Supplementary Information (SI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2024

Supporting Information for "Data-driven analysis of text-mined seed-mediated growth AuNP syntheses"

Llama-2 finetuning details

The code base for Llama-2 finetuning used in this work can be found at https://github.com/lbnlp/nerre-llama,[1] and the fine-tuned weights can be found at https://github.com/slee-lab/AuNP-seedmed-recipes.[2] Base model of Llama-2-13B-hf[3] was used with 8 bit quantization, with batch_size_training=4, micro_batch_size=4, num_epochs=4.

Precursors extracted

The pre-defined precursors were "AuCl4-", "AgNO3", "BH4-", "Cit", "CTAB", "xTA", "CTAC", "TOAB", "BSA", "PVP", "PDDA", "TEOS", "BDAC", "5-BrSA", "TritonX", "NH2OH", "LSB", "SDS", "xAC", "xAB", "OlA", "Thiol", "PEI", "Glu", "HEPES", "C18N3", "GSH", "Tann", "Cl", "Br", "I", "AA", "H2O2", "HQ", "DMF", "othersolvent", "H2O", "HCl", "NaOH", "HNO3", and "H2SO4".

For determining precursor category, a longer-length category is prioritized. For instance, "Cetrimonium bromide" matches with both "CTAB" and "Br" but would be detected as "CTAB".

"Cit" is patterns involving citrate, including "citric acid" and "C6H5O7". "HQ" is hydroquinone or quinol and "Glu" is glucose. "Cl", "Br", "I" are any patterns of "*um

chloride/bromide/iodide" or binary chemical formula involving halides (e.g. KBr). "xAC", "xAB" are any patterns with "ammonium chloride/bromide", and "xTA" covers any patterns with "trimethyl ammonium". "othersolvent" is liquid solvents outside of DMF and H2O, including alcohol, glycerine, toluene, acteone. "TritonX" is "TritonX-100".

'5-BrSA', 'TritonX', 'LSB', 'SDS', and 'NH2OH' were detected during the LLM extraction of partial and hard recipes, thereby later added only for post-processing.

Manual validation statistics

Hard (651) and partial (326) datasets were further filtered for complete recipes without hallucinations for manual validation. Still, potential bipyramid, cube, star recipes were first hand validated or corrected before filtering out, while rod recipes were filtered out and only validated (not necessarily corrected). For manual validation, 221 hard recipes and 145 partial recipes were checked for correct extraction and were corrected if possible.

Table S1: Accuracy statistics from manual validation of the datasets. "Correct" indicates where all entities of the parser (precursors and amounts) were correctly parsed.

Category	Clean (search-based)	Partial (LLM)	Hard (LLM)	Sum
Correct	272 (78%)	66 (46%)	87 (39%)	425
Total validated	350	145	221	716

Full description of precursor sets and axes in Figure 3

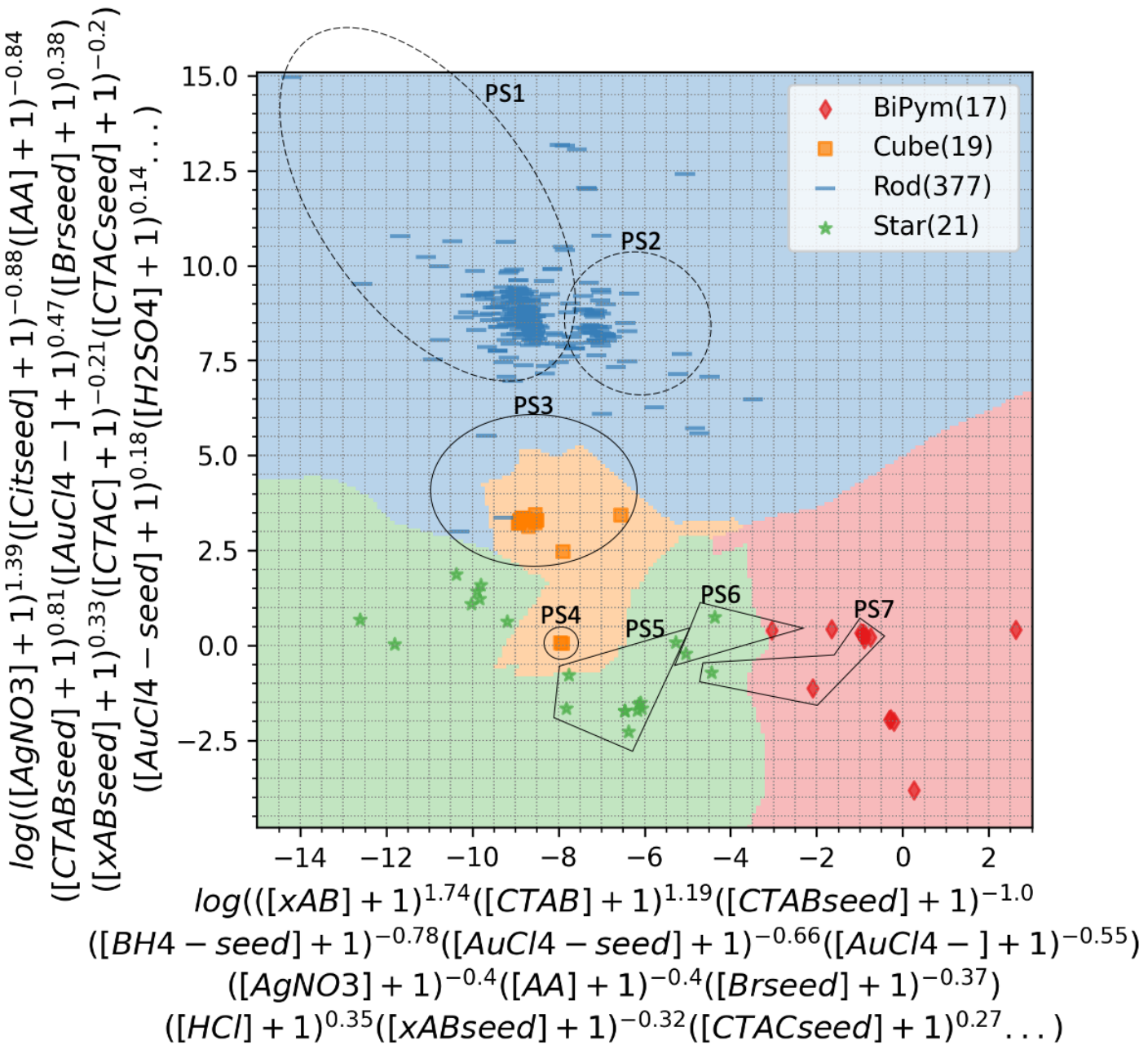

Figure S1: Figure 3 axes expanded and showing precursor sets used. Solid line circles or polygons indicate that all data points in them have the corresponding precursor set, while dashed line circles indicate that most of the data points in the circle have the corresponding precursor set. PS1 is ('AuCl4-seed', 'BH4-seed','CTABseed','AA','AgNO3','AuCl4-','CTAB'), PS2 is ('AuCl4-seed','BH4-seed','CTABseed','AA','AgNO3','AuCl4-','CTAB','HCl'), PS3 is ('AuCl4-seed','BH4-seed','CTABseed','AA','AuCl4-','CTAB'), PS4 is ('AuCl4-seed', 'BH4-seed','CTACseed','AA','AgNO3','AuCl4-','CTAC','HCl'), PS5 is any recipes that use Turkevich method seeds ('AuCl4-seed','Citseed'), PS6 is ('AuCl4-seed','BH4-seed','Citseed','AA','AgNO3','AuCl4-','CTAB'), PS7=('AuCl4-seed','BH4-seed','Citseed','AA','AuCl4-','CTAB','HCl','AgNO3').

Figure 3 uses learned weight matrix of shape (number of features, number of classes) obatined from multinomial logistic regression. The exponent values of each axis are as follows. x-axis: [('xAB', '1.735e+00'), ('CTAB', '1.192e+00'), ('CTAB seed', '-1.002e+00'), ('BH4- seed', '-7.765e-01'), ('AuCl4- seed', '-6.585e-01'), ('AuCl4-', '-5.467e-01'), ('AgNO3', '-4.039e-01'), ('AA', '-3.979e-01'), ('Br seed', '-3.658e-01'), ('HCl', '3.478e-01'), ('xAB seed', '-3.171e-01'), ('CTAC seed', '2.712e-01'), ('CTAC', '-2.455e-01'), ('HEPES', '-2.247e-01'), ('Cit seed', '-1.893e-01'), ('NH2OH', '-1.328e-01'), ('H2SO4', '-5.242e-02'), ('OlA', '-1.007e-02'), ('TritonX', '7.694e-03'), ('TritonX seed', '7.586e-03'), ('H2O2', '-4.701e-03'), ('BSA', '4.118e-03'), ('PDDA', '-3.663e-03'), ('PDDA seed', '-3.660e-03'), ('5-BrSA', '-2.476e-03'), ('BDAC', '-1.947e-03'), ('LSB', '1.220e-03'), ('LSB seed', '1.217e-03'), ('SDS', '7.421e-04'), ('HQ', '5.438e-04'), ('HNO3', '-4.477e-04'), ('NaOH seed', '2.838e-04'), ('NaOH', '1.478e-04'), ('Cit', '0'), ('PVP', '0'), ('Br', '0'), ('I', '0'), ('HCl seed', '0')]

y-axis: [('AgNO3', '1.388e+00'), ('Cit seed', '-8.818e-01'), ('AA', '-8.411e-01'), ('CTAB seed', '8.098e-01'), ('AuCl4-', '4.671e-01'), ('Br seed', '3.792e-01'), ('xAB seed', '3.284e-01'), ('CTAC', '-2.089e-01'), ('CTAC seed', '-2.023e-01'), ('AuCl4- seed', '1.798e-01'), ('H2SO4', '1.354e-01'), ('TritonX', '-1.240e-01'), ('CTAB', '1.234e-01'), ('TritonX seed', '-1.228e-01'), ('LSB', '-1.149e-01'), ('LSB seed', '-1.147e-01'), ('xAB', '-9.460e-02'), ('HCl', '-9.202e-02'), ('PDDA', '-5.765e-02'), ('PDDA seed', '-5.760e-02'), ('BH4- seed', '-2.835e-02'), ('5-BrSA', '2.017e-02'), ('HNO3', '1.665e-02'), ('BDAC', '1.476e-02'), ('HEPES', '-1.461e-02'), ('OlA', '1.193e-02'), ('H2O2', '8.749e-03'), ('NH2OH', '-8.633e-03'), ('BSA', '5.228e-03'), ('SDS', '-3.306e-03'), ('HQ', '-1.603e-03'), ('NaOH seed', '-6.126e-04'), ('NaOH', '-3.304e-04'), ('Cit', '0'), ('PVP', '0'), ('Br', '0'), ('I', '0'), ('HCl seed', '0')]

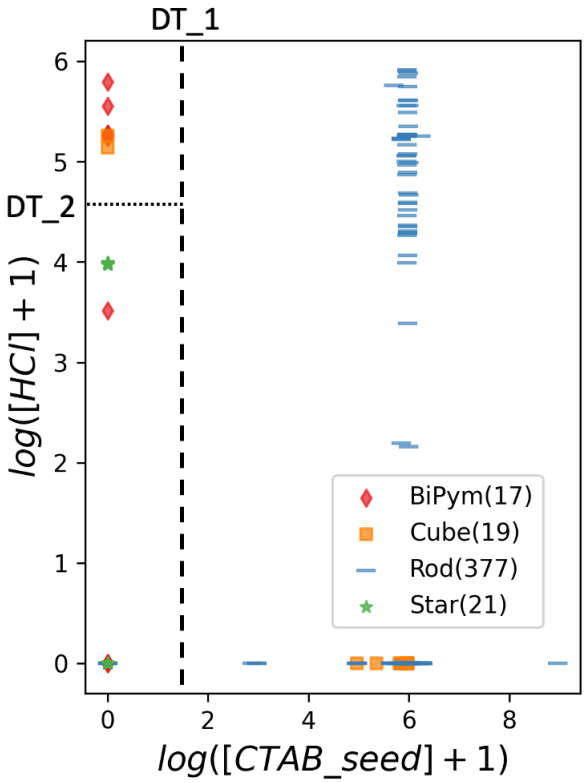

Figure S2: The first two decision boundaries from the decision tree of depth 3 for morphology classification. The first and second criterion are shown as the dashed line and the dotted line, respectively.

Correlation between AuNR Sizes and $AgNO_3$

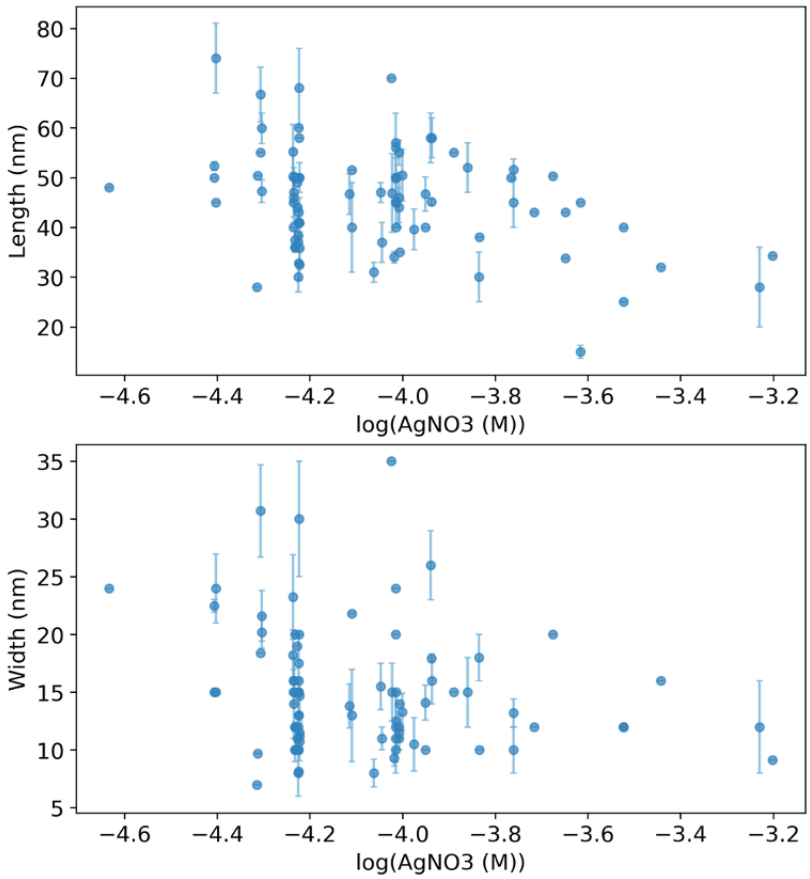

Figure S3: Lengths and widths of the synthesized AuNRs using seed solution of $AuCl_4^-$, CTAB, BH_4^- and growth solution of $AuCl_4^-$, CTAB, $AgNO_3$ and ascorbic acid. Scatter with error bar show range (or variance) of sizes indicated in each publication. All data points are manually validated.

Identical recipes in full detail

Table S2: Apparently identical recipes with different outcomes

Index	Group	HAuCl4(seed)	CTAB(seed)	NaBH4(seed)	CTAB	HAuCl4	AgNO3	AA	Seed/growth	AR	Reference
1	A	5 mL, 0.05 mM	5 mL, 0.20 M	0.6 mL, 0.01 M	20 mL, 0.20 M	20 mL, 1 mM	1 mL, 4 mM	0.28 mL, 0.0788 M	0.001161	3.3	[4]
2	A	0.5 mM, 5 mL	0.2 M, 5 mL	0.01 M, 0.6 mL	0.2 M, 5.0 mL	1.0 mM, 5.0 mL	4.0 mM, 0.25 mL	78.8 mM, 70 µL	0.001161	3.61	[5]
3	A	0.5 mM, 5 mL	0.2 M, 5 mL	0.01 M, 600 µL	0.2 M, 5.00 mL	1 mM, 5.00 mL	4 mM, 250 µL	78.8 mM, 70 µL	0.001161	2.375	[6]
4	B	0.25 mL, 0.01 M	7.5 mL, 0.1 M	0.6 mL, 0.01 M	4.75 mL, 0.10 M	0.200 mL, 0.01 M	30 µL, 0.01 M	32 µL, 0.10 M	0.001991	3-4	[7]
5	B	250 µL, 10 mM	7.5 mL, 100 mM	600 µL, 10 mM	47.5 mL, 100 mM	2 mL, 10 mM	300 µL, 10 mM	320 µL, 100 mM	0.001991	5	[8]
6	B	250 µL, 0.01 M	7.5 mL, 0.1 M	600 µL, 0.01 M	42.75 mL, 0.1 M	1.8 mL, 0.01 M	270 µL, 0.01 M	0.1 M, 288 µL	0.001991	2.69	[9]
7	B	0.25 mL, 0.01 M	7.5 mL, 0.1 M	0.6 mL, 0.01 M	4.75 mL, 0.10 M	0.20 mL, 0.01 M	0.03 mL, 0.01 M	0.032 mL, 0.10 M	0.001991	3.15	[10]
8	B	0.25 mL, 0.01 M	7.5 mL, 0.1 M	0.6 mL, 0.01 M	4.75 mL, 0.10 M	0.2 mL, 0.01 M	30 µL, 0.01 M	32 µL, 0.10 M	0.001991	3.2	[11]
9	C	5.0 mL, 0.0005 M	5 mL, 0.2 M	0.6 mL, 0.01 M	5 mL, 0.2 M	5.0 mL, 0.001 M	0.2 mL, 0.004 M	70 µL, 0.0788 M	0.001167	2.6 ± 0.3	[12]
10	C	5 mL, 0.5 mM	5 mL, 0.20 M	0.60 mL, 0.01 M	5 mL, 0.20 M	5 mL, 1 mM	0.2 mL, 4 mM	70 , 0.0788 M	0.001167	4.0	[13]
11	C	0.5 mM, 5 mL	0.2 M, 5 mL	0.01 M, 0.6 mL	0.2 M, 5 mL	1 mM, 5 mL	4 mM, 0.2 mL	0.0788 M, 0.07 mL	0.001167	4	[14]
12	D	0.25 mL, 0.01 M	7.5 mL, 0.1 M	0.6 mL, 0.01 M	9.44 mL, 0.10 M	0.4 mL, 0.01 M	0.06 mL, 0.01 M	0.06 mL of 0.10 M	0.004	2-3	[15]
13	D	0.125 mL, 0.01 M	3.75 mL, 0.1 M	0.3 mL, 0.01 M	33.04 mL, 0.10 M	1.4 mL, 0.01 M	0.21 mL, 0.01 M	0.21 mL of 0.10 M	0.004	3.21	[16]
14	D	0.125 mL, 0.01 M	3.75 mL, 0.1 M	0.3 mL, 0.01 M	47.2 mL, 0.1 M	2 mL, 0.01 M	0.3 mL, 0.01 M	0.3 mL, 0.1 M	0.004	3.6 ± 0.5	[17]

Feature transformations and Interpretability of ML models

It is interesting to note the significance of the choice of feature embeddings in the synthesis space. For instance, working with log of the concentrations as the features was used not only because the dataset was positive-skew, but also because it improved classifier performance with regularization.

In linear regression or other models that are based on linear predictors, the model predictions are based on linear combinations of the features ((S1),(S2)).

$$y_{model} = \sum_{featurei} w_i x_i \tag{S1}$$

$$y_{model} = \sum_{precursori} w_i[X_i] = w_0 + w_1[HAuCl_4] + w_2[CTAB] + ... \tag{S2}$$

where x is precursor vector from one data point (i.e. recipe) and w is coefficient vector. We propose feature processing by log transformation not only to address the skewness of the dataset but also for better interpretability. While linear combinations of the precursor concentrations physically can be interpreted as balance equations, linear combination of the logarithms of precursor concentrations is closely related to kinetics (reaction rate) or thermodynamics (reaction quotient Q), hence likely to provide more insight into the mechanism.

$$y_{model} = \sum_i w_i log[X_i] = w_0 + w_1 log[X_1] + w_2 log[X_2] + ... = log(k[X_1]^{w_1}[X_2]^{w_2}\dots) \tag{S3}$$

However, the sparseness of the concentrations should be noted and addressed for log transformation, as 0 maps to $-\infty$. First solution to address the sparseness is to featurize by the precursor's role, in this case specifically, gold source (gold chloride), capping agent (e.g. CTAB, citrate), reducing agent (e.g. borohydride, ascorbic acid). This way, a data point would be featurized as a fixed-size vector consisted of each role's material embedding and

its quantifier. However, in AuNP synthesis, there are often cases involving "additives" that have unclear roles (e.g. acids, metal cations), have more than one capping agents, or have a chemical that serves as both reducing agent and capping agent.

Another potential solution would be to separate the dataset into many datasets by the combinations of precursors used, to simply eliminate the precursor columns that are not used and ensure every value is nonzero. This can be useful for the analysis of recipes using a fixed set of precursors, assuming that the size of each separated dataset is large enough for meaningful analysis.

$$y_{model} = \sum_i w_i log([X_i] + \frac{1}{C}) \simeq \sum_{[X_i]\neq 0} w_i log[X_i] - logC \cdot \sum_{[X_i]=0} w_i \quad (\text{if } C \gg 1) \tag{S4}$$

The alternative solution we considered is the featurization scheme in (S4), which essentially replaces zero values with very small constant values. This constant was chosen considering the range of precursor concentrations, the minimum being 1×10^{-8}M for seed precursors after converting to the final concentration in the final seed-growth solution mixture. When trained for the whole dataset, using sufficiently large C, either the model prediction simply deviates from the interpretable model in (S6), or the coefficients of the precursors that were not used in any data point would end up very small during the training process. This essentially means that HAuCl4, which is nonzero in all of the data points, would have a considerably large coefficient than any of the other precursors. This is counter-intuitive, as we expect different mechanisms to take place depending on the precursors that are present.

$$y_{model} = \sum_i w_i log(C \cdot [X_i] + 1) \simeq \sum_{[X_i]\neq 0} w_i log[X_i] + logC \cdot \sum_{[X_i]\neq 0} w_i \quad (\text{if } C \gg 1) \tag{S5}$$

$$y_{model} \simeq \sum_{[X_i]\neq 0} w_i log[X_i] = log(k \cdot \prod_{[X_i]\neq 0} [X_i]^{w_i}) \quad (\text{initial-reaction-rate-like form}) \tag{S6}$$

Therefore, we used featurization scheme (S5), which scales the concentration features with a large enough value and shifts the features by +1 to map 0 M to 0 when log-transformed.

As we want to interpret in the form of (S6), using regularization like (S8) can penalize large coefficients w's.

$$loss = \sum_{datumd} (y - y_{model})_d^2 \quad \text{(Least Squares)} \tag{S7}$$

$$loss = \sum_{datumd} (y - y_{model})_d^2 + \lambda \cdot \sum_{precursori} w_i^2 \quad \text{(Ridge)} \tag{S8}$$

References

(1) Lee, S. `https://github.com/lbnlp/nerre-llama`, 2023.

(2) Lee, S. `https://github.com/slee-lab/AuNP-seedmed-recipes`, 2024.

(3) Touvron, H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023; `https://arxiv.org/abs/2307.09288`.

(4) Fateixa, S.; Correia, M. R.; Trindade, T. Resizing of Colloidal Gold Nanorods and Morphological Probing by SERS. *The Journal of Physical Chemistry C* **2013**, *117*, 20343–20350.

(5) Yang, Z.; Lin, Y.-W.; Tseng, W.-L.; Chang, H.-T. Impacts that pH and metal ion concentration have on the synthesis of bimetallic and trimetallic nanorods from gold seeds. *J. Mater. Chem.* **2005**, *15*, 2450–2454.

(6) Fernando, D.; Nigro, T. A.; Dyer, I.; Alia, S. M.; Pivovar, B. S.; Vasquez, Y. Synthesis and catalytic activity of the metastable phase of gold phosphide. *Journal of Solid State Chemistry* **2016**, *242*, 182–192, Solid State Chemistry of Energy-Related Materials.

(7) Zhang, C.; Lee, J. Y. Synthesis of Au Nanorod@Amine-Modified Silica@Rare-Earth Fluoride Nanodisk Core–Shell–Shell Heteronanostructures. *The Journal of Physical Chemistry C* **2013**, *117*, 15253–15259.

(8) Kang, K.; Jang, H.; Kim, Y.-K. The influence of polydopamine coating on gold nanorods for laser desorption/ionization time-of-flight mass spectrometric analysis. *Analyst* **2017**, *142*, 2372–2377.

(9) Ukaegbu, M.; Enwerem, N.; Bakare, O.; Sam, V.; Southerland, W.; Vivoni, A.; Hosten, C. Probing the adsorption and orientation of 2,3-dichloro-5,8-dimethoxy-1,4-naphthoquinone on gold nano-rods: A SERS and XPS study. *Journal of Molecular Structure* **2016**, *1114*, 197–205.

(10) Zhang, J.; Feng, Y.; Mi, J.; Shen, Y.; Tu, Z.; Liu, L. Photothermal lysis of pathogenic bacteria by platinum nanodots decorated gold nanorods under near infrared irradiation. *Journal of Hazardous Materials* **2018**, *342*, 121–130.

(11) Shou, Q.; Ebara, M.; Wang, J.; Wang, Q.; Liang, X.; Liu, H.; Aoyagi, T. Preparation of phase diagram of gold nanorods in mixture solvent of DMSO and water and its application for efficient surface-modification. *Applied Surface Science* **2018**, *457*, 264–270.

(12) Chandrasekar, G.; Mougin, K.; Haidara, H.; Vidal, L.; Gnecco, E. Shape and size transformation of gold nanorods (GNRs) via oxidation process: A reverse growth mechanism. *Applied Surface Science* **2011**, *257*, 4175–4179.

(13) Chen, H.-Q.; Yuan, F.; Wang, S.-Z.; Xu, J.; Zhang, Y.-Y.; Wang, L. Near-infrared to near-infrared upconverting NaYF4:Yb3+,Tm3+ nanoparticles-aptamer-Au nanorods light resonance energy transfer system for the detection of mercuric(ii) ions in solution. *Analyst* **2013**, *138*, 2392–2397.

(14) Khaletskaya, K.; Reboul, J.; Meilikhov, M.; Nakahama, M.; Diring, S.; Tsujimoto, M.; Isoda, S.; Kim, F.; Kamei, K.-i.; Fischer, R. A.; Kitagawa, S.; Furukawa, S. Integration of Porous Coordination Polymers and Gold Nanorods into Core–Shell Mesoscopic

Composites toward Light-Induced Molecular Release. *Journal of the American Chemical Society* **2013**, *135*, 10998–11005, PMID: 23672307.

(15) Parab, H. J.; Chen, H. M.; Lai, T.-C.; Huang, J. H.; Chen, P. H.; Liu, R.-S.; Hsiao, M.; Chen, C.-H.; Tsai, D.-P.; Hwu, Y.-K. Biosensing, Cytotoxicity, and Cellular Uptake Studies of Surface-Modified Gold Nanorods. *The Journal of Physical Chemistry C* **2009**, *113*, 7574–7578.

(16) Park, J.-H.; Byun, J.-Y.; Shim, W.-B.; Kim, S. U.; Kim, M.-G. High-sensitivity detection of ATP using a localized surface plasmon resonance (LSPR) sensor and split aptamers. *Biosensors and Bioelectronics* **2015**, *73*, 26–31.

(17) Chung, S.; Lee, H.; Kim, H.-S.; Kim, M.-G.; Lee, L. P.; Lee, J. Y. Transdermal thiol–acrylate polyethylene glycol hydrogel synthesis using near infrared light. *Nanoscale* **2016**, *8*, 14213–14221.