Electronic Supplementary Material (ESI) for Nanoscale Horizons.

Supporting Information

Incorporating Plasmonic Featurization with Machine Learning to Achieve Accurate and Bidirectional Prediction of Nanoparticle Size and Size Distribution

Emily Xi Tan,[1#] Yichao Chen,[2#] Yih Hong Lee,[1#] Yong Xiang Leong,[1] Shi Xuan Leong, [1] Chelsea Violita Stanley,[1] Chi Seng Pun,[2*] Xing Yi Ling[1*]

[1] Division of Chemistry and Biological Chemistry, School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore 637371.

[2] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore 637371.

[#] These authors contributed equally.

* Correspondence to: xyling@ntu.edu.sg; cspun@ntu.edu.sg

Supporting Information

Experimental Procedures

Chemicals. Gold chloride trihydrate ($HAuCl_4 \cdot 3H_2O$, 99%), sodium borohydride ($NaBH_4$, 98%), ascorbic acid (99%) and cetyltrimethylammonium chloride (CTAC) in water (25 wt%) were purchased from Sigma-Aldrich. cetyltrimethylammonium bromide (CTAB, 98%) was purchased from Alfa Aesar. Ethanol (ACS, ISO, Reag. Ph Eur) was obtained from Merck. Milli-Q water (> 18.0 MΩ. cm) was purified with a Sartorius Arium® 611 UV ultrapure water system. All reagents, unless otherwise stated, were used without further purification.

Preparation of Small Au NS Seeds. Small Au NS seeds were synthesised via seed-mediated method[1], in which Au^{3+} is reduced to Au using $NaBH_4$ as a reducing agent. Firstly, a 0.25 mL of 0.01 M $HAuCl_4$ solution was made up to 10 mL with 0.1 M of CTAB solution, followed by rapid injection of 0.6 mL ice-cold 0.01 M $NaBH_4$ solution under stirring. The resultant solution was gently stirred for 3 hours at room temperature. Subsequently, 0.12 mL of the seed solution was then injected into a growth solution made of 9.75 mL 0.1M CTAB, 190 mL water, 4mL 0.01M $HAuCl_4$ and 15 mL 0.1 M ascorbic acid. The reaction mixture was shaken gently and left overnight at room temperature. The resultant small Au NS seed sample was washed and concentrated by four times by centrifugation and redispersion in water.

Preparation of Large Gold Nanospheres. Larger Au NS are synthesised by growing the smaller Au NS seeds into larger Au nanopolyhedrons, which are then transformed into rounded Au NS by a mild oxidation process in presence of $HAuCl_4$ and CTAB. To prepare large Au nanopolyhedrons, a varying volume of the seed solution ranging from 0.025 mL to 4 mL was added into 30 mL of 0.025 M cetyltrimethylammonium chloride (CTAC) solution. For the seed solution at volumes less than 0.2mL, the seed solution was first diluted by four times with water. 0.75 mL of 0.1 M ascorbic acid was added to the mixture solution, followed by 1.5mL of 0.01 M $HAuCl_4$. The mixture solution was placed in an air-bath shaker at ~ 45 °C for 3 hours. The obtained Au nanopolyhedrons were centrifuged and redispersed in 30 mL of 0.02 M CTAB solution. Finally, to obtain the large, rounded Au NSs, the prepared Au nanopolyhedrons were mildly oxidised with 0.2 mL of 0.01 M $HAuCl_4$ solution. The resultant solution was kept in the air-bath shaker at 45 °C for another 2 hours. The obtained large Au NSs were centrifuged and redispersed in water.

TEM, SEM and UV-vis characterization of the synthesized Au NS. The extinction spectra were taken on an Agilent Technologies Cary 60 UV/visible spectrophotometer. The synthesized Au NSs were also subjected to transmission electron microscopic (TEM) imaging using JEOL JEM 1400 electron microscope at an accelerating voltage of 120 kV and scanning electron microscopic (SEM) imaging using JEOL JSM-7600F Schottky field emission electron microscope at an accelerating voltage of 5 kV. Smaller Au NSs sized between 20 and 40 nm were imaged using TEM, while larger Au NSs >40 nm was visualised using SEM. Measurements were taken at 5 different spots on the TEM/SEM substrate containing the as-synthesized Au NSs to get a representative group of images for each sample.

Supporting Information

For each sample of Au NSs, 100 randomly selected nanoparticles were measured on their TEM/SEM images using the ImageJ software to obtain the average Feret diameters, size distributions, and elongation factors (**Figure S1**).

Feature engineering and model building. Feature-engineering is performed using an automated spectral deconvolution routine in R, with an initial guess of the peak position(s). The fitting process automatically adjusts the estimates to find the most optimal fit, given certain restrictions. The machine learning algorithms used are basis-spline (Bspline), random forest (RF), and extreme gradient boosting (XGB). (**Supplementary Information 2 and 4**).

If the input data are transformed based on either the dipole peak all 94 datasets will be considered, in which 4 will be randomly selected and set aside as training data, while the remaining 90 will be split into 5 (4 train 1 test) sets of 18. If both the dipole and quadrupole are required as input data, the dataset is reduced to 46, of which 1 set will be randomly picked and placed in the training set while the other 45 will be split into 5 (4 train 1 test) sets of 9. With the number and types of data balanced in the 80:20 train-test split, cross-validation is performed until all five sets have been considered as test data. The cycle was repeated 100 times and the mean error of all validations (80:20 train-test, cross-validation, 100 iterations) was used as a metric of precision.

For the prediction of size, the relative squared error, L_1, is used to calculate the mean error:

$$L_1 = \left(Y_e/Y_o - 1\right)^2,$$

where Y_e is the predicted value and Y_o is the true value.

For the prediction of size distribution, the corresponding Y_o can vary widely (from 1.06 nm for the smallest Au NPs to 14.63 nm for the largest Au NPs) in our dataset and the relative error would be abnormally large for small nanoparticles and vice versa, skewing and inflating the mean error. Hence, the (absolute) squared error, L_2, was chosen instead:

$$L_2 = (Y_e - Y_o)^2.$$

Supporting Information

Supplementary Information 1. Additional details on particle image analysis.

The Feret diameter is a measure of a particle diameter between 2 parallel axes and the ratio between 2 perpendicular diameters will give the elongation ratio (**Figure S1**).[2]

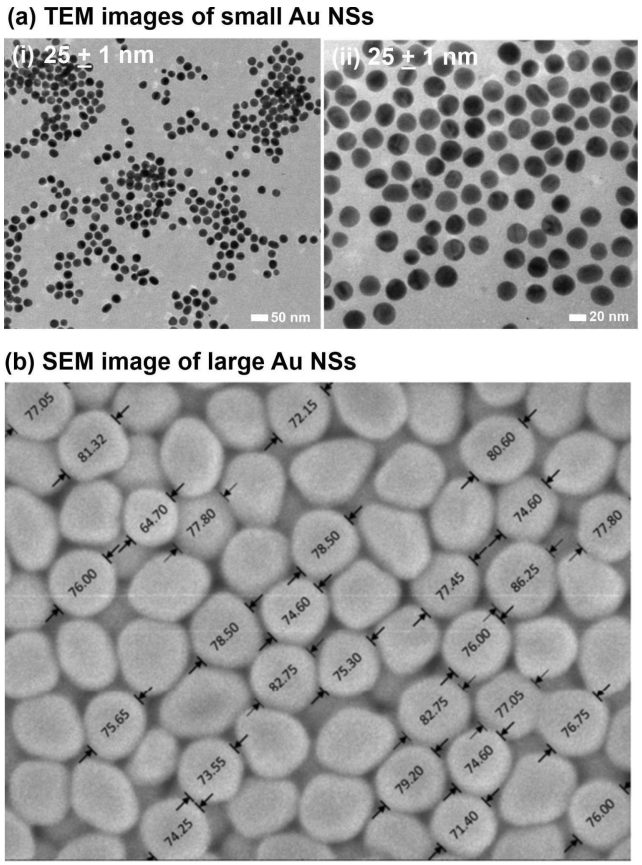


Figure S1 Electron microscopy of Au NS. (a) Transmission electron microscope (TEM) images of small Au nanospheres (25 ± 1 nm) and (b) Scanning electron microscope (SEM) images of large Au nanospheres with examples of nanoparticle measurements.

Supplementary Information 2. Additional details on feature engineering

The extinction spectra are in the form of functional data while measured at finite discrete points across the wavelength (nm). To feed them into the machine learning algorithms, we need to process feature engineering to extract the key structured information from them. We perform this using Gaussian curve fitting (GCF) that leverages the domain knowledge to identify the pair(s) of the localized surface plasmon resonance (LSPR) position and the full width at half maximum (FWHM) as the key features (inputs of machine learning algorithms) for our predictive tasks.

Gaussian Curve Fitting

It is known that the LSPR position (in wavelength) and the FWHM value are key indicators of the nanoparticle size and size distribution.[3, 4] To identify them from the extinction spectrum, there are two noteworthy insights: 1) the extinction spectrum is bell-shaped and thus the peak and the standard deviation of the bell curve can serve as proxies of the LSPR position and the FWHM; 2) there could be two LSPRs resulting in a spectrum of a mixture of two bell curves. The first insight motivates us to consider GCF, while the second insight reminds us of the possibility of the mixture of Gaussian curves. Additional chemical domain knowledge is required to determine the number of peaks (LSPRs) in a spectrum and provide the initial rough estimates on the peak positions. The rule of thumb for the former is that nanoparticles with average size more than 120 nm normally have two LSPRs in their extinction spectrum.[4-6] The fitting process automatically adjusts the estimates. In what follows, we elaborate the GCF process for two cases of one LSPR and two LSPRs.

2.1 The Case of One LSPR

Suppose that there is only one peak in the extinction spectra, which can be approximated by a Gaussian curve. A Gaussian curve can be described by a parametrized function $f(x)$ with the wavelength x:

$$G(x;c,\mu,\sigma) = c \cdot e^{-(x-\mu)^2/\sigma^2},$$

where μ is the position of the peak, σ is the standard deviation of the curve that is supposedly proportional to the FWHM, and c is the magnitude of the curve.

Technical details of fitting: The GCF aims to identify the parameters (c,μ,σ) for each sample curve that minimizes the mean of the squared error (MSE) between the sample curve and the predicted Gaussian curve $G(x;c,\mu,\sigma)$ over a range of wavelengths. It should be noted that the extinction curves were measured at discrete finite points (wavelengths), the mean is essentially the averaging in practice. The range of wavelength, i.e., the number of points, for calculating the errors will affect the fitting while reflecting the domain knowledge. We used the initial estimate of μ as the center of the range and considered different numbers of neighbouring points which are symmetric around μ to form the range

of wavelengths for fitting. Among the choices of 40, 80, and 120 points in our preliminary experiments, picking 40 points (neighbours around the center) can consistently fit the curve as well as its tip region; see (**Figure S2**). We prioritise the fitting of the tip region because it is the most relevant to the actual LSPR position and the FWHM. The GCF as a loss minimization problem can be solved numerically with the initial values of (c,μ,σ), where the initial value μ is chosen as the maximum extinction value of the curve and the initial values of (c,σ) are chosen as some positive constants (e.g., 1).

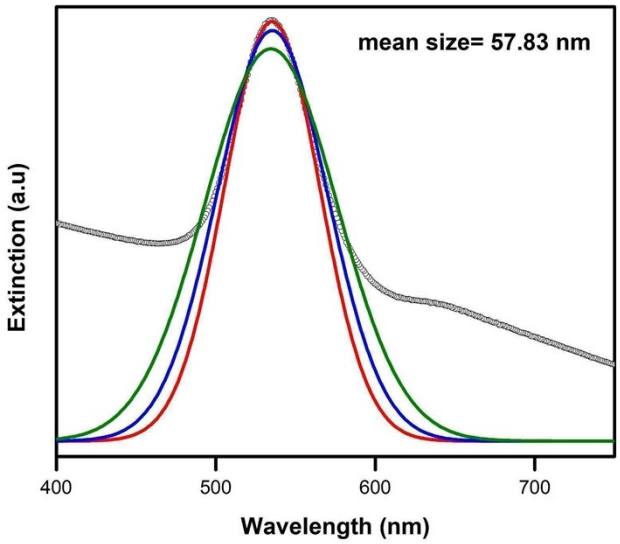


Figure S2. Examples of the experimental data (hollow black circles) and the fitted gaussian function for extinction spectra of Au nanospheres of mean size of 57.83 nm when interpolated with 40 (red), 80 (green) and 120 (blue) points between the peaks.

Following the process above, we can identify a pair of (μ,σ) for each sample curve. Consequently, these two structured features can then be inputted into ML algorithms.

2.2 The Case of Two LSPRs

For spectra displaying a secondary quadrupole peak, we should consider the mixture of Gaussian curves. The corresponding fitting function comprises the sum of two Gaussian functions:

$$mG(x) = c_1 \cdot e^{-(x-\mu_1)^2/\sigma_1^2} + c_2 \cdot e^{-(x-\mu_2)^2/\sigma_2^2},$$

where the parameters (c_1,μ_1,σ_1) and (c_2,μ_2,σ_2) are interpreted similarly as (c,μ,σ) in the previous section for two different Gaussian curves.

Technical details of fitting: Similarly, the mixture of Gaussian curves fitting (mGCF) aims to identify the parameters $(c_1,\mu_1,\sigma_1,c_2,\mu_2,\sigma_2)$ for each sample curve that minimizes the MSE between the sample curve and the predicted mixture of Gaussian curves $mG(x;c_1,\mu_1,\sigma_1,c_2,\mu_2,\sigma_2)$ over a range

of wavelengths. In this case, we need rough estimates of the two peaks' positions (μ_1,μ_2) while the initial values of $(c_1, \sigma_1, c_2, \sigma_2)$ do not make a significant impact in fitting. To determine the range of wavelength for calculating the errors, we consider the lower bound of 500 nm, which is a common inflection point that prompted the right bound of the first Gaussian function; see (**Figure S3**. The upper bound is selected from the set of the larger peak position shifted to the right for 20, 30, 40, 50, or 60 points of wavelength, such that the sum of the squared errors between the sample curve and the predicted curve over the wavelength interval of two peaks' positions is minimized. Here, for more complicated extinction spectra presenting two plasmon resonances, the same principle of preferential fitting of the tip regions was applied.

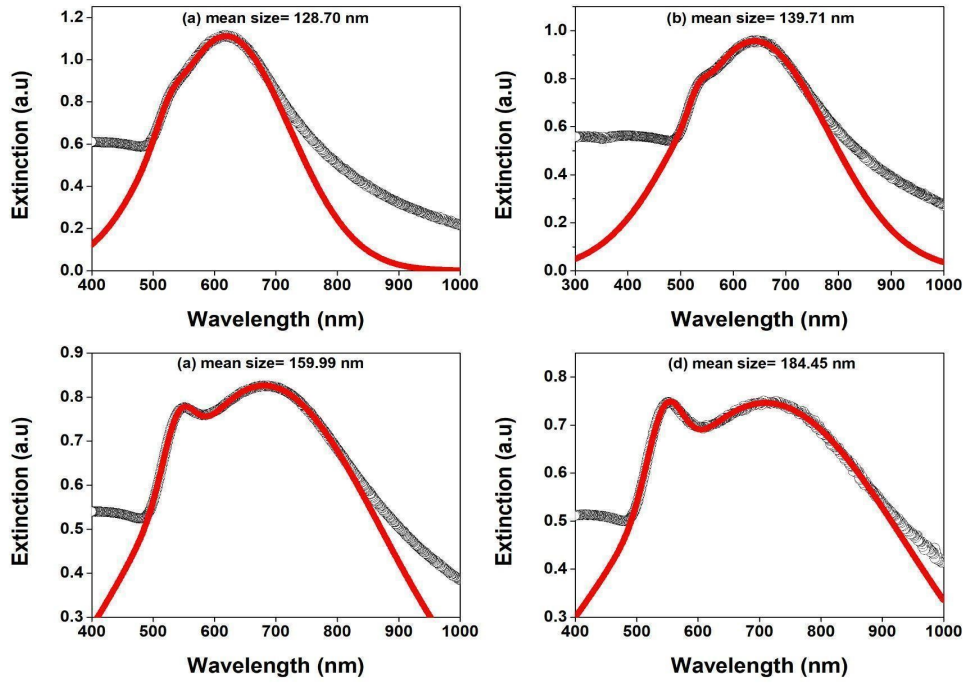


Figure S3. Examples of the experimental data (hollow black circles) and the fitted sum of two gaussian functions (red) for extinction spectra of Au nanospheres of mean sizes= (a) 128.70 nm, (b) 139.71 nm, (c) 159.99 nm and (d) 184.45 nm.

As a result, each of the extinction spectra with two LSPRs can be represented by $(\mu_1, \sigma_1, \mu_2, \sigma_2)$, which can also form a structured input matrix for the use of ML algorithms.

In our experiments, we separately study one-peak- and two-peak-fitting. For the former, one-peak-fitting (GCF) is applied to all 94 data (including those with two peaks) and it does not require additional domain knowledge (on the peaks). Hence, in this case, each of the 94 spectra is represented by two features (μ, σ). For the latter, two-peak-fitting (mGCF) is applied only to the spectra with two clear peaks, which normally refer to the case with mean nanoparticle size of greater than 120 nm. Hence, in

this case, the sample size is smaller and each of the spectra is represented by four features $(\mu_1, \sigma_1, \mu_2, \sigma_2)$.

Supplementary Information 3. LSPR vs size and size distribution function fitting

To develop a more comprehensive and complete understanding of size-dependencies of dipole and quadrupole resonance peak position and FWHM, we quantitatively analyzed their correlations with particle size and size distributions over a very wide size range (20-220 nm). We first fit various functions to determine the optimal curve describing their relationships and calculate the goodness-of-fit and errors between predicted and true values.

Like our approach for machine learning prediction, the percentage error (%) is used to calculate the error for size:

$$Percentage\ error\ (\%) = \frac{V_A - V_T}{V_T} \times 100\%$$

Where V_A is the approximate (measured) value and V_T is the **true** value

For the calculation of the error for size distribution, the corresponding V_T can vary widely (from 1.06 nm for the smallest Au NPs to 14.63 nm for the largest Au NPs) in our dataset and the percentage error would be abnormally large for small nanoparticles and abnormally small for larger nanoparticles, which will skew and inflate the value. Hence, the relative error (nm), was chosen instead:

$$relative\ error\ (nm) = V_A - V_T$$

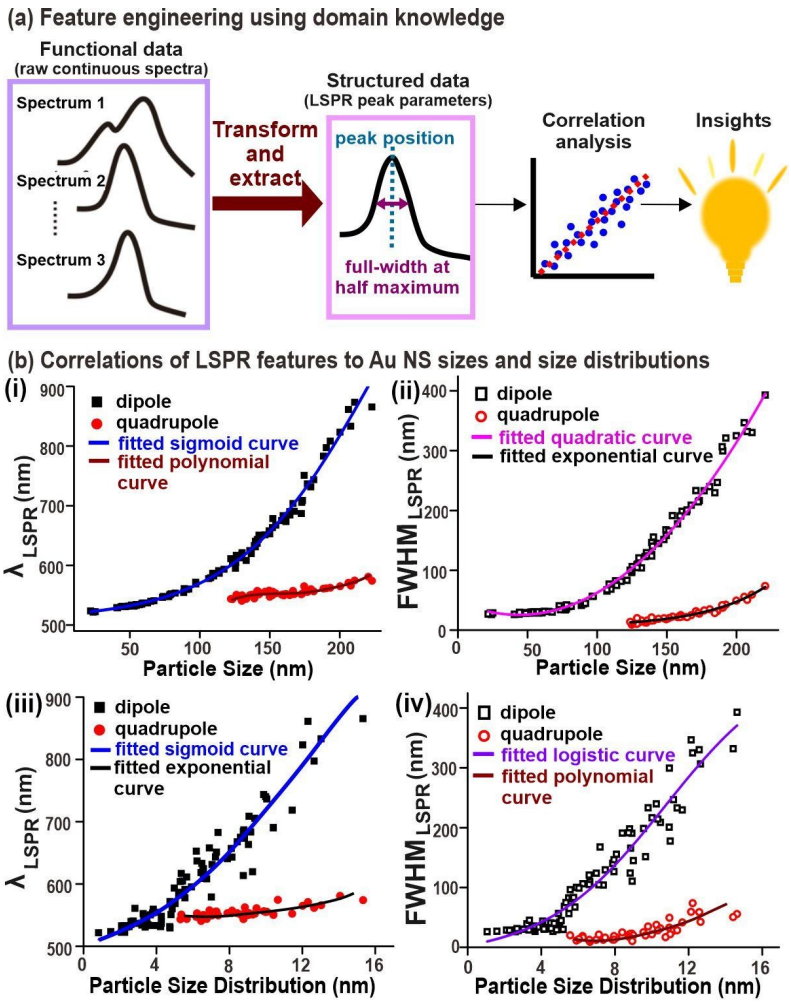
The relationship between dipole peak position (R^2 = 0.99, 5% percentage error) and particle size, is best represented by a sigmoid function which contrasts with the exponential trend previously reported for Au NSs between a smaller size range of 35-100 nm[4] (**Figures S5b(i), Figure S6**). According to our results, the sigmoidal curve can be categorized into 3 stages: 1) the initial segment is approximately exponential (geometric) between 20-150 nm, 2) followed by a linear (arithmetic) trend between 150-200 nm, and 3) the gradient decreases between 200-220 nm. By analyzing the dipole position of Au NSs beyond 100 nm, we discovered the latter two stages, which confirms that the relationship between dipole position and size will not follow an exponential trend indefinitely. Meanwhile, a quadratic function gives the best fit to describe the relationship between dipole peak FWHM and particle size (R^2 = 0,99, 6% percentage error) (**Figures S5b(ii), Figure S6**).

Since the LSPR positions and FWHMs are inextricably linked, they will be discussed together. Our results indicates that as the size of Au NSs increases beyond 20 nm, higher multipole terms that

incorporate the size, medium, and dielectric properties of the metal become increasingly important, resulting in a modest redshift in the LSPR.[5, 6] Beyond 100 nm, the resonance continues to redshift but becomes significantly broadened due to 1) contributions from higher multipole resonances and 2) resonance radiative damping as the scattering cross-section increases rapidly compared to the size.[5-8] The confluence of the two factors results in a rapid rise in nanoparticle LSPR positions and FWHMs between 100 and 200 nm, before the magnitude of the gradient decreases between 200-220 nm when phase retardation effects become significant in larger nanostructures.

Furthermore, we determine that the dipole peak position is sigmoidal-correlated (R^2 =0.91, 1.27 nm relative error) to size distribution, while the dipole FWHM can be logit-correlated (R^2 = 0.94, 1.19 nm relative error) to the size distribution (**Figures S5b (iii and iv), Figure S7**).

Besides dipoles, larger Au NSs also display quadrupole resonance peaks. We observe that the quadrupole FWHM is exponentially correlated with particle size (R^2=0.97), however the best fit curve still carries a percentage error of 8% (**Figure S5b(ii), Figure S6**). By correlating the LSPR features to the particle size and size distributions, we reveal and quantify the size and size distribution dependencies of these plasmon resonance modes over a wide size range (20-220 nm) and confirm their best-fit single function relationships. These quantitative analyses are important as they will guide the selection of machine learning algorithm types and architecture.



(c) Correlation analysis results

	Features	R^2	Percentage error (%)
Au NS size	λ_{Dipole}	0.99	5
	$\lambda_{Quadrupole}$	0.84	28
	FWHM Dipole	0.99	6
	FWHM Quadrupole	0.97	8

	Features	R^2	relative error (nm)
Au NS size distribution	λ_{Dipole}	0.91	1.27
	$\lambda_{Quadrupole}$	0.61	3.30
	FWHM Dipole	0.94	1.19
	FWHM Quadrupole	0.76	2.55

Figure S5. Domain-knowledge-based feature engineering and subsequent correlative analysis of extracted spectral features with Au NS size and size distribution, respectively. (a) Schematic illustration of the feature-engineering process where functional data (raw and continuous UV-Vis extinction spectra) is transformed into structured data consisting of LSPR peak positions and full width at half maximum (FWHM). (b) (i and ii) Correlation of LSPR positions and FWHMs derived from the automated Gaussian peak fitting routine with particle sizes. Best-fitted sigmoid, polynomial, quadratic and exponential functions in blue, brown, magenta and black respectively. (iii and iv) Correlation of LSPR positions and FWHMs with particle size distributions. Best-fitted sigmoid, exponential, logistic and polynomial functions in blue, black, violet and brown respectively. (c) Results of correlation analysis including the goodness-of-fit (R^2) and percentage error (%) for size and relative error (nm) for size distribution using best-fit single functions.

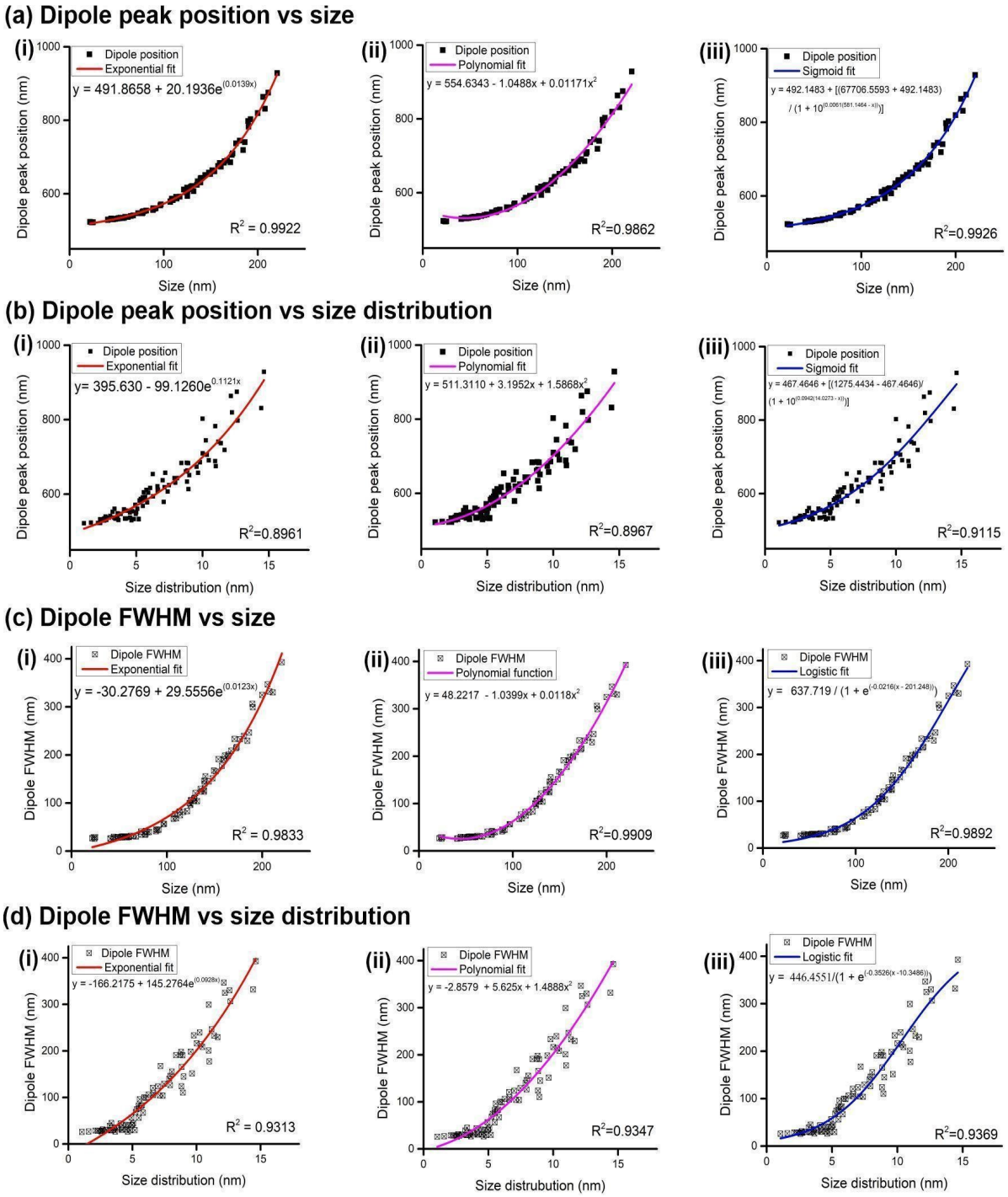


Figure S6. (a-b) Correlation of dipole plasmon resonance positions derived from the Gaussian fits with particle sizes and size distributions. Best-fit exponential, polynomial and sigmoid functions in red, magenta and blue respectively. (c-d) Correlation of dipole plasmon resonance FWHMs derived from the Gaussian fits with particle sizes and size distributions. Best-fit exponential, polynomial and logistic functions plotted in red, magenta, and blue respectively.

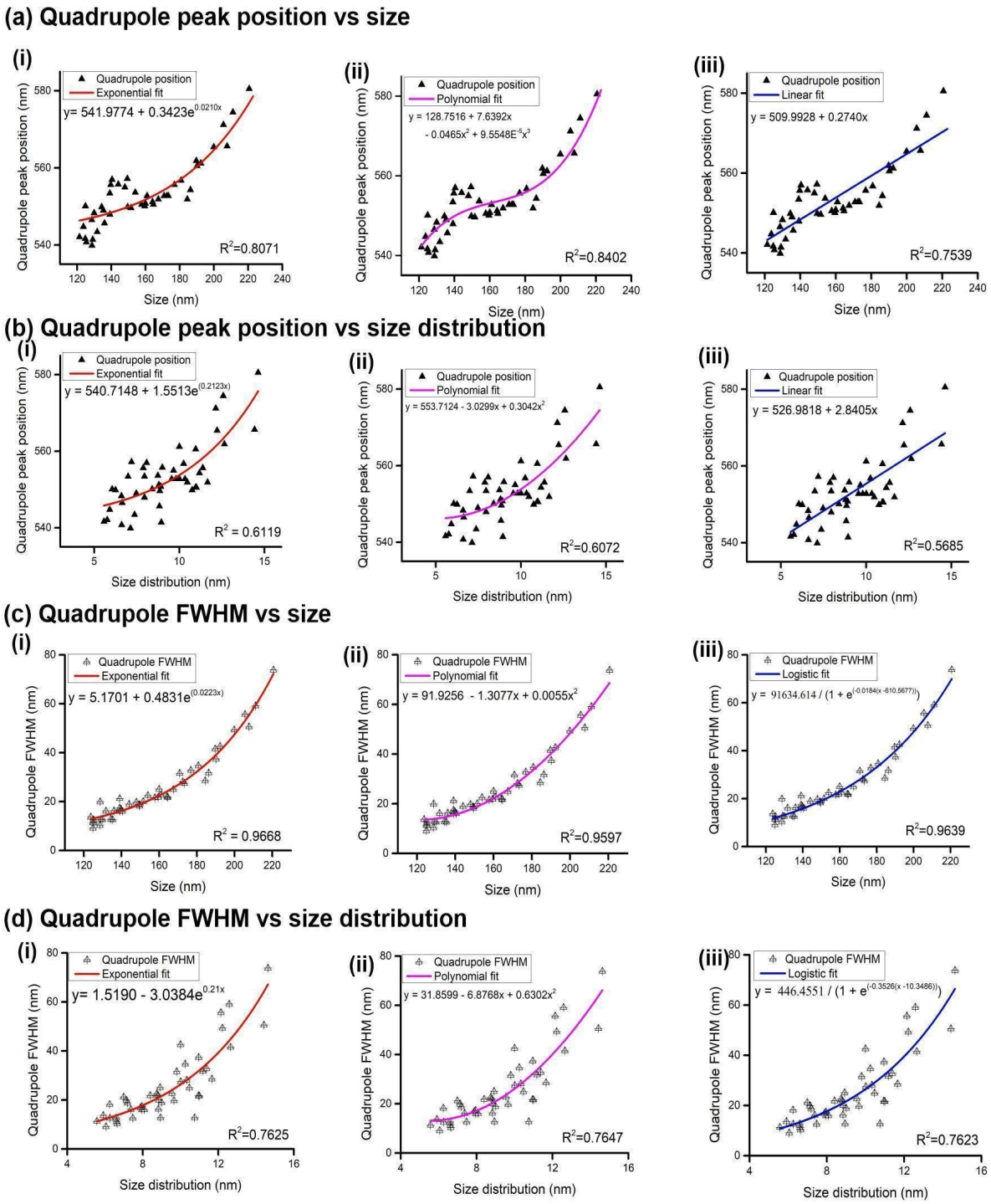


Figure S7. (a-b) Correlation of quadrupole plasmon resonance positions derived from the Gaussian fits with particle sizes and size distributions. Best-fit exponential, polynomial, and linear functions in red, magenta and blue respectively. (c-d) Correlation of quadrupole plasmon resonance FWHMs derived from the Gaussian fits with particle sizes and size distributions. Best-fit exponential, polynomial and logistic functions plotted in red, magenta, and blue respectively.

Supporting Information

Supplementary Information 4. Machine learning (ML) model building

Basis-spline (Bspline) Regression

A Bspline regression model works by generating piecewise step functions to fit different segments of the dataset instead of imposing an overarching linear or polynomial function as the global structure which makes it less prone to over- and under-fitting.[7, 8] To prepare the data for a Bspline model, a basis function was applied to each feature (peak position or FWHM) to generate n features (where n=degrees of freedom-3) as input data.

Random Forest (RF)

Random forest is an ensemble technique where a user-specified number of trees are constructed during the training phase and an averaged prediction of all the trees is returned, making it less prone to overfit compared to simple decision trees models.[9, 10] For a random forest model, two parameters can be tuned- **ntree** (number of trees constructed) and **mtry** (number of features randomly sampled at each split) to affect the final accuracy. In this experiment, we fixed the ntree=1000, and varied the mtry from 1 to 4 depending on the total number of input features. For example, when both dipole and quadrupole position and FWHMs are available, we can afford to sample up to 4 different features at each split (mtry=4).

Extreme Gradient Boosting (XGB)

Extreme gradient boosting (XGB) is another decision-tree-based ensemble machine learning technique that employs a gradient descent algorithm (boosting) to minimize errors in sequential models. Parallel processing and regularization, which considerably improve computation efficiency and buffer against overfitting or bias, are two additional advantages of XGB, making it a robust and versatile approach ideal for both classification and regression.[11] For the XGB model exploration, the **nround** (number of iterations) was fixed at 25, the **eta** (step size or learning rate) was varied from 0.1 to 0.5 in 0.2 increments, while the **max depth** (depth or trees) was varied between 1 to 10. Typically, the prediction results will start to stabilize at the optimal maximum depth with minimal fluctuation in results when the depth of trees is increased. This phenomenon is a result of regularization inherent to XGB which serves to shrink the learned estimated coefficients towards zero to 1) increase generalization of the model and 2) prevent overfitting.

Supplementary Information 5. Model optimization results.

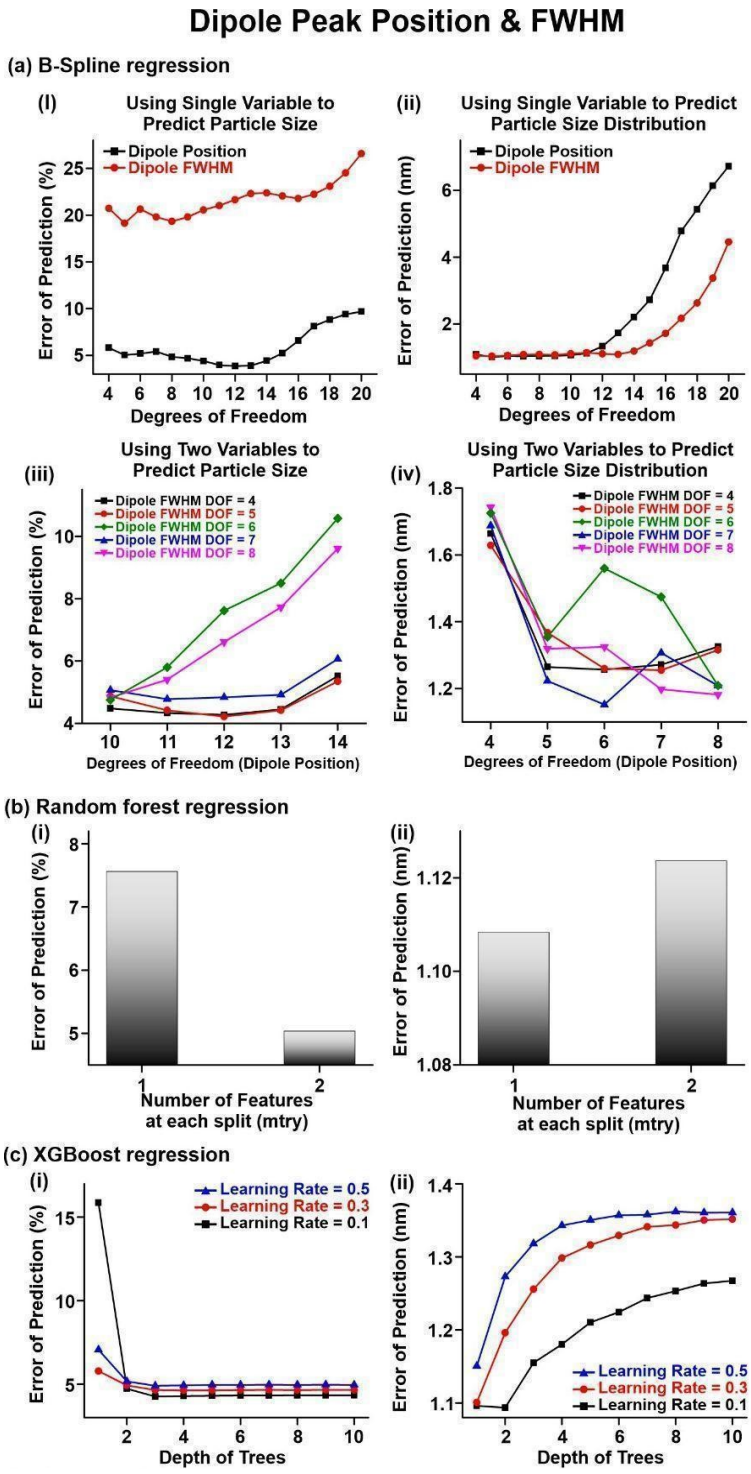


Figure S8. Evaluating the predictive capabilities of various machine learning algorithms in predicting Au nanoparticle size and size distribution using dipole peak position and FWHM as input data by comparing the (a) Error of prediction for (i) and (iii) nanoparticle size (%) and (ii) and (iv) size distribution (nm) using Bspline regression. (b) Error of prediction for (i) nanoparticle size (%) and (ii) size distribution (nm) using random forest model. (c) Error of prediction for (i) nanoparticle size (%) and (ii) size distribution (nm) using XGB regressor.

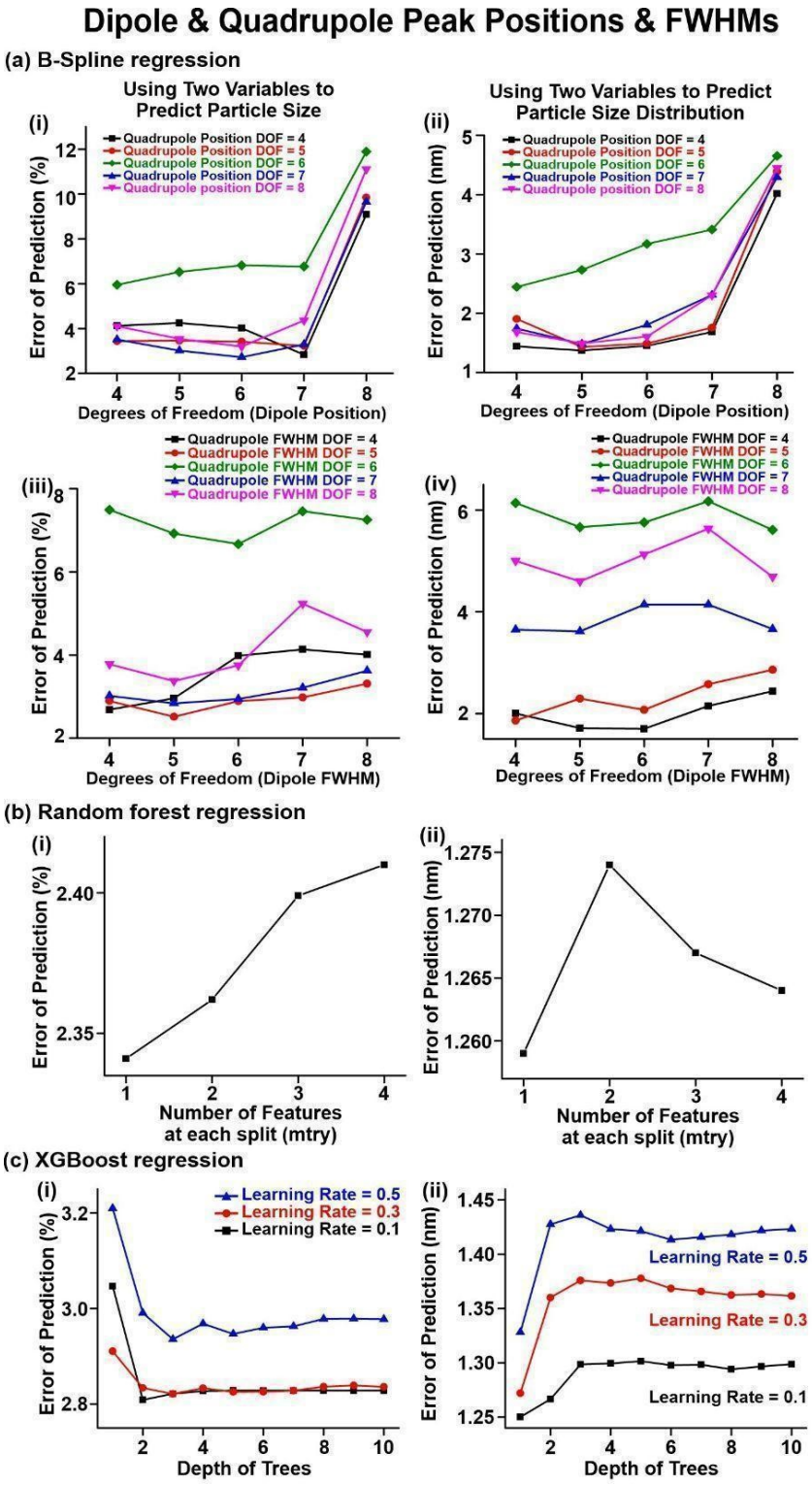


Figure S9. Evaluating the predictive capabilities of various machine learning algorithms in predicting Au nanoparticle size and size distribution using both dipole and quadrupole peak positions and/or 2 FWHMs as input data by comparing the (a) Error of prediction for (i) and (iii) nanoparticle size (%) and (ii) and (iv) size distribution (nm) using Bspline regression. (b) Error of prediction for (i) nanoparticle size (%) and (ii) size distribution (nm) using random forest model. (c) Error of prediction for (i) nanoparticle size (%) and (ii) size distribution (nm) using XGB regressor.

Supplementary Information 6. Summary tables of the best prediction results of the 3 machine-learning algorithms

Table S1. Best prediction results of different machine learning algorithms using one peak position and FWHM as input. The lowest errors are indicated in red.

ML algorithm	Minimum relative error of size prediction (%)	Minimum error of size distribution prediction (nm)
Bspline regression	3.85	1.02
Random forest	5.04	1.11
XGBoost	4.26	1.09

Table S2. Best prediction results of different machine learning algorithms using two peak positions and FWHMs, for data with two peaks only as input. The lowest errors are indicated in red.

ML algorithm	Minimum relative error of size prediction (%)	Minimum error of size distribution prediction (nm)
Bspline regression	2.52	1.15
Random forest	2.34	1.26
XGBoost	2.81	1.25

Supporting Information

Bspline regression using dipole features only

Table S3. Relative error of particle size (%) and error of size distribution (nm) predictions of Bspline regression using single variable peak position or FWHM individually as input. The lowest errors are indicated in red.

Degrees of freedom	Relative error of peak position to size prediction (%)	Relative error of peak FWHM to size prediction (%)	Error of peak position to size distribution prediction (nm)	Error of peak FWHM to size distribution prediction (nm)
4	5.8155	20.7288	1.085046	1.044668
5	5.0330	19.1313	1.015088	1.031579
6	5.1926	20.6385	1.036172	1.058209
7	5.4081	19.7952	1.030635	1.082546
8	4.8412	19.3343	1.038474	1.088715
9	4.6827	19.7966	1.046837	1.072774
10	4.4027	20.5546	1.065320	1.114015
11	3.9655	21.0191	1.123040	1.140260
12	3.8510	21.6475	1.331572	1.108857
13	3.9043	22.3019	1.729465	1.089095
14	4.4224	22.3677	2.205546	1.188371
15	5.2374	22.0458	2.724038	1.430756
16	6.5709	21.7731	3.677463	1.716486
17	8.1240	22.2341	4.786034	2.167821
18	8.8224	23.0852	5.434744	2.628076
19	9.3949	24.5275	6.139898	3.371843
20	9.6806	26.5783	6.724159	4.454638

The prediction results from using both the peak position and FWHM as input is displayed below. Systematic testing by varying the degrees of freedom (4 to 20) of both inputs were conducted. Only the important parts of the result are shown due to space limitations.

Table S5. Relative error of particle size (%) predictions of Bspline regression using 2 variables, including both peak position (first column) and FWHM (first row) as input. The lowest error is indicated in red.

Degrees of freedom	4	5	6	7	8

10	4.4776	4.8699	5.0642	4.8249	4.7523
11	4.3314	4.4189	4.7765	5.4014	5.7999
12	4.2715	4.2163	4.8369	6.6110	7.6173
13	4.4459	4.4161	4.9179	7.7239	8.4983
14	5.5193	5.3461	6.0622	9.6087	10.5756

Table S5. Error of particle size distribution (nm) predictions of Bspline regression using both peak position (first column) and FWHM (first row) as input. The lowest error is indicated in red.

Degrees of freedom	**4**	**5**	**6**	**7**	**8**
4	1.663913	1.62835	1.687359	1.742602	1.725226
5	1.264328	1.367106	1.222653	1.318704	1.354107
6	1.256248	1.259441	1.151529	1.324534	1.559567
7	1.271100	1.254946	1.306539	1.197414	1.474128
8	1.325061	1.315125	1.208156	1.18149	1.208615

Random Forest using dipole features only

Table S6. Relative error of particle size (%) and error of size distribution (nm) predictions of random forest regression using peak position and FWHM as input. The lowest errors are indicated in red.

'Mtry'	Relative error of size prediction (%)	Error of size distribution prediction (nm)
1	7.561	1.1083
2	5.035	1.1236

Extreme gradient boosting algorithm using dipole features only

Table S7. Relative error of particle size (%) and error of size distribution (nm) predictions of XGB regression using peak position and FWHM as input. The lowest errors are indicated in red.

Max depth	Relative error of size prediction (%)			Max depth	Error of size distribution prediction (nm)		
	Eta=0.1	Eta=0.3	Eta=0.5		Eta=0.1	Eta=0.3	Eta=0.5
1	15.8647	5.7870	7.0559	1	1.095994	1.100750	1.150423
2	4.7288	4.9219	5.1660	2	1.093583	1.195985	1.272854
3	4.2568	4.6450	4.9005	3	1.154843	1.255738	1.318141
4	4.2912	4.6339	4.9273	4	1.180191	1.298309	1.342922
5	4.3062	4.6325	4.9584	5	1.210222	1.316131	1.350410
6	4.3185	4.6481	4.9558	6	1.224242	1.329210	1.356862
7	4.3193	4.6500	4.9651	7	1.243565	1.341122	1.357734
8	4.3245	4.6468	4.9575	8	1.253122	1.343393	1.361938
9	4.3270	4.6513	4.9677	9	1.263490	1.350160	1.360462
10	4.3279	4.6523	4.9586	10	1.267137	1.351568	1.360523

Bspline regression using both dipole and quadrupole features

Table S8. Relative error of particle size (%) predictions of Bspline regression using both dipole (first column) and quadrupole (first row) peak positions as input, for data with two peaks only. The lowest error is indicated in red.

Degrees of freedom	4	5	6	7	8
4	4.12449	3.440235	3.511857	4.103009	5.952497
5	4.249787	3.461954	3.016202	3.540414	6.527864
6	4.017791	3.41734	2.723840	3.195863	6.821350
7	2.834473	3.230956	3.290557	4.35887	6.767945
8	9.091974	9.846198	9.657211	11.10754	11.89529

Table S9. Error of particle size distribution (nm) predictions of Bspline regression using both dipole (first column) and quadrupole (first row) peak positions as input, for data with two peaks only. The lowest error is indicated in red.

Degrees of freedom	4	5	6	7	8
4	1.441054	1.901171	1.740996	1.682099	2.441559
5	1.368532	1.428408	1.481865	1.489281	2.729603
6	1.449892	1.487307	1.801296	1.603756	3.167285
7	1.683162	1.755971	2.306755	2.302142	3.409261
8	4.017013	4.387913	4.290749	4.448702	4.651867

Table S10. Relative error of particle size (%) predictions of Bspline regression using both dipole (first column) and quadrupole (first row) peak FWHMs as input, for data with two peaks only. The lowest error is indicated in red.

Degrees of freedom	4	5	6	7	8
4	2.6857	2.8964	3.0198	3.7807	7.4930
5	2.9609	2.5164	2.8386	3.3749	6.9223
6	3.9831	2.8928	2.9433	3.7523	6.6684
7	4.1362	2.9815	3.2127	5.2364	7.4575
8	4.0094	3.3108	3.6231	4.5539	7.2538

Table S11. Error of particle size distribution (nm) predictions of Bspline regression using both dipole (first column) and quadrupole (first row) peak FWHMs as input, for data with two peaks only. The lowest error is indicated in red.

Degrees of freedom	**4**	**5**	**6**	**7**	**8**
4	2.002933	1.863187	3.650194	5.001058	6.139825
5	1.713988	2.295165	3.616040	4.597188	5.663667
6	1.699825	2.073149	4.142663	5.127308	5.756428
7	2.148693	2.576066	4.140247	5.634330	6.172683
8	2.438475	2.862251	3.660585	4.687414	5.609820

Random Forest using both dipole and quadrupole features

Table S12. Relative error of particle size (%) and error of size distribution (nm) predictions of random forest regression using both peak positions and both FWHMs as input (total 4 features), for data with two peaks only. The lowest errors are indicated in red.

'Mtry'	Relative error of size prediction (%)	Error of size distribution prediction (nm)
1	2.341	1.259
2	2.362	1.274
3	2.399	1.267
4	2.410	1.264

Extreme gradient boosting algorithm using both dipole and quadrupole features

Table S13. Relative error of particle size (%) and error of size distribution (nm) predictions of XGB regression using both peak positions and both FWHMs as input (total 4 features), for data with two peaks only. The lowest errors are indicated in red.

Max depth	Relative error of size prediction (%)			Max depth	Error of size distribution prediction (nm)		
	Eta=0.1	Eta=0.3	Eta=0.5		Eta=0.1	Eta=0.3	Eta=0.5
1	3.0467	2.9106	3.2096	**1**	1.250113	1.271958	1.328052
2	2.8089	2.8338	2.9907	**2**	1.266544	1.360114	1.42753
3	2.8215	2.8213	2.9356	**3**	1.298572	1.375809	1.435943
4	2.8273	2.8332	2.9683	**4**	1.299402	1.373418	1.423063
5	2.8282	2.8256	2.9467	**5**	1.301521	1.377703	1.421304
6	2.8281	2.8258	2.9598	**6**	1.297725	1.368457	1.413312
7	2.8280	2.8280	2.9627	**7**	1.298270	1.365655	1.415719
8	2.8282	2.8364	2.9782	**8**	1.294053	1.362448	1.418144
9	2.8281	2.8389	2.9786	**9**	1.296639	1.363365	1.421776
10	2.8281	2.8360	2.9776	**10**	1.298715	1.361587	1.423217

Supplementary Information 7. Inverse prediction

Unlike the forward prediction, the inverse b-spline model predicts only (μ, σ) or $(\mu_1, \sigma_1, \mu_2, \sigma_2)$, where μ is the mean which indicates the position of the peak, σ is the standard deviation of the curve that is proportional to the FWHM (peak width). Consequently, we use the similarity between LSPR position and FWHM to evalute the accuracy of inverse prediction.

For the mathematical reconstruction of the spectra, the 2(μ, σ) or $4^{(\mu_1, \sigma_1, \mu_2, \sigma_2)}$ predicted variables are then substituted in the following Gaussian functions to bypass the need for **c** (magnitude in Gaussian) and only use (μ, σ) to achieve accurate predictions of position and FWHM and reconstruct the extinction spectra.

For single Gaussian curve:

$$G(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\left(\frac{1(x-\mu)^2}{2 \quad \sigma^2}\right)}$$

where (μ, σ) where μ is the mean or position of the peak, σ is the standard deviation of the curve that is proportional to the FWHM (peak width).

For a mixture of Gaussian curves:

$$MG(x;\mu_1,\sigma_{1,},\mu_2,\sigma_{2,}) = \frac{1}{\sigma_1\sqrt{2\pi}} \cdot e^{-\left(\frac{1(x-\mu_1)^2}{2 \quad {\sigma_1}^2}\right)} + \frac{1}{\sigma_2\sqrt{2\pi}} \cdot e^{-\left(\frac{1(x-\mu_2)^2}{2 \quad {\sigma_2}^2}\right)}$$

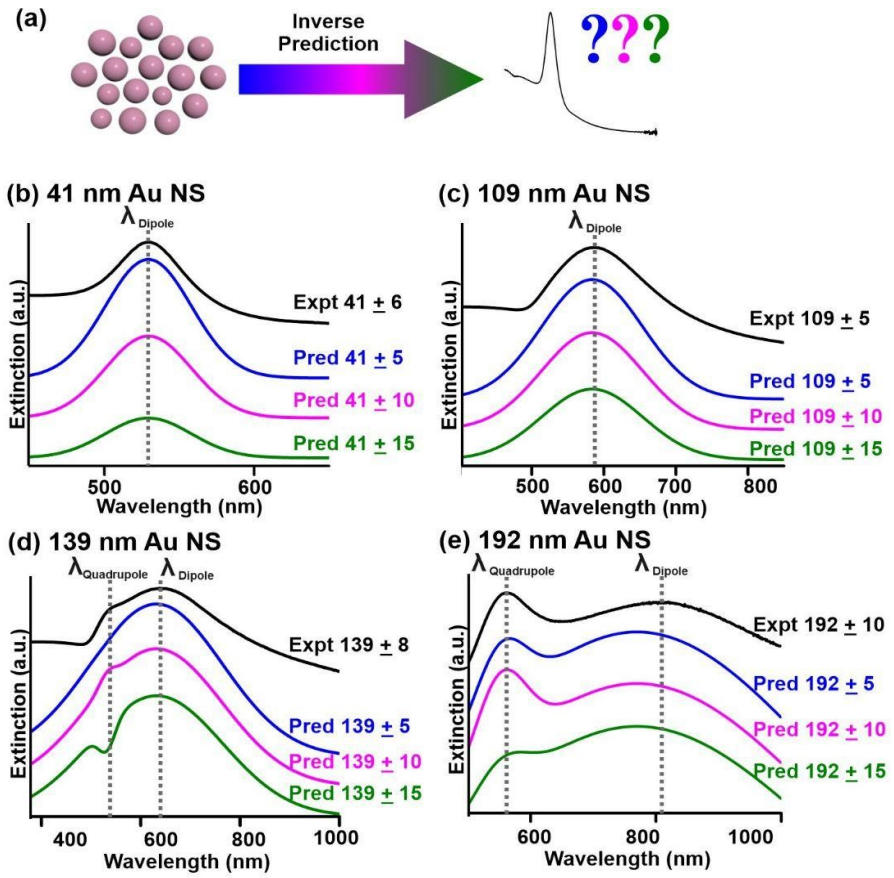
Where $(\mu_1,\sigma_{1,},\mu_2,\sigma_{2,})$ is interpreted similarly as (μ, σ) above.

The purpose is to make the inverse model more generalisable and concentration independent. This is to eliminate the variable **c** (magnitude in Gaussian) also commonly referred to as absorbance (A) of the extinction spectra which dependent on the concentration (C) according to the Beer-Lambert law:

$$A=\varepsilon bC$$

where A is the absorbance, ε is the molar absorptivity, b is the path length of light and C is the concentration.

This is because if we take **c** into account, the number and permutations of **c** available will make the number of predicted curves to be astronomical to calculate and it distracts the purpose of understanding the 2 more important variables/parameters which is the LSPR position and FWHM.



(f) Quantitative comparison of experimental and predicted spectra

(i) 41 nm Au NS

		Exp 41±6 nm	41±5 nm	41±10 nm	41±15 nm	Best match	Mean error of best match (%)
Dipole	λ (nm)	529	529 (0%)	529 (0%)	529 (0%)	41±5 nm	0
	FWHM (nm)	68	68 (0%)	66 (3%)	50 (26%)		

(i) 109 nm Au NS

		Exp 109±5 nm	109±5 nm	109±10 nm	109±15 nm	Best match	Mean error of best match (%)
Dipole	λ (nm)	584	584 (0%)	584 (0%)	584 (0%)	109±5 nm	1
	FWHM (nm)	165	162 (2%)	155 (6%)	153 (7%)		

(iii) 139 nm Au NS

		Exp 139±8 nm	139±5 nm	139±10 nm	139±15 nm	Best match	Mean error of best match (%)
Dipole	λ (nm)	638	632 (1%)	635 (0%)	632 (1%)	139±10 nm	2
	FWHM (nm)	501	460 (8%)	464 (7%)	260 (48%)		
Quadrupole	λ (nm)	555	534 (4%)	555 (0%)	502 (10%)		
	FWHM (nm)	57	58 (2%)	57 (0%)	36 (37%)		

(iv) 192 nm Au NS

		Exp 192±10 nm	192±5 nm	192±10 nm	192±15 nm	Best match	Mean error of best match (%)
Dipole	λ (nm)	810	770 (5%)	770 (5%)	770 (5%)	192±10 nm	5
	FWHM (nm)	330	367 (11%)	358 (8%)	392 (18%)		
Quadrupole	λ (nm)	563	567 (1%)	563 (0%)	584 (4%)		
	FWHM (nm)	84	79 (6%)	80 (5%)	70 (17%)		

Figure S10. Robust spectral regeneration based on inverse prediction. (a-d) Comparison of extinction spectra of specific sizes with 3 different prespecified size distributions (5, 10 and 15 nm) generated from our Bspline regression model with the actual experimental extinction spectra for (a) 41 ± 6 nm (b) 109 ± 5 nm (c) 139 ± 8 nm and (d) 192 ± 10 nm Au nanospheres. (e) Quantitative comparison of peak features of experimental and predicted spectra.

Supporting Information

Example of percentage error calculation for 109 nm Au NSs containing dipole peak only,

$$Percentage\ error\ (\%) = \frac{|V_A - V_T|}{V_T} \times 100\%$$

Where V_A is the approximate (measured) value and V_T is the **true** value.

$$Percentage\ error\ of\ \lambda_{Dipole}\ (\%) = \frac{|162 - 165|}{165} = 2\%$$

$$Mean\ error\ (\%) = \frac{Percentage\ error\ of\ \lambda_{Dipole} + Percentage\ error\ of\ FWHM_{Dipole}}{2} \text{x}100$$

$$= \frac{\frac{|584 - 584|}{584} + \frac{|162 - 165|}{165}}{2} \text{x}100 = 1\%$$

References

1. Ruan, Q.; Shao, L.; Shu, Y.; Wang, J.; Wu, H., Growth of monodisperse gold nanospheres with diameters from 20 nm to 220 nm and their core/satellite nanostructures. *Adv. Opt. Mater.* **2014**, *2* (1), 65-73.

2. Haiss, W.; Thanh, N. T.; Aveyard, J.; Fernig, D. G., Determination of size and concentration of gold nanoparticles from UV−Vis spectra. *Anal. Chem.* **2007**, 79 (11), 4215-4221.

3. Saleh, N. M.; Aziz, A. A. In *Simulation of surface plasmon resonance on different size of a single gold nanoparticle*, *J. Phys. Conf. Ser. IOP Publishing*: **2018**; p 012041.

4. Shopa, M.; Kolwas, K.; Derkachova, A.; Derkachov, G., Dipole and quadrupole surface plasmon resonance contributions in formation of near-field images of a gold nanosphere. *Opto-Electron. Rev.* **2010,** *18* (4), 421-428.

5. Derkachova, A.; Kolwas, K., Size dependence of multipolar plasmon resonance frequencies and damping rates in simple metal spherical nanoparticles. *Eur. Phys. J.: Spec. Top.* **2007,** *144* (1), 93-99.

6. Eustis, S.; El-Sayed, M. A., Why gold nanoparticles are more precious than pretty gold: noble metal surface plasmon resonance and its enhancement of the radiative and nonradiative properties of nanocrystals of different shapes. *Chem. Soc. Rev*. **2006**, 35 (3), 209-217.

7. Park, S.; Lee, J., Multivariate L\'evy Adaptive B-Spline Regression. *arXiv preprint arXiv:2108.11863* **2021**.

8. Wang, Z.; Yu, J.; Su, B.; Xie, X. In *An improved algorithm for surface fitting based on B spline function*, 2011 Fourth International Conference on Information and Computing, IEEE: 2011; pp 80-82.

9. Oshiro, T. M.; Perez, P. S.; Baranauskas, J. A. In *How many trees in a random forest?*, International workshop on machine learning and data mining in pattern recognition, *Springer*: 2012; pp 154-168.

10. Segal, M. R., Machine learning benchmarks and random forest regression. **2004**.

11. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H., Xgboost: extreme gradient boosting. *R package version 0.4-2* **2015,** *1* (4), 1-4.