Electronic Supplementary Information

SERS multiplexing of methylxanthine drug isomers via host-guest size matching and machine learning

Weng-I Katherine Chio,[a,b,c] Jia Liu,[a,b] Tabitha Jones,[a,b] Jayakumar Perumal,[c] Dinish U. S,[c] Ivan P. Parkin,[b] Malini Olivo,[c] and Tung-Chun Lee*[,a,b]

[a] Institute for Materials Discovery, University College London (UCL), London WC1H 0AJ, U.K.

[b] Department of Chemistry, University College London (UCL), London WC1H 0AJ, U.K.

[c] Translational Biophotonics Laboratory, Institute of Bioengineering and Bioimaging (IBB), Agency for Science Technology and Research (A*STAR), Singapore 138667, Singapore

Table of Contents

1. State-of-the-art SERS detection of MeX

Table S1. SERS detection of MeX reported in literature.

Substrate	MeX (Limit of detection)	Ref.
Ag NPs	CAF (1 µM), TBR (1 µM), PRX (1 µM)	1
Ag NPs	CAF (1.4 µM), TBR (280 nM), TPH (1.4 µM)	19
Ag NPs	CAF (250 µM)	20
Ag NPs	CAF (5.7 mM)	21
Ag NPs	CAF (5.7 mM)	25
Ag NPs	CAF (1 mM)	26
Au NP: CB7 nanoaggregates	TBR (500 nM), TPH (50 nM), CAF (5 µM)	This study
Au NP: CB8 nanoaggregates	TBR (50 nM), TPH (100 nM), CAF (1 µM)	

**PRX = paraxanthine (1,7-dimethylxanthine)*

2. NMR and energy-minimised model of [CB7-TBR-H]$^+$

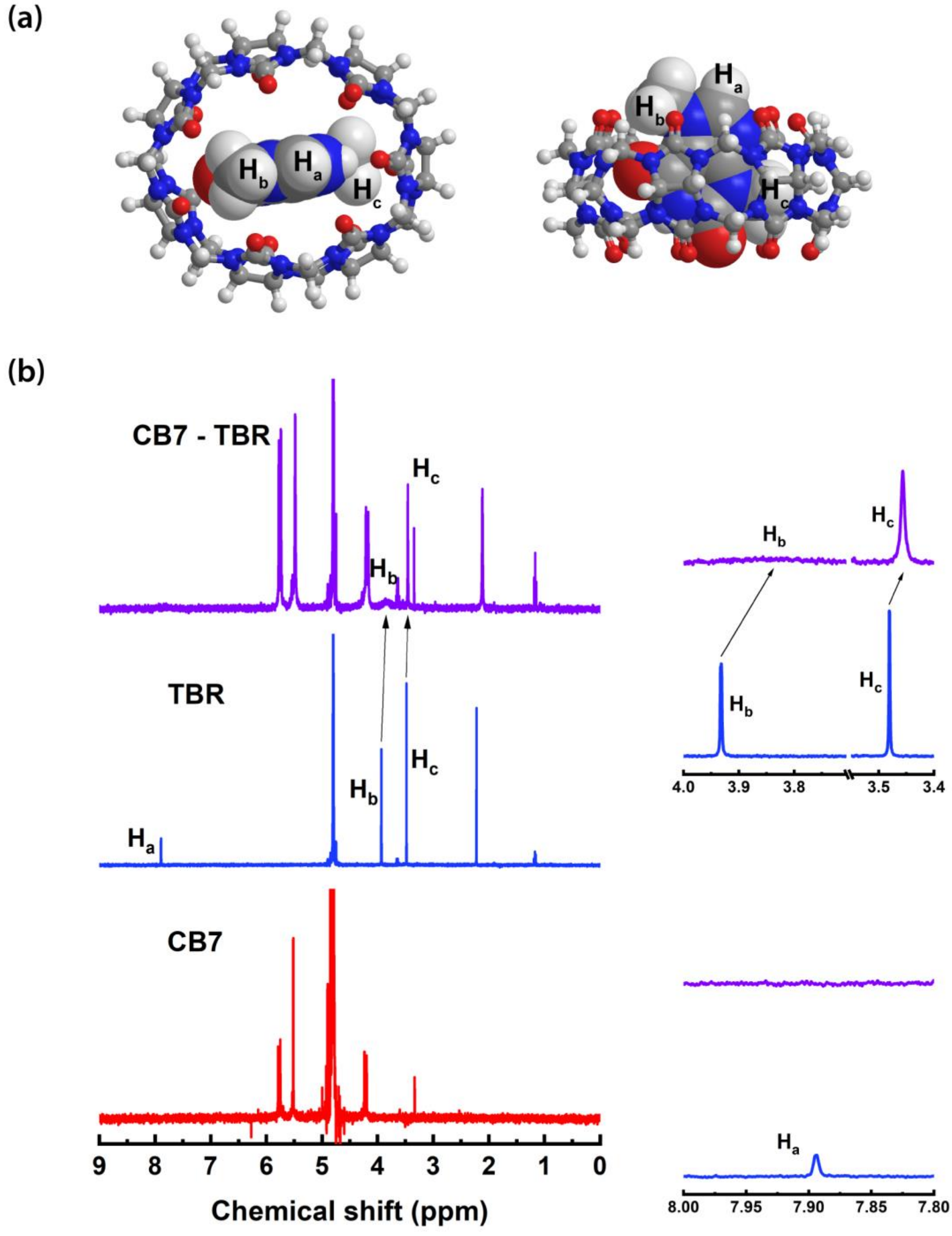


Figure S1. (a) Energy-minimised molecular model of a [CB7-TBR-H]$^+$ host-guest complex at CPCM/wB97X-D/6-31G* level of theory. CPCM implicit water model was used to approximate the solvent effects. (b) ^{1}H NMR spectra of CB7, TBR and 1:1 CB7-TBR host-guest complex in D_2O. Inset: Zoom-in NMR spectra.

3. NMR and energy-minimised model of $[CB8\text{-}TBR\text{-}H]^+$

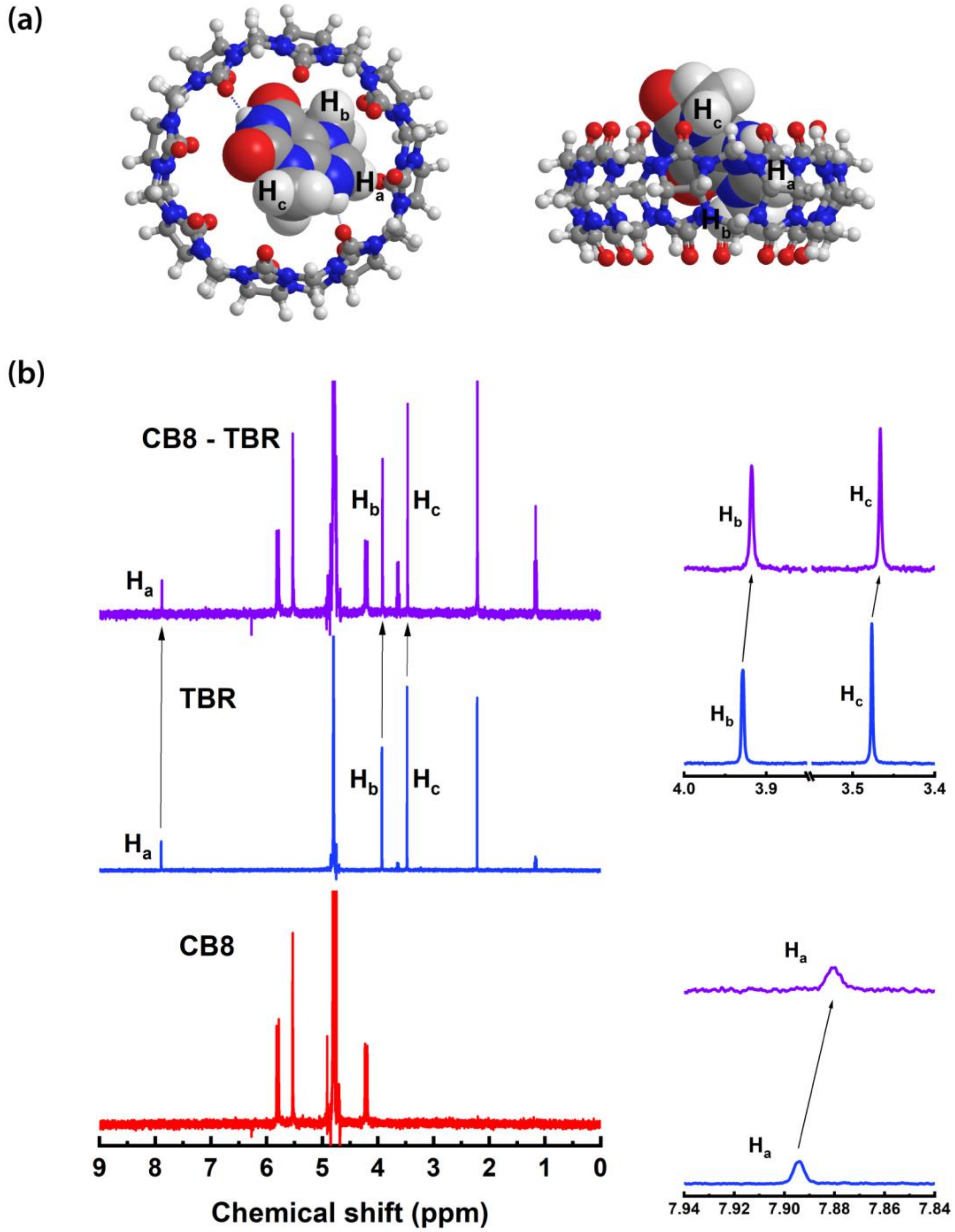


Figure S2. (a) Energy-minimised molecular model of a $[CB8\text{-}TBR\text{-}H]^+$ host-guest complex at CPCM/wB97X-D/6-31G* level of theory. CPCM implicit water model was used to approximate the solvent effects. (b) ^1H NMR spectra of CB8, TBR and 1:1 CB8-TBR host-guest complex in DCl. Inset: Zoom-in NMR spectra.

4. NMR and energy-minimised model of [CB7-TPH-H]$^+$

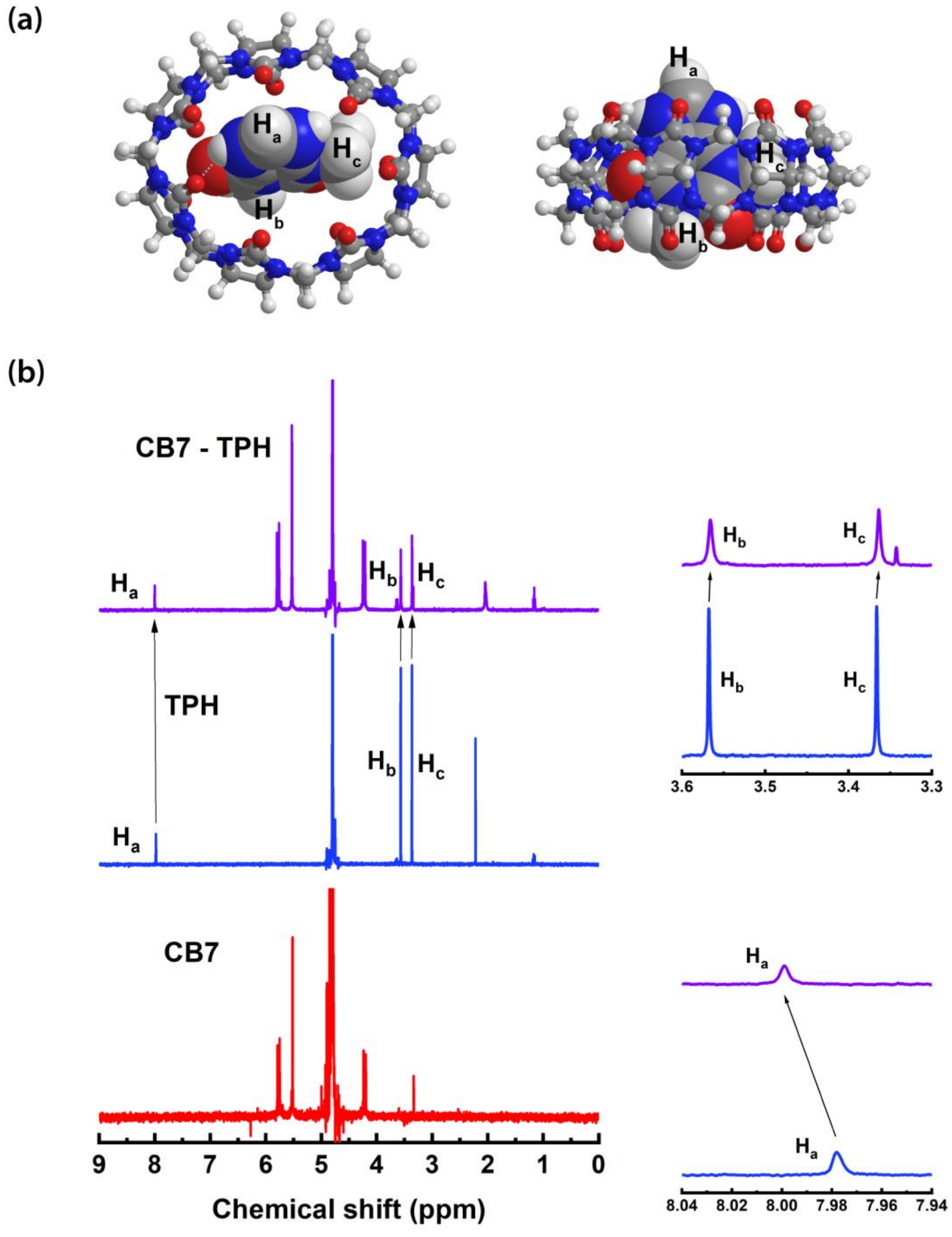


Figure S3. (a) Energy-minimised molecular model of a [CB7-TPH-H]$^+$ host-guest complex at CPCM/wB97X-D/6-31G* level of theory. CPCM implicit water model was used to approximate the solvent effects. (b) ^{1}H NMR spectra of CB7, TPH and 1:1 CB7-TPH host-guest complex in D_2O. Inset: Zoom-in NMR spectra.

5. NMR and energy-minimised model of $[CB8\text{-}TPH\text{-}H]^+$

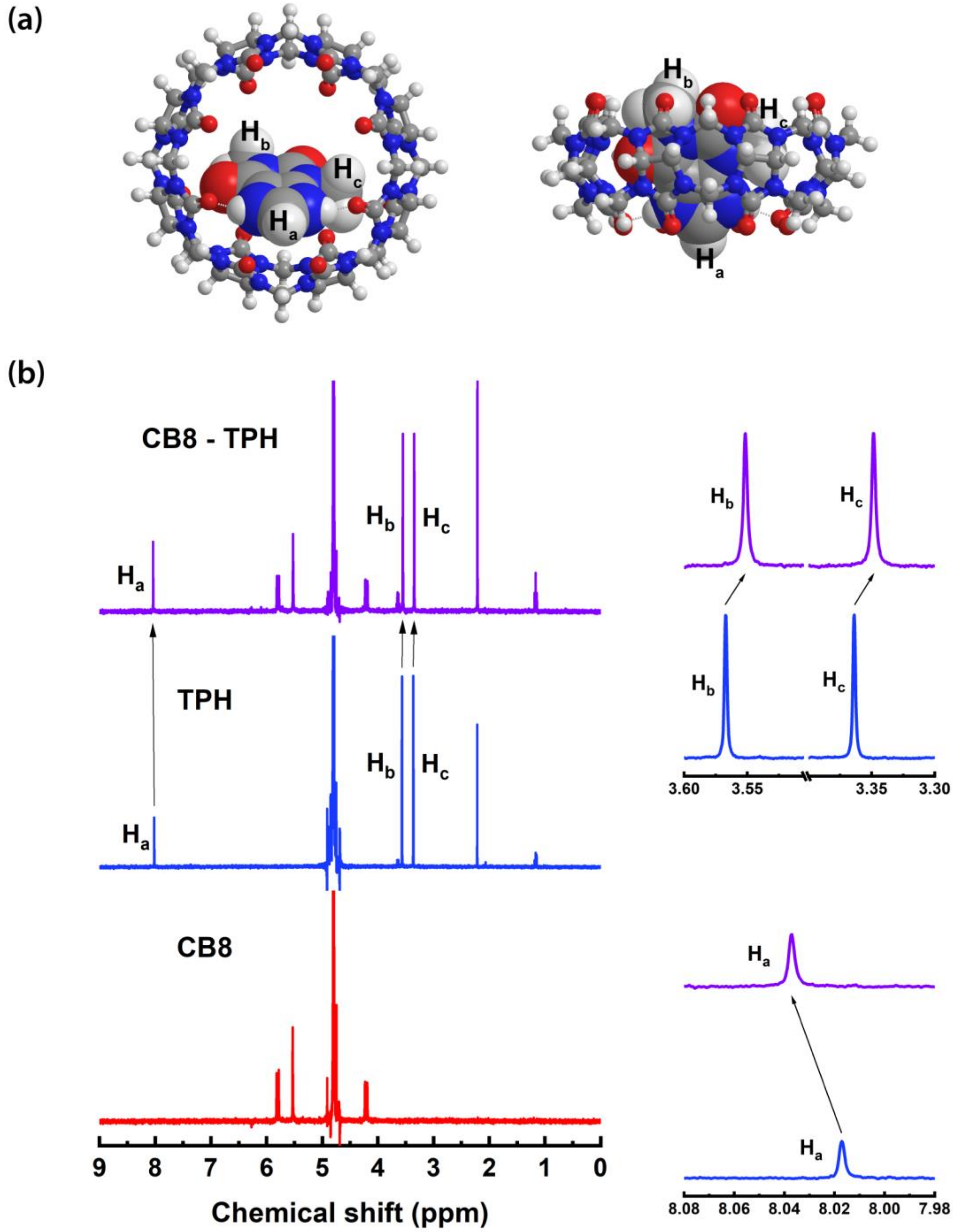


Figure S4. (a) Energy-minimised molecular model of a $[CB8\text{-}TPH\text{-}H]^+$ host-guest complex at CPCM/wB97X-D/6-31G* level of theory. CPCM implicit water model was used to approximate the solvent effects. (b) ^1H NMR spectra of CB8, TPH and 1:1 CB8-TPH host-guest complex in DCl. Inset: Zoom-in NMR spectra.

6. NMR and energy-minimised model of $[CB7\text{-}CAF\text{-}H]^+$

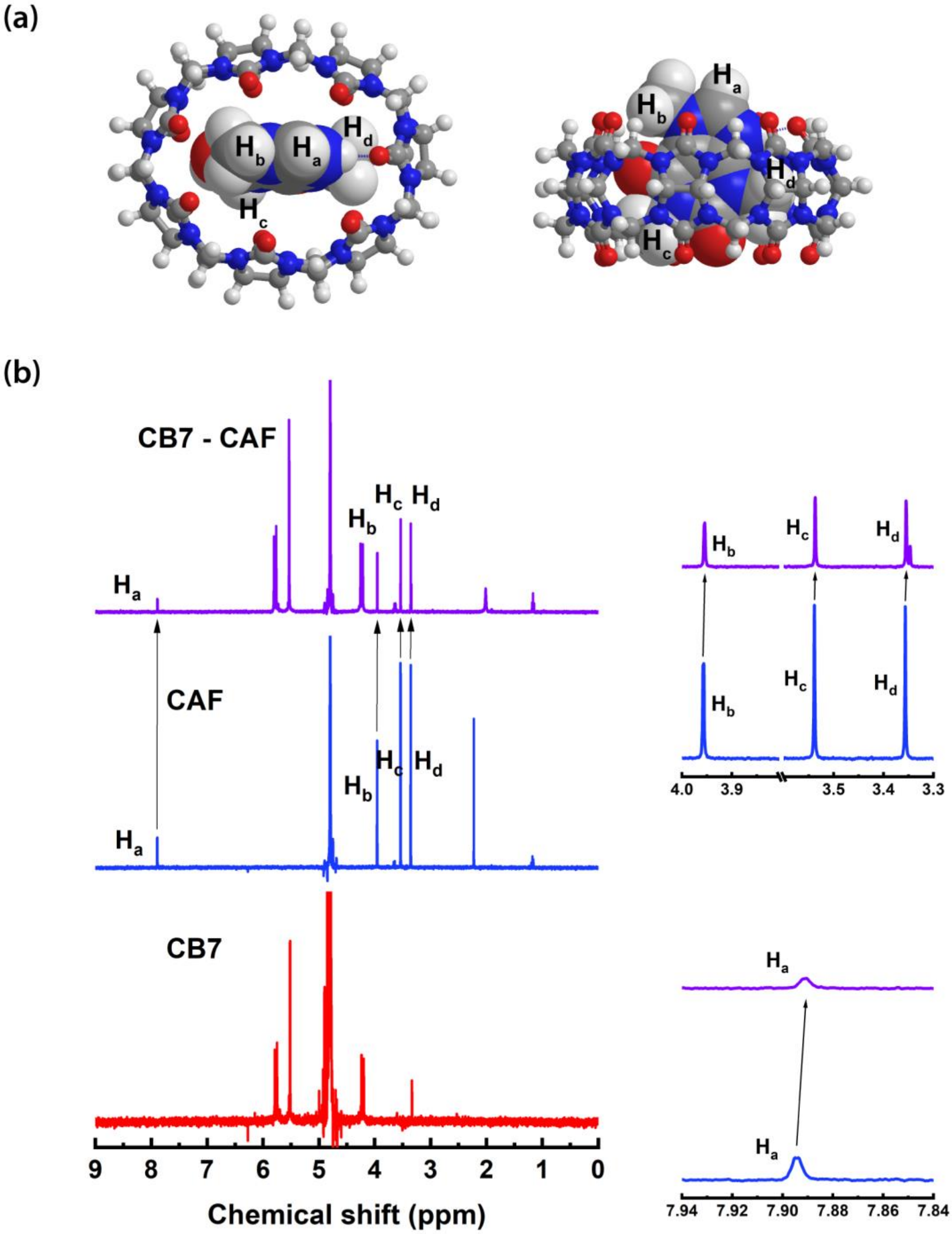


Figure S5. (a) Energy-minimised molecular model of a $[CB7\text{-}CAF\text{-}H]^+$ host-guest complex at CPCM/wB97X-D/6-31G* level of theory. CPCM implicit water model was used to approximate the solvent effects. (b) ^1H NMR spectra of CB7, CAF and 1:1 CB7-CAF host-guest complex in D_2O. Inset: Zoom-in NMR spectra.

7. NMR and energy-minimised model of [CB8-CAF-H]$^+$

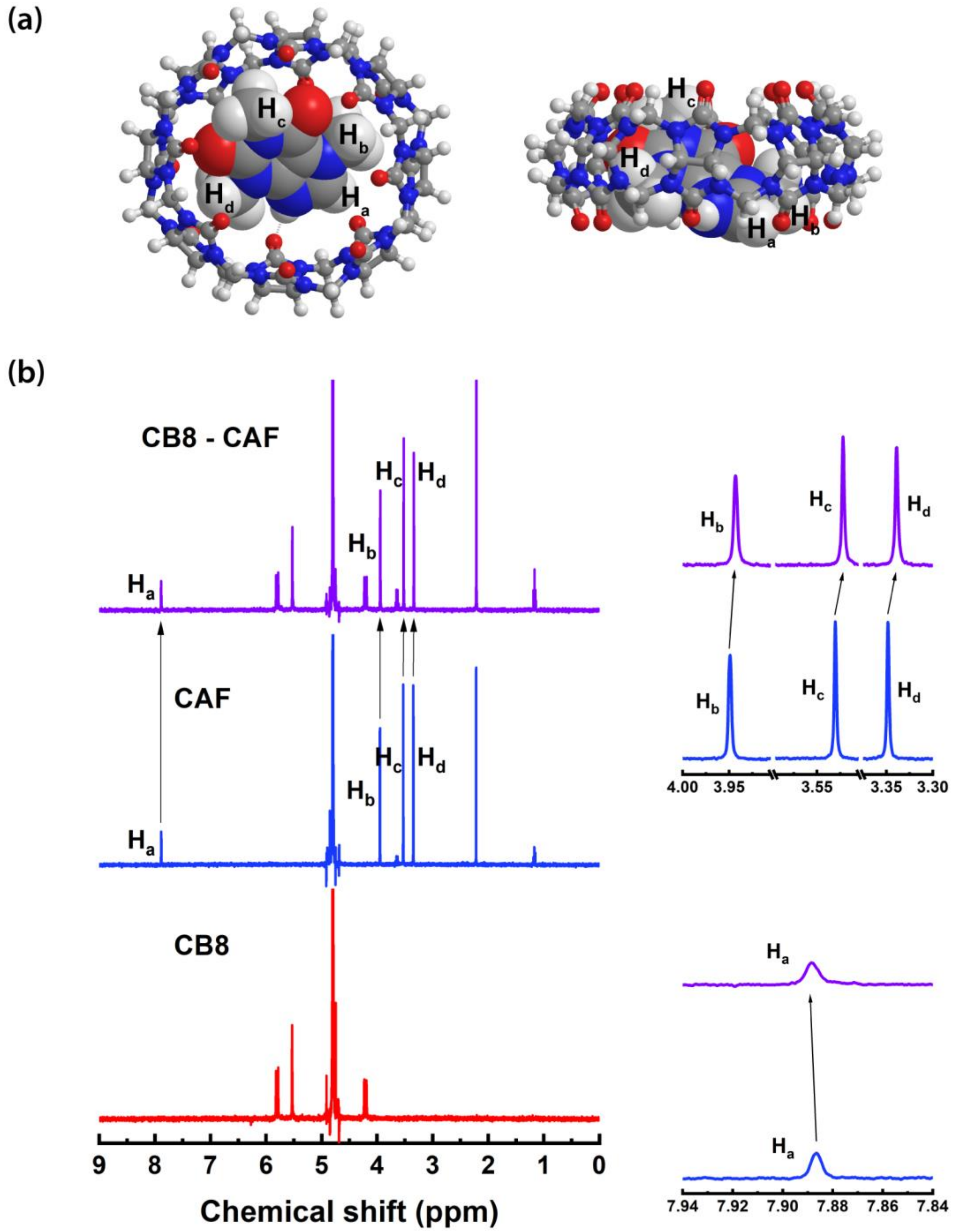
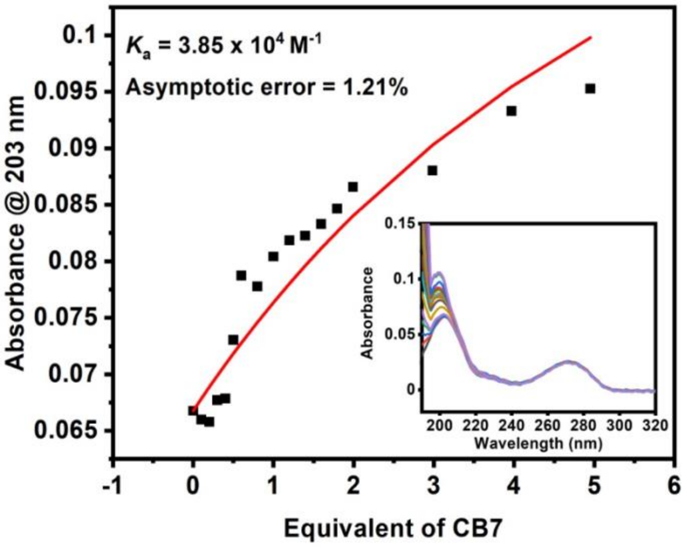


Figure S6. (a) Energy-minimised molecular model of a [CB8-CAF-H]$^+$ host-guest complex at CPCM/wB97X-D/6-31G* level of theory. CPCM implicit water model was used to approximate the solvent effects. (b) ^{1}H NMR spectra of CB8, CAF and 1:1 CB8-CAF host-guest complex in DCl. Inset: Zoom-in NMR spectra.

8. UV-Vis titration of $[CB\text{-}TPH\text{-}H]^+$

(a)



(b)

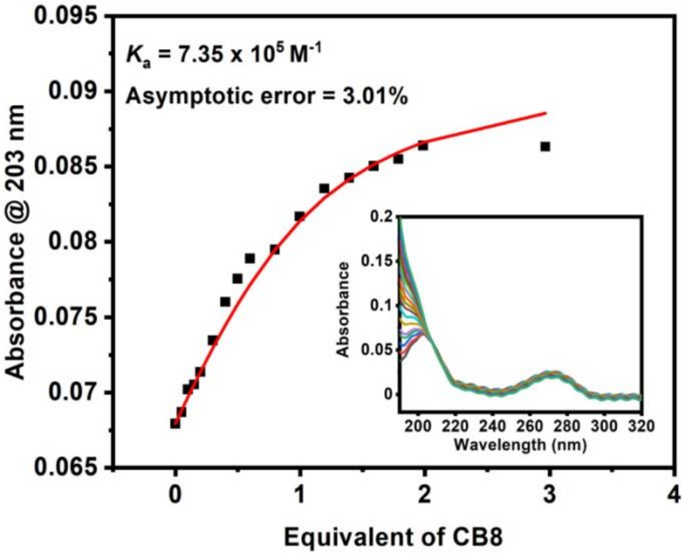


Figure S7. UV-Vis titration of 4 µM TPH with (a) CB7 and (b) CB8. Insets: UV-Vis spectra of TPH upon stepwise addition of CB7 or CB8. The binding curves were fitted by assuming 1:1 binding model from which the binding constants were derived.

9. UV-Vis titration of $[CB\text{-}CAF\text{-}H]^+$

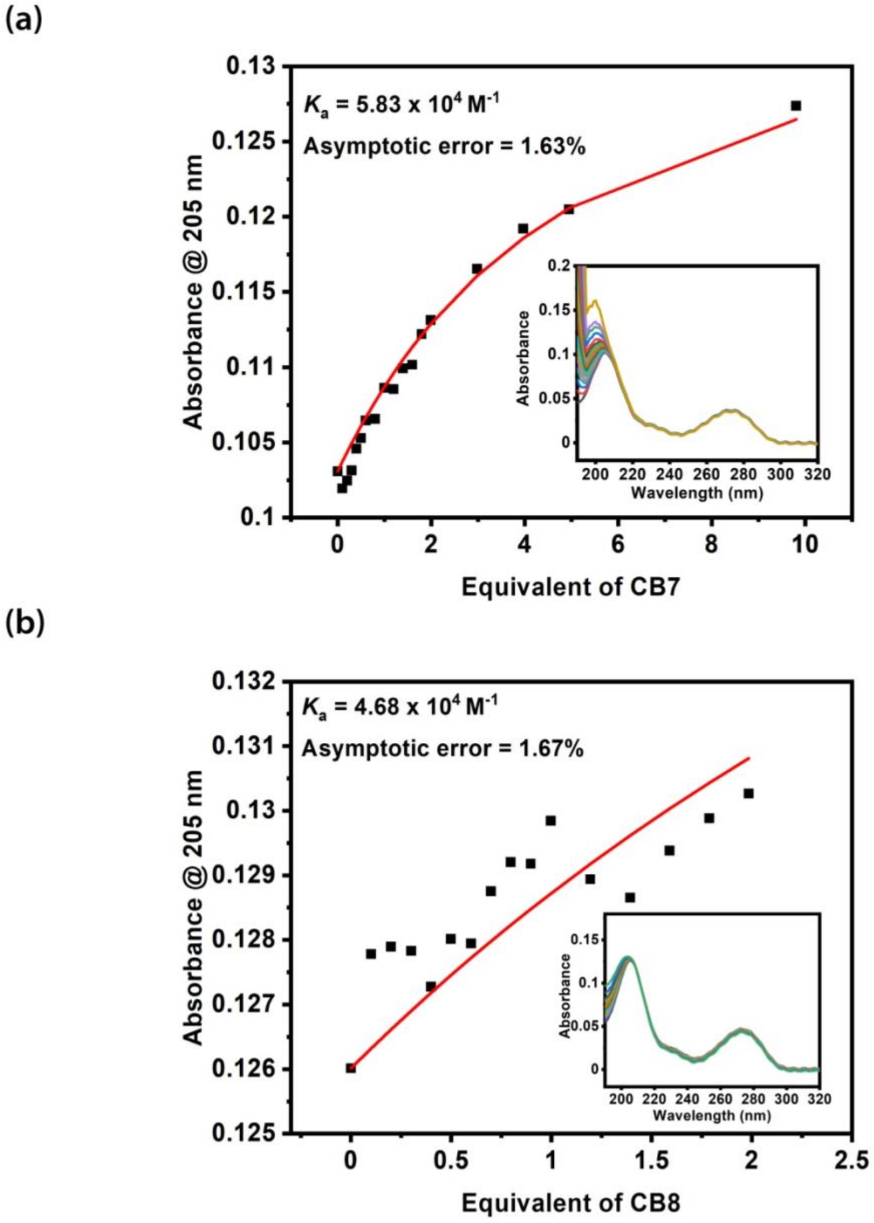


Figure S8. UV-Vis titration of 4 µM CAF with (a) CB7 and (b) CB8. Insets: UV-Vis spectra of CAF upon stepwise addition of CB7 or CB8. The binding curves were fitted by assuming 1:1 binding model from which the binding constants were derived.

10. Binding parameters of the $[CB\text{-}MeX\text{-}H]^+$ inclusion complexes

Table S2. Packing coefficients of the $[CB\text{-}MeX\text{-}H]^+$ inclusion complexes.

Structure	**Inner cavity volume / Å^3**	**Molecular volume / Å^3**	**Packing coefficient**	**Binding constant / M^{-1}**
CB7	242[1]			
CB8	367[1]			
$[TBR\text{-}H]^+$		168.86		
$[TPH\text{-}H]^+$		168.23		
$[CAF\text{-}H]^+$		188.11		
$[CB7\text{-}TBR\text{-}H]^+$			0.70	2.08×10^4
$[CB7\text{-}TPH\text{-}H]^+$			0.70	3.85×10^4
$[CB7\text{-}CAF\text{-}H]^+$			0.78	5.83×10^4
$[CB8\text{-}TBR\text{-}H]^+$			0.46	1.05×10^5
$[CB8\text{-}TPH\text{-}H]^+$			0.46	7.35×10^5
$[CB8\text{-}CAF\text{-}H]^+$			0.51	4.68×10^4

11. Binding energies of the [CB-MeX-H]$^+$ inclusion complexes

Table S3. Binding energies, in kcal mol^{-1}, of the [CB-MeX-H]$^+$ inclusion complexes in gas phase, optimised at the wB97X-D/6-31G* level of theory.

Structure	Energy / Hartree	Energy / kcal mol^{-1}	Binding energy / kcal mol^{-1}
CB7	-4211.1343	-2642524.6517	
CB8	-4812.7175	-3020023.5319	
[CAF-H]$^+$	-680.5313	-427039.5438	
[TBR-H]$^+$	-641.2288	-402376.8512	
[TPH-H]$^+$	-641.2215	-402372.2503	
[CB7-CAF-H]$^+$	-4891.8067	-3069652.7048	-88.5093
[CB7-TBR-H]$^+$	-4852.5047	-3044990.3687	-88.8658
[CB7-TPH-H]$^+$	-4852.4856	-3044978.3738	-81.4718
[CB8-CAF-H]$^+$	-5493.3988	-3447157.1832	-94.1075
[CB8-TBR-H]$^+$	-5454.0728	-3422479.7843	-79.4012
[CB8-TPH-H]$^+$	-5454.0750	-3422481.1624	-85.3801

Table S4. Binding energies, in kcal mol^{-1}, of the [CB-MeX-H]$^+$ inclusion complexes in water, optimised at the CPCM/wB97X-D/6-31G* level of theory.

Structure	**Energy / Hartree**	**Energy / kcal mol^{-1}**	**Binding energy / kcal mol^{-1}**
CB7	-4211.2727	-2642611.4972	
CB8	-4812.8780	-3020124.2609	
[TBR-H]$^+$	-641.3206	-402434.4394	
[TPH-H]$^+$	-641.3159	-402431.4858	
[CAF-H]$^+$	-680.6192	-427094.6582	
[CB7-TBR-H]$^+$	-4852.6554	-3045084.9433	-39.0066
[CB7-TPH-H]$^+$	-4852.6576	-3045086.2958	-43.3128
[CB7-CAF-H]$^+$	-4891.9562	-3069746.5395	-40.3841
[CB8-TBR-H]$^+$	-5454.2537	-3422593.2585	-34.5581
[CB8-TPH-H]$^+$	-5454.2600	-3422597.2184	-41.4718
[CB8-CAF-H]$^+$	-5493.5596	-3447258.0938	-39.1747

12. Raman spectra of MeX, and SERS spectra of CB*n*

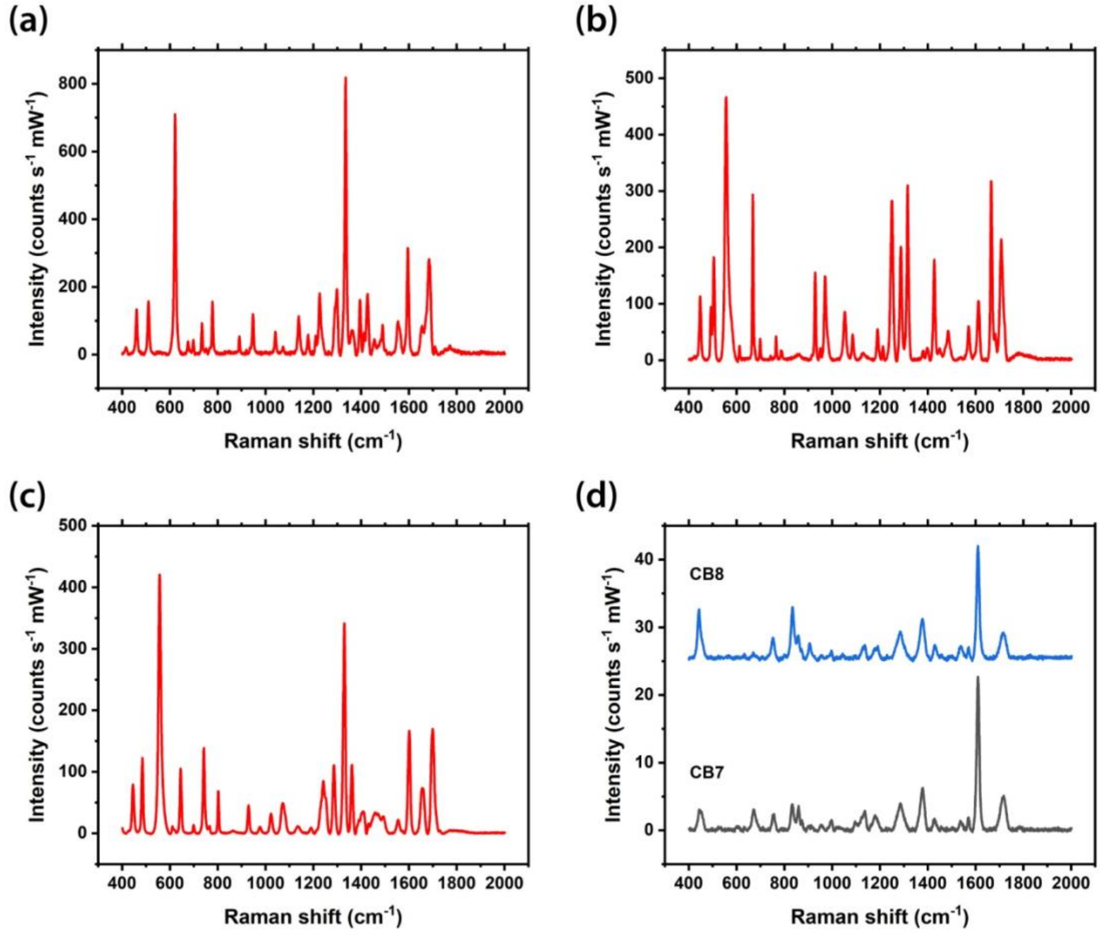


Figure S9. (a-c) Raman spectra of (a) TBR, (b) TPH and (c) CAF powder respectively. (d) SERS spectra of 10 µM CB7 and 5 µM CB8.

13. SERS spectra of CB7-TBR complexes

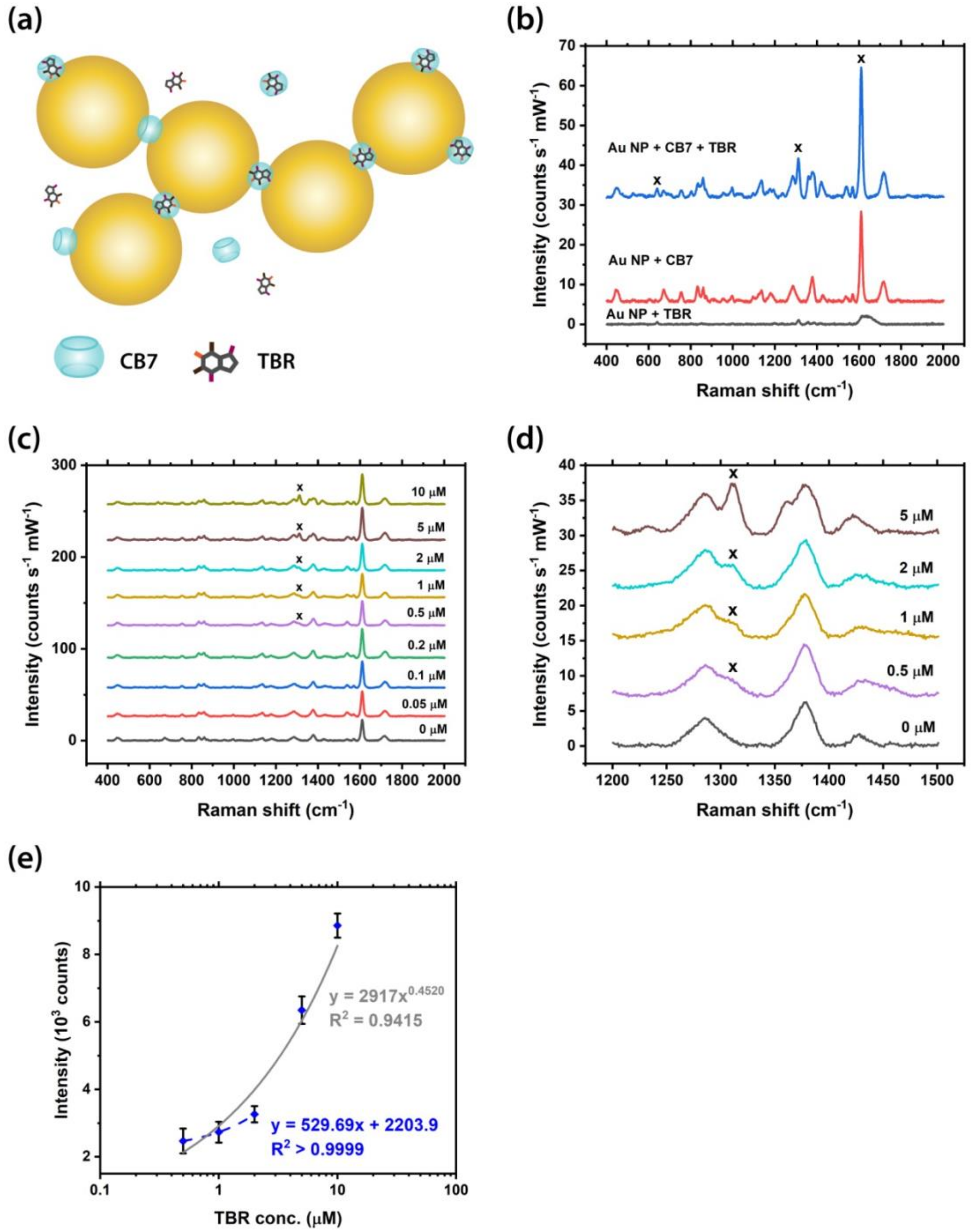


Figure S10. (a) Schematic illustration of the precise plasmonic hotspots within Au NP: CB7 nanoaggregates for TBR detection (not to scale). (b) SERS spectra of TBR in the presence or absence of CB7. (c) Full-range and (d) zoom-in SERS spectra of TBR with different concentrations from 0 to 10 µM. Main Raman peak of TBR at 1312 cm^{-1} is marked by x. Spectra were baseline corrected and offset for clarity. (e) Corresponding plot of SERS intensity of the main TBR peak (marked by x in (d)) against TBR concentration. (Note: x-axis is plotted in log-scale to even out the spread of the data points for better illustration. The linear region at low concentration had been identified and fitted linearly, while the full range fitted well by power law)

14. SERS spectra of CB7-TPH complexes

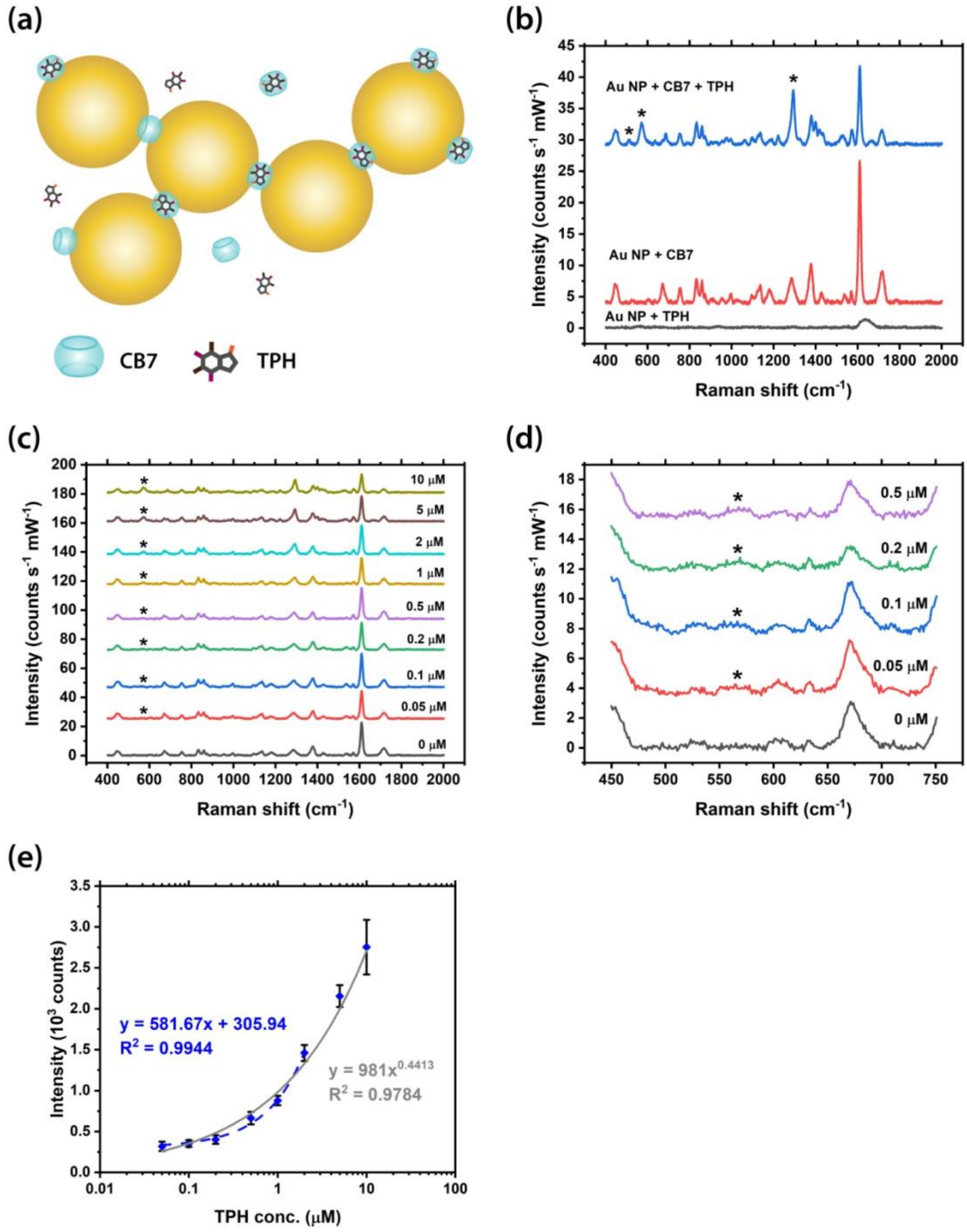


Figure S11. (a) Schematic illustration of the precise plasmonic hotspots within Au NP: CB7 nanoaggregates for TPH detection (not to scale). (b) SERS spectra of TPH in the presence or absence of CB7. (c) Full-range and (d) zoom-in SERS spectra of TPH with different concentrations from 0 to 10 µM. Main Raman peak of TPH at 557 cm^{-1} is marked by *. Spectra were baseline corrected and offset for clarity. (e) Corresponding plot of SERS intensity of the main TPH peak (marked by * in (d)) against TPH concentration. (Note: x-axis is plotted in log-scale to even out the spread of the data points for better illustration. The linear region at low concentration had been identified and fitted linearly, while the full range fitted well by power law)

15. SERS spectra of CB7-CAF complexes

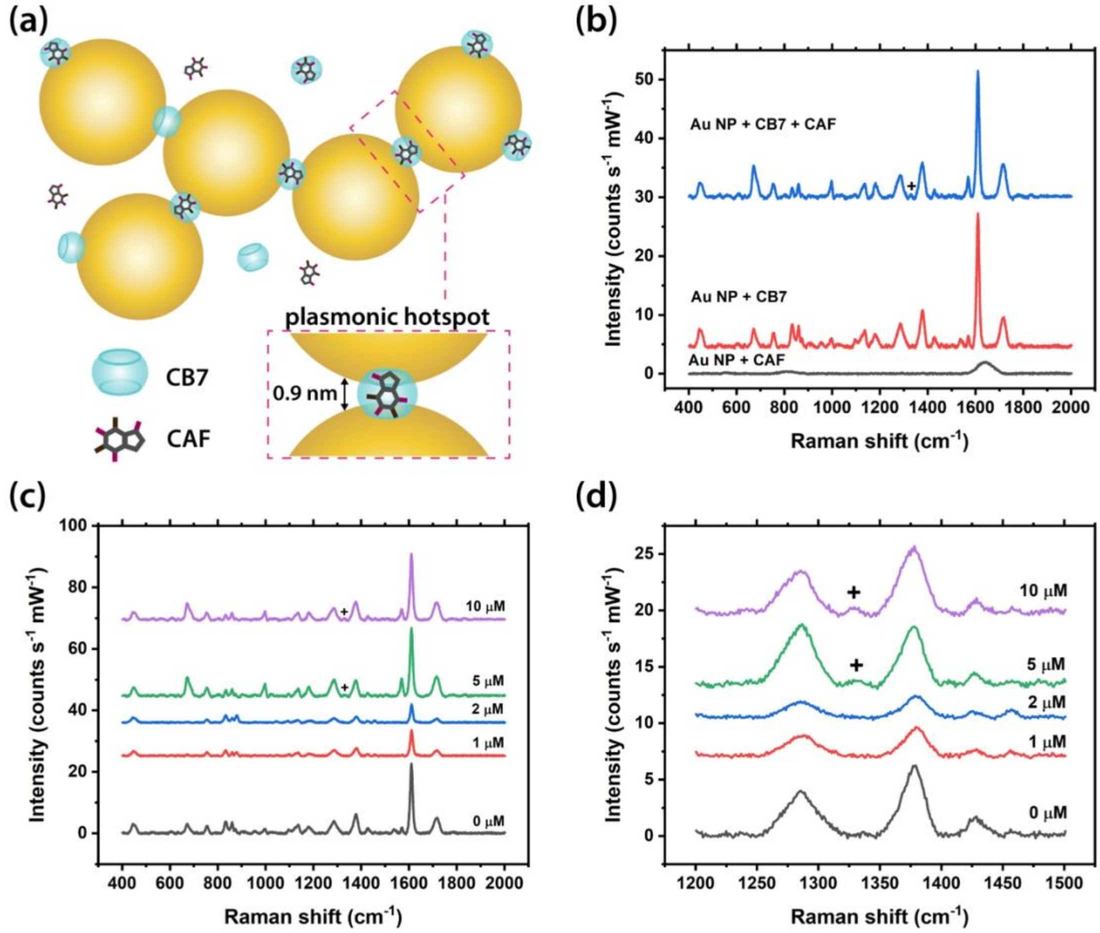


Figure S12. (a) Schematic illustration of the precise plasmonic hotspots within Au NP: CB7 nanoaggregates for CAF detection (not to scale). (b) SERS spectra of CAF in the presence or absence of CB7. (c) Full-range and (d) zoom-in SERS spectra of CAF with different concentrations from 0 to 10 μM. Main Raman peak of CAF at 1330 cm^{-1} is marked by +. Spectra were baseline corrected and offset for clarity.

16. SERS spectra of CB8-CAF complexes

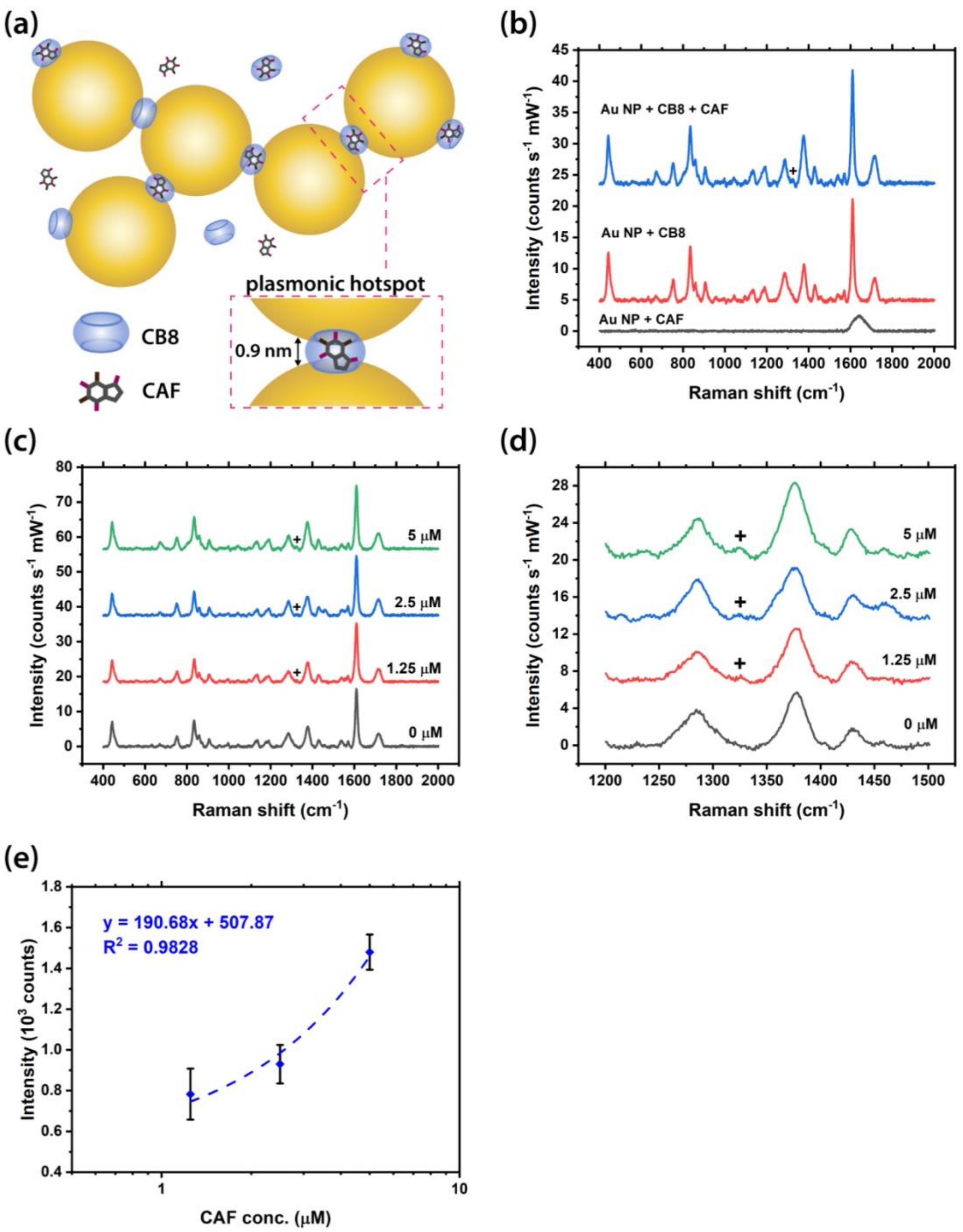


Figure S13. (a) Schematic illustration of the precise plasmonic hotspots within Au NP: CB8 nanoaggregates for CAF detection (not to scale). (b) SERS spectra of CAF in the presence or absence of CB8. (c) Full-range and (d) zoom-in SERS spectra of CAF with different concentrations from 0 to 5 μM. Main Raman peak of CAF at 1330 cm^{-1} is marked by +. Spectra were baseline corrected and offset for clarity. (e) Corresponding plot of SERS intensity of the main CAF peak (marked by + in (d)) against CAF concentration. (Note: x-axis is plotted in log-scale to even out the spread of the data points for better illustration.)

17. Multiplexed quantification using machine learning techniques

A. Partial Least Squares Regression

Partial Least Squares Regression (PLSR) is a well-established multivariate regression technique related to Principle Component Analysis (PCA).[2,3] In PLSR, the dimensionality of the predictor (X) variables are reduced by finding new latent variables which best describe the variation in the response (Y) variables.[3] To perform the analysis, the dataset is split into a *m* x *n* **X** matrix and a *p* x *n* **Y** matrix where *n* is the number of spectra, *m* is the number of wavenumber shifts in the SERS spectra after pre-processing, in this case, 1167 (from 500 - 1800 cm^{-1}, equally spaced) and *p* is the number of analytes, which is 2.[3,4]

The **X** and **Y** matrices are decomposed according to:[2,3]

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{t}} + \mathbf{E} \qquad \textit{Equation 1}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{t}} + \mathbf{F} \qquad \textit{Equation 2}$$

where $\mathbf{T}$ and $\mathbf{U}$ are the $\mathbf{X}$ and $\mathbf{Y}$ scores and $\mathbf{P}$ and $\mathbf{Q}$ are the $\mathbf{X}$ and $\mathbf{Y}$ loadings. $\mathbf{E}$ and $\mathbf{F}$ are the residuals. The decomposition is performed in a way that maximises the covariance of $\mathbf{T}$ and $\mathbf{U}$. The scores are related by:

$$\mathbf{U} = \mathbf{B}\mathbf{T} \qquad \textit{Equation 3}$$

where $\mathbf{B}$ is a matrix of the PLSR coefficients.

In this work, the Python class 'sklearn.cross_decomposition.PLSRegression', which utilizes the NIPALS algorithm[2], was used to perform PLSR on the pre-processed spectra. The number of PLSR components used in the model was chosen to minimise the mean squared error.

B. Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational systems inspired by the structure of biological brains.[5] Their ability to recognise patterns and classify data has led to their widespread use in recent years.[6] They are made up of many connected nodes called artificial neurons which take in multiple inputs and compute a single output value.[5,7] ANNs are built by connecting these neurons into larger networks. Multilayer perceptrons (MLPs) are one of the simplest and most popular classes of feed-forward ANNs.[5] They consist of at least three layers of neurons – an input layer, one or more hidden layers and an output layer.[5,8] Since multilayer perceptrons are fully connected, each neuron in one layer connects with every neuron in the following layer.[9] Initially, the connections between neurons are assigned random weights.[8] When the network is presented with a training dataset, back-propagation is used to iteratively adjust the weights to reduce the difference between the output results and the actual results.[5,8,9] This process is repeated until the error is below an acceptable level. The resulting trained network can then be used to determine the output for a new unseen input dataset.

In this work, the Python class ‘sklearn.neural_network.MLPRegressor’ was used to build an MLP and predict the concentration of analytes from unseen spectra. The parameters used in this work are detailed in Table S5.

Table S5. Optimised parameters used for the multilayer perceptrons.

Parameter	Value
Number of input nodes	1167
Number of output nodes	2
Number of nodes in the hidden layers (analyte concentrations ≤ 1 µM)	16-128
Number of nodes in the hidden layers (analyte concentrations ≤ 5 µM)	32-32-128-16
Activation function	ReLU (Rectified Linear Function)
Algorithm for weight optimisation	Limited-memory BFGS
Maximum number of iterations	200

C. Pre-processing

To prepare the dataset for machine learning, the spectra were truncated to eliminate the noisy regions close to the edge of the spectral detection band of the spectrometer.[8] Measurements at Raman shifts below 500 cm^{-1} and above 1800 cm^{-1} were removed. After trimming the spectra, asymmetric least squares (ALS) baseline correction was applied.[10] Then, standard normal variate normalisation was performed to give each spectrum a mean intensity of 0 and a standard deviation of 1.[8,11] For the artificial neural network models, the analyte concentrations were also scaled via min-max normalization so that all of the values were transformed into the range [0,1].

D. Bootstrapping random resampling

The models were evaluated using the bootstrapping random resampling procedure. In this method, the training set is created by randomly selecting *n* observations from the dataset with replacement.[11] The fact that the selections are replaced after they are chosen means that the same observation can appear in the training set more than once. The number of selections, *n*, is equal to the number of observations in the original dataset. The test set is made up of any observations that were never selected and are therefore not in the training set. Once the two groups have been created, the model is built using the training set and evaluated using the test set. 1000 bootstrapping iterations were performed to evaluate the models built in this work.

E. Relationship between SERS characteristic peak intensity and analyte concentration

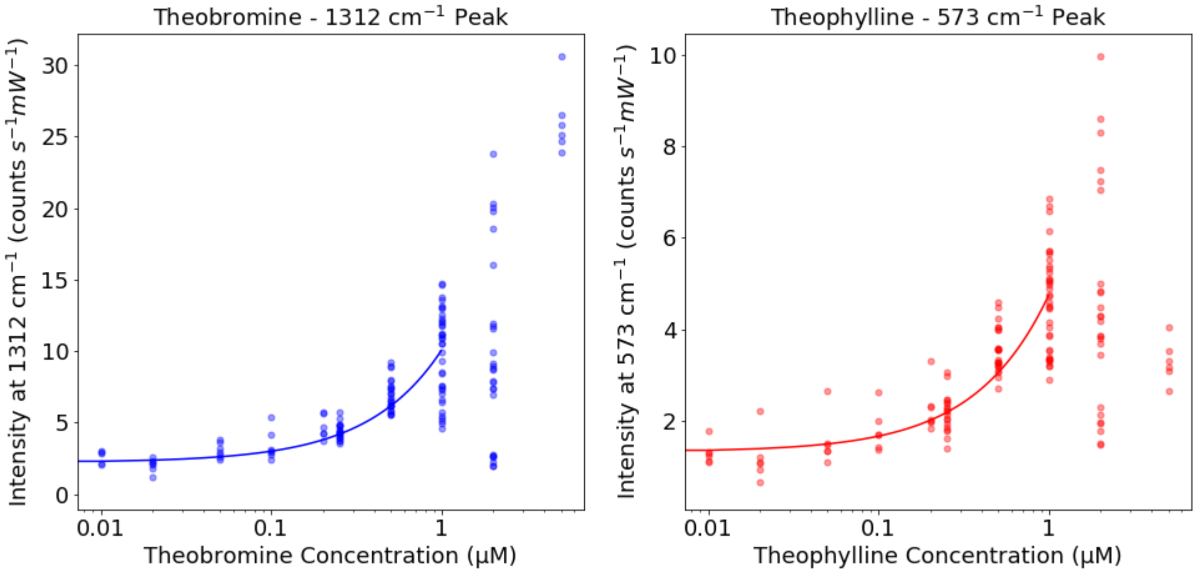


Figure S14. The intensities of the theobromine (blue) and theophylline (red) characteristic SERS peaks plotted against analyte concentration for the Au NP: CB8 dataset. A linear relationship between the analyte concentration and the peak intensity is present up to 1 μM. Above 1 μM, the relationship between peak intensity and concentration is no longer linear.

F. PLSR Predictions (≤ 1 μM)

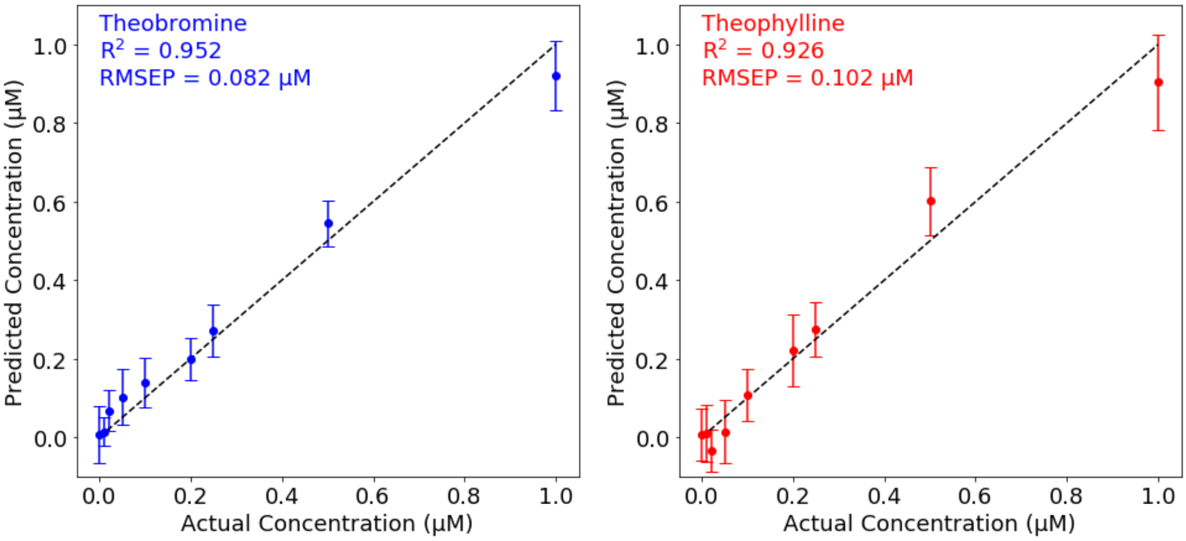


Figure S15. Predictions of the theobromine (blue) and theophylline (red) concentrations made using the PLSR model trained with SERS spectra of solutions with analyte concentrations ≤ 1 μM from the Au NP: CB8 dataset. The points are the mean values calculated from 1000 bootstrapping iterations and the error bars show the standard deviation of the predictions.

G. ANN Predictions (≤ 1 μM)

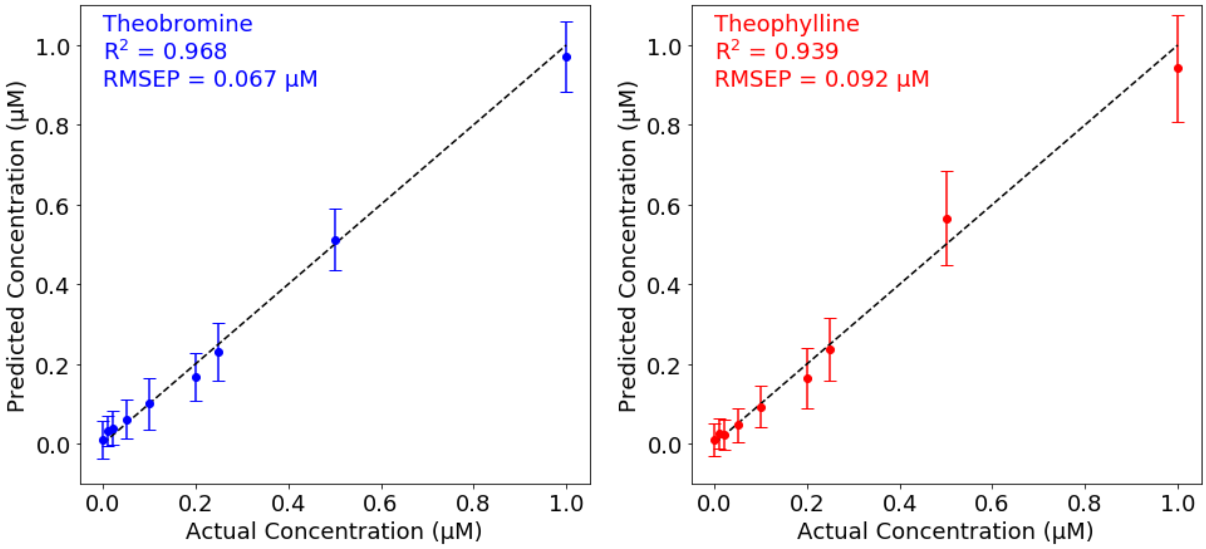


Figure S16. Predictions of the theobromine (blue) and theophylline (red) concentrations made using the ANN model trained with SERS spectra of solutions with analyte concentrations ≤ 1 μM from the Au NP: CB8 dataset. The points are the mean values calculated from 1000 bootstrapping iterations and the error bars show the standard deviation of the predictions.

H. PLSR Predictions (≤ 5 μM)

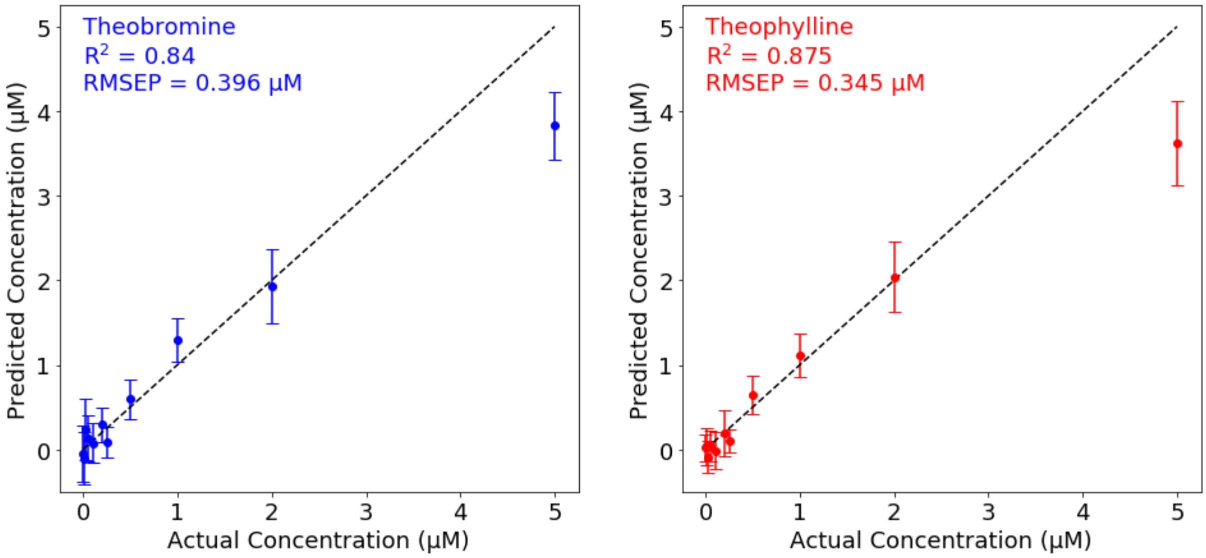


Figure S17. Predictions of the theobromine (blue) and theophylline (red) concentrations made using the PLSR model trained with SERS spectra of solutions with analyte concentrations ≤ 5 μM from the Au NP: CB8 dataset. The points are the mean values calculated from 1000 bootstrapping iterations and the error bars show the standard deviation of the predictions.

I. ANN Predictions (≤ 5 µM)

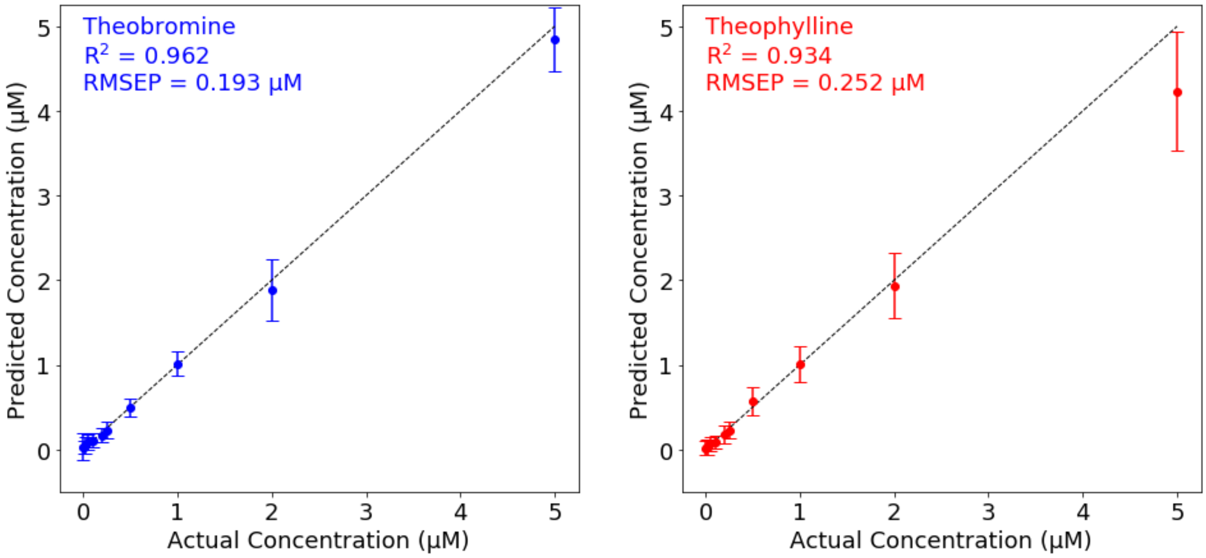


Figure S18. Predictions of the theobromine (blue) and theophylline (red) concentrations made using the ANN model trained with SERS spectra of solutions with analyte concentrations ≤ 5 µM from the Au NP: CB8 dataset. The points are the mean values calculated from 1000 bootstrapping iterations and the error bars show the standard deviation of the predictions.

J. X-loadings for the components in a PLSR model

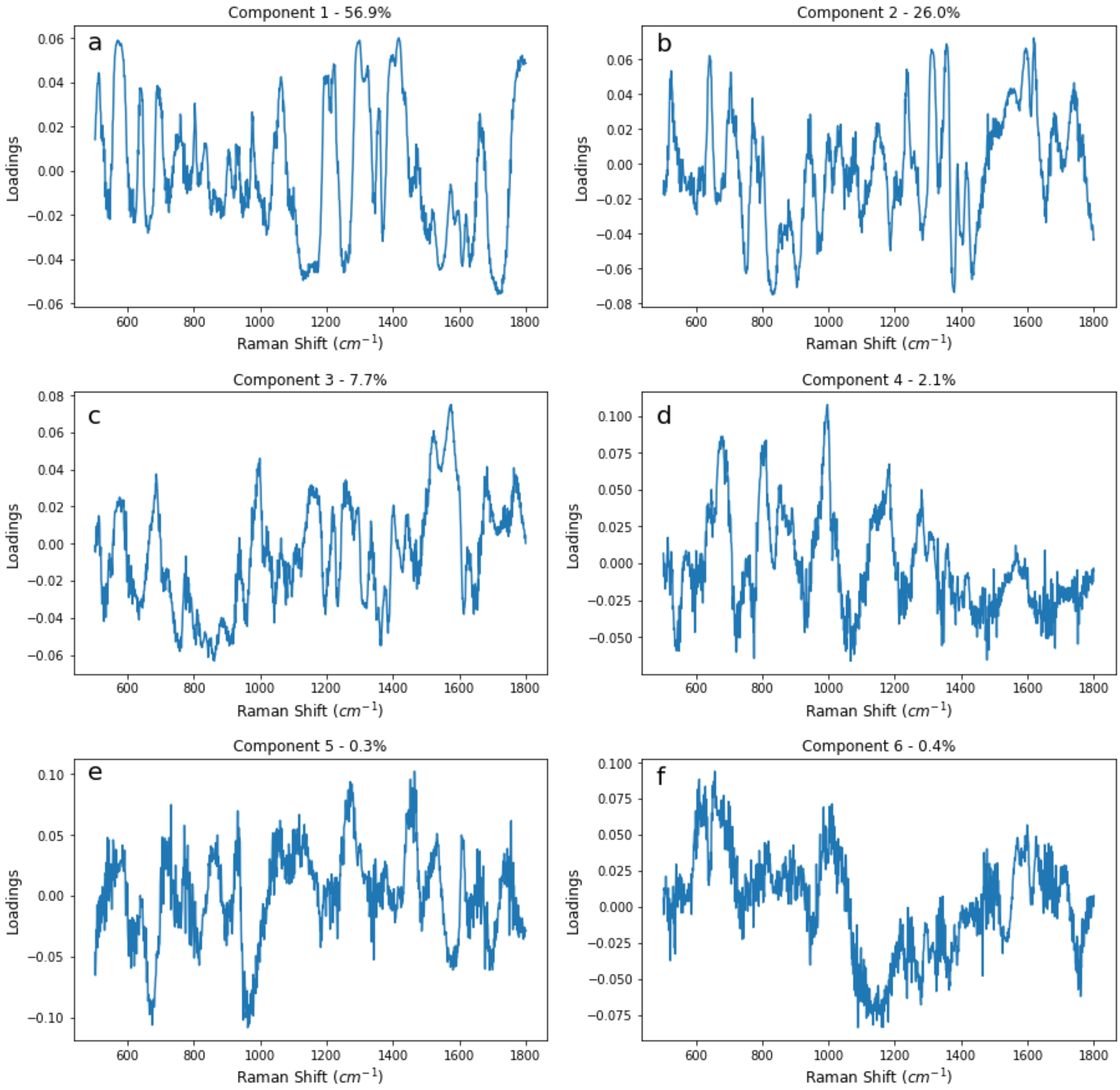


Figure S19. The X-loadings for each component in a PLSR model built with 6 components. The model was trained with SERS spectra of solutions with analyte concentrations ≤ 1 µM from the Au NP: CB8 dataset. The percentage of the variance explained (R^2) by each component is stated in the title of the subplot.

K. Building models without the characteristic SERS peaks

PLSR and ANN models were built using spectra in which the characteristic TBR and TPH peaks had been removed. All of the spectra were modified so the intensities at the wavenumbers surrounding the characteristics peaks, 642 cm^{-1} and 1312 cm^{-1} for theobromine and 573 cm^{-1} and 1298 cm^{-1} for theophylline, were changed to zero.

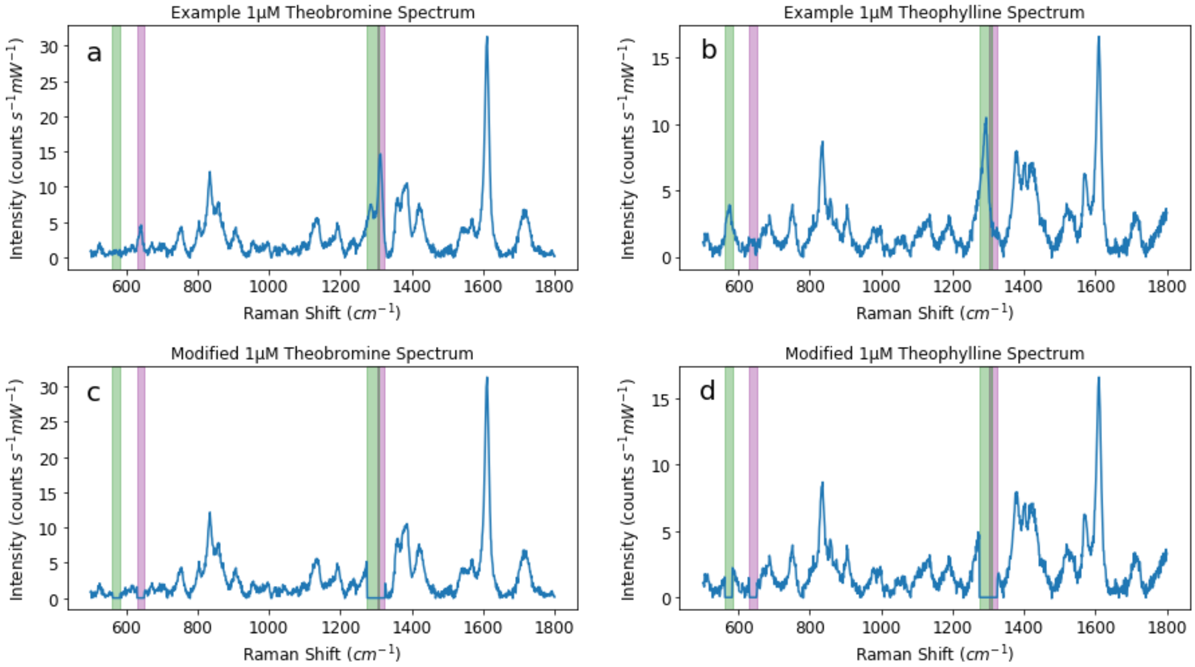


Figure S20. Examples of the original (a and b) and modified (c and d) SERS spectra from the Au NP: CB8 dataset. Spectra a and c are from a solution containing 1 µM TBR and 0 µM TPH and spectra b and c are from a solution containing 1 µM TPH and 0 µM TBR. The intensities at the wavenumbers surrounding the characteristic peaks, 642 cm^{-1} and 1312 cm^{-1} for TBR and 573 cm^{-1} and 1298 cm^{-1} for TPH, were changed to zero. The edited areas are highlighted in green and purple for the TBR and TPH peaks respectively.

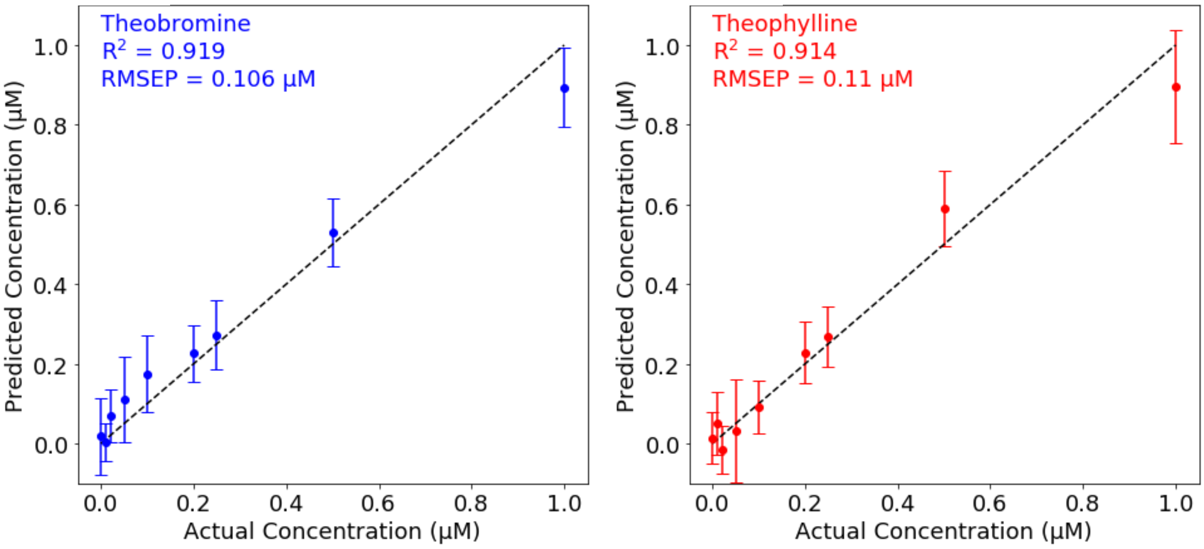


Figure S21. Predictions of the theobromine (blue) and theophylline (red) concentrations made using the PLSR model trained with the modified SERS spectra of solutions with analyte concentrations ≤ 1

µM from the Au NP: CB8 dataset. The points are the mean values calculated from 1000 bootstrapping iterations and the error bars show the standard deviation of the predictions.

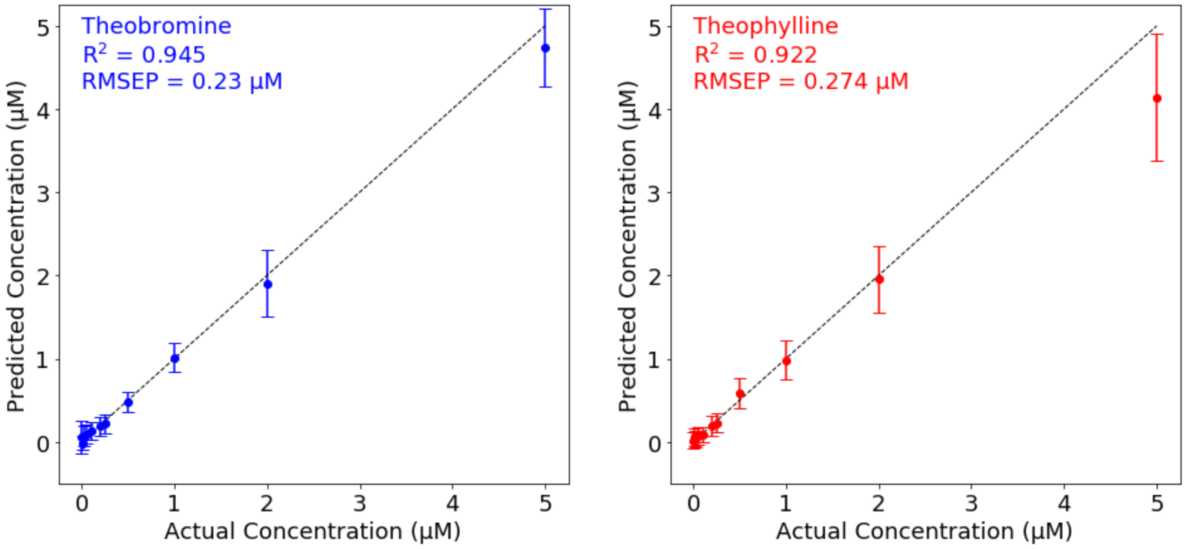


Figure S22. Predictions of the theobromine (blue) and theophylline (red) concentrations made using the ANN model trained with the modified SERS spectra of solutions with analyte concentrations ≤ 5 µM from the Au NP: CB8 dataset. The points are the mean values calculated from 1000 bootstrapping iterations and the error bars show the standard deviation of the predictions.

18. References

1. W. M. Nau, M. Florea and K. I. Assaf, *Isr. J Chem.,* 2011, **51**, 559.
2. S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. Syst.,* 2001, **58**, 109.
3. P. K. Kreeger, *Sci. Signal.,* 2013, **6**, tr7.
4. L. Eriksson, T. Byrne, E. Johansson, J. Trygg and C. Vikström, Multi- and Megavariate Data Analysis Basic Principles and Applications, *Umetrics Academy*, 2013.
5. J. S. Hallinan, Methods in Microbiology, eds. C. Harwood and A. Wipat, *Academic Press*, 2013, **vol. 40**, pp. 1–37.
6. C. M. Bishop, P. of N. C. C. M. Bishop, Neural Networks for Pattern Recognition, *Clarendon Press*, 1995.
7. C. Q. Nguyen, *UC Irvine*, 2018.
8. S. Kasera, L. O. Herrmann, J. del Barrio, J. J. Baumberg and O. A. Scherman, *Sci. Rep*., 2015, **4**, 6785.
9. F. Lussier, V. Thibault, B. Charron, G. Q. Wallace and J.-F. Masson, *TrAC Trends Anal. Chem.,* 2020, **124**, 115796.
10. P. H. C. Eilers, *Anal. Chem.,* 2003, **75**, 3631.
11. G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning, *Springer New York, New York, NY*, 2013, **vol. 103**.