLS spectrum.*

Sample	Effective Cu concentration[a] [mM]	TMAH[b] [µl]	Time[c] [min]	Temperatur[d] [°C]	Size ± STDEV[e] [nm]
MOTB241202_1	5.06	36.0	8	196	252.50 ± 59.37
MOTB241202_2	4.21	29.0	4	188	228.40 ± 41.47
MOTB241202_3	2.91	21.8	6	198	148.80 ± 25.56
MOTB241202_4	2.00	14.6	7	184	99.90 ± 14.07
MOTB241202_5	8.82	65.4	4	180	183.30 ± 25.42
MOTB241202_6	3.11	21.8	1	194	160.70 ± 27.62
MOTB241206_9	2.80	21.8	17	182	192.90 ± 28.88
MOTB241206_10	6.91	51.0	20	194	237.90 ± 30.29
MOTB241206_11	6.81	51.0	14	186	167.00 ± 31.15
MOTB241206_14	1.10	7.2	15	198	102.90 ± 13.51
MOTB241206_15	2.10	14.6	10	196	114.40 ± 17.64
MOTB241210_16	3.91	29.0	18	200	157.90 ± 22.70
MOTB241210_17	3.91	29.0	14	192	304.50 ± 52.21
MOTB241210_19	2.10	14.6	17	176	188.10 ± 26.04
MOTB241212_21	5.91	43.6	7	178	312.90 ± 55.82
MOTB241212_22	6.91	51.0	3	176	271.60 ± 45.13
MOTB241212_23	9.02	65.4	13	178	127.60 ± 18.55
MOTB241212_25	5.01	36.0	2	190	257.60 ± 48.85

[a)]Effective Cu concentration of $Cu(OAc)_2$ used in 5 ml ethylene glycol; [b)]Volume of 2.75 M TMAH in H_2O added to 5 ml ethylene glycol in µl; [c)]Time in minutes reaction mixture is kept on defined temperature; [d)]Temperature in °C of reaction mixture for defined reaction time; [e)] Size of the NPs and the standard deviation (STDEV) in nm as measured using DLS.

Table S5 *- Resulting sizes and SPR peaks of the different synthesis samples.*

Sample	Size_1 ± STDEV[a] [nm]	Percentage_1[b] [%]	Size_2 ± STDEV[a] [nm]	Percentage_2[b] [%]	Size_3 ± STDEV[a] [nm]	Percentage_3[b] [%]	SPR[d] [nm]
MOTB241202_1	252.50 ± 59.37	99.9	5414.00 ± 647.80[c]	0.1			613
MOTB241202_2	228.40 ± 41.47	99.9	5135.00 ± 775.60[c]	0.1			625
MOTB241202_3	148.80 ± 25.56	100					612
MOTB241202_4	99.90 ± 14.07	100					699
MOTB241202_5	183.30 ± 25.42	100					597
MOTB241202_6	160.70 ± 27.62	100					699
MOTB241202_7	64.65 ± 8.80	29.1	204.00 ± 34.54	70.9			592
MOTB241206_8	40.85 ± 5.20	33.1	222.30 ± 31.62	66.9			590
MOTB241206_9	192.90 ± 28.88	100					609
MOTB241206_10	237.90 ± 30.29	100					591
MOTB241206_11	167.00 ± 31.15	98.9	5164.00 ± 744.50[c]	1.1			636
MOTB241206_12	60.10 ± 6.88	27.1	123.70 ± 22.78	14.1	269.50 ± 60.33	58.80	622
MOTB241206_13	66.02 ± 11.94	53	126.20 ± 18.36	4.7	259.90 ± 45.83	42.30	611
MOTB241206_14	102.90 ± 13.51	100					618
MOTB241206_15	114.40 ± 17.64	100					619
MOTB241210_16	157.90 ± 22.70	100					593
MOTB241210_17	304.50 ± 52.21	100					616
MOTB241210_18	59.09 ± 6.13	30.4	184.90 ± 26.97	69.6			605
MOTB241210_19	188.10 ± 26.04	99.7	5532.00 ± 608.70[c]	0.3			626
MOTB241212_20	50.36 ± 8.16	63.5	233.10 ± 37.46	36.5			595
MOTB241212_21	312.90 ± 55.82	100					609
MOTB241212_22	271.60 ± 45.13	99.9	5380.00 ± 654.70[c]	0.1			601
MOTB241212_23	127.60 ± 18.55	100					595
MOTB241212_24	58.69 ± 10.61	56.4	247.00 ± 39.81	43.6			600
MOTB241212_25	257.60 ± 48.85	100					597

[a)] Hydrodynamic diameter in nm of the Cu NPs and the standard deviation (STDEV) measured using DLS in ethanol; [b)] Percentage contribution of the respective peak to the DLS spectrum; [c)] Due to their irrelevant size and minimal contribution to the DLS spectrum, these peaks will not be considered during the further study; [d)] Location of the SPR peak in nm as measured in ethylene glycol using UV-vis.

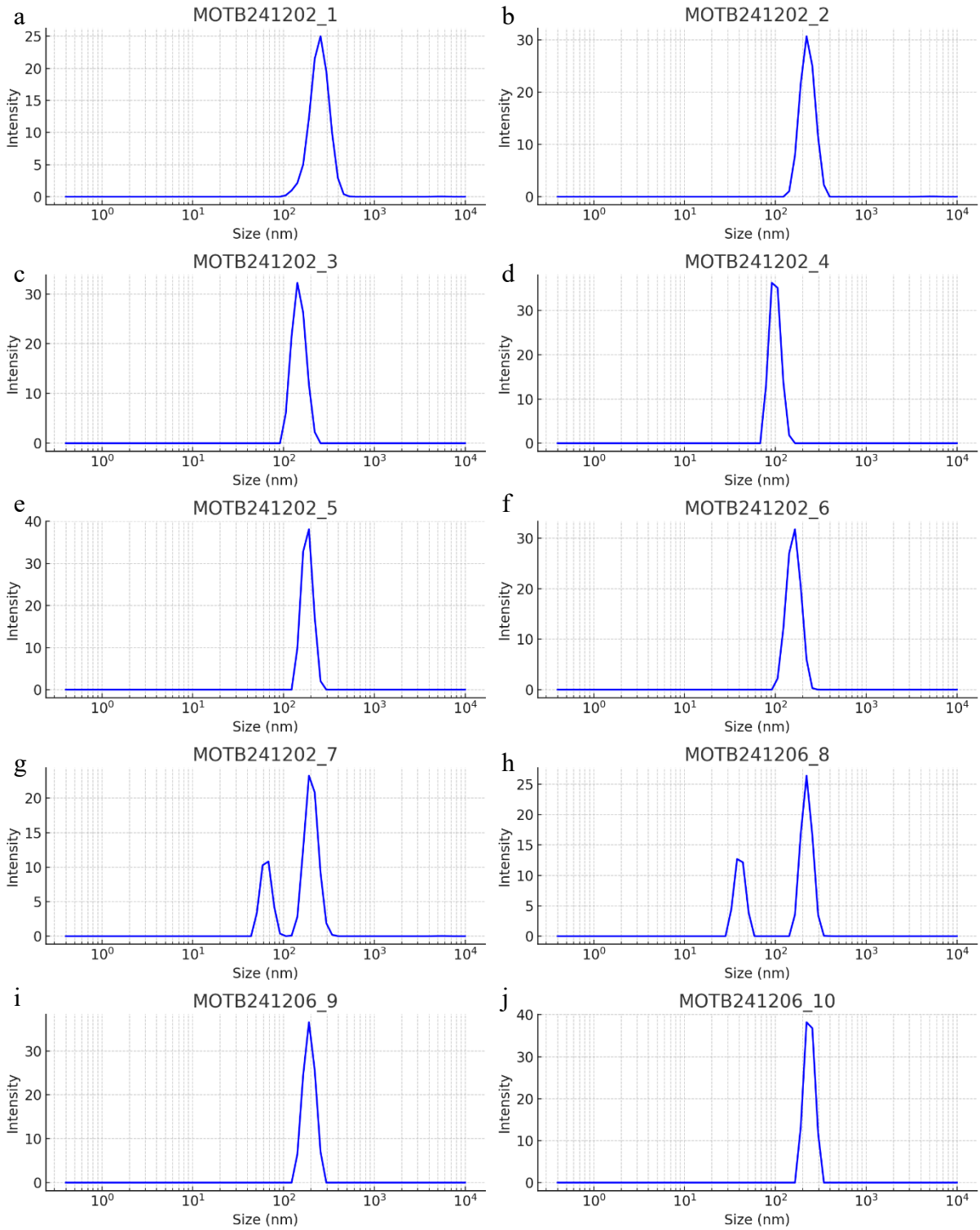

Figure S1 *– DLS spectra of experimental synthesis samples a)MOTB241202_1; b)MOTB241202_2; c)MOTB241202_3; d)MOTB241202_4; e)MOTB241202_5; f)MOTB241202_6; g)MOTB241202_7; h)MOTB241206_8; i)MOTB241206_9; j)MOTB241206_10.*

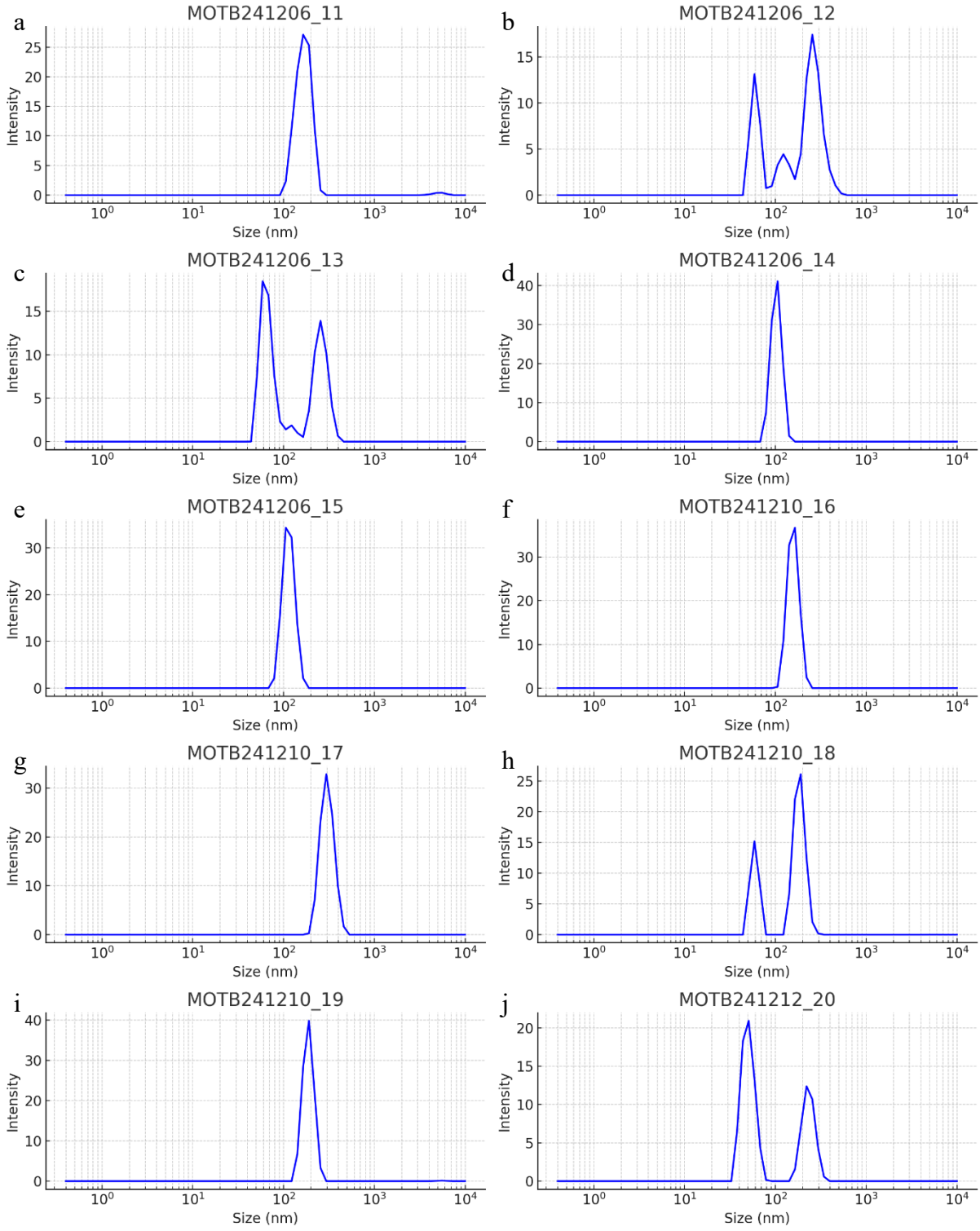

Figure S2 *- DLS spectra of experimental synthesis samples a)MOTB241206_11; b)MOTB241206_12; c)MOTB241206_13; d)MOTB241206_14; e)MOTB241206_15; f)MOTB241210_16; g)MOTB241210_17; h)MOTB241210_18; i)MOTB241210_19; j)MOTB241212_20.*

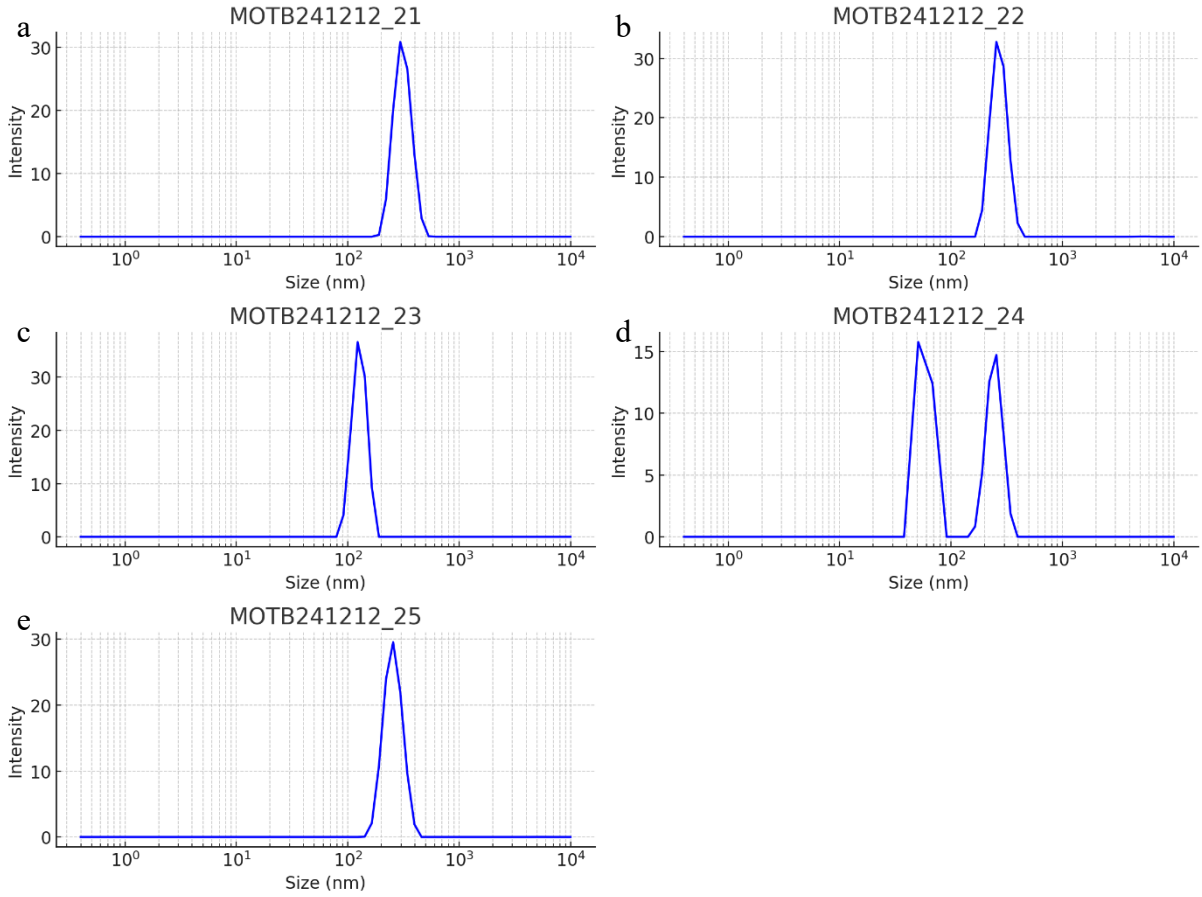

Figure S3 *- DLS spectra of experimental synthesis samples a)MOTB241212_21; b)MOTB241212_22; c)MOTB241212_23; d)MOTB241212_24; e)MOTB241212_25.*

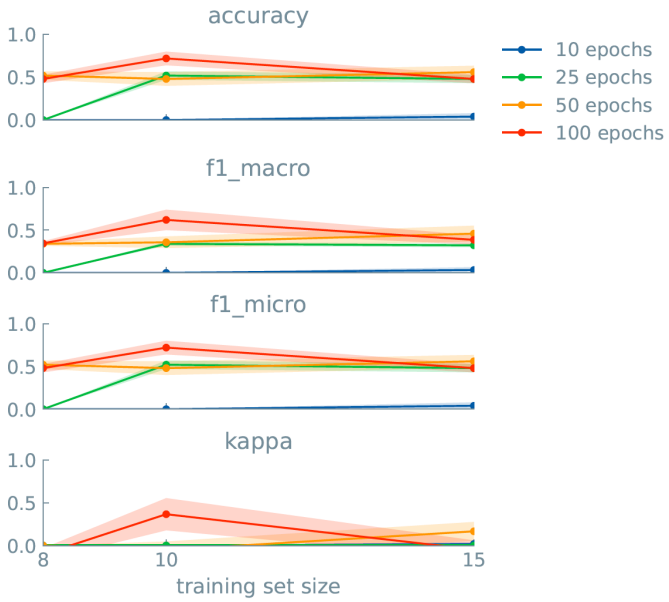

Figure S4 *- Learning curve analysis of predictions for the binary class of the size of nanoparticles. Models fine-tuned with 10 (blue), 25 (green), 50 (orange) and 100 (red) epochs were validated. GPT-J was used as the base model.*

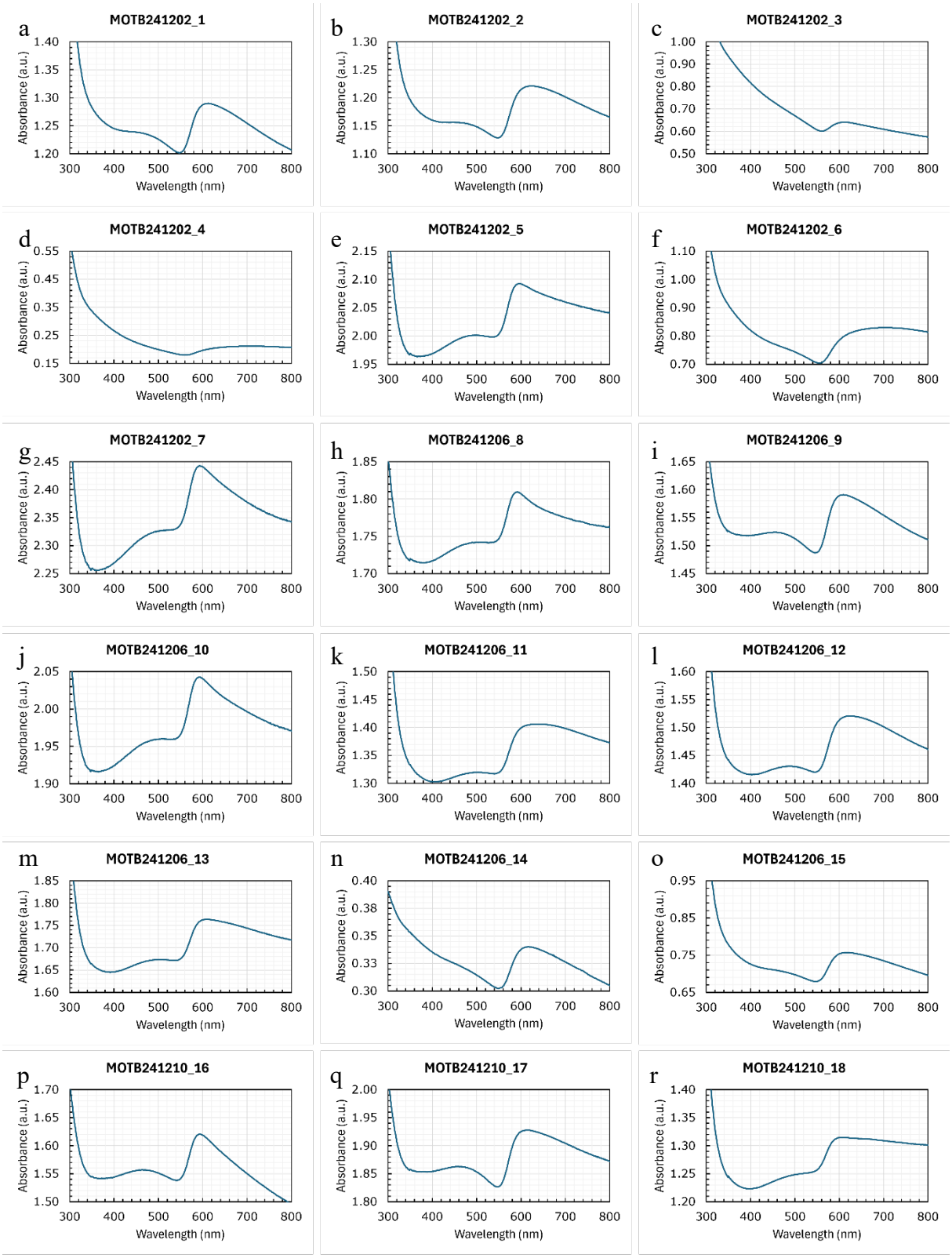

Figure S5 *- UV-vis spectra of samples a)MOTB241202_1; b)MOTB241202_2; c)MOTB241202_3; d)MOTB241202_4; e)MOTB241202_5; f)MOTB241202_6; g)MOTB241202_7; h)MOTB241206_8; i)MOTB241206_9; j)MOTB241206_10; k)MOTB241206_11; l)MOTB241206_12; m)MOTB241206_13; n)MOTB241206_14; o)MOTB241206_15; p)MOTB241210_16; q)MOTB241210_17; r)MOTB241210_18.*

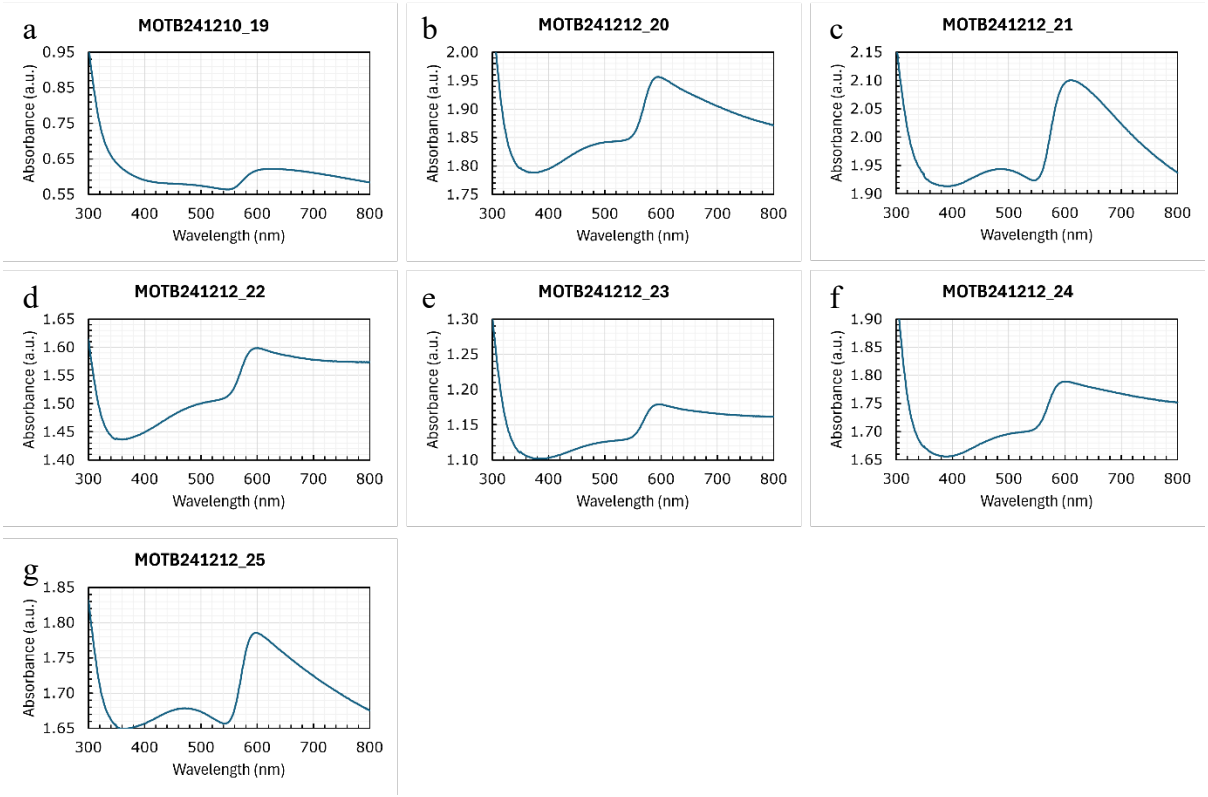

Figure S6 *- UV-vis spectra of samples a)MOTB241210_19; b)MOTB241212_20; c)MOTB241212_21; d)MOTB241212_22; e)MOTB241212_23; f)MOTB241212_24; g)MOTB241212_25.*