CHEMISTRY OF MATERIALS

pubs.acs.org/cm

Article

Machine Learning and Small Data-Guided Optimization of Silica Shell Morphology on Gold Nanorods

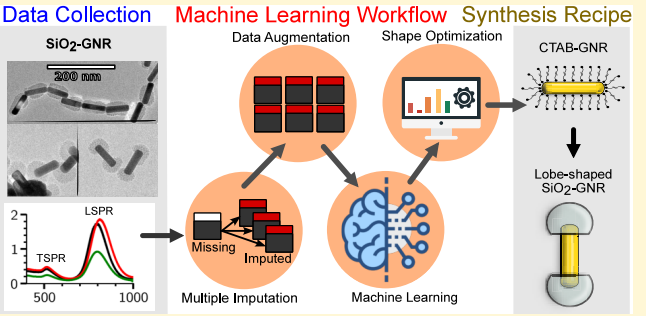
Akhlak U. Mahmood, Melanie M. Ghelardini, Joseph B. Tracy, and Yaroslava G. Yingling*

Cite This: *Chem. Mater.* 2024, 36, 9330−9340

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Anisotropic plasmonic nanorods offer a wide range of applications in photovoltaics, energy conversion, sensing, and surface-enhanced Raman spectroscopies. However, achieving control over the size and shape of the surface overcoating on these nanorods remains a challenge due to the complexity arising from the multistep wet chemical processes involved in their experimental synthesis. Here, we show that by employing data imputation and data augmentation methods, we can minimize the limitations of a small experimental data set and successfully train supervised machine learning models that can optimize the experimental synthesis. Using a small data set collected from 30 multistep syntheses of silica-overcoated gold nanorods (GNRs) characterized by optical extinction spectroscopy and transmission electron microscopy, we trained complementary supervised models to predict the overcoating shape of the nanorods using optical spectral features. The effects of experimental parameters and measurements made during different stages of the synthesis were analyzed. Our approach enabled us to design an experimental synthesis recipe to yield a target SiO_2 overcoating shape on GNRs employing inverse design optimization. The developed workflow can be extended to other plasmonic nanoparticles and multistage synthesis experiments, where a limited data set is available to understand the effects of synthesis parameters and to establish correlations between measurements and synthetic yields.



INTRODUCTION

Gold nanorods (GNRs) are anisotropic nanomaterials that exhibit unique plasmonic properties under the influence of incident light.[1,2] Due to electromagnetic excitation, the quasi-free electrons located on the metallic surface collectively oscillate and induce a local electric field around the particles.[3] The surface plasmon resonances for GNRs generally lay between the visible (vis) and near-infrared (NIR) ranges. Enhancement of the electric field produces large scattering cross-section GNRs. An enhanced electric field is crucial for numerous applications of GNRs in Raman scattering sensors, plasmonic sensors, and photocatalysis.[4,5] Two resonance modes are generally observed in the vis−NIR spectroscopic measurements of GNRs, namely, the longitudinal surface plasmon resonance (LSPR) along the length and the transverse surface plasmon resonance (TSPR) along the width of the nanorods.[6] The intensity and width of the LSPR band are greater and more sensitive to the size, shape, and surfactants of GNRs than those of the TSPR band because of the concentrated electric field at the two ends of the nanorods during longitudinal excitation. Hence, driving the absorption of analytes to the sharp ends of GNRs, rather than the sides, is of great interest to improve the detection capabilities of GNR-based plasmonic sensors.[7] Moreover, mesoporous silica (SiO_2) can be deposited on the surface of GNRs using chemical reactions to drive absorption of molecules as well as to retain colloidal stability, improve biocompatibility, and use the nanorods as molecular cargo for *in vivo* applications by loading the SiO_2 pores with drug molecules.[8−10] In addition, the sensitivity of surface-enhanced Raman scattering using SiO_2-overcoated GNRs is improved by depositing thin layers of SiO_2 on the ends of the GNRs, to which analytes can absorb.[7]

Despite its potential for applications, highly precise control over the deposition of mesoporous SiO_2 shells on GNRs remains challenging. A commonly used method to deposit silica on the GNRs using tetraethyl orthosilicate (TEOS) in an aqueous environment can yield both fully coated GNRs with uniform covering of the surface and lobe-shaped overcoating that covers the ends of the GNRs (Figure 1).[11] The complex chemical synthesis procedure involves a series of steps and depends on the careful selection and control of several reaction variables including pH; concentration and amount of TEOS; concentration of alcohol; temperature, time, and intensity of centrifugation; quality of the initial uncoated GNRs; and native

Received: December 15, 2023
Revised: August 6, 2024
Accepted: August 7, 2024
Published: August 22, 2024

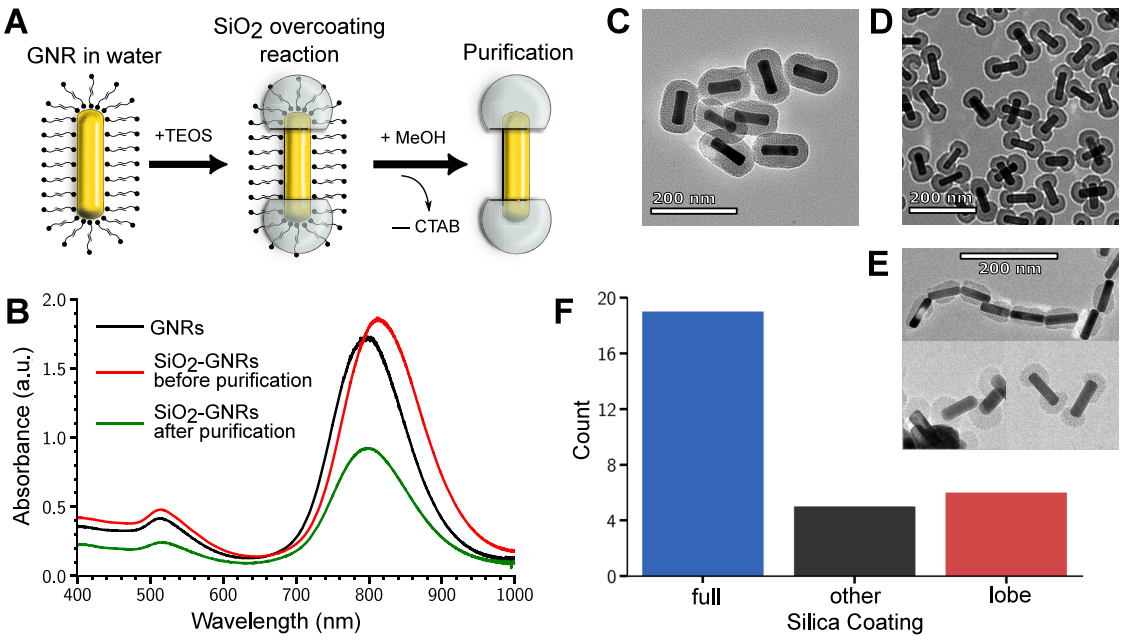


Figure 1. (A) Schematic representation of three stages of lobe-shaped SiO_2-GNR synthesis. (B) Representative optical absorption spectra were measured after each stage. (C,D,E) Representative TEM images of the obtained SiO_2-GNRs labeled as "full", "lobe", and "other", respectively. All scale bars are 200 nm. (F) Distribution of different morphologies of SiO_2 overcoating obtained in 30 syntheses.

stabilizing surfactants on the GNRs such as cetyltrimethylammonium bromide (CTAB). At the end of the synthesis procedure, expensive and time-consuming transmission electron microscopy (TEM) is needed to determine the shape of the product, which is a bottleneck for further functionalization or applications. As a result, consistently producing uniform or lobe-shaped silica overcoating on GNRs suffers from numerous issues including poor reproducibility, aggregation between particles, nonuniform overcoatings, and production of a heterogeneous mixture of nanorods with different types of overcoatings. Adelt et al. recently reported a high yield of lobe-shaped SiO_2-GNRs by carefully controlling the aspect ratio, the concentration of CTAB, and the reaction temperature.[12] Meyer and Murphy reported that the position of the LSPR band exhibits a redshift for both fully coated and lobe-shaped SiO_2-GNRs when exposed to dye molecules.[7] However, the relationship between the anisotropic deposition of silica and the observed properties of the resonance bands is not fully understood.

Machine learning (ML) techniques offer the ability to forecast the outcome of a complex process involving a large number of variables, provide insights about the relationships between the variables, and facilitate an automated decision-making process during synthesis.[13−15] ML algorithms trained on experimental and simulated extinction spectra were used to predict the geometry of nanostructures.[16] Random forest models trained on 186 spectra of gold nanospheres have been successful in predicting the outcome of synthesis and explaining the effects of chemical parameters through SHAP (Shapley Additive exPlanations) analysis, such as concentration of ascorbic acid and the ratio of gold precursor in seed solution.[17] Leveraging Latin hypercube sampling techniques to systematically vary synthesis parameters, dimensionality reduction techniques, and decision trees trained on extinction spectra of 30 syntheses could predict the time of formation and shape of gold nanoparticles with an accuracy of approximately 80%.[18]

Due to the inherent complexity and time-consuming nature of a typical experimental synthesis procedure, high-throughput characterization of GNRs can be quite costly. To overcome the problem of data scarcity, recent studies have employed synthetic data generated from numerical simulation methods for GNRs.[19] The size and aspect ratio of GNRs could be predicted with 90% accuracy using decision tree models trained on 450 scattering spectra obtained from experiments and finite-difference time-domain simulations.[20] However, unlike the uncoated GNRs, there is currently no analytical model that can correctly correlate the anisotropy of silica overcoating to the optical properties of the GNRs, such as the energy or the width of the LSPR and TSPR bands in the extinction spectra. This absence of an analytical model leads to the lack of a numerical simulation method, preventing one from obtaining synthetic data to supplement the ML data set. Thus, the analysis and predictions of SiO_2-coated GNR recipes must depend on smaller data sets collected from experiments.

It is well known that ML models trained on a small data set generally lead to overfitting and fail to accurately predict conditions not encountered during training.[21−23] Although techniques like transfer learning and generative modeling can mitigate these challenges by pretraining on larger related data sets,[14,24−26] this is seldom feasible for chemical reactions and syntheses where input features or descriptors can drastically differ across various domains. For simpler material processes where there are direct relationships between a few of the input features and outcomes, sparse modeling can be used to reduce the complexity of the model by carefully eliminating the noninfluential features and thereby improving the predictive capability of models trained on small data.[27,28] Dimensionality reduction techniques, such as principal component analysis, can improve the model accuracy, albeit often at the expense of interpretability.[23] For more complex processes, where the

https://doi.org/10.1021/acs.chemmater.3c03204
Chem. Mater. 2024, 36, 9330−9340

important features are not known *a priori* or cannot be easily distinguished, ensemble modeling techniques can significantly improve the performance of the ML models trained on small data sets.[29]

Data augmentation through oversampling a small data set can improve the accuracy of ML models, which has been demonstrated by numerous studies in imaging and computer vision-related problems.[24,30−32] This technique also addresses the issue of class imbalance in classification problems, thereby boosting predictive accuracy when the data set has an unequal number of samples for each class. Therefore, data augmentation could facilitate the effective use of small data by reducing overfitting and eliminating class imbalance. Moreover, given the complex nature of chemical synthesis and the challenges in characterization, experimental data sets often contain missing values for multiple key features that might be important for correct prediction.[33] Multiple imputations, a well-established statistical method, can be used to estimate these missing values of the data set while minimizing statistical biases and thereby efficiently using the available features.[34,35] Although data augmentation and multiple imputation techniques hold promise for improving the predictive capabilities of ML models trained on small data sets, their effectiveness remains relatively unexplored in the field of materials informatics.

In this work, using a data set containing vis−NIR spectra collected at different stages of syntheses of SiO_2-coated GNRs, along with corresponding TEM micrographs of the GNRs, we trained decision tree-based classifier models to predict optimal spectral features to yield either fully coated or lobe-shaped GNRs. The initial data set contained characterization data from 30 independent experiments. To expand this data set, we applied multiple imputation using chained equations (MICE) algorithm to estimate the missing values[36] and utilize various popular data augmentation algorithms. Additionally, we introduced a quality factor-based data augmentation algorithm to prioritize syntheses yielding higher-quality SiO_2 overcoating on GNRs. We assessed the model's classification accuracy on a held-out test data set, assessed the impact of multiple imputation and augmentation methods on model performance, and identified the relative importance of spectral parameters to better understand the structure−property relationships of SiO_2-overcoated GNRs. Ultimately, we present a workflow that leverages optimal vis−NIR spectral features to guide the synthesis of lobe-shaped SiO_2-GNRs.

MATERIALS AND METHODS

ML. The vis−NIR spectral features were utilized as inputs for various regression and classification algorithms to predict the likelihood or class of silica shape overcoating, as determined from the TEM images. The training data set comprised the peak position, the bandwidth of TSPR and LSPR peaks, the volume of TEOS used, and the concentration of TEOS employed during synthesis. Depending on the type of algorithm used, regression, or classification, the models' output was the probability or class of "full", "lobe", or "other" shapes. The raw data set underwent cleaning, standardization, and multiple imputation to predict missing values. The resulting data set was small and thus augmented to mitigate overfitting due to the limited data set size. To assess the impacts of imputation and data augmentation, we selected a variety of algorithms, specifically kernel-based Gaussian process regression, decision tree-based XGBoost algorithm, nearest neighbors-based KNN, and support vector-based SVM algorithm. Prior research on GNRs has demonstrated the enhanced predictive capacity of decision tree-based algorithms in predicting the shape and aspect ratio of GNRs. Our study concentrated on a broad spectrum of algorithms to elucidate the influence of imputation and data augmentation. The evaluation of the classification models was conducted using precision and recall metrics, which were derived from a comparison of predicted and actual overcoating shapes present in the test set. The true shapes of the overcoat were ascertained from TEM images. Precision was computed by determining the ratio of true positives (cases where both the predicted and actual shapes were positive) to the aggregate of true positives and false positives (instances where the predicted shape was positive but the actual shape was negative). Conversely, recall was calculated by determining the ratio of true positives to the sum of true positives and false negatives (situations where the predicted shape was negative but the actual shape was positive). The F1-score, a metric that combines precision and recall, was subsequently computed as the harmonic mean of the two metrics.

Feature Extraction, Preprocessing, and Model Training. For each sample of the GNR, three optical absorption spectra were measured throughout the process: after the synthesis and purification of CTAB-coated GNRs, after the overcoating reaction using TEOS, and after purification with methanol. Longitudinal and transverse peak positions and full-width half-max values for each peak were extracted from each spectrum using the SciPy package (v1.7.3).[45] Despite the presence of artifacts in a few of the experimental spectra (Table S2), these did not interfere with the extraction of distinct TSPR and LSPR peak values used to train the ML models.

Though a relatively small number of values (4%) are missing in our data set, training most ML models requires complete matrices. While the most straightforward solution to missing data is to repeat the experiments, the synthesis procedure of SiO_2-GNRs is complex and expensive to perform multiple times. An alternative and simpler approach is to drop the rows containing one or more missing columns during ML model training; however, it would significantly reduce the size of the already small data set (e.g., 22 rows are completely observed out of 30 rows). Missing values can be replaced by zeros, mean, or an average of similar values via data imputation, but such an approach can statistically bias the data set as the correlation between the features and their distribution are ignored in the process. This problem has been studied in statistics, and multiple imputation algorithms have been developed to overcome the potential biases, where the missing values are estimated multiple (usually five) times using ML algorithms. In this work, the MICE algorithm implemented in the mice package (v3.14) of R (v4.1.2)[46] was used for multiple imputations using the default hyperparameter values. All of the features were used as the predictor features of MICE.

Similar to multiple imputations, the data augmentation algorithm assumes that the actual relationship and correlation of the features are sufficiently captured by the distribution of the available data, and thus, additional values can be sampled from the distribution to minimize overfitting. For data augmentation, the imbalance-learn (v0.10.1) Python package was used with the default hyperparameters for each algorithm. For a small data set, it could be beneficial for an ML model to prioritize the data that correspond to the products with high quality to reduce the uncertainty of model prediction while still accounting for the data of low-quality products to improve accuracy. The quality factor was only used to oversample the data via bootstrapping to prioritize the data with higher quality and was not used as a feature for ML models. Quality-based augmentation method was implemented using Python (v3.9.16) and numpy (v1.21.5), with the frequency hyperparameter $F = 2$ and the scaling parameter $s = 0.3$. The algorithm of the method is provided in the Supporting Information.

The XGBoost package (v1.6.1)[47] was used for the eXtreme Gradient Boosting (XGB) algorithm, and the scikit-learn package (v1.2.1)[48] was used for the other ML models. After data augmentation, standard scaling was performed on all features by removing the mean of the data set and scaling by the variance. Three types of feature selection were performed: (1) selection of all features, (2) selection of only the noncollinear features, and (3) selection of features by first training a decision tree classifier on all available features and then using the recursive feature elimination technique implemented by scikit-learn. For hyperparameter tuning, GridSearchCV was used for KNN and SVM classifiers and Random-

https://doi.org/10.1021/acs.chemmater.3c03204
Chem. Mater. 2024, 36, 9330−9340

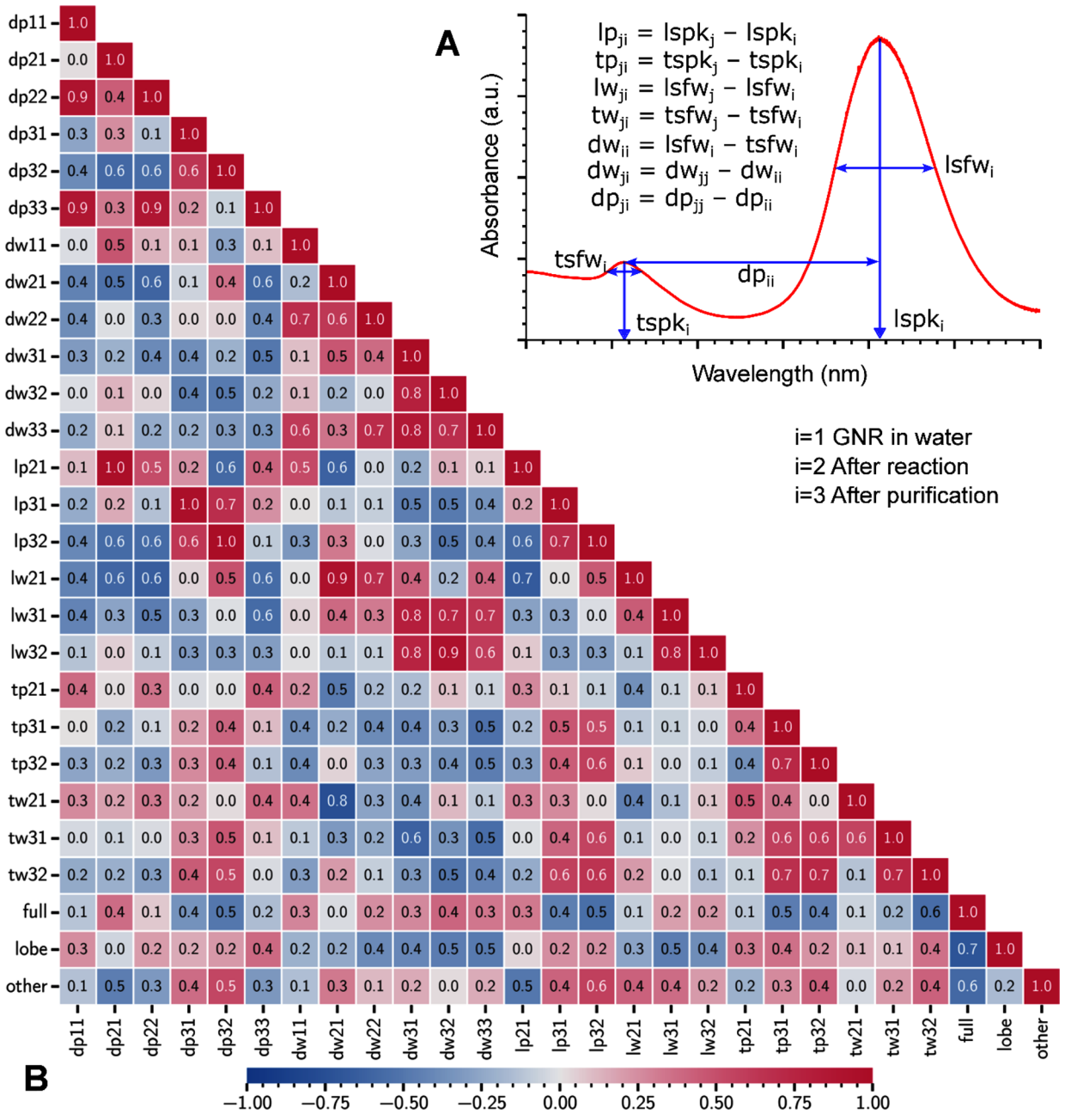


Figure 2. (A) Representation of the features extracted from the absorption spectra measured after three stages of SiO_2-GNRs synthesis. (B) Pearson's correlation matrix of peak and width changes for the LSPR and TSPR plasmon bands. The relative fractions of overcoating shapes (full, lobe, and other) were determined from TEM images. The numbers in the cells denote absolute values, while blue and red colors, respectively, represent negative and positive correlations between the features.

izedSearchCV was used for the XGB classifier; both methods were implemented in scikit-learn. A combination of a radial basis kernel and a white noise kernel was used for the GP classifier, and no additional hyperparameter search was performed for GP since scikit-learn automatically optimizes the kernel parameters during the fitting procedure.

Synthesis and Characterization of Silica-Coated GNRs. GNRs were prepared according to a previously reported method, yielding 190 mg of CTAB-coated GNRs with an LSPR of approximately 800−820 nm in 1 L of water.[6] Several batches of GNRs were prepared with a higher aspect ratio and a more redshifted initial LSPR of approximately 840 nm. SiO_2 was deposited onto the CTAB-GNRs according to Rowe et al. with minor modifications.[11] The aqueous CTAB-coated GNRs were gently heated at 30 °C for 1 h to dissolve excess CTAB, and then isolated by centrifugation (10,000 rpm for 20 min) of two 50 mL aliquots. Approximately 48 mL of supernatant was discarded, and then the volume was brought back up to 50 mL using deionized water. After a second round of centrifugation and supernatant removal using the same conditions, the volume was brought back up to 50 mL using 0.9 mM CTAB water. A final round of centrifugation was performed, 48 mL of supernatant was removed, and the total volume was brought to 10 mL using 0.9 mM CTAB water. The pH was adjusted to 10.0−10.4 using 0.1 M NaOH, and the mixture was heated at 29 °C with gentle stirring. A 20 vol % solution of TEOS:MeOH was then injected into the mixture over 5 min. For 10 out of 30 syntheses, 125 μL of this TEOS solution was used, while the volume for the other syntheses was varied from 50 μL up to 500 μL to control the thickness of the SiO_2 shell, with larger volumes causing deposition of more SiO_2. The mixture was stirred for an additional 30 min and then allowed to age

https://doi.org/10.1021/acs.chemmater.3c03204
Chem. Mater. 2024, 36, 9330−9340

undisturbed at 29 °C for 20 h before undergoing purification via centrifugation. 10 mL of unpurified solution was evenly divided among four centrifuge tubes, and the volume of each tube was brought to 40 mL by adding MeOH. The mixture was centrifuged at 8500 rpm for 10 min, and as much supernatant was removed as possible before adding MeOH to bring the volume back to 40 mL. Three additional rounds of purification by redispersion in MeOH and centrifugation were completed. After the final round the volume of each tube was brought to 2.5 mL by adding MeOH, and the contents of the tubes were combined, giving a total sample volume of 10 mL.

RESULTS AND DISCUSSION

Our experimental procedure of synthesizing silica-coated GNRs (SiO_2-GNRs) closely follows the method reported by Rowe et al.[11] and involves three stages shown in Figure 1A. First, GNRs coated in a CTAB bilayer are synthesized in water. The CTAB-GNRs are approximately 80 nm in length and 20 nm in width. Figure 1B shows vis-NIR absorption spectra measured at different stages of the SiO_2 shell deposition. The peaks of the LSPR and TSPR bands for GNRs are located near 800 and 510 nm, respectively, which are typical for CTAB-GNRs with similar sizes and aspect ratios.[6,37] The LSPR is more sensitive to chemical and environmental changes compared to the TSPR due to the enhancement of the induced local electric field along the length of the GNRs. The SiO_2 overcoating is incorporated onto the GNR surface through hydrolysis of TEOS dispersed in MeOH, and the thickness of the SiO_2 shell is controlled by the amount of TEOS added. Similar to previous reports, a redshift of the LSPR band is observed due to an increase in the refractive index of the medium as SiO_2 deposits on the GNR surface.[9,38] In the third and final stage of the procedure, the SiO_2-deposited CTAB-GNRs are purified four or more times by washing with methanol. The relatively smaller methanol molecules can traverse through the pores of the SiO_2 shell and dissolve the CTAB ligands from the surface of the nanorods to produce pure SiO_2-GNRs. A blueshift is observed in the LSPR peak after purification, moving the peak back to near its initial wavelength observed for CTAB-GNRs dispersed in water. Following this procedure, 30 independent SiO_2-GNRs synthesis experiments were performed and the optical extinction spectra after each stage, namely, (1) after synthesizing CTAB-GNRs and before performing the TEOS overcoating reaction, (2) after the TEOS overcoating reaction, and (3) after purification with methanol, were acquired, yielding a total of 90 individual spectra. Data for seven syntheses were collected from Rowe et al., and 23 syntheses were performed in this work. Details of the experimental procedure are provided in the Materials and Methods section.

It is well known that morphological control of SiO_2 shells on GNRs can be challenging to control.[7,11,12] Different forms of SiO_2 overcoating were obtained in different experimental batches. To quantify the extent of each overcoating type, we manually identified the SiO_2 overcoating shape of each GNR and counted their relative fractions in the representative TEM images for each batch (Figure 1F). By analyzing multiple TEM images, we determined the total number of GNRs with each shape and manually counted the full-, lobe-, and other-shaped silica coatings from the TEM images to calculate each overcoating shape's fraction. During the process, we strictly labeled the SiO_2-GNRs as full or lobe that have SiO_2 overcoating without any visible artifacts (e.g., Figure 1D is labeled as lobe vs the lower-right panel of Figure 1E is labeled as other) while GNRs showing lobes on both nanorod sides were classified as lobe. In contrast, we classified any asymmetric shapes, aggregation, or one-sided overcoating as the other shape. Additional representative TEM images are provided in the Supporting Information.

In principle, one can employ classification and regression algorithms to analyze optical spectroscopy data to predict the three SiO_2 shapes and their relative fractions. For the training of the classification model, we labeled the outcome for each independent experiment as full, lobe, or other SiO_2 morphology based on the maximum fraction of the observed shapes. Out of the 30 independent experiments, 19 yielded SiO_2-GNRs with a majority of full overcoating (Figure 1C), 6 syntheses yielded lobes (Figure 1D), and 5 syntheses yielded a heterogeneous mixtures including aggregates of GNRs, SiO_2 coating on one side, lobe at one end, and chain-like structures (Figure 1E), which we labeled as the other type of SiO_2-GNRs and included in the training data set. The inclusion of negative data is expected to improve the accuracy of the ML models.

Properties of the Optical Extinction Data. To systematically analyze the collected optical extinction spectra data and for ML model training, we identified various spectroscopic features from the extinction spectra and determined the peak positions, widths, and relative distances of the LSPR and TSPR bands as well as the peak shifts in the last two stages of synthesis. We extracted 12 original features corresponding to peak position and bandwidth (measured as the full width at half of the maximum) from the TSPR and LSPR bands that constituted the raw data set. Additionally, 20 derivative features were computed based on the differences between these original features (Figure 2A and Table 1). Peak height, measured in terms of extinction or intensity, was excluded from the analysis due to its arbitrary unit.

Exploratory data analysis reveals a general trend of a redshift in LSPR after overcoating reaction and a blueshift after purification. The distributions of the relevant features after the overcoating reaction and purification grouped by the SiO_2

Table 1. Selected Spectral Features Used for Data Analysis and ML (shown in Figure 2A)[a]

feature	description	type
$tsfw_i$	TSPR bandwidth (full-width half-max)	absolute
$tspk_i$	TSPR peak position	absolute
$lsfw_i$	LSPR bandwidth (full-width half-max)	absolute
$lspk_i$	LSPR peak position	absolute
dp_{ii}	relative distance between TSPR and LSPR peaks	absolute
dw_{ii}	relative size of TSPR and LSPR bandwidths	absolute
lp_{ji}	LSPR peak shift	change
tp_{ji}	TSPR peak shift	change
lw_{ji}	LSPR bandwidth change	change
tp_{ji}	TSPR bandwidth change	change
dp_{ji}	change of relative peak distance	change
dw_{ji}	change of relative size of bandwidths	change

[a]The indices i and j correspond to the measurements at different stages of the synthesis of SiO_2-GNRs (i = 1 for uncoated GNRs, i = 2 for SiO_2-GNRs after the overcoating reaction, and i = 3 for SiO_2-GNRs after purification). Similarly, ii corresponds to differences in feature values at the same stage and ji corresponds to changes between two stages, for example, before and after overcoating reaction.

https://doi.org/10.1021/acs.chemmater.3c03204
Chem. Mater. 2024, 36, 9330−9340

overcoating shapes are shown in Figures S1 and S2. The TSPR bandwidth increases after the overcoating reaction, and the LSPR bandwidth increases after purification for all three SiO_2 shapes. However, for fully overcoated SiO_2-GNRs, one or two syntheses show the opposite trends. Since all SiO_2 overcoating shapes exhibit similar trends in peak shifts and peak broadening, it was not possible to distinguish their morphology based on any of these features. This is further confirmed by the Pearson correlation matrix of the features and fractions of SiO_2 overcoating shapes, shown in Figure 2B. While many features of the absorption spectra are highly correlated with each other, as expected, they do not show a strong correlation with the SiO_2 overcoating shapes. All overcoating shapes (i.e., the fractions of full, lobe, and other GNRs) exhibit a low to moderate correlation with the features corresponding to changes in LSPR and TSPR peaks with a maximum correlation of 0.6. The low correlations between the shapes and spectral features suggest that complex, nonlinear relationships likely exist between them. This validates the use of ML algorithms to capture intricate relationships. Further analyses involving peak positions, widths, and TEOS volume also show similar low to moderate correlations (see Figure S3).

Multiple Imputation and Data Augmentation. Before training the supervised ML algorithms, we performed extensive preprocessing of the collected optical extinction spectra data set to minimize the effects of its small size. In the multistage synthesis procedures of nanoparticles, agglomeration of the nanoparticles at later stages of the synthesis or during purification can impede effective characterization. In our experiments, due to difficulties in experimental characterization and particle agglomeration in later stages of several syntheses, 4% of the values in the collected data were missing (Figure S4A) such as values after the SiO_2 overcoating reaction and purification and the volume of TEOS used during the reaction. This phenomenon, although commonly encountered in experimental synthesis, is seldom reported in the scientific literature. The existence of these missing values within the spectral descriptors necessitated the application of imputation methods in our study. We performed data imputation using the MICE algorithm to more accurately estimate the missing values in the data set (see Materials and Methods). The imputation method generated five complete data sets by estimating the missing features. The Kolmogorov−Smirnov (KS) test was used on the imputed data to quantify the uncertainty of the estimated values. The imputation method estimated six of the nine features related to the peak positions of LSPR and TSPR within a 5% confidence interval, and three features had slightly higher uncertainties (see Figure S4B).

In addition to missing values, the collected data set is imbalanced with a predominance of fully overcoated SiO_2 shell (Figure 1F). This imbalance could potentially bias an ML model to favor the more frequently occurring classes during prediction.[39] To address this, we performed data augmentation to account for the class imbalance along with a regularization technique to prevent overfitting. Numerous data augmentation algorithms have been established within ML research. For example, the random oversampling introduces new data points for the less frequent classes by randomly sampling existing data via a bootstrapping technique to add slight variations to the sample distribution.[40] The Synthetic Minority Oversampling Technique (SMOTE)[39] and the Adaptive Synthetic (ADASYN)[41] methods generate new samples via interpolation of the existing data points, using the *K*-nearest neighbors (KNN) classifier. The ADASYN method oversamples the data, where KNN struggles to classify the response accurately; however, SMOTE does not differentiate between the correct and incorrect predictions when generating synthetic samples. The BorderlineSMOTE method, a variant of SMOTE, attempts to generate samples near the decision boundary of the classifier.[42] All these algorithms aim to balance the data set by oversampling under-represented classes until each class has an approximately equal representation. This balanced, augmented data set helps to mitigate any algorithmic biases introduced in ML models that could arise due to over-representation of certain classes. The representative distributions of the oversamples generated by the algorithms are shown in Figure S5.

It is important to note that the abovementioned data augmentation algorithms assume that all available data are of equal importance or quality. However, this assumption may not hold true in many chemical experiments, where the quality of the product can vary within batches. For example, an expert might consider the batches of SiO_2-GNRs shown in Figure 1C,D to be of higher quality than those where a mixture of different SiO_2 shapes occurs. Therefore, when data augmentation techniques are applied in such contexts, it may be beneficial to introduce a quality factor to prioritize high-quality data points. We defined a "quality" column in the data set, where each batch of synthesized nanorods was manually rated by a domain expert on a scale of 1 to 10 to distinguish the higher-quality nanorods using the TEM images and extinction spectra. We implemented a quality-based data augmentation algorithm that replicates the available data based on this "quality" factor following the bootstrapping technique similar to the random oversampling algorithm. Additional details of multiple imputation, data augmentation, distribution of the generated samples, and effects of hyperparameters are provided in Figures S6 and S7.

Training of Supervised Models. To assess the impact of data imputation and augmentation on the performance of ML models, we selected and trained classification models from different families of ML algorithms. We chose relatively simpler algorithms, such as Gaussian process (GP), KNN, and support vector machine (SVM), to minimize model complexity. A slightly more sophisticated XGB classifier was chosen because depending on hyperparameters, it can approximate the random forest and decision tree classifiers, with and without boosting; has potential for lower model bias than bagging algorithms such as ExtraTrees; and has suitability for low-noise and complex data structures.[43] We stratified the data into test and training sets before performing imputation and data augmentation and randomly selected six data points for the test set with a condition to choose at least one data point from each of the overcoating shapes to reliably calculate the accuracy of the predictions. The held-out test data set was kept fixed to compare the performance across the algorithms. Contrary to the test data set, the size of the training data set varied from 185 to 805 for the imputed data and from 28 to 135 for the completely observed data depending on the imputation and data augmentation algorithms. The fluctuation in the size of the training data set is due to data augmentation and imputation methods. Initially, there were 30 experiments, with 22 being fully observed and 8 having missing values. Without multiple imputations, the ML algorithms are able to use the 22 complete data points. However, with the application of multiple imputation, the data from all 30 experiments

https://doi.org/10.1021/acs.chemmater.3c03204
Chem. Mater. 2024, 36, 9330−9340

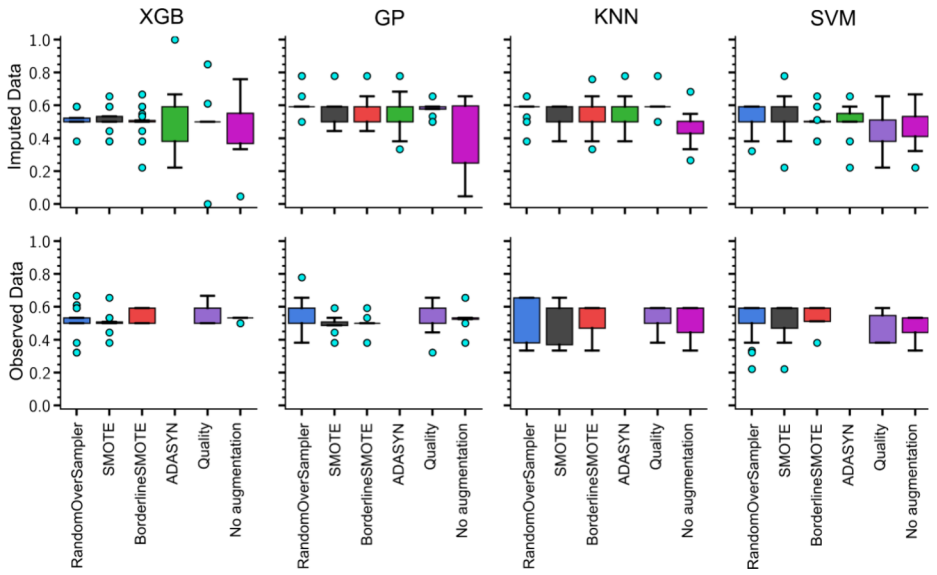


Figure 3. Distribution of F1-scores grouped by imputation, data augmentation, and classification algorithms obtained by training the algorithms with different combinations of scaling, feature selection, and hyperparameters tuning via 1440 ML pipelines. The boxplots are colored by data augmentation methods, and the green circles show the outliers. The outliers indicate superior performances of certain pipelines over the others.

became usable for training the ML models, because the missing values were filled. Similarly, data augmentation algorithms were employed to oversample the 30 data points, thereby generating more training data. This led to a varying quantity of training data sizes, as each of the 30 data points could be oversampled several times, depending on the algorithm used.

Different combinations of imputation and augmentation methods as well as standard scaling, feature selection, and hyperparameter tuning were performed on 1440 individual pipelines prior to fitting different ML models (Figure 3). We employed all features extracted from extinction spectra from the three synthesis stages as inputs for the ML models to predict the class of the silica overcoating shape on GNRs. The TEM images were independently acquired after synthesis and purification. We found that model performance, quantified using the weighted F1-score calculated from the harmonic mean of precision and recall, was significantly affected by a combination of data imputation and augmentation methods, scaling, choice of features, and hyperparameters. The distribution of the obtained F1-scores of predictions on the test set is shown in Figure 3.

The boxplot outliers in the F1-score distributions represent the maximum performance obtained for a combination of model and imputation methods. All algorithms performed similarly when trained on the completely observed data, regardless of the method used to augment the data set. The GP model attained the highest F1-score of 0.78 (obtained as the outliers). The ADASYN algorithm failed to augment the small training set with 16 rows due to an insufficient number of nearest neighbors and therefore could not be evaluated without performing imputation. With multiple imputations, the classification accuracy of the algorithm slightly improved. The maximum score of 1.00 was obtained by the XGB classifier for ADASYN augmentation, and a score of 0.85 was obtained twice when trained on the quality-augmented data. The GP and KNN classifiers performed similarly, with the maximum F1-scores varying between 0.76 and 0.78. The SVM classifier also obtained an accuracy of 0.78 when trained on the data augmented by SMOTE. While XGB showed the best performance in specific cases, it also demonstrated a wide range of F1-scores due to overfitting with a nonoptimal combination of hyperparameters. We were not able to reproduce the 1.00 accuracy score in the subsequent runs of the XGB classifier, which indicates that the randomness of the oversampling method may have contributed to such perfect accuracy. Overall, the plot showed the importance of a hyperparameter when dealing with small data sets, as well as the potential improvement that could be gained by multiple imputations and data augmentation. The details of the model training, feature selection method, and hyperparameters are discussed in the Materials and Methods section.

Feature Importance. Different sets of features were chosen by the recursive feature elimination methods used in each of the 1440 ML pipelines tested for model training. The most frequently selected features across the pipelines are listed in Figure 4A. Surprisingly, only the features with the absolute values of the LSPR or TSPR bands were consistently chosen regardless of the scaling, data imputation, or augmentation methods applied. Among the top five most frequently selected features, the volume and volume % of TEOS were chosen the most, followed by optical properties. Such sensitivity of overcoating shape to the volume and concentration of TEOS and methanol have been previously reported.[11] The plot also shows that both TSPR and LSPR features, for example, the TSPR bandwidth after the SiO_2 overcoating reaction ($tsfw_2$), followed by the LSPR bandwidth after purification ($lsfw_3$) were important to distinguish between the SiO_2 overcoating shapes. Furthermore, the TSPR bandwidth of the initial CTAB-GNRs ($tsfw_1$) also played a role to discriminate between the SiO_2 shapes.

Inverse Design via Multiobjective Optimization. To understand the role of each of the most important features in determining the SiO_2 overcoating shapes and to guide future experiments, we utilized the best-performing ML pipeline (multiple imputations, quality-augmentation, and the XGB algorithm) to maximize the probability of either lobe or full shape and simultaneously minimize the probability of obtaining the other shapes predicted by the ML model. The

https://doi.org/10.1021/acs.chemmater.3c03204
Chem. Mater. 2024, 36, 9330−9340

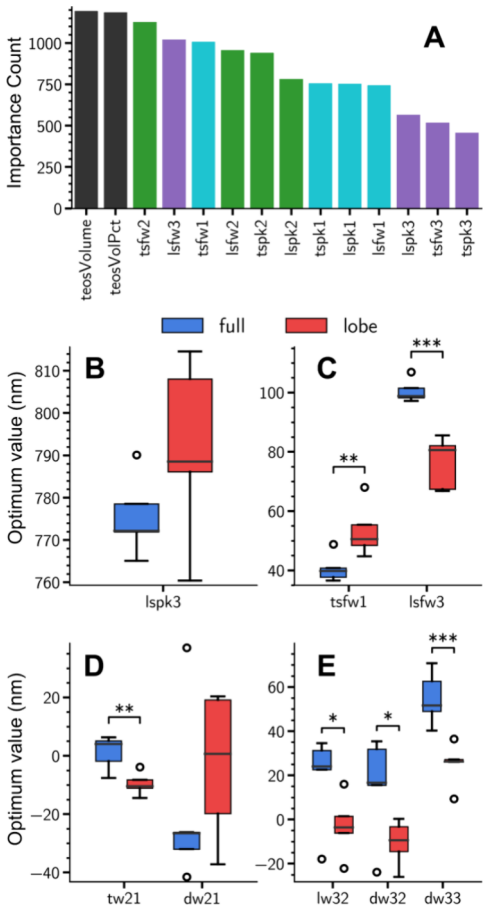


Figure 4. (A) Frequency of each feature selected by recursive feature elimination method in 1440 pipelines tested for ML model training. Features from the same stage of synthesis are colored similarly. (B–E) Distribution of the optimum values of the statistically relevant features predicted by the XGB algorithm to obtain either full- or lobe-shaped SiO_2-GNRs. The model was individually trained on each of the five multiply imputed complete data sets. The brackets and stars are statistical annotations indicative of p values (*0.1, **0.05, ***0.01) calculated using Welch's t test. The circles represent outliers in the predicted optimum values.

optimization process produced the optimum values for each of the important features corresponding to a target shape (i.e., full or lobe) while different sets of optimum values were obtained for the five imputed data sets. Only six features (namely, $lsfw_3$, dw_{33}, $tsfw_1$, tw_{21}, lw_{32}, and dw_{32}) were found to be statistically significant, as shown in Figure 4 to EB, which had large differences in the distributions of optimum values depending on the fully coated or lobe-shaped SiO_2 overcoating on GNRs. Among the obtained features, one feature related to the energy of the band (the LSPR peak position after purification, $lspk_3$) was found to have a distinguishable effect on the SiO_2 morphology. While the difference was not statistically significant, the ML model suggested that the position of the LSPR peak was expected to be higher for lobe-shaped SiO_2-GNRs than for fully coated rods.

All other optimal feature values that serve to differentiate among various SiO_2 morphologies are associated with the widths of the plasmon bands. The width of the LSPR band after purification ($lsfw_3$) and the difference between the width of the TSPR and LSPR bands (dw_{33}) have highly significant (p value = 0.01) shape-distinguishing effects. Furthermore, a broader TSPR peak of CTAB-GNRs ($tsfw_1$) and peak shrinking by approximately 10 nm after the overcoating reaction (tw_{21}) are suggested by the model to obtain lobe-shaped SiO_2 overcoating. The bandwidth of the freshly synthesized CTAB-GNRs generally indicates the polydispersity of the GNRs. The reduction of the TSPR bandwidth after the SiO_2 overcoating reaction indicates a damping process of the collective oscillation of the electrons due to changes in the local dielectric environment along the width of the nanorods.[44] Such a reduction in TSPR bandwidth in the lobe-shaped nanorods (compared to the fully coated ones) is expected because the lobe-shaped SiO_2-GNRs are not coated along the width of the nanorods. Overall, the approach identifies several important correlations between optimal feature values and SiO_2 overcoating shapes, which are crucial for distinguishing between various overcoating shapes based on optical spectral measurements.

Based on the obtained distributions of optimum values, we formulated a synthesis recipe for producing lobe-shaped SiO_2-GNRs. To synthesize an NP with the target lobe overcoating, optical extinction spectra should be measured after each stage of the synthesis process and compared against the recommended values from the ML model. Figure 5A summarizes the developed ML workflow for predicting the set of optimum values for the target SiO_2 shapes. Initially, optical extinction spectra were collected at different stages of the experimental process, used as training data, and then passed through the multiple imputation models. Each batch of syntheses was weighted in the data augmentation procedure using a quality factor set by a domain expert. A supervised classification model (or regression model, as applicable) was then trained using the optical spectra as input features to predict the shape of SiO_2-GNRs (or fraction of each shape) obtained in the TEM images of the synthesized SiO_2-GNRs. Finally, using the best-performing ML model, a genetic algorithm-based multi-objective shape optimization algorithm was employed to predict the set of optimum values for achieving the target shapes.

The measurement of the extinction spectra provides a way to guide the synthesis of a target overcoating shape. Figure 5B shows the conditions set by the distributions of the optimum values as depicted in Figure 4B–E. For example, to obtain lobe-shaped SiO_2-GNRs, the condition $tsfw_1 > 50$ nm indicates that after the first stage of synthesis of CTAB-GNRs, the measured spectra should exhibit a TSPR bandwidth greater than 50 nm. Likewise, the condition $tw_{21} < -10$ nm indicates that the bandwidth should decrease by at least 10 nm following the SiO_2 overcoating reaction in the second stage. For the third and final stage of synthesis (i.e., purification), the model proposes four conditions to increase the likelihood of yielding lobe-shaped SiO_2-GNRs: (1) the LSPR bandwidth should be below 80 nm, (2) the difference between the bandwidths of LSPR and TSPR should be less than 40 nm, (3) the LSPR peak should decrease by 5 nm, and (4) the difference between the LSPR and TSPR bandwidths should decrease after the purification stage. While criteria 1 and 2 depend on the underlying properties of the uncoated GNRs, the ML predictions indicate that failure to meet these conditions increases the probability of obtaining a fully overcoated SiO_2-GNR.

Because the developed ML pipeline attempts to predict the optimum values of the optical properties for a target SiO_2 shape based on the available data, the prediction accuracy of

https://doi.org/10.1021/acs.chemmater.3c03204
Chem. Mater. 2024, 36, 9330−9340

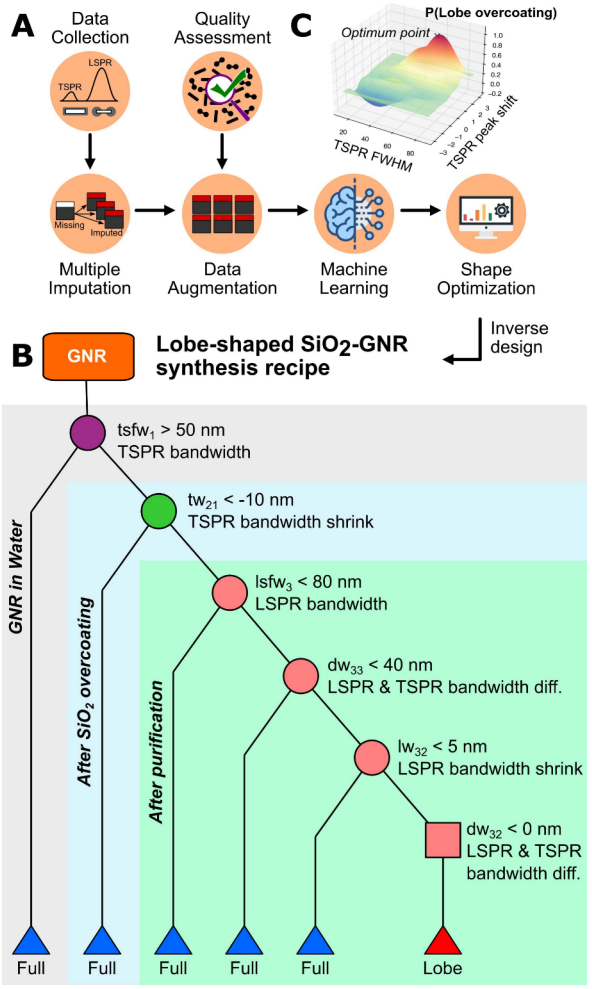


Figure 5. (A) Overall workflow of the small data-driven ML procedure to synthesize SiO_2-GNRs. (B) Decision tree representation of the optimum values of the most important features. The conditions imposed by each of the features guide the synthesis of the lobe-shaped SiO_2-GNRs at each of the three different stages. (C) Representative response surface of two TSPR-related features that shows the position of the most optimum feature values to maximize the probability for obtaining lobe-shaped SiO_2 overcoating.

the model could be iteratively improved by adding new experimental data to the pipeline and by following an active learning scheme to efficiently determine the optimal synthesis protocol to produce the target GNRs. Data imputation and augmentation reduce the tendency of the ML model to overfit the small available data. However, similar to any other statistical sampling techniques, the imputation and augmentation methods would be limited if the available training data is not representative of the true nature of the underlying physical process. Additionally, the model's predictive ability is highly sensitive to the quality of the optical extinction and TEM data. While ML tools could be applied to understanding the role of different SiO_2 shell thicknesses, the focus here is on the shape of the SiO_2 shell. ML predictions for significantly different shell thicknesses may have large deviations from those of experiments.

CONCLUSIONS

We investigated the relationship between the optical extinction spectra and the morphology of SiO_2 overcoating of plasmonic GNRs using multiple imputation and data augmentation algorithms on a small experimental data set consisting of 30 syntheses. Our ML model accelerated the characterization of SiO_2-overcoated GNRs by deducing whether the SiO_2 shell was lobed or uniform from extinction spectra instead of TEM. To overcome the challenges like overfitting, data quality, and class imbalance inherent to the small data set, we implemented ML pipelines featuring multiple imputation and data augmentation techniques. The top-performing ML model, validated on a held-out test data set, identified several statistically significant correlations between the plasmon resonance bands and morphology of SiO_2 overcoating on GNRs. The model proposed specific conditions in the optical extinction spectra that could guide the synthesis process to yield either fully overcoated or lobe-shaped SiO_2-GNRs. Furthermore, we demonstrated that multiple imputations can be used to deal with the missing values often encountered in experimental data sets due to difficulties in multistep synthesis and characterizations. Overall, our study identified optimal feature values, such as the width of the LSPR band and the difference between the widths of the TSPR and LSPR bands, which were key in distinguishing between various SiO_2 morphologies. These values revealed important correlations between these features and the shape of the SiO_2 overcoating, providing a basis for distinguishing between lobed and uniform shapes solely on the basis of optical extinction spectra.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge via the Internet. The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemmater.3c03204.

Python and R code, additional analyses and plots including distribution of features, correlations, samples produced by data augmentation algorithms, uncertainty of the imputation method, and raw experimental TEM images and spectra (PDF)

Data set containing missing values and ML descriptors (XLSX)

AUTHOR INFORMATION

Corresponding Author

Yaroslava G. Yingling − *Department of Materials Science and Engineering, NC State University, Raleigh, North Carolina 27695, United States;* orcid.org/0000-0002-8557-9992; Email: yara_yingling@ncsu.edu

Authors

Akhlak U. Mahmood − *Department of Materials Science and Engineering, NC State University, Raleigh, North Carolina 27695, United States;* orcid.org/0000-0002-5607-2885

Melanie M. Ghelardini − *Department of Materials Science and Engineering, NC State University, Raleigh, North Carolina 27695, United States*

Joseph B. Tracy − *Department of Materials Science and Engineering, NC State University, Raleigh, North Carolina 27695, United States;* orcid.org/0000-0002-3358-3703

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemmater.3c03204

https://doi.org/10.1021/acs.chemmater.3c03204
Chem. Mater. 2024, 36, 9330−9340

Author Contributions

M.M.G. and J.B.T. designed and performed the synthesis and measurements of SiO_2-GNRs. A.U.M. and Y.G.Y. performed the data augmentation and multiple imputation, trained the ML models, analyzed the data, and discussed the results. All the authors wrote, edited, revised, and approved the final manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was funded by National Science Foundation (NSF) grant number CMMI-1763025 and Science and Technology and by the Science and Technologies for Phosphorus Sustainability (STEPS) Center, an NSF Science and Technology Center (CBET-2019435). This work was performed in part at the Analytical Instrumentation Facility (AIF) at North Carolina State University, which is supported by the State of North Carolina and the National Science Foundation (ECCS-2025064). This work made use of instrumentation at AIF acquired with support from the National Science Foundation (DMR-1726294). The AIF is a member of the North Carolina Research Triangle Nanotechnology Network (RTNN), a site in the National Nanotechnology Coordinated Infrastructure (NNCI).

■ REFERENCES

(1) Liang, Z.; et al. Gold Nanorods@Mesoporous SiO2@Hyaluronic Acid Core−Shell Nanoparticles for Controlled Drug Delivery. *ACS Appl. Nano Mater.* **2022**, *5* (5), 7440−7448. May

(2) Greybush, N. J.; et al. Dynamic Plasmonic Pixels. *ACS Nano* **2019**, *13* (4), 3875−3883. Apr.

(3) Fong, K. E.; Yung, L.-Y. L. Localized surface plasmon resonance: a unique property of plasmonic nanoparticles for nucleic acid detection. *Nanoscale* **2013**, *5* (24), 12043.

(4) Yang, H.; He, H.; Tong, Z.; Xia, H.; Mao, Z.; Gao, C. The impact of size and surface ligand of gold nanorods on liver cancer accumulation and photothermal therapy in the second near-infrared window. *J. Colloid Interface Sci.* **2020**, *565*, 186−196. Apr.

(5) Zheng, J.; et al. Gold Nanorods: The Most Versatile Plasmonic Nanoparticles. *Chem. Rev.* **2021**, *121* (21), 13342−13453. Nov.

(6) Kozek, K. A.; Kozek, K. M.; Wu, W.-C.; Mishra, S. R.; Tracy, J. B. Large-Scale Synthesis of Gold Nanorods through Continuous Secondary Growth. *Chem. Mater.* **2013**, *25* (22), 4537−4544.

(7) Meyer, S. M.; Murphy, C. J. Anisotropic silica coating on gold nanorods boosts their potential as SERS sensors. *Nanoscale* **2022**, *14* (13), 5214−5226. Mar.

(8) Castillo, R. R.; Vallet-Regí, M. Functional Mesoporous Silica Nanocomposites: Biomedical Applications and Biosafety. *Int. J. Mol. Sci.* **2019**, *20* (4), 929. Art. no. 4

(9) Fernández-Lodeiro, A.; Djafari, J.; Fernández-Lodeiro, J.; Duarte, M. P.; Muchagato Mauricio, E.; Capelo-Martínez, J. L.; Lodeiro, C.; et al. Synthesis of Mesoporous Silica Coated Gold Nanorods Loaded with Methylene Blue and Its Potentials in Antibacterial Applications. *Nanomaterials* **2021**, *11* (5), 1338. Art. no. 5

(10) Pellas, V.; et al. Gold Nanorods for LSPR Biosensing: Synthesis, Coating by Silica, and Bioanalytical Applications. *Biosensors* **2020**, *10* (10), 146. Art. no. 10

(11) Rowe, L. R.; Chapman, B. S.; Tracy, J. B. Understanding and Controlling the Morphology of Silica Shells on Gold Nanorods. *Chem. Mater.* **2018**, *30* (18), 6249−6258. Sep.

(12) Adelt, M.; MacLaren, D. A.; Birch, D. J. S.; Chen, Y. Morphological Changes of Silica Shells Deposited on Gold Nanorods: Implications for Nanoscale Photocatalysts. *ACS Appl. Nano Mater.* **2021**, *4* (8), 7730−7738. Aug.

(13) Revignas, D.; Amendola, V. Artificial Neural Networks Applied to Colorimetric Nanosensors: An Undergraduate Experience Tailorable from Gold Nanoparticles Synthesis to Optical Spectroscopy and Machine Learning. *J. Chem. Educ.* **2022**, *99* (5), 2112−2120. May

(14) Kim, Y.; Kim, Y.; Yang, C.; Park, K.; Gu, G. X.; Ryu, S. Deep learning framework for material design space exploration using active transfer learning and data augmentation. *Npj Comput. Mater.* **2021**, *7* (1), 140. Art. no. 1

(15) Liu, Z.; Zhu, D.; Raju, L.; Cai, W. Tackling Photonic Inverse Design with Machine Learning. *Adv. Sci.* **2021**, *8* (5), 2002923.

(16) Malkiel, I.; Mrejen, M.; Nagler, A.; Arieli, U.; Wolf, L.; Suchowski, H. Plasmonic nanostructure design and characterization via Deep Learning. *Light Sci. Appl.* **2018**, *7* (1), 60. Art. no. 1

(17) Schletz, D.; Breidung, M.; Fery, A. Validating and Utilizing Machine Learning Methods to Investigate the Impacts of Synthesis Parameters in Gold Nanoparticle Synthesis. *J. Phys. Chem. C* **2023**, *127* (2), 1117−1125. Jan.

(18) Guda, A. A.; et al. Machine Learning Analysis of Reaction Parameters in UV-Mediated Synthesis of Gold Nanoparticles. *J. Phys. Chem. C* **2023**, *127* (2), 1097−1108. Jan.

(19) Pashkov, D. M.; et al. Quantitative Analysis of the UV−Vis Spectra for Gold Nanoparticles Powered by Supervised Machine Learning. *J. Phys. Chem. C* **2021**, *125* (16), 8656−8666. Apr.

(20) Shiratori, K.; et al. Machine-Learned Decision Trees for Predicting Gold Nanorod Sizes from Spectra. *J. Phys. Chem. C* **2021**, *125* (35), 19353−19361. Sep.

(21) Mbaye, M. T.; Pradhan, S. K.; Bahoura, M. Data-driven thermoelectric modeling: Current challenges and prospects. *J. Appl. Phys.* **2021**, *130* (19), 190902. Nov.

(22) Lookman, T.; Balachandran, P. V.; Xue, D.; Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *Npj Comput. Mater.* **2019**, *5* (1), 21. Art. no. 1

(23) Saal, J. E.; Oliynyk, A. O.; Meredig, B. Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches. *Annu. Rev. Mater. Res.* **2020**, *50* (1), 49−69.

(24) Barnard, A. S.; Opletal, G. Predicting structure/property relationships in multi-dimensional nanoparticle data using t-distributed stochastic neighbour embedding and machine learning. *Nanoscale* **2019**, *11* (48), 23165−23172. Dec.

(25) Pathak, Y.; Juneja, K. S.; Varma, G.; Ehara, M.; Priyakumar, U. D. Deep learning enabled inorganic material generator. *Phys. Chem. Chem. Phys.* **2020**, *22* (46), 26935−26943. Dec.

(26) Jha, D.; Gupta, V.; Liao, W.; Choudhary, A.; Agrawal, A. Moving closer to experimental level materials property prediction using AI. *Sci. Rep.* **2022**, *12* (1), 11953. Art. no. 1

(27) Haraguchi, Y.; Igarashi, Y.; Imai, H.; Oaki, Y. Sparse modeling for small data: case studies in controlled synthesis of 2D materials. *Digit. Discovery* **2022**, *1* (1), 26−34. Feb.

(28) Zhang, Y.; Ling, C. A strategy to apply machine learning to small datasets in materials science. *Npj Comput. Mater.* **2018**, *4* (1), 25. Art. no. 1

(29) Vanpoucke, D. E. P.; van Knippenberg, O. S. J.; Hermans, K.; Bernaerts, K. V.; Mehrkanoon, S. Small data materials design with machine learning: When the average model knows best. *J. Appl. Phys.* **2020**, *128* (5), No. 054901. Aug.

(30) Miyake, Y.; Kranthiraja, K.; Ishiwari, F.; Saeki, A. Improved Predictions of Organic Photovoltaic Performance through Machine Learning Models Empowered by Artificially Generated Failure Data. *Chem. Mater.* **2022**, *34* (15), 6912−6920. Aug.

(31) Lin, B.; Emami, N.; Santos, D. A.; Luo, Y.; Banerjee, S.; Xu, B.-X. A deep learned nanowire segmentation model using synthetic data augmentation. *Npj Comput. Mater.* **2022**, *8* (1), 88. Art. no. 1

(32) Shorten, C.; Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6* (1), 60. Jul.

(33) Sturluson, A.; Raza, A.; McConachie, G. D.; Siderius, D. W.; Fern, X. Z.; Simon, C. M. Recommendation System to Predict

https://doi.org/10.1021/acs.chemmater.3c03204
Chem. Mater. 2024, 36, 9330−9340

Missing Adsorption Properties of Nanoporous Materials. *Chem. Mater.* **2021**, *33* (18), 7203−7216. Sep.
(34) Little, R.; Rubin, D.*Statistical Analysis with Missing Data*; John Wiley & Sons, 2019, pp 223−232.
(35) van Ginkel, J. R.; Linting, M.; Rippe, R. C. A.; van der Voort, A. Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. *J. Pers. Assess.* **2020**, *102* (3), 297−308. May
(36) White, I. R.; Royston, P.; Wood, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **2011**, *30* (4), 377−399.
(37) Wu, W.-C.; Tracy, J. B. Large-Scale Silica Overcoating of Gold Nanorods with Tunable Shell Thicknesses. *Chem. Mater.* **2015**, *27* (8), 2888−2894. Apr.
(38) Liz-Marzán, L. M.; Giersig, M.; Mulvaney, P. Synthesis of Nanosized Gold−Silica Core−Shell Particles. *Langmuir* **1996**, *12* (18), 4329−4335. Jan.
(39) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321−357. Jun.
(40) Menardi, G.; Torelli, N. Training and assessing classification rules with imbalanced data. *Data Min. Knowl. Discovery* **2014**, *28* (1), 92−122. Jan.
(41) He, H.; Bai, Y.; Garcia, E. A.; Li, S.ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*; IEEE Jun. 2008; pp 1322−1328..
(42) Han, H.; Wang, W. Y.; Mao, B. H.; Huang, D. S.; Zhang, X. P.; Huang, G. B.; H., Han; W.-Y., Wang; B.-H., Mao; D.-S., Huang; X.-P., Zhang; G.-B., HuangBorderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning " in *Advances in Intelligent Computing*, Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2005; pp 878−887. doi: DOI: .
(43) Li, J.; et al. AI Applications through the Whole Life Cycle of Material Discovery. *Matter* **2020**, *3* (2), 393−432. Aug.
(44) Foerster, B.; Spata, V. A.; Carter, E. A.; Sönnichsen, C.; Link, S. Plasmon damping depends on the chemical nature of the nanoparticle interface. *Sci. Adv.* **2019**, *5* (3), No. eaav0704. noe. Mar.
(45) Virtanen, P.; et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17* (3), 261. Art. no. 3
(46) Buuren, S. v.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1. Art. no. 1
(47) Chen, T.; Guestrin, CXGBoost: A Scalable Tree Boosting System " in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16*; Association for Computing Machinery: New York, NY, USA, Aug. 2016, pp 785−794. .
(48) Pedregosa, F; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn, Res* **2011**, 2825.

https://doi.org/10.1021/acs.chemmater.3c03204
Chem. Mater. 2024, 36, 9330−9340