Full-Spectrum Mapping of the Synthesis Landscape for Chiral Gold Nanoparticles

Zhi-Bo Yang[1], Jia-Tong Li[1], Xue-Yao Wang[1], Zi-Xuan Zhang[1], Tianyi Wu[2], Qiang Fei[3], Guodong Feng[3], Ning-Ning Zhang[1*], Eugenia Kumacheva[2,4,5,6*], Kun Liu[1*]

[1]State Key Laboratory of Supramolecular Structure and Materials, College of Chemistry, Jilin University, Changchun, Jilin, 130012, China

[2]Department of Chemistry, University of Toronto, Toronto, ON, M5S 3H6, Canada

[3]College of Chemistry, Jilin University, Changchun, Jilin, 130012, China

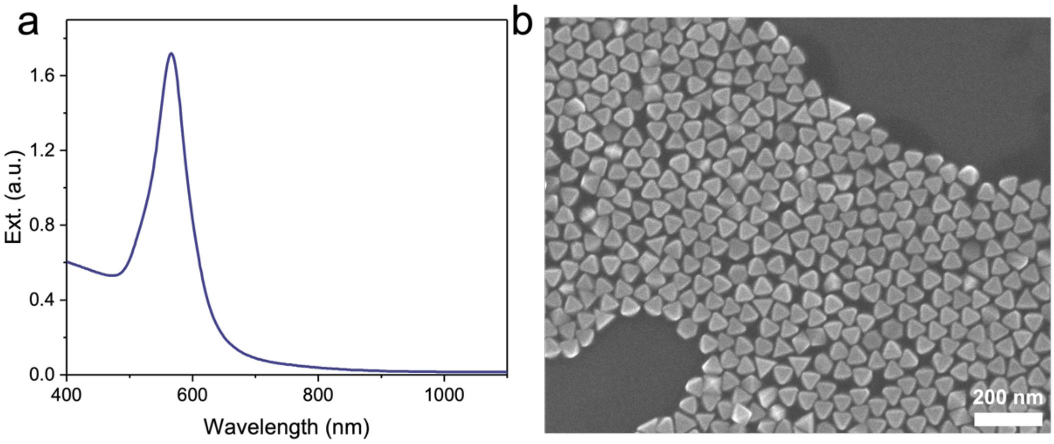
[4]Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON M5S 3E5, Canada

[5]Acceleration Consortium, University of Toronto, Toronto, ON M5S 3H6, Canada

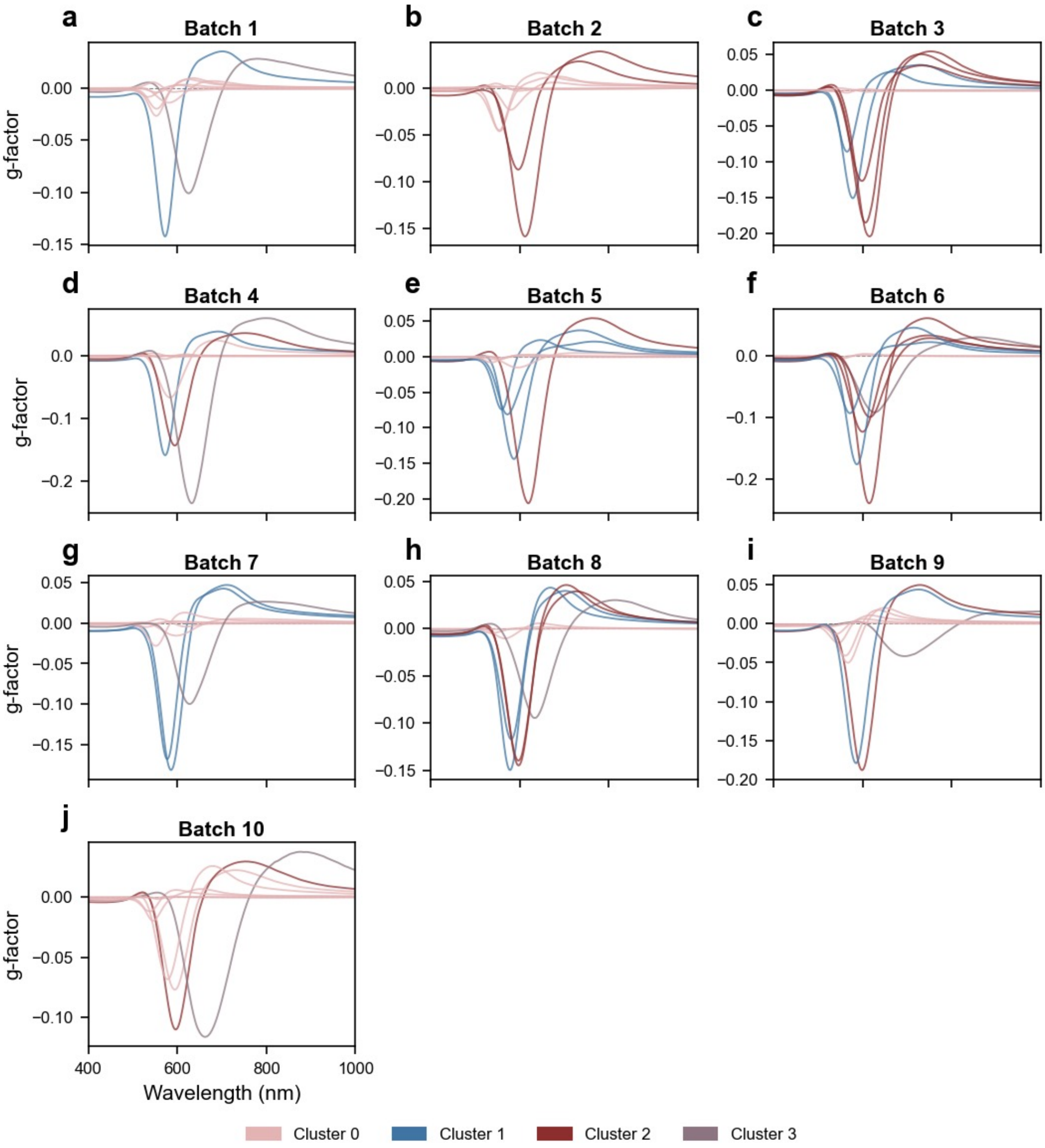
[6]Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON M5S 3G9, Canada

Table of Contents

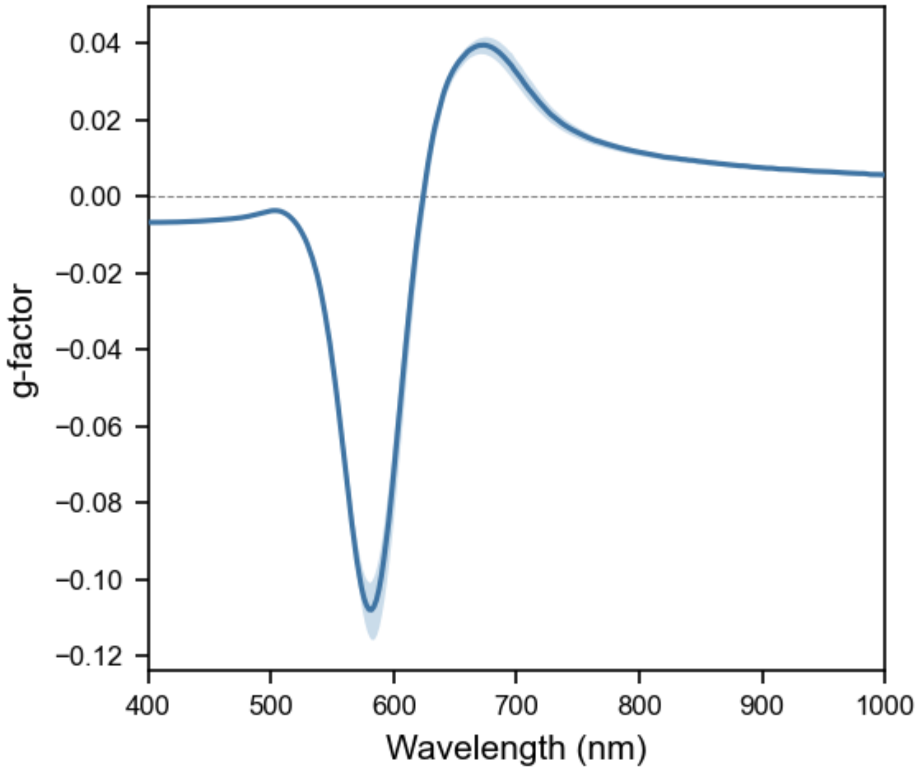
simulations.



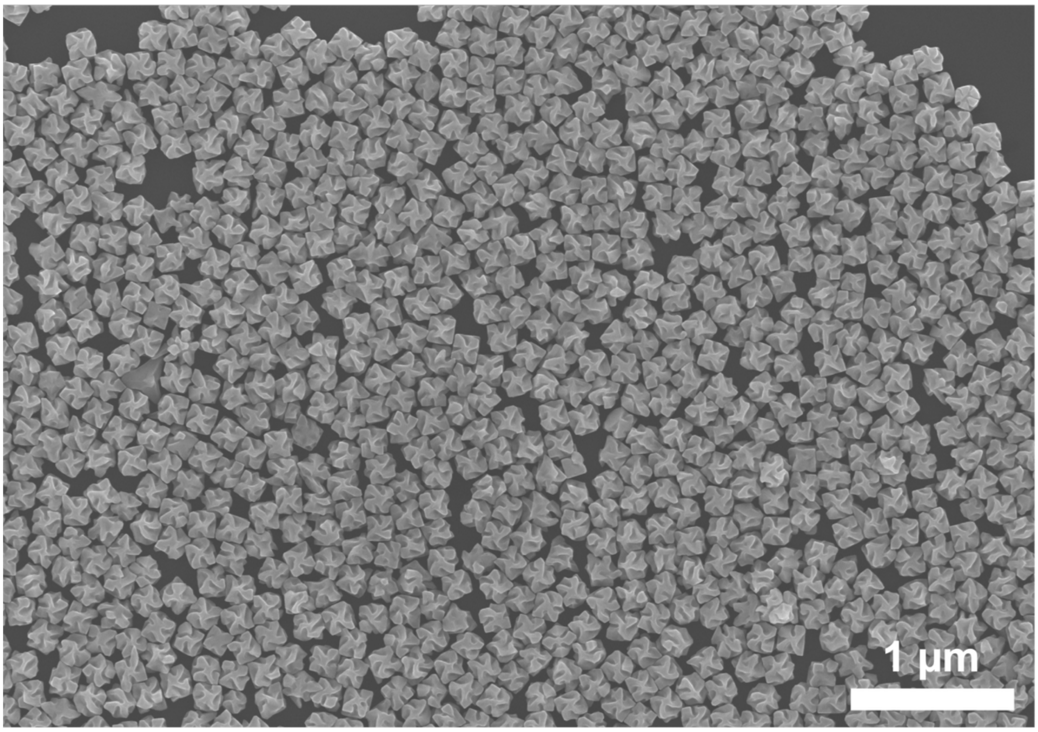
Supplementary Fig. 1. (a) UV–vis absorption spectrum and (b) SEM image of the gold nano-octahedra.



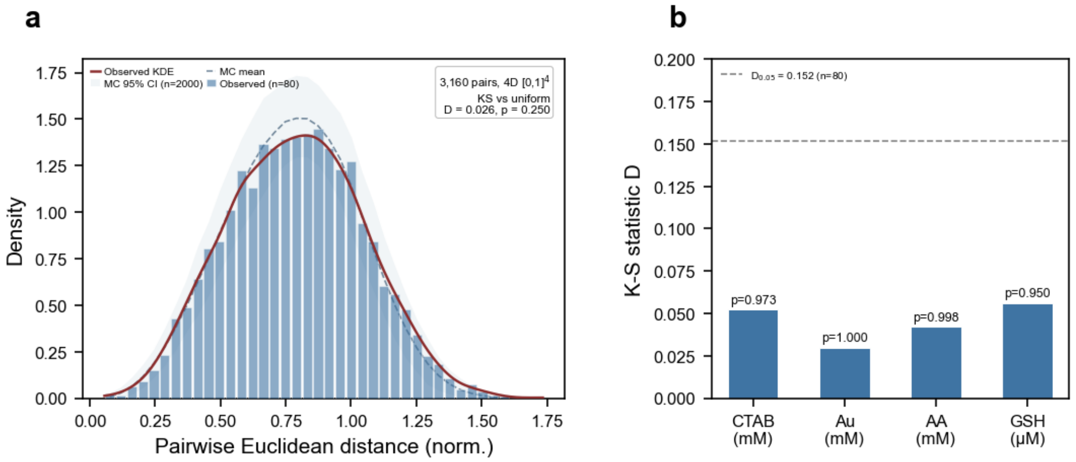
Supplementary Fig. 2. The *g*-factor spectra of all 80 samples arranged by batch. Different colors represent different cluster classifications.



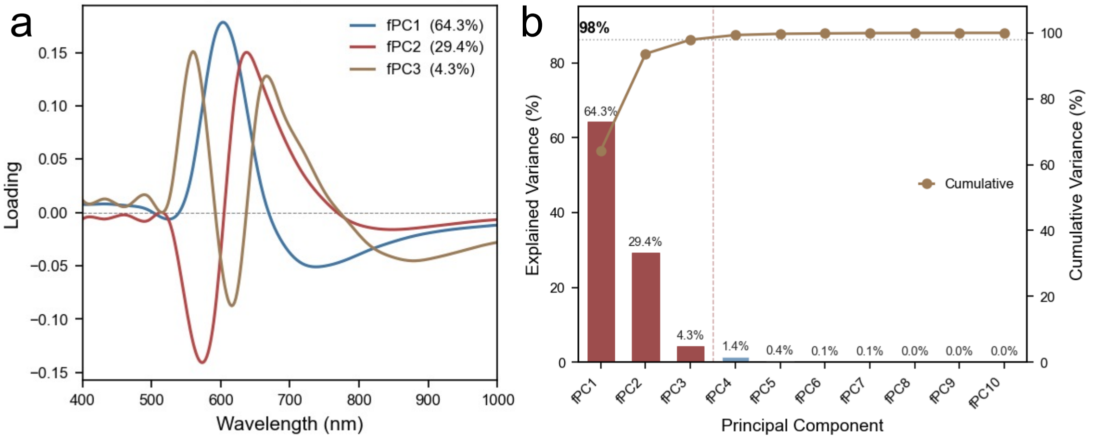
Supplementary Fig. 3. Mean *g*-factor spectrum of the internal control formulation across ten batches. Solid line: mean; shaded region: ±1 s.d. The low spectral variance (mean $|g\text{-factor}|_{\max} = 0.1083 \pm 0.0073$) confirms batch-to-batch reproducibility of the automated synthesis platform.



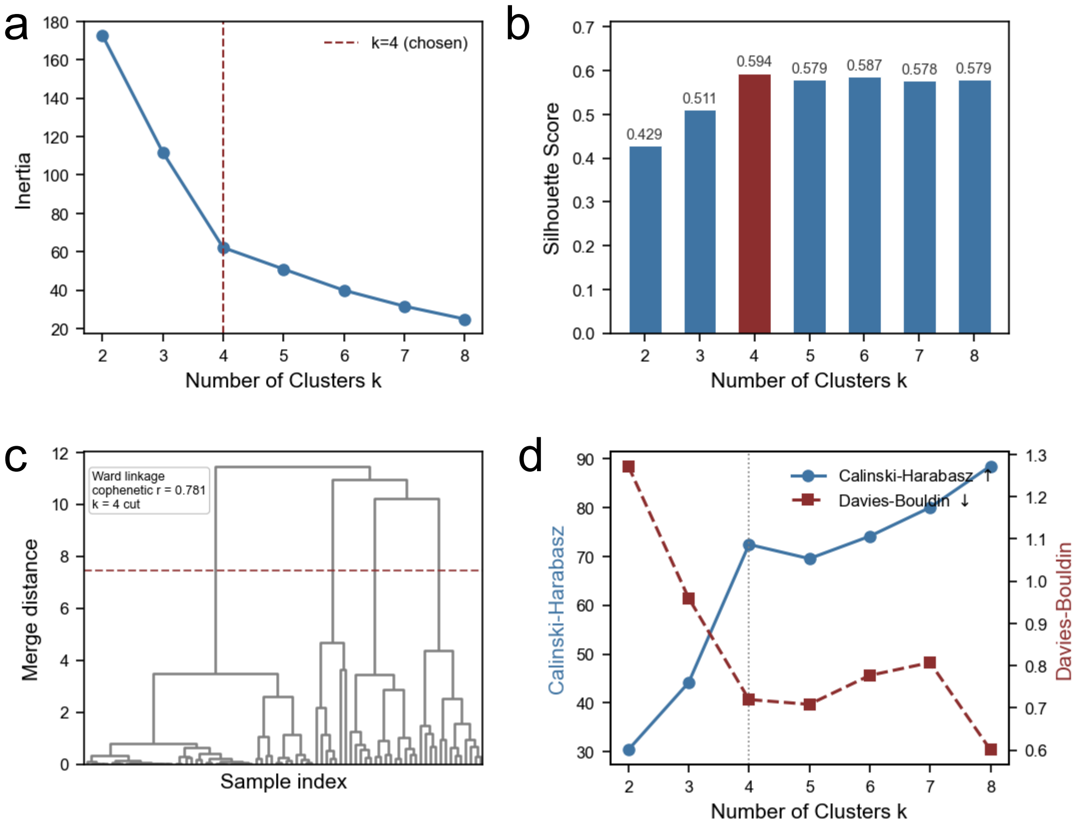
Supplementary Fig. 4. Low-magnification SEM image of the ML-identified high-|*g*-factor| sample (Batch 6, Well 6).



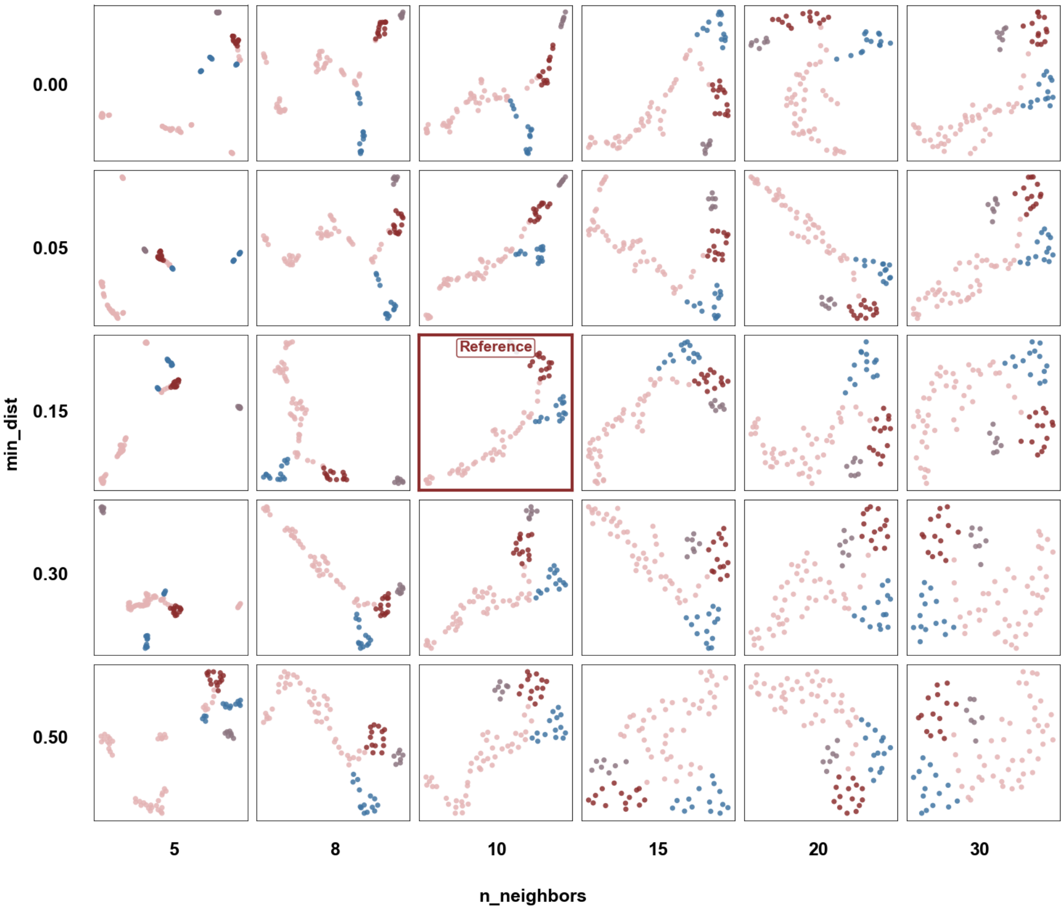
Supplementary Fig. 5. Uniformity assessment of the 80 data points in the four-dimensional synthesis parameter space. **a,** Distribution of all 3,160 pairwise Euclidean distances among the 80 experimental points in min–max normalised $[0,1]^4$ space (light blue bars) overlaid with the kernel density estimate (dark red curve). The light blue band represents the bin-wise 2.5th–97.5th percentile range obtained from 2,000 Monte Carlo draws of 80 uniform random points in $[0,1]^4$; the navy dashed line denotes the Monte Carlo mean histogram. A two-sample Kolmogorov–Smirnov test against the pooled uniform baselines returns D and *p*-values annotated in the upper-right inset. **b,** Marginal Kolmogorov–Smirnov test for each of the four synthesis parameters after min–max normalisation. Bar heights denote the K-S statistic D; the grey dashed line marks the critical value $D_{0.05} = 0.152$ for $n = 80$. Deep blue bars indicate $p > 0.05$ (uniformity not rejected). Experimentally measured *p*-values are annotated above each bar. All four parameters pass the marginal uniformity test, corroborating the conclusion from panel a that the 80 data points provide approximately unbiased coverage of the four-dimensional chemical space.



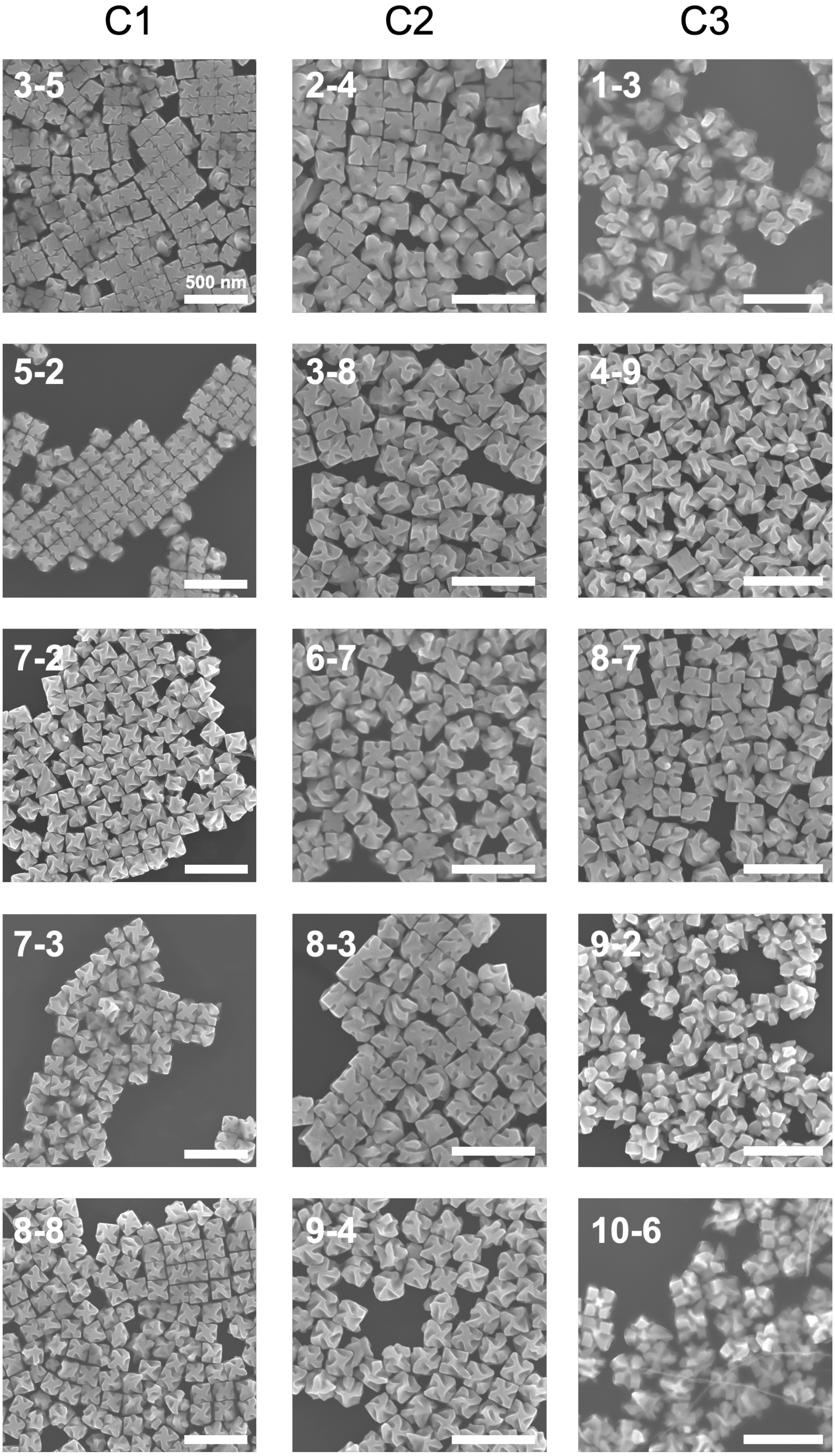
Supplementary Fig. 6. Loading curves (**a**) and cumulative variance explained (**b**) for the first three principal components of fPCA. The first three fPC loading curves (eigenfunctions) plotted against wavelength. fPC1 (64.3% variance) captures the overall amplitude modulation and peak shift; fPC2 (29.4% variance) captures the asymmetry and zero-crossing variation; fPC3 (4.3% variance) captures fine spectral shape features. The three components collectively explain 98.0% of the total spectral variance.



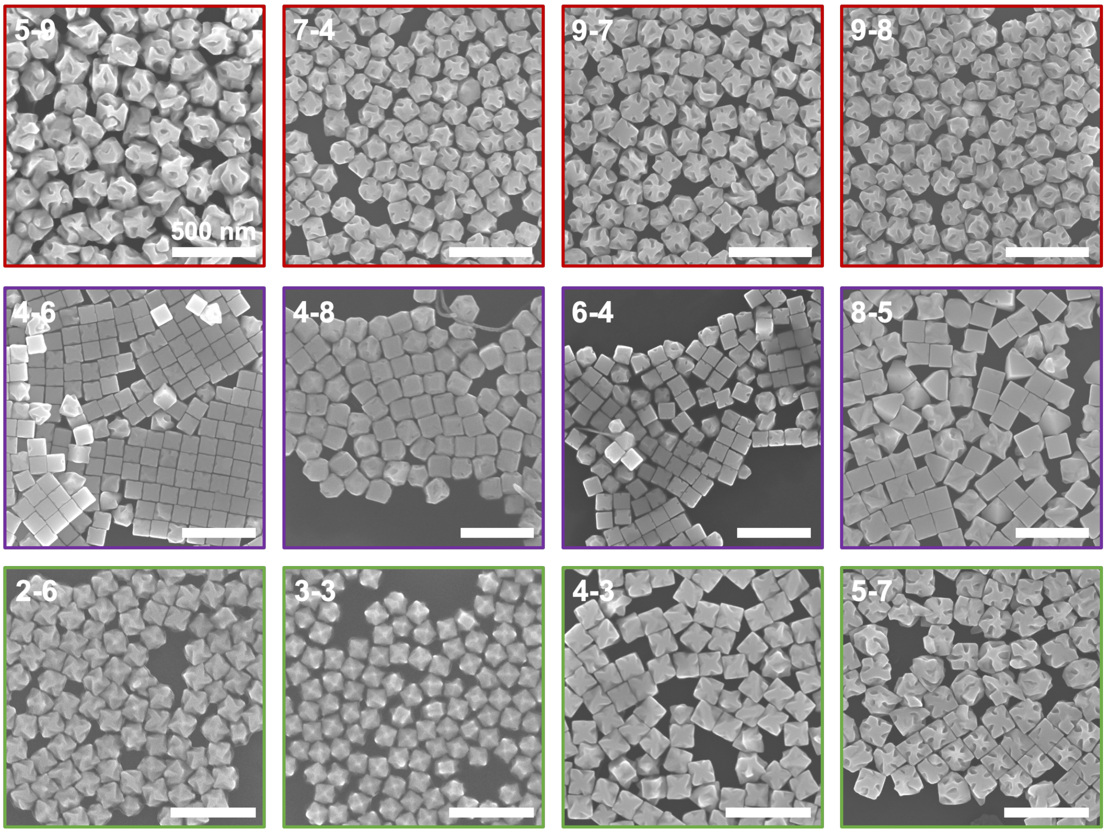
Supplementary Fig. 7. Validation of the optimal cluster number $k = 4$. **a,** Elbow method: within-cluster inertia as a function of k ($k = 2$–8). **b,** Silhouette score as a function of k, $k = 4$ yields the highest score (0.594). **c,** Ward hierarchical clustering, cophenetic correlation = 0.781. **d,** Calinski-Harabasz (blue, higher is better) and Davies-Bouldin (red, lower is better) indices. **All data show that $k = 4$ produces the most spectrally distinct and physically interpretable cluster separation.**



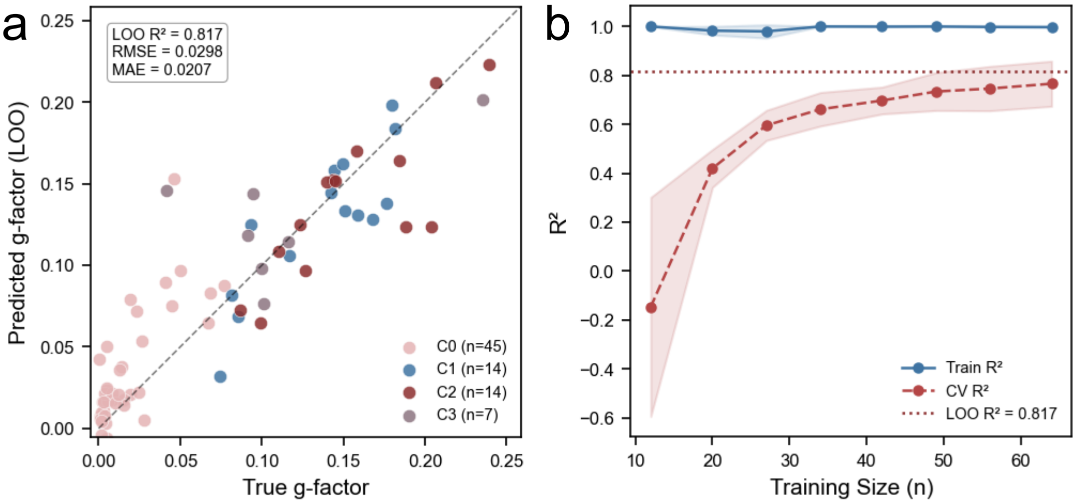
Supplementary Fig. 8. UMAP embedding robustness to hyperparameter variation. Each panel shows the UMAP embedding of 80 samples colored by fixed fPC-based cluster labels, for varying n_neighbors (columns) and min_dist (rows). The four-branch topology remains stable across n_neighbors ≥ 8 and min_dist ≤ 0.30. **The reference parameter set (n_neighbors = 10, min_dist = 0.15) is highlighted in red.**



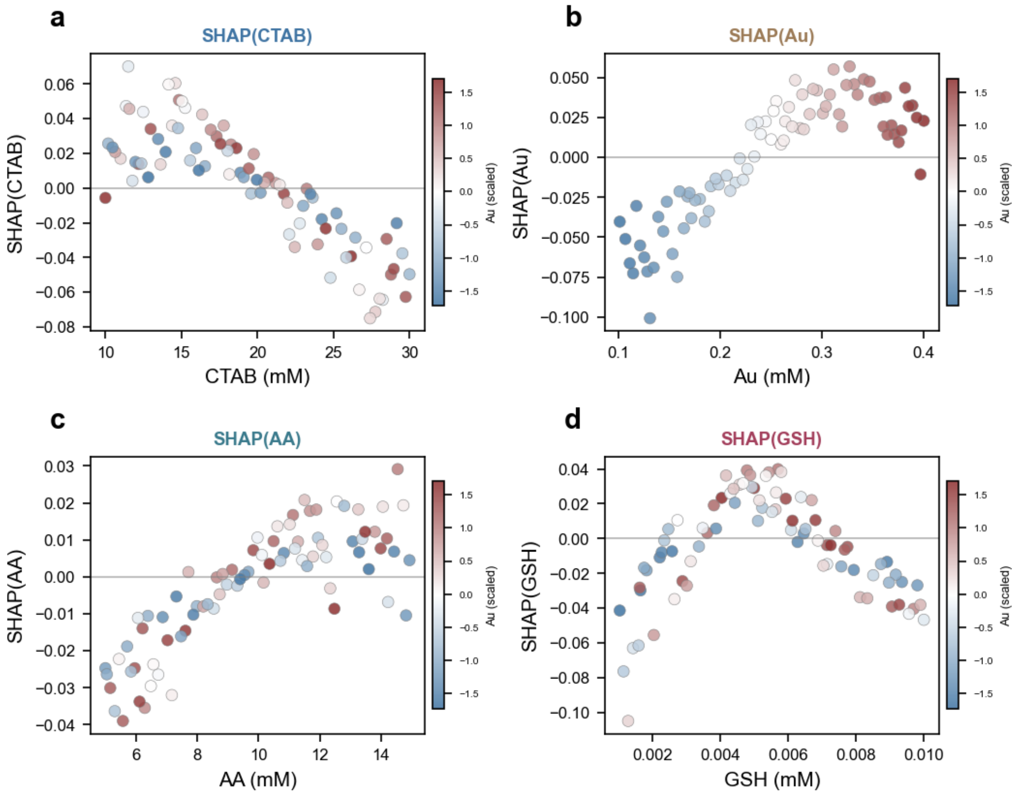
Supplementary Fig. 9. SEM images of representative samples from cluster (C1, C2, C3), all scale bars are 500 nm.



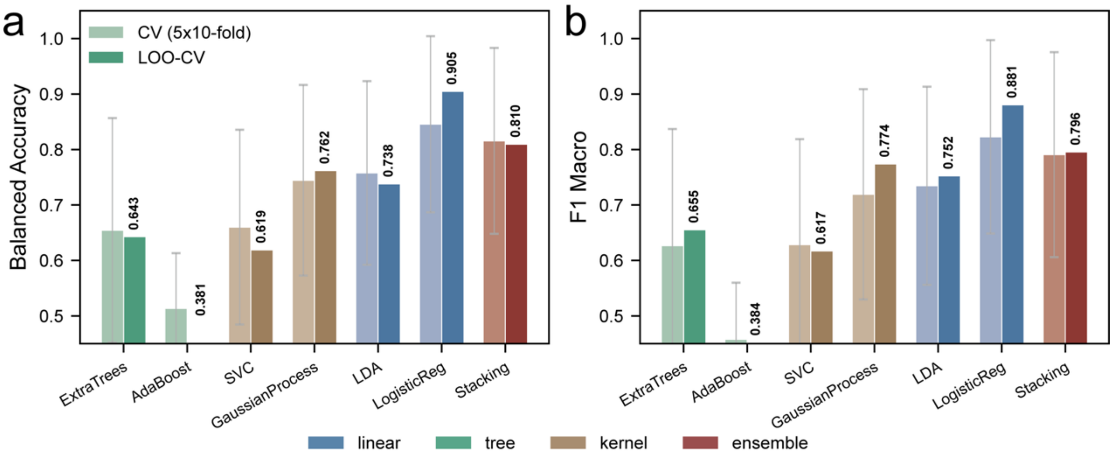
Supplementary Fig. 10. SEM images of representative samples from C0 cluster. The red box indicates a chiral rhombic dodecahedron, the purple box indicates a structure similar to a cube, and the green box indicates other structures. All scale bars are 500 nm.



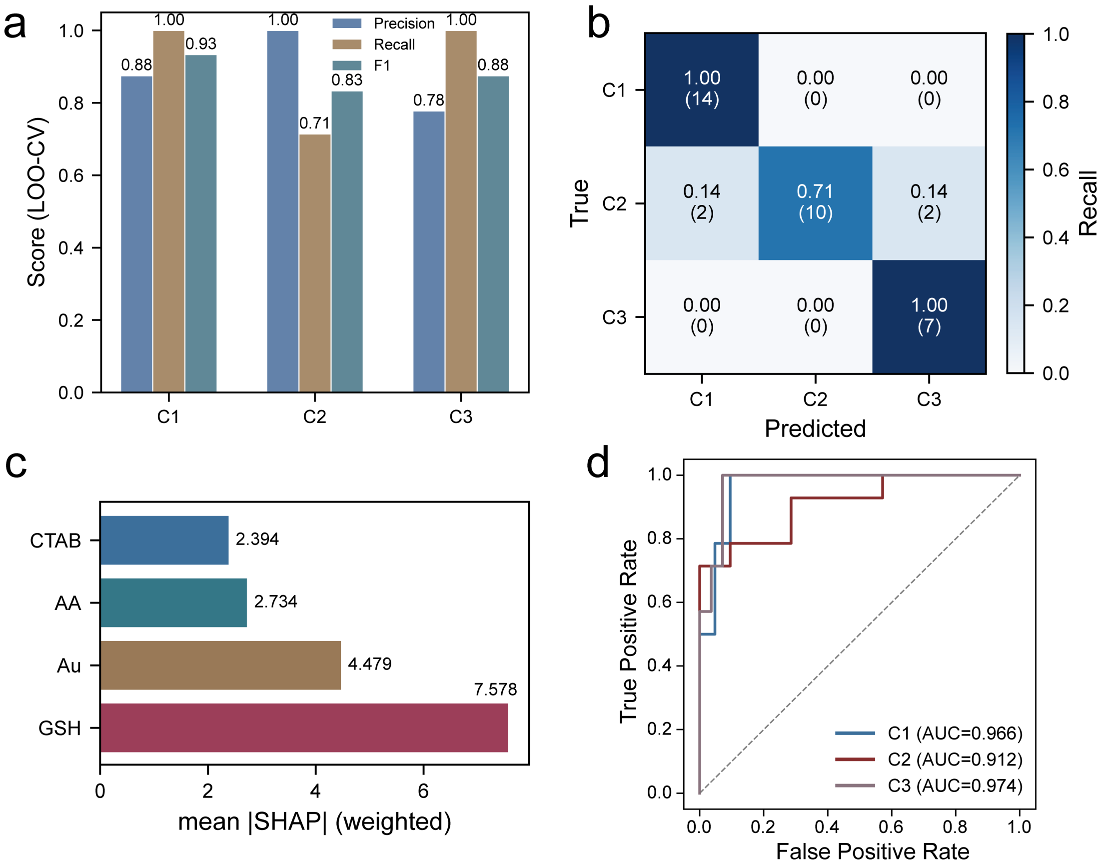
Supplementary Fig. 11. a, Leave-one-out (LOO) cross-validation of the Gaussian process regression model.[1] Predicted *vs.* true *g*-factor values for 80 samples, colored by spectral cluster. The dashed diagonal line represents perfect prediction. LOO $R^2 = 0.817$, RMSE = 0.0298, MAE = 0.0207. **b,** Gaussian process regression model learning curve. Training R^2 (blue) remains near unity across all training set sizes, indicating sufficient model flexibility. Five-fold cross-validation R^2 (red) increases monotonically with training size from near zero to approximately 0.79 at $n = 80$ without plateauing, confirming that additional data would likely further improve generalisation. The train–CV gap of ~0.21 is stable and non-divergent, ruling out severe overfitting. The grey dotted line marks the leave-one-out R^2 of 0.817 as the best estimate of generalisation performance. Shaded bands denote ±1 s.d. across folds.



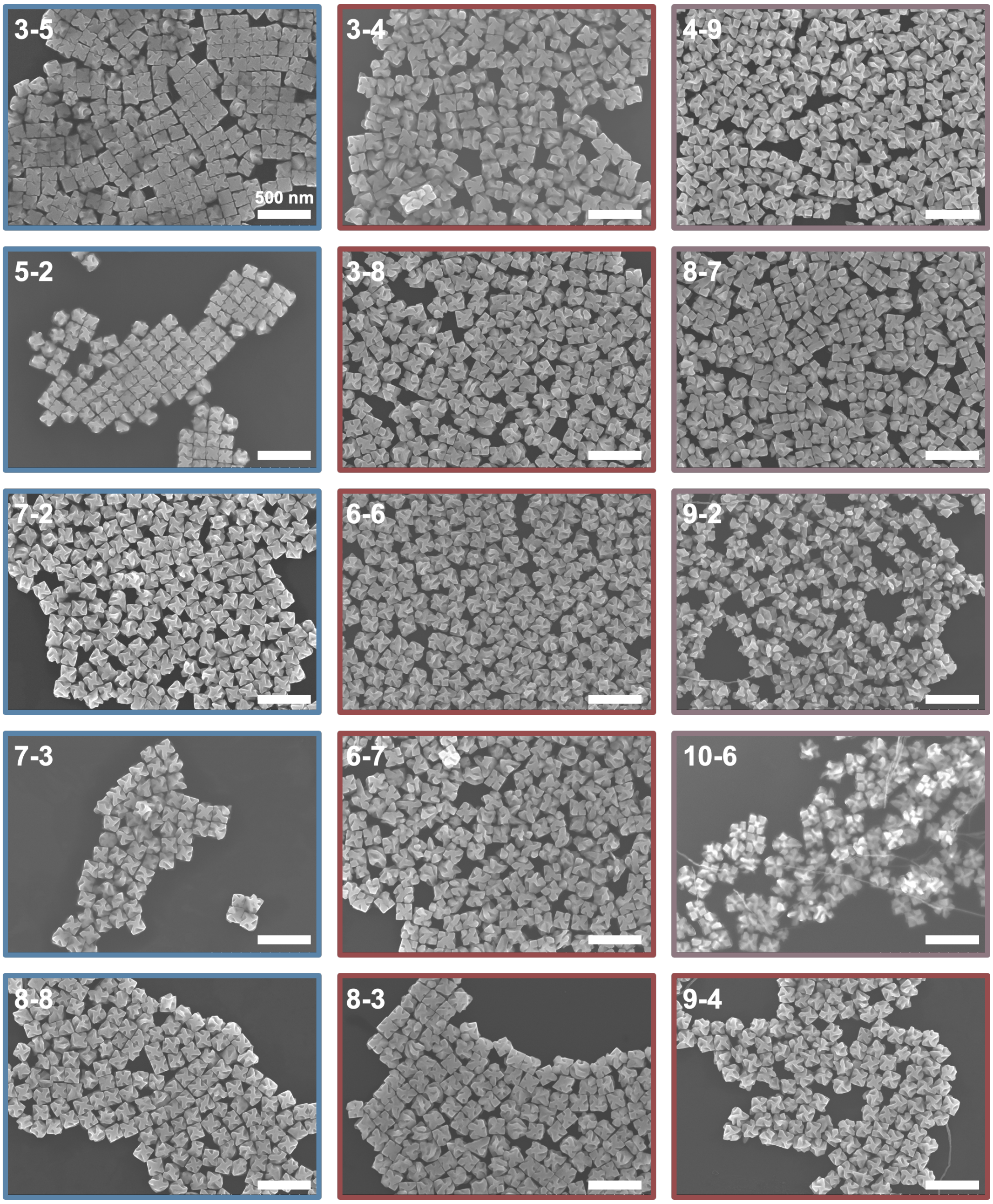
Supplementary Fig. 12. Complete SHAP dependence plots showing the interaction between the four synthetic parameters and the primary parameter Au. **a,** CTAB. **b,** $HAuCl_4$. **c,** AA. **d,** GSH.



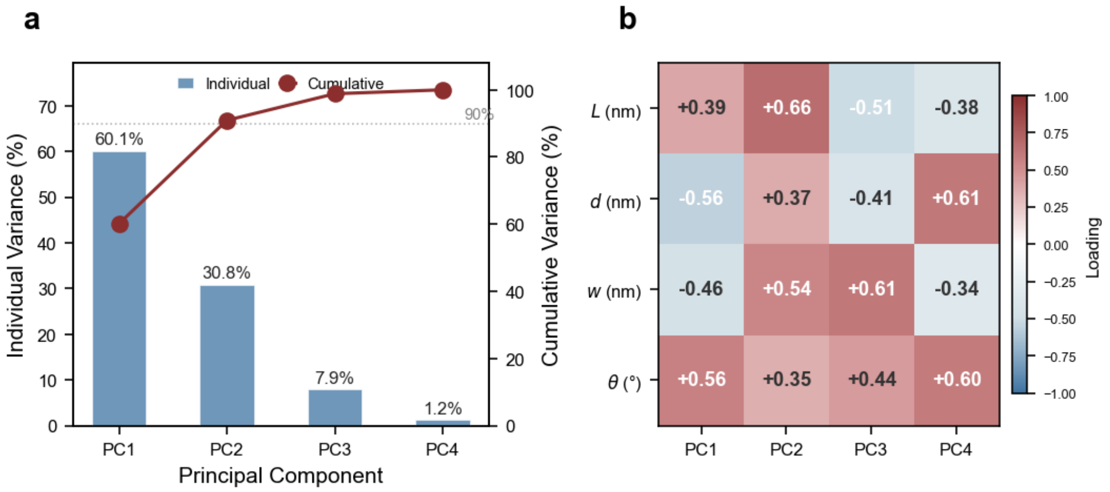
Supplementary Fig. 13. Classification model comparison for C1, C2, and C3 (*n* = 35, excluding C0). **a,** LOO balanced accuracy (BAcc) (dark bars) and 5 × 10-fold repeated stratified cross-validation balanced accuracy (light bars with ±1 s.d. error bars). **b,** Corresponding F1-macro scores. Bar colours indicate model category: green for tree-based, brown for kernel-based, blue for linear, and red for ensemble. Logistic regression achieves the highest LOO BAcc (0.905) and F1-macro (0.881), followed by the stacking ensemble (0.810), Gaussian process classifier (0.762), and linear discriminant analysis (0.738). The systematic superiority of linear models over their non-linear counterparts confirms that C1, C2, and C3 are linearly separable in the four-dimensional synthesis parameter space, **ruling out complex non-linear coupling as a prerequisite for structural pathway identification.**



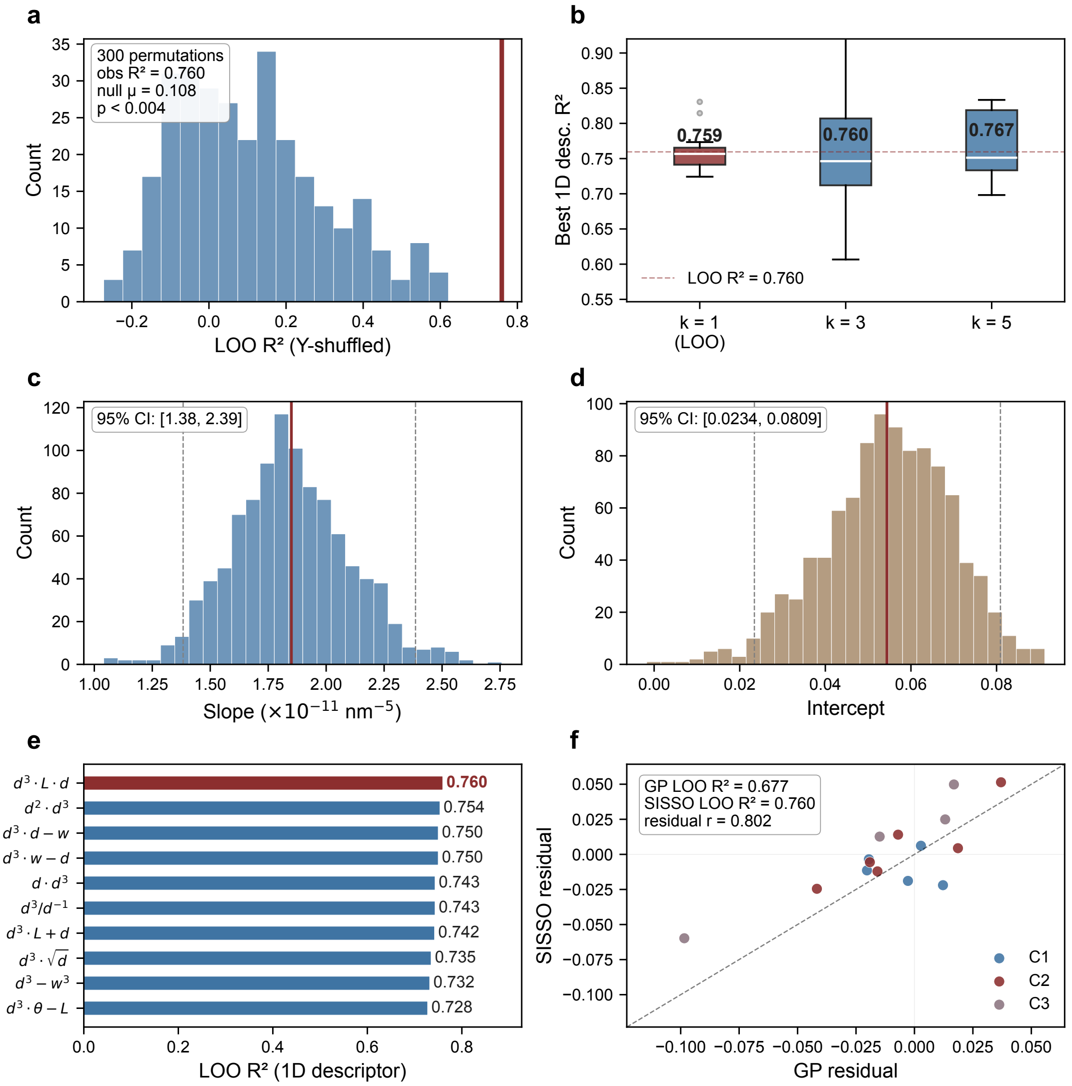
Supplementary Fig. 14. Logistic regression classification model performance for C1, C2, and C3.[2] **a,** Precision, recall, and F1 score per class from LOO cross-validation of the logistic regression model (C = 55, L1 regularization). **b,** LOO confusion matrix (normalized by row). **c,** Global weighted mean |SHAP| feature importance, averaged across the three classes with class-frequency weighting. GSH is the dominant discriminative feature, followed by $HAuCl_4$. **d,** One-versus-rest ROC curves and corresponding AUC values for each class.



Supplementary Fig. 15. SEM images of 15 representative chiral NPs from 3 clusters (C1, C2 and C3) used for the measurement of four geometric parameters. The blue boxes indicate C1 (*n* = 5), the red boxes indicate C2 (*n* = 6) and the brown boxes indicate C3 (*n* = 4). All scale bars are 500 nm.

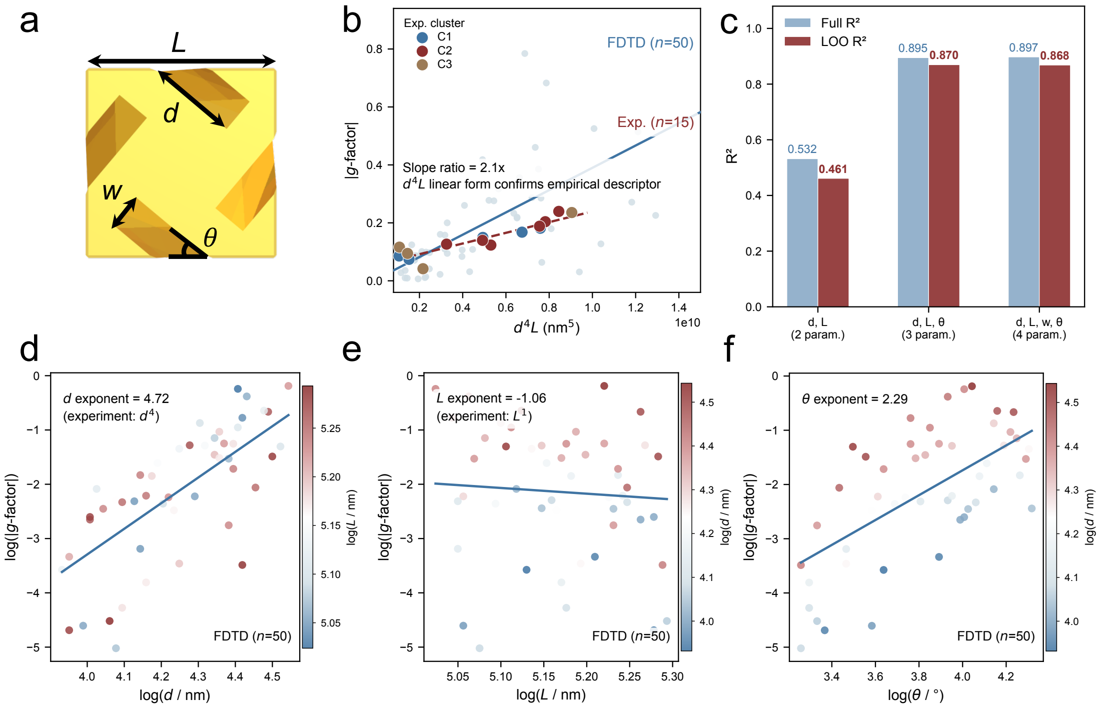


Supplementary Fig. 16. PCA of four morphological parameters (*L*, *d*, *w*, *θ*) across 15 particles from C1, C2, and C3. **a,** Scree plot. Blue bars: individual explained variance per principal component. Red curve (right axis): cumulative variance. PC1 and PC2 together account for 90.9% of total morphological variation. **b,** Loadings matrix. Red cells denote positive correlation between a morphological parameter and a principal component; blue cells denote negative correlation. PC1 is dominated by gap depth *d* (−0.56) and gap angle *θ* (+0.57) with opposite signs, defining the primary axis of gap geometry development. PC2 is dominated by particle edge length *L* (+0.66) and gap width *w* (+0.54) with the same sign, defining the particle size axis. The remaining PCs (3–4, 9.1% total) capture residual shape variation and individual particle differences.

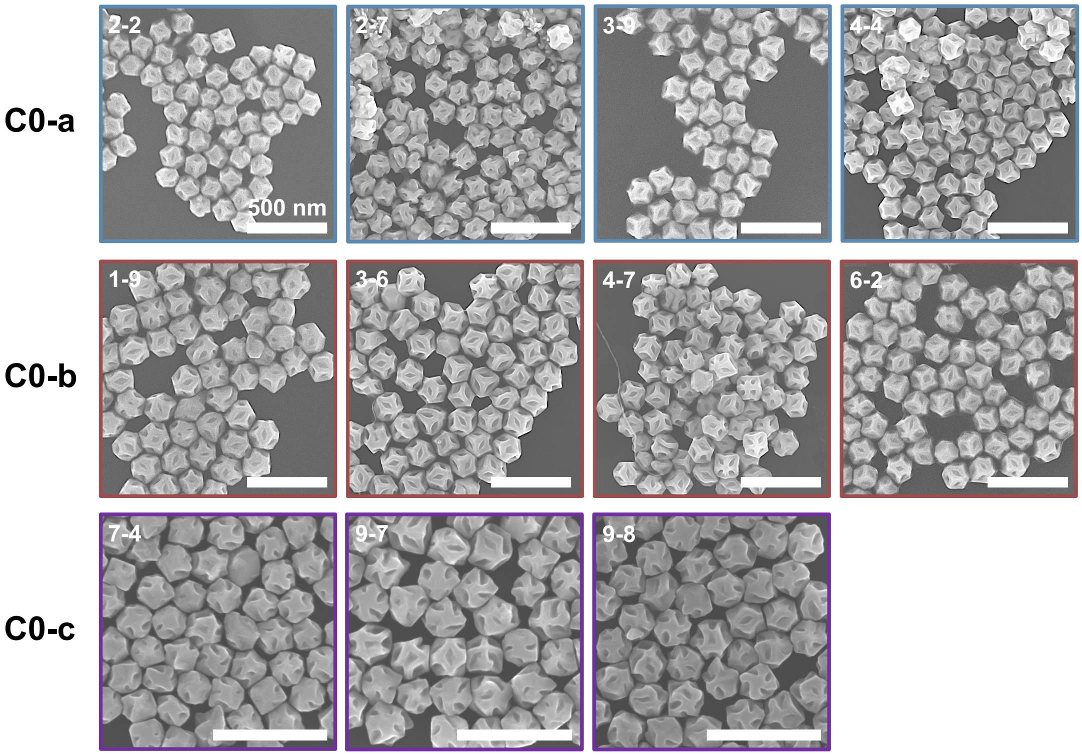


Supplementary Fig. 17. Comprehensive robustness validation of the SISSO-derived d^4L empirical descriptor.[3] **a,** Y-randomization test (300 permutations). The *g*-factor vector is randomly shuffled while preserving the structural parameter matrix; the full SISSO pipeline (SIS screening → 1D SO search) is re-executed on each shuffled dataset. The null distribution of the best LOO R^2 (blue bars, $\mu = 0.108$) is entirely separated from the observed value (red line, $R^2 = 0.760$, $p < 0.004$), ruling out the possibility that the d^4L descriptor emerged from multiple-comparison artefacts among the ~4600 candidate features. **b,** Leave-*k*-out stability. The optimal 1D descriptor R^2 does not degrade when progressively larger fractions of the data are held out during both SIS and SO stages (LOO = 0.759; L3O = 0.762; L5O = 0.767), demonstrating that the d^4L signal is carried by the dataset as a whole rather than by individual influential points. **c,** Bootstrap distribution of the d^4L slope coefficient (1000 resamples). The 95% confidence interval $[1.38, 2.39] \times 10^{-11}$ is compact and does not cross zero, confirming that the positive superlinear dependence is a robust feature of the data rather than an artefact of the particular 15-particle sample. **d,** Bootstrap distribution of the intercept. The 95% CI [0.023, 0.081] is similarly well-constrained, ruling out a zero-intercept model and corroborating the physical expectation of a finite baseline *g*-factor for particles with vanishing gap depth. **e,** Top-10 SIS candidate descriptors ranked by LOO R^2. The optimal descriptor ($d^3 \cdot L \cdot d$, equivalent to d^4L,

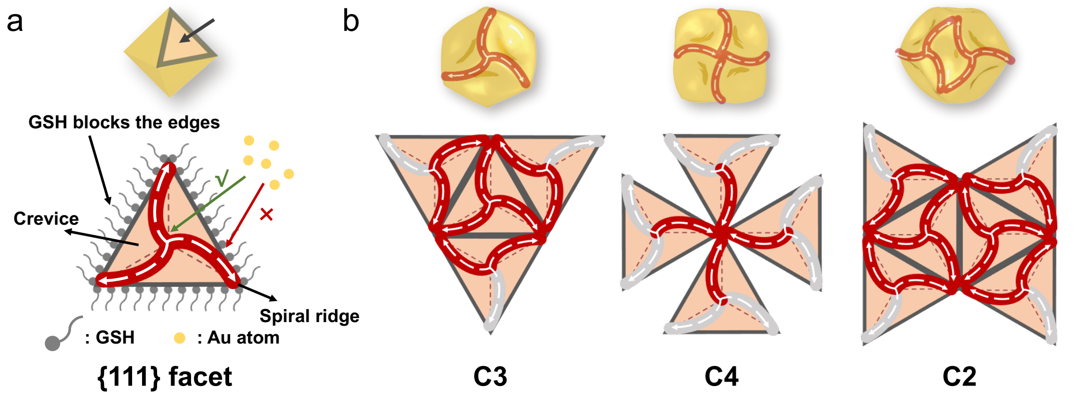
$R^2 = 0.760$, red) is surrounded by a family of highly degenerate alternatives (next-best $R^2 = 0.754$), indicating that the structure–performance signal is sufficiently strong that numerous mathematically distinct but physically equivalent functional forms capture it with comparable fidelity. **f,** Leave-one-out residual correlation between Gaussian process regression (Matérn 5/2 kernel) and SISSO d^4L on the same 15 samples. The two methods—one a non-parametric Bayesian kernel machine, the other an L0-regularised linear symbolic regression—yield a residual Pearson correlation of $r = 0.802$, meaning they systematically agree on which samples are harder or easier to predict. This consistency across fundamentally different modelling paradigms demonstrates that the remaining prediction error is not attributable to SISSO methodology but reflects genuine physical variance not captured by the four geometric parameters (L, d, w, θ). **Collectively, these six independent validation procedures establish that the d^4L scaling law is not a statistical artefact of the small sample size, a consequence of descriptor over-selection, or a peculiarity of the SISSO algorithm—it is a robust, reproducible structural descriptor of chiroptical performance in this system.**



Supplementary Fig. 18. a, The base geometry is a cube (edge length L) with all 12 edges chamfered at 45° to approximate the rounded-edge morphology of real particles. Twelve parametrised twisted slots—each a rectangular cross-section cavity (depth d, width w) rotated by angle θ about the face normal (chirality fixed at −1)—are carved into each face via CSG (Constructive Solid Geometry) Boolean subtraction. Starting from three canonical experimental configurations (C1/C2/C3), Latin Hypercube Sampling produced 50 independent structures spanning $L \in [150, 200]$ nm, $d \in [40, 95]$ nm, $w \in [25, 45]$ nm, $\theta \in [25°, 75°]$. Each structure was exported as STL and its |g-factor| computed via FDTD solution of Maxwell's equations (**Supplementary Table 3**). **b,** Cross-scale validation of the d^4L descriptor. FDTD-simulated structures (blue scatter, $n = 50$) and experimental measurements (three-colour scatter, $n = 15$, coloured by C1/C2/C3 cluster) are fitted independently with linear regressions of |g-factor| against d^4L. The FDTD fit (solid blue line, slope = 3.85×10^{-11}) and experimental fit (dashed red line, slope = 1.84×10^{-11}) share the same linear functional form, with a slope ratio of 2.1× attributable to systematic differences in dielectric function values, geometric idealisation, and measurement conditions rather than to a functional-form difference. **c,** Effect of predictor count on LOO R^2 in log(|g|) linear models $\log(|g|) \sim \alpha\cdot\log(d) + \beta\cdot\log(L) + \gamma\cdot\log(\theta) + \partial\cdot\log(w)$. The 2-parameter model ($\log(d) + \log(L)$) is insufficient to describe the FDTD chiroptical response (LOO $R^2 = 0.461$). Adding $\log(\theta)$ raises LOO R^2 to 0.870 (+0.41), confirming θ as an indispensable variable in the structure–performance mapping. Further addition of $\log(w)$ yields a slight decrease to LOO $R^2 = 0.868$ ($\Delta = -0.002$), confirming w as a redundant predictor. All log–log power-law analyses therefore employ the 3-parameter model (d, L, θ), which achieves the optimal generalisation–complexity trade-off. **d,** $\log(d)$ as primary variable, coloured by $\log(L)$: the optimal d-exponent is $\alpha = 4.72$, quantitatively consistent with the exponent of 4.0 in the experimental d^4L descriptor and ruling out linear ($\alpha = 1$) or quadratic ($\alpha = 2$) dependencies. **e,** $\log(L)$ as primary variable, coloured by $\log(d)$: the partial regression coefficient is $\beta = -1.06$. Because d and L are not independent in the FDTD parameter space (larger particles tend toward deeper slots), this partial effect does not correspond to the L exponent in the d^4L product descriptor (which is + 1). **f,** log(θ) as primary variable, coloured by log(d): the θ-exponent is $\gamma = 2.29$, confirming a superlinear contribution of twist angle. The absence of θ from the experimental SISSO descriptor is consistent with the d–θ anti-correlation in the experimental parameter space, which allows d's high-order terms to absorb θ's variance. Across all three panels, the d-exponent of ~4.7 is insensitive to model complexity (panel c) and provides independent numerical validation of the power-law form adopted in the experimental d^4L descriptor.



Supplementary Fig. 19. SEM images of representative chiral NPs from C0 (C0-a, C0-b and C0-c) that can be identified as a novel chiral rhombic dodecahedron. All scale bars are 500 nm.



Supplementary Fig. 20. Growth path of the achiral octahedral seed into a chiral rhombic dodecahedron, along with schematic diagrams of the plane configurations of the various chiral centres. **a,** In the early stages of growth, the gold octahedron is entirely covered by CTAB; the 12 high-curvature edges have fewer ligands compared to the eight {111} faces, the thiol group of GSH tends to anchor preferentially on the edges rather than the {111} faces; this blocks the subsequent deposition of gold atoms, leading to preferential deposition on the {111} faces and the formation of three clockwise-rotating ridges. **b,** When the various faces of a chiral rhombic dodecahedron are unfolded onto a plane, a significant observation is that if the C3 rotation axes are oriented clockwise, their four C3 rotation axes will intersect at a single point, forming a C4 rotation axis that must necessarily be counter-clockwise; this indicates that the structure of this chiral rhombic dodecahedron naturally exhibits racemisation. **This provides a structural explanation for the phenomenon of chiral inversion.**

Supplementary Table 1 Chemical space of 432 helicoid III

Reagents	Concentration range
CTAB	10-30 mM
Au	0.10-0.40 mM
AA	5.0-15 mM
GSH	1.0-10 μM

Supplementary Table 2 Complete formula and |*g*-factor| for 80 data points

Batch	Well	CTAB (mM)	Au (mM)	AA (mM)	GSH (mM)	|*g*-factor|
1	2	14.386912	0.267475	10.618938	0.003123	0.14259
1	3	18.260235	0.334969	8.637146	0.009702	0.10105
1	4	29.53271	0.190743	11.944931	0.002343	0.02647
1	5	11.777981	0.21919	14.233681	0.001422	0.00819
1	6	19.585796	0.18879	5.30593	0.005234	0.01962
1	7	29.751007	0.368227	6.214054	0.007143	0.01379
1	8	13.962822	0.114453	7.319064	0.008914	0.00583
1	9	28.234304	0.23352	11.428969	0.007146	0.00395
2	2	24.229244	0.125381	13.07571	0.006276	0.00192
2	3	22.456345	0.30084	10.165047	0.008117	0.02316
2	4	26.157228	0.4	10.361018	0.004037	0.0869
2	5	15.960346	0.157739	9.671894	0.007565	0.00483
2	6	10.185624	0.171617	7.98742	0.003366	0.04499
2	7	25.198912	0.13917	14.931929	0.009354	0.00215
2	8	21.200634	0.31981	13.800185	0.002035	0.04614
2	9	10.998981	0.280687	12.343538	0.004563	0.15815
3	2	17.312541	0.351583	9.136357	0.003611	0.15059
3	3	12.817019	0.10129	13.60326	0.00103	0.00153
3	4	21.739428	0.381539	7.02957	0.005928	0.12646
3	5	27.156643	0.250062	12.559454	0.002734	0.08563
3	6	25.842171	0.206286	8.969851	0.008508	0.00336
3	7	17.782202	0.311357	8.195677	0.006691	0.18453
3	8	23.240148	0.341049	14.545627	0.004783	0.20393
3	9	16.550536	0.16429	5.710887	0.009173	0.00305
4	2	15.246807	0.244599	6.727258	0.004293	0.15884
4	3	27.754726	0.28787	13.260798	0.005625	0.06679
4	4	18.881014	0.128394	10.827485	0.007934	0.00398
4	5	18.616789	0.388778	7.614081	0.004994	0.14347
4	6	20.207115	0.147544	5	0.001823	0.00085
4	7	22.783295	0.226726	12.21199	0.01	0.0053
4	8	26.697591	0.259701	5.44417	0.006876	0.00105
4	9	14.841342	0.346728	11.126955	0.005698	0.23559
5	2	23.629824	0.18003	12.799918	0.002464	0.0741
5	3	20.722483	0.356521	10.493836	0.003807	0.14411
5	4	19.756943	0.327082	11.679464	0.005429	0.20638
5	5	12.181577	0.37533	9.830433	0.00288	0.08122
5	6	13.46935	0.130977	9.543899	0.006458	0.00387
5	7	15.574217	0.198474	9.307595	0.001586	0.00803
5	8	25.5244	0.16785	6.388588	0.008735	0.00143
5	9	28.78301	0.360358	14.195902	0.00906	0.0152
6	2	24.800045	0.215106	8.540798	0.006084	0.00423

6	3	12.976096	0.371094	5.958953	0.007746	0.09158
6	4	29.158638	0.117623	7.871654	0.00221	0.00278
6	5	20.45402	0.307672	11.290313	0.008332	0.09922
6	6	14.600508	0.273635	11.020296	0.005793	0.24000
6	7	16.929473	0.331601	6.28677	0.007392	0.12327
6	8	10.642766	0.316782	8.831228	0.003009	0.09299
6	9	21.969938	0.295073	11.504072	0.004166	0.17622
7	2	12.485008	0.236907	9.968219	0.003478	0.16786
7	3	14.159187	0.255044	6.568134	0.005342	0.18142
7	4	19.096849	0.153139	7.471007	0.004425	0.02804
7	5	30	0.175335	8.340229	0.005108	0.00118
7	6	16.144636	0.106846	9.400242	0.001641	0.00471
7	7	11.973752	0.144028	14.820657	0.00981	0.00505
7	8	28.962764	0.37792	13.991194	0.009273	0.01543
7	9	11.359476	0.239598	10.110768	0.009562	0.09978
8	2	27.406518	0.278469	12.085642	0.007038	0.01028
8	3	23.962132	0.323947	11.888529	0.004879	0.13966
8	4	18.046782	0.209793	13.385621	0.004932	0.11696
8	5	28.49593	0.364348	5.158616	0.001619	0.00047
8	6	13.627653	0.282111	11.781183	0.001278	0.00468
8	7	24.492054	0.393098	13.48325	0.006125	0.09447
8	8	15.033428	0.252413	6.484268	0.00467	0.14967
8	9	21.463613	0.263877	14.731789	0.005643	0.1447
9	2	10	0.396716	12.479392	0.007253	0.04171
9	3	10.46946	0.134645	6.870562	0.00387	0.02479
9	4	11.506174	0.230299	6.065136	0.006376	0.18832
9	5	12.363319	0.195563	10.716364	0.001142	0.00334
9	6	23.482895	0.121276	14.427805	0.002269	0.01307
9	7	22.090374	0.222439	11.193609	0.005523	0.05009
9	8	14.777369	0.184872	5.838168	0.006509	0.04109
9	9	16.36553	0.291179	8.7325	0.004387	0.17948
10	2	19.433201	0.362729	5.572227	0.007665	0.07692
10	3	26.442986	0.169818	11.589325	0.008973	0.00162
10	4	17.567311	0.385187	6.11557	0.006803	0.11
10	5	19.95926	0.111154	13.285738	0.002573	0.01181
10	6	11.573731	0.303811	7.711353	0.0099	0.11614
10	7	18.153799	0.262345	7.170869	0.002644	0.01951
10	8	28.040096	0.270657	14.061792	0.005148	0.06823
10	9	23.010865	0.160046	5.048018	0.001749	0.00142

Supplementary Table 3 Table of 50 geometric parameters and |g-factor| for FDTD simulations

Name	*L* (nm)	*d* (nm)	*w* (nm)	*θ* (°)	|*g*-factor|
LHS_001	161	77	41	70	0.3171
LHS_002	179	78	35	68	0.3561
LHS_003	157	73	42	37	0.1083
LHS_004	169	51	28	38	0.028
LHS_005	183	52	45	49	0.0356
LHS_006	177	81	38	43	0.2854
LHS_007	167	56	42	63	0.1242
LHS_008	163	67	32	40	0.0946
LHS_009	185	79	42	49	0.2868
LHS_010	174	59	35	51	0.101
LHS_011	193	55	39	54	0.0706
LHS_012	172	57	40	75	0.087
LHS_013	177	60	30	27	0.0139
LHS_014	175	69	31	36	0.0838
LHS_015	181	77	36	46	0.2331
LHS_016	155	83	26	43	0.4605
LHS_017	172	75	43	52	0.276
LHS_018	159	80	35	73	0.2168
LHS_019	186	68	27	42	0.1065
LHS_020	176	64	27	27	0.0222
LHS_021	193	89	28	69	0.5142
LHS_022	160	59	37	26	0.0066
LHS_023	188	63	29	71	0.1603
LHS_024	180	70	43	32	0.0314
LHS_025	168	88	41	64	0.5254
LHS_026	190	86	40	31	0.1274
LHS_027	188	81	26	38	0.1798
LHS_028	157	54	40	36	0.01
LHS_029	199	58	29	28	0.0109
LHS_030	197	90	41	35	0.2257
LHS_031	170	70	27	74	0.2602
LHS_032	164	74	26	63	0.4182
LHS_033	185	94	30	57	0.8282
LHS_034	173	66	41	61	0.2001
LHS_035	189	64	44	52	0.1096
LHS_036	178	65	44	59	0.1576
LHS_037	176	78	40	44	0.2195
LHS_038	196	55	39	55	0.0741
LHS_039	187	57	31	56	0.0861
LHS_040	156	63	27	32	0.0412

LHS_041	194	72	34	66	0.2777
LHS_042	158	84	37	51	0.6827
LHS_043	166	82	42	47	0.3856
LHS_044	196	52	30	29	0.0092
LHS_045	152	82	28	55	0.7844
LHS_046	198	83	44	26	0.0306
LHS_047	165	92	25	33	0.2714
LHS_048	190	60	43	58	0.0973
LHS_049	156	62	41	45	0.099
LHS_050	187	80	33	28	0.0636

References

1. Beckers, T. An Introduction to Gaussian Process Models. Preprint at https://doi.org/10.48550/arXiv.2102.05497 (2021).
2. Starbuck, C. Logistic Regression. in *The Fundamentals of People Analytics* 223–238 (Springer International Publishing, Cham, 2023).
3. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* **2**, 083802 (2018).