Supporting information for paper Nitroaromatic Explosives detection and quantification using Attention-based Transformer on surface-enhanced Raman spectroscopy maps

Bo Li*[a], Giulia Zappalà[b], Elodie Dumont[b], Anja Boisen[b], Tomas Rindzevicius[b], Mikkel N. Schmidt[a], Tommy S. Alstrøm[a]

July 27, 2023

A SERS maps simulation and experimental results

We first demonstrated the performance of our proposed methods with simulated datasets. We simulate three types of SERS maps in total, including 1) SERS maps without any contaminants, 2) SERS maps with contaminants at a fixed Raman shift location, and 3) SERS maps with random contaminants at random Raman shift locations. We explain the SERS map simulation and show the SERS maps below. The implementation of the SERS map generation is publicly available.

A.1 SERS maps simulation

In this section, we explain the process for simulating a SERS map with the shape of $[S_h, S_w, N_w]$. We define a SERS map that has $S_h S_w$ spectra, and each spectrum has N_w wavenumbers. Each spectrum S_{ij} at location i, j in the SERS map S is a summation of a background spectrum B_{ij}, a pure Raman spectrum R_{ij}, and the noise spectrum N_{ij} as shown in Eq. A.1.

$$S_{ij} = B_{ij} + R_{ij} + N_{ij} \tag{A.1}$$

A.1.1 Background simulation

We first simulate the background spectrum B_{ij} by multiplying a base background spectrum b that follows an AR(1) process and a location-specific coefficient τ_{ij} that is drawn from a Beta distribution. Given a constant c and the autoregression coefficient ψ, we simulate the base background spectrum b_w at Raman shift w using Eq. A.2a. We then normalize the entire base spectrum b such that its intensity is between [0, 1] using Eq. A.2b. To add variation on the background spectrum B_{ij} at different locations in the SERS map, we multiply the base background spectrum b with a coefficient τ_{ij} that is drawn from a beta distribution $\mathrm{Beta}(\alpha, \beta)$.

$$b_w = c + \psi b_{w-1} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1), w \in \{2, ..., N_w\} \tag{A.2a}$$

$$b = b - \min(b) \quad b = \frac{b}{\max(b)} \tag{A.2b}$$

$$B_{ij} = b\tau_{i,j}, \quad \tau_{i,j} \sim \mathrm{Beta}(\alpha, \beta) \tag{A.2c}$$

A.1.2 Signal simulation

We assume the Raman signal is a combination between the hotspot map S_{hotspot} that has the shape of $[S_h, S_w]$ and the Raman spectrum V. We assume the hotspot map S_{hotspot} has N_h hotspots and it can show the spatial distribution of the enhancement factors[1]. The Raman spectrum V is assumed to be a summation of N_p spectra where each of the spectrum V_p is a combination between a Lorentzian distribution and a Gaussian distribution[2]. We will explain the generation of the hotspot map S_{hotspot} and Raman spectrum V below.

Hotspot map S_{hotspot}

We first simulate the hotspot map S_{hotspot} that has the shape of $[S_h, S_w]$ and N_h hotspots. Since the statistics such as the location μ, size r, and intensity a of each hotspot can be different, we assume that the map S_{hotspot} is a summation of N_h hotspot maps as shown in Eq. A.3c. We draw each hotspot map H_h that has the shape of $[S_h, S_w]$ following a Gaussian distribution that is parameterised by the location of the hotspot μ_h and the size of the hotspot r_h as shown in Eq. A.3b. We draw the location μ_h for hotspot h from an uniform distribution and the size of the hotspot r_h from a Gaussian distribution as shown in Eq. A.3a

[a] *Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Lyngby, Denmark; E-mail: blia@dtu.dk*

[b] *Center for Intelligent Drug Delivery and Sensing Using Microcontainers and Nanomechanics (IDUN), Department of Health Technology, Technical University of Denmark, 2800 Lyngby, Denmark*

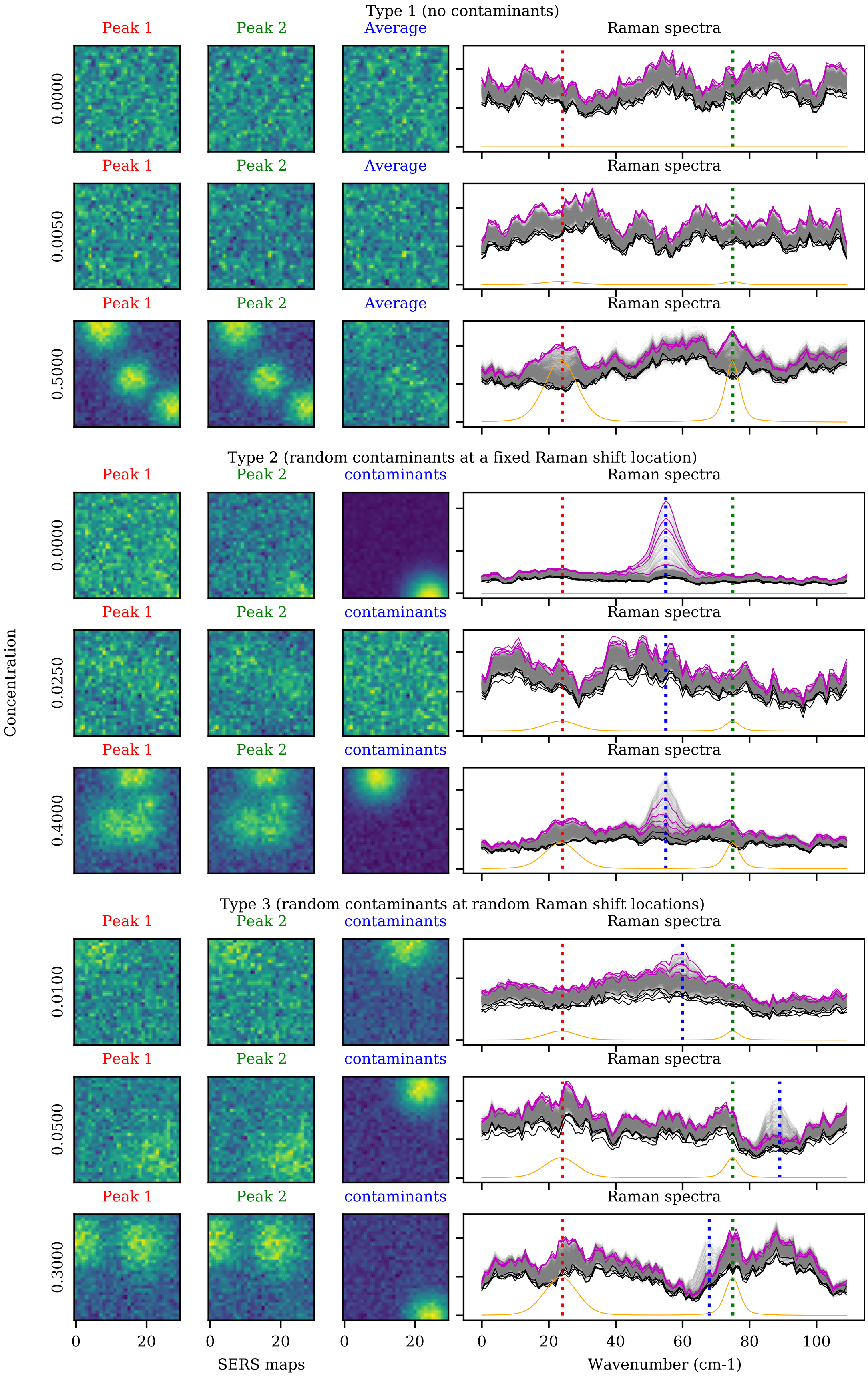

Fig. A.1 Example SERS maps and the corresponding spectra. The first two columns are the SERS maps at the signal peak locations. The third column is either the averaged SERS map over the wavenumber dimension or the SERS map at the contaminated Raman shift locations. The last column shows all the spectra from the SERS map and the pseudo-Voigt profile (V). We highlight ten spectra that have the highest and lowest peak intensities

$$\mu_h \sim [\text{Unif}(0, S_h - 1), \text{Unif}(0, S_w - 1)]$$
$$r_h \sim \mathcal{N}(h_s, 1), a_h \sim \text{Unif}(0.5, 0.6) \tag{A.3a}$$

$$H_{h,ij} = a_h \exp(-\frac{(i - \mu_h[0])^2 + (j - \mu_h[1])^2}{r_h^2}) \tag{A.3b}$$

$$S_{\text{hotspot}} = \sum_{h=0}^{N_h - 1} H_h \tag{A.3c}$$

Raman spectrum V

We next simulate the spectrum V for the Raman signal. We assume the Raman spectrum V contains N_p peaks, and each of the peak corresponds to a Raman spectrum V_p that is drawn from a combination of the Lorentzian distribution and the Gaussian distribution using Eq. A.4c. The Lorentzian distribution and Gaussian distribution are parameterised by the centre location of a peak c_p and the half-width-at-half-maximum γ_p as shown in Eq. A.4a and A.4b. We draw the centre location c_p for peak p in the Raman shift domain of the spectrum following a uniform distribution that ranges from $v = 20$ to $v = N_w - 20$. We sample γ_p following a Gaussian distribution $\mathcal{N}(5, 1)$ and the ratio of the Lorentzian distribution ρ is sampled following a uniform distribution Unif(0, 1). Each spectrum V_p is normalised such that the maximum intensity is one using Eq. A.4d. We then take the sum of the spectrum V_p over p peaks as the Raman spectrum V as shown in Eq. A.4e.

$$L_p = \frac{1}{1 + \frac{(v - c_p)^2}{\gamma_p^2}} \tag{A.4a}$$

$$G_p = \exp(-\frac{(v - c_p)^2}{2\gamma_p^2}) \tag{A.4b}$$

$$V_p = \rho_p L_p + (1 - \rho_p) G_p \tag{A.4c}$$

$$V_p = \frac{V_p}{\max V_p} \tag{A.4d}$$

$$V = \sum_{p=0}^{N_p - 1} V_p \tag{A.4e}$$

Raman signal R_{ij}

Given the hotspot map S_{hotspot}, the Raman spectrum V, and the background B, we can calculate the amplitude of the peaks with respect to the signal-to-background ratio *SBR* following Eq. A.5. For simplicity, we assume the amplitude for all the peaks are same. We then take the product between the hotspot map S_{hotspot}, amplitude A, and the Raman spectrum V to get the final Raman signal R_{ij} at a specific location using Eq. A.6.

$$A = \sqrt{\text{SBR} \frac{\sum_{i=0}^{S_h - 1} \sum_{j=0}^{S_w - 1} \sum_{w=0}^{N_w - 1} B_{ijw}^2}{\sum_{i=0}^{S_h - 1} \sum_{j=0}^{S_w - 1} \sum_{w=0}^{N_w - 1} (S_{\text{hotspot},ij} V_w)^2} \frac{1}{N_p^2}} \tag{A.5}$$

$$R_{ij} = S_{\text{hotspot},ij} A V \tag{A.6}$$

A.1.3 Noise simulation

We simulate each noise spectrum at Raman shift w using eq. A.7 where σ_n is defined as the degree of the noises.

$$N_{ij,w} = \sigma_n^2 \varepsilon_n, \quad \text{where} \quad \varepsilon_n \sim \mathcal{N}(0, 1) \tag{A.7}$$

A.2 Simulated dataset

We explain three different types of simulated SERS maps in this section. We assume that the molecule of interest contains two peaks ($N_p = 2$) and the fingerprint characteristics, such as the peak locations c_p, half-width-at-half-maximum γ_p, and the ratio of the Lorentzian distribution ρ_p, are the same for all three types of maps. There are 900 spectra in each SERS map $S_h = 30, S_w = 30$ and the length of the wavenumbers is 110 ($N_w = 110$). We assume that the relation between the signal-to-background ratio (SBR) and the concentration is non-linear, as can be seen in Fig. A.2.To make the simulated maps more realistic, we propose including contaminants. These contaminants can appear either at a fixed Raman shift location or at random Raman shift locations. We summarize the simulated SERS maps into three types:

- Type 1: SERS maps without any contaminants.

- Type 2: SERS maps with contaminants that appear at a fixed Raman shift location. This contaminant may appear because of other Raman active molecules in the substrate in a real-world scenario[3]. We assume that 20% of the SERS maps contain contaminants, and the peak location of the contaminants is fixed. Other characteristics, such as γ and ρ for these contaminants, are randomly drawn.

- Type 3: SERS maps with contaminants that are at random Raman shift locations. These contaminants can be caused by the preparation of the substrates, environmental contaminants, and lighting[3]. We assume that 50% of the SERS maps contain such contaminants and every characteristic, such as c, γ, and ρ of these contaminants, are randomly drawn at each time.

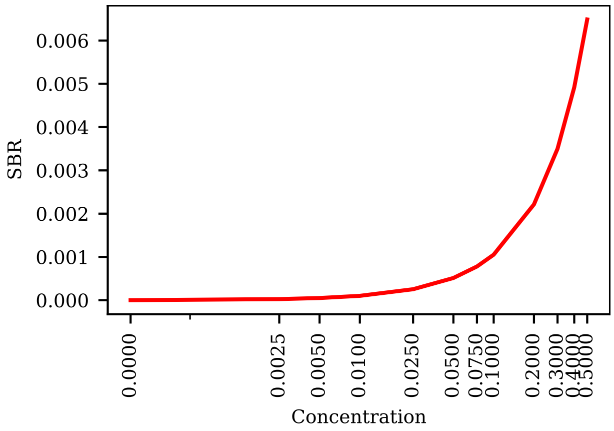

Fig. A.2 Relation between concentration and signal-to-background ratio (SBR). As the concentration increases, the signal-to-background ratio (SBR) is also increasing non-linearly

Depending on the task, we simulate a different number of SERS maps per concentration, as can be seen in Table. A.1. For the detection task, the goal is to detect whether a SERS map contains a molecule of interest (1/0). Therefore, we simulate the same number of maps for class 0 and class 1 to mimic the perfect class-balance scenario. Similarly, we simulate the same amount of SERS maps per concentration for the quantification task. We simulate the same amount of SERS maps for training, validating, and testing the model.

Table A.1 The number of training/validation/testing maps per concentration for detection and quantification tasks. To simulate the perfect class-balance scenario, we simulate the same number of maps for class 0 and class 1 for the detection task and the same number of maps per concentration for the quantification task

Concentration (label)	0 (0)	0.0025 (1)	0.0050 (1)	0.0100 (1)	0.0250 (1)	0.0500 (1)	0.0750 (1)	0.1000 (1)	0.2000 (1)	0.3000 (1)	0.4000 (1)	0.5000 (1)
Detection	550	50	50	50	50	50	50	50	50	50	50	50
Quantification	50	50	50	50	50	50	50	50	50	50	50	50

Some example SERS maps at the signal peak locations and the corresponding spectra can be seen in Fig A.1. We highlight ten spectra with the highest and lowest intensities at the signal peak locations. The SERS maps at low and zero concentrations are hardly distinguishable. Besides, the spectra with the highest intensities at signal peak locations are less similar to the corresponding pseudo-Vogit profile due to the background and noise. However, we can quickly identify the hotspots at high concentrations, and the differences between the highest and lowest intensity in fingerprint regions are more obvious.

We further show the relationship between the intensities at the peak locations and concentrations to explore the differences between the spectra at different concentrations. We extract all the spectra from each SERS map and rank them based on the intensities at the signal peak locations. We then select a subset of the spectra, such as the top 10%, and calculate the average of the peak intensities over them. The average and the standard deviation of the peak intensities over 50 SERS maps per concentration are shown in Fig. A.3. The trend that a higher concentration is more likely to correspond to a higher peak intensity is more visible when we only select and aggregate a subset of spectra per map.

A.3 Experimental results on simulated datasets

We report and compare the performance between our proposed approach (ViT) that takes SERS maps as input and the methods that take the averaged spectra as input in this section. For the spectra-based models, except for ranking and selecting the spectra based on the peak intensities Eq. A.8a, we also experimented with using mean Eq. A.8b, standard deviation Eq. A.8c, and the difference of the intensity Eq. A.8d between the neighbouring wavenumbers for selecting the spectra as shown in:

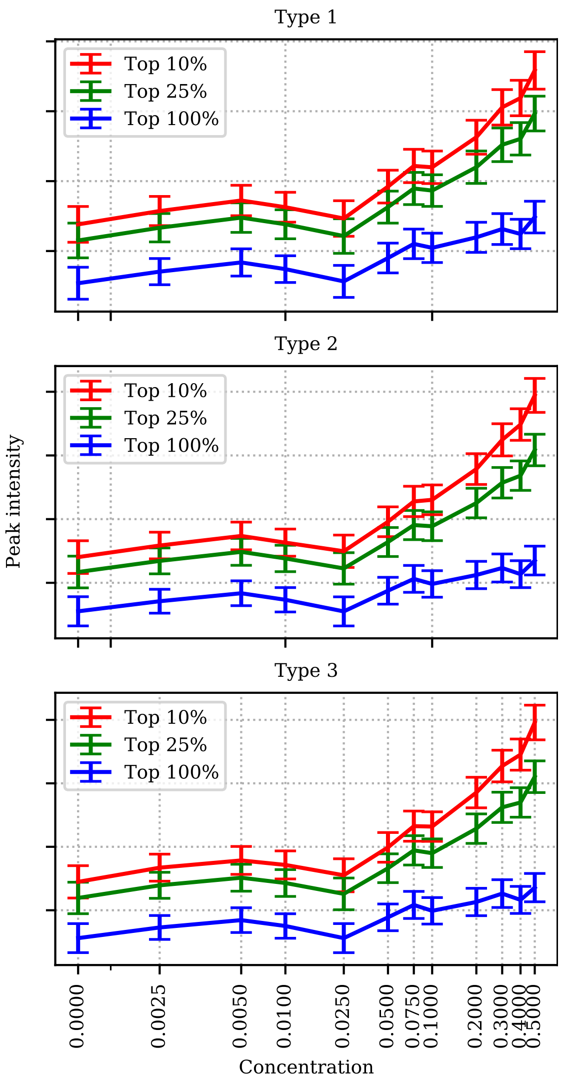

Fig. A.3 The correlation between the averaged peak intensity and concentration for different datasets. The errorbar represents the standard deviation of the averaged peak intensities over 50 SERS maps per concentration. The correlation is more obvious when we extract the peak intensity with fewer spectra per map, i.e., the top 10% of the spectra.

$$P_{x,y} = \sum_{w \in \mathscr{P}} X_{x,y,w}, \tag{A.8a}$$

$$M_{x,y} = \frac{\sum_{w=1}^{N_w} X_{x,y,w}}{N_w} \tag{A.8b}$$

$$S_{x,y} = \frac{\sum_{w=1}^{N_w} (X_{x,y,w} - M_{x,y})^2}{N_w} \tag{A.8c}$$

$$D_{x,y} = \sum_{w=1}^{N_w-1} ||X_{x,y,w+1} - X_{x,y,w}|| \tag{A.8d}$$

We quantitatively show the detection and quantification performance on three different simulated datasets: Type 1 without contaminants, Type 2 with random contaminants at a fixed Raman shift location, and Type 3 with random contaminants at random Raman shift locations. We also provide a qualitative explanation of the performance difference between the models.

A.3.1 Detection

For the methods that take the spectra as the input, we select the most appropriate combination between α_{train} and α_{eval} following the procedure in the manuscript. The heatmap of the validation loss is shown in Fig. A.7. For each of these three datasets, averaging a spectrum from a subset of spectra leads to lower validation loss compared to averaging the entire SERS map. Besides, we benefit

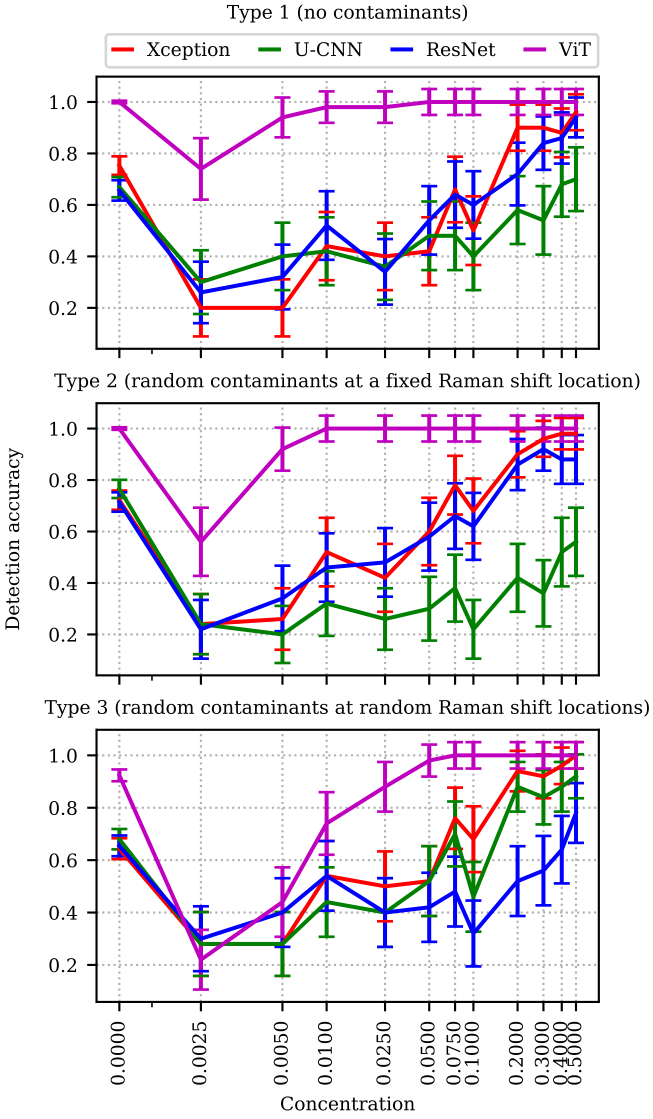

Fig. A.4 Detection accuracy with a 95% confidence interval of the correct detection rate over five runs on different datasets. Our method significantly improves the detection accuracy, especially at low concentrations, compared to the methods that use spectra as the input. When the SERS maps contain more complicated contaminates, the detection accuracy drops.

by selecting a similar amount of spectra during training and validating since we tend to get lower validation loss near the diagonal. When the peak locations are known, selecting spectra based on the peak intensities leads to lower validation loss. We then use the best combination between α_{train} and α_{eval} for each of the datasets on the test dataset to demonstrate the detection accuracy.

Fig. A.4 shows the averaged detection accuracy over five runs with the 95% confidence interval of the correct detection rate ($\pm 1.96\frac{\hat{p}(1-\hat{p})}{54}$, $\hat{p} = TP + TN + 2$, $TP + TN + FP + FN = 50$). In general, we obtain a relatively lower detection accuracy when the concentration of the molecule of interest is lower. We observe little performance difference between using Xception, U-CNN, and ResNet. Compared to the methods that use the averaged spectra as the input, our proposed approach (ViT) achieves much higher accuracy, especially when the concentration of the molecule of interest is low. The accuracy of ViT drops when the dataset contains more complicated contaminants.

We calculate the validation loss at each concentration level to understand what prevents the spectra-based models from obtaining good accuracies at high concentrations. Fig A.6 shows the result on the Type 1 dataset when we select spectra based on the peak intensities. As the concentration increases, the modality of the accuracy is also changing. The model benefits more from selecting fewer spectra at 0 concentration. However, we get a lower loss when more spectra are selected for calculating the averaged spectrum at low concentrations ($0.0025 \sim 0.0500$). The models produce similar detection results regardless of the percentage of selected spectra at high concentrations ($0.3 \sim 0.5$). These changes in the modalities make it harder to obtain a single selection criterion appropriate for all the concentrations. We need to sacrifice the accuracies at low or high concentrations. However, this can be dangerous in a real-world scenario. For example, we need to detect the chemical hazards to a concentration as low as possible while also being able to signal

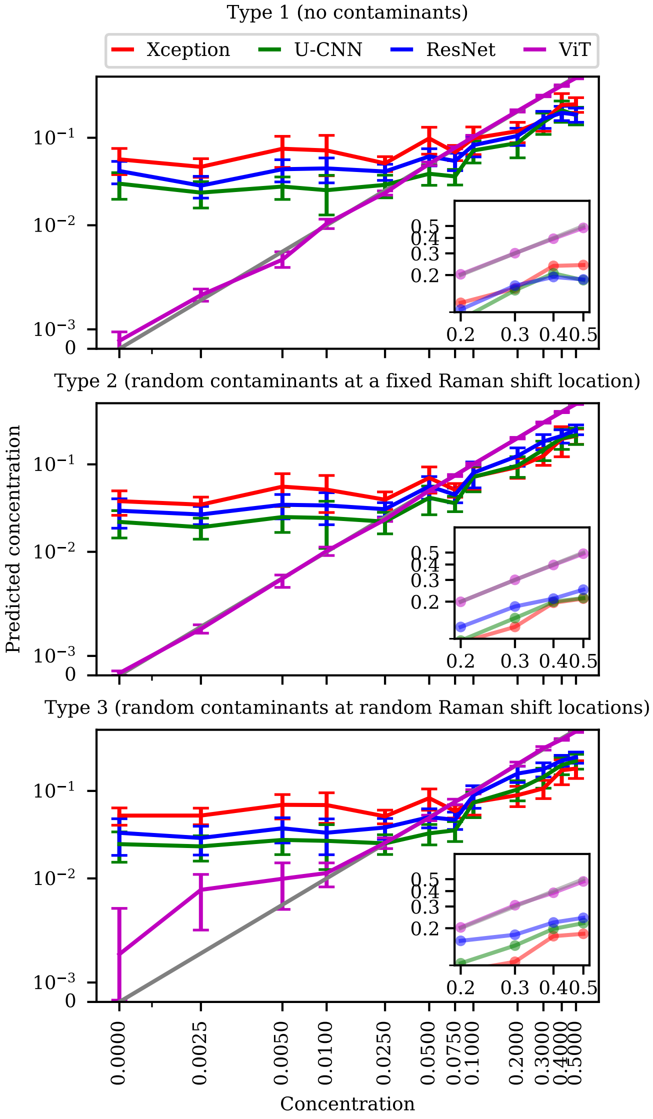

Fig. A.5 Concentration prediction and a 95% confidence interval over 50 maps per concentration. The predicted concentration is an average of over five runs. Our method can predict the concentration much more accurately than the models that take the spectra as the input, especially at low concentrations. The quantification accuracy drops when we include more complicated contaminants in the SERS maps.

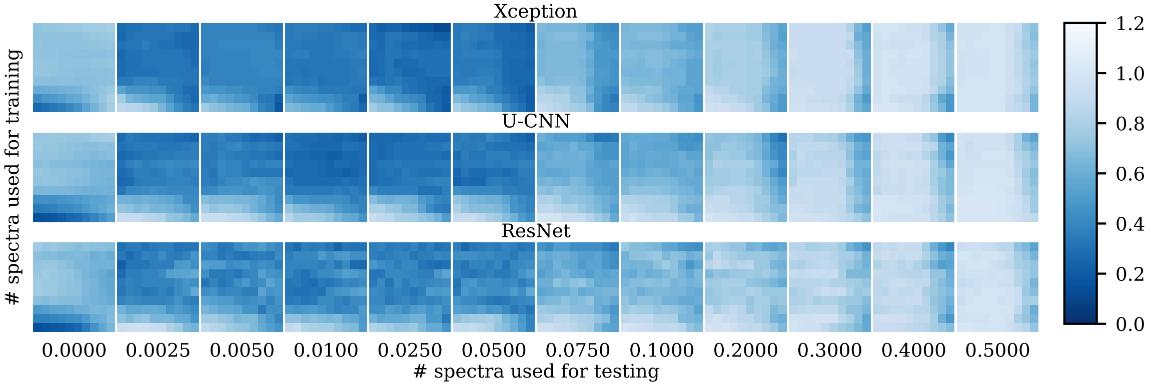

Fig. A.6 Binary cross-entropy loss on the validation dataset at each concentration level for Type 1 SERS maps where there are no contaminants and the spectra are selected based on the peak intensities. The best combination of the number of spectra used for training and testing is different across different concentrations, which makes it challenging to select a single combination for all the concentrations

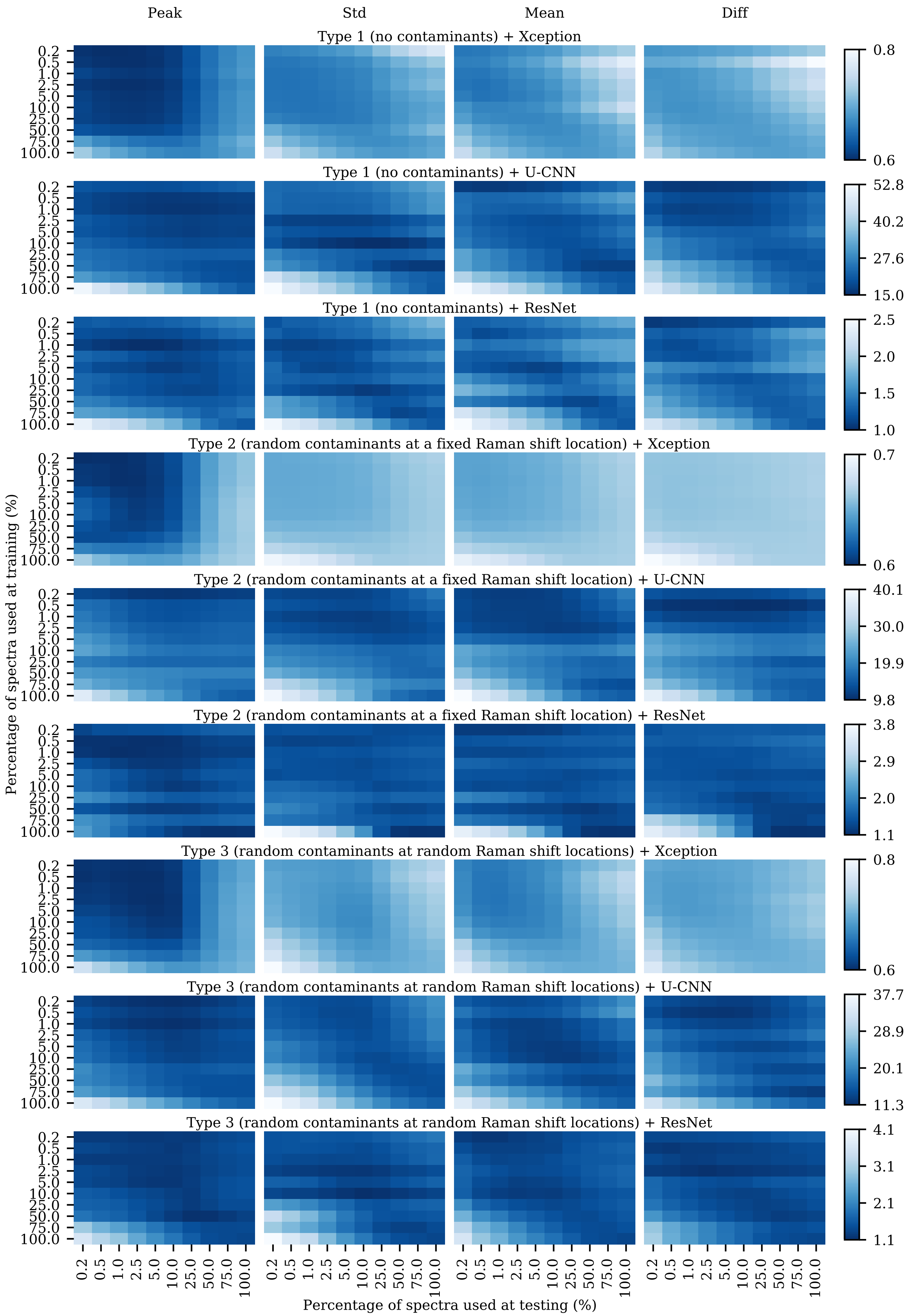

Fig. A.7 Detection performance on the validation datasets (binary cross-entropy loss). When the peak location is known, selecting spectra based on the peak intensities leads to a lower loss compared to other selection criteria. Preparing the input by averaging a subset of spectra from each map leads to better detection performance than averaging the full SERS map.

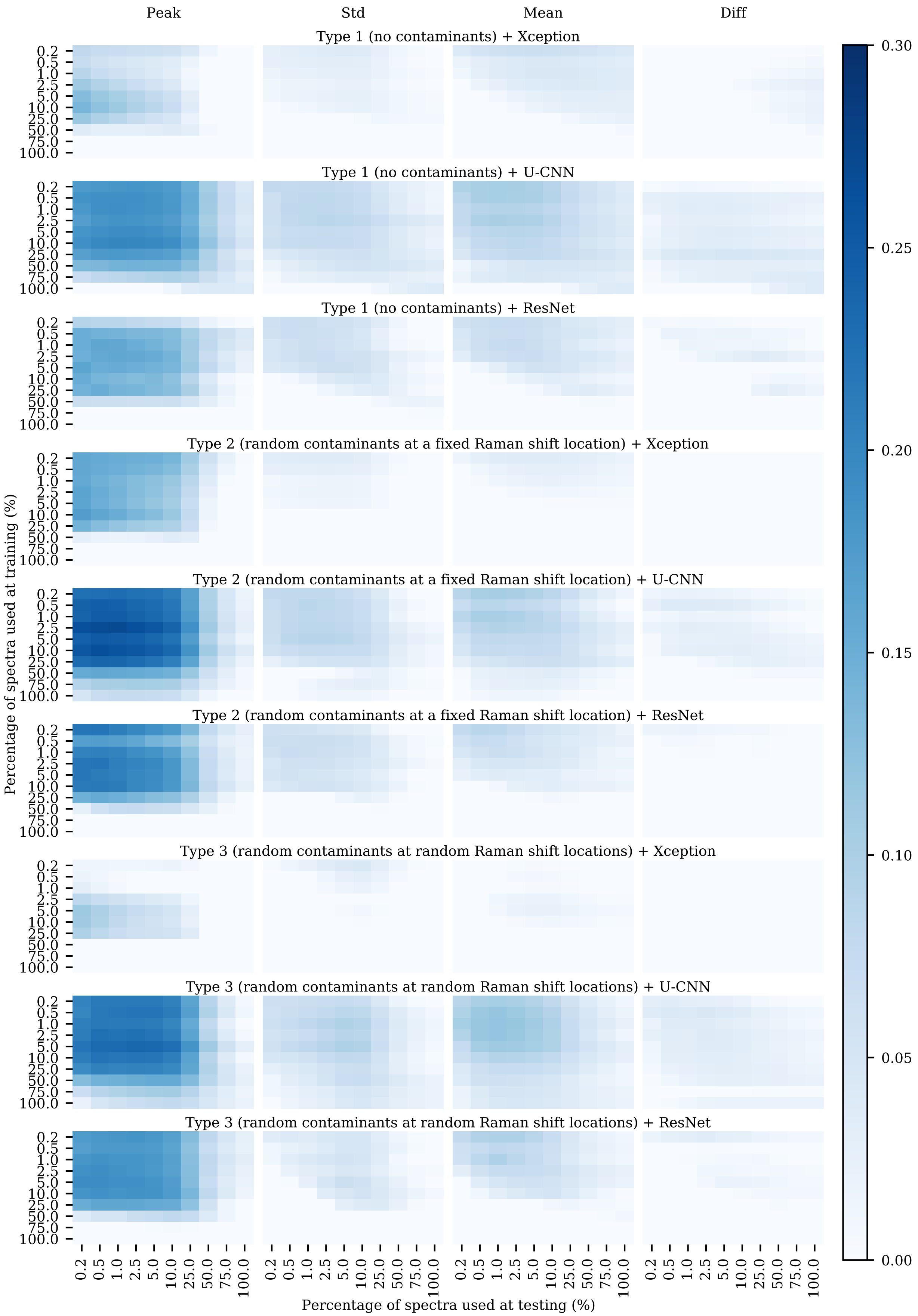

Fig. A.8 Quantification performance $R^2(\log c)$ on the validation datasets. When the peak location is known, selecting and aggregating spectra based on the peak intensities leads to a better quantification performance than other selection criteria. Selecting a subset of spectra gives better quantification performance compared to averaging the whole SERS map

when the concentration is high.

A.3.2 Quantification

We follow the same process to report the quantification performance. Fig. A.8 shows the R^2 on the log-transformed concentrations on the validation dataset. We observe a similar trend as shown in the detection task. Selecting a subset of spectra and a similar number of spectra during training and validation gives better quantification accuracy against other combinations. This phenomenon can attribute to the consistency between the peak intensities when selecting the spectra during training and validating with the same selection parameters. We then use the combination that gives the highest R to demonstrate the spectra-based models' performance on the test dataset.

The averaged predicted concentration with a 95% confidence interval ($\pm \frac{1.96\sigma_{\hat{c}}}{\sqrt{50}}$) where $\sigma_{\hat{c}}$ is the standard deviation of the predicted concentration within each concentration level using different models for each dataset is shown in Fig. A.5. There is no significant difference between the spectra-based models Xception, U-CNN, and ResNet. They all can give relatively accurate and non-overlapping concentration predictions down to a concentration of 0.2. The model tends to fail to differentiate between the spectra when the concentration is below 0.2. However, our proposed model ViT which takes the SERS map as input, can achieve much more accurate quantification results. We can differentiate the SERS maps when their concentrations are larger than 0.01. We obtain slightly worse prediction results, especially at lower concentrations (≤ 0.01), as we add more complicated contaminants to the dataset.

B Addition details for preparing the SERS maps

B.1 Fabrication process

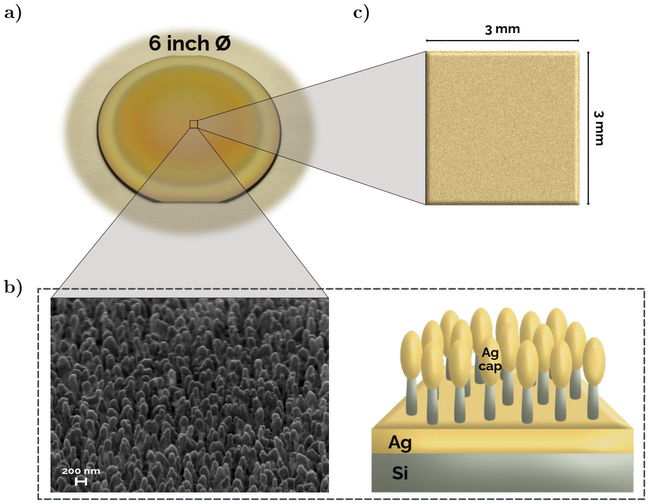

Fig. B.1 SERS substrates obtained from a two-step fabrication process. A maskless RIE of Si and e-beam evaporation of Ag were performed on a polished 6-inch black Si wafer (a). The resulting structures are vertically standing Ag-capped Si NPs, as shown in (b) from a SEM image and a graphical representation. 3x3 mm^2 chips, shown in (c), were then cut using a laser micromachining tool.

B.2 Additional figures for the spectra

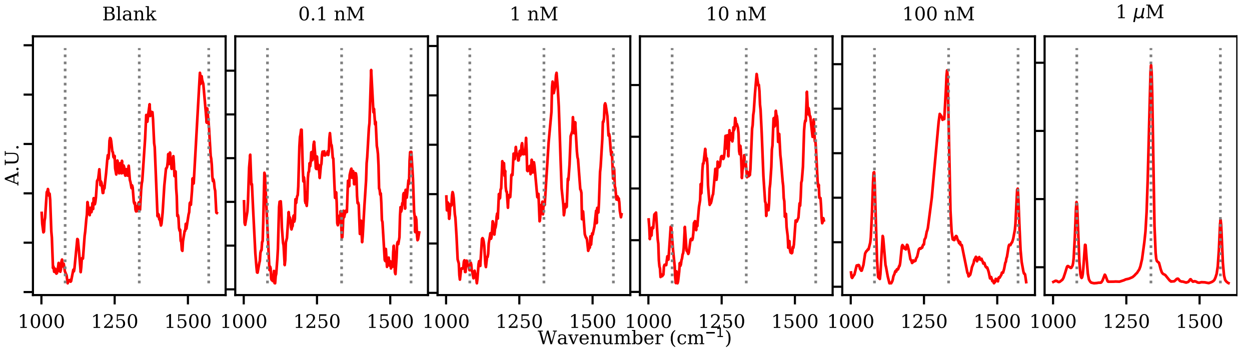

Fig. B.2 Example 4-NBT spectra at different concentration levels. We here select the top 1% of spectra with the highest fingerprint peak intensities from each SERS map and show the normalised averaged spectrum. We normalize it by subtracting the mean of the intensities and dividing by the standard deviation of the intensities.

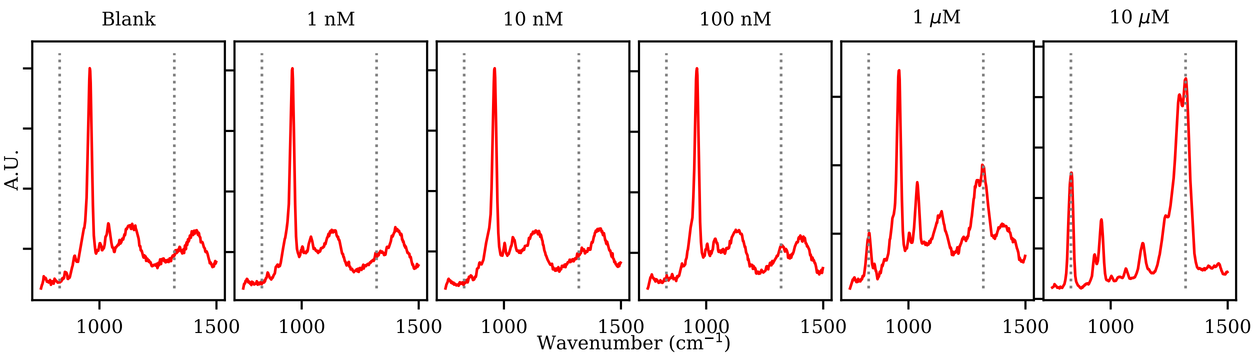

Fig. B.3 Example DNP spectra at different concentration levels. We here select the top 1% of spectra with the highest fingerprint peak intensities from each SERS map and show the normalised averaged spectrum. We normalize it by subtracting the mean of the intensities and dividing by the standard deviation of the intensities.

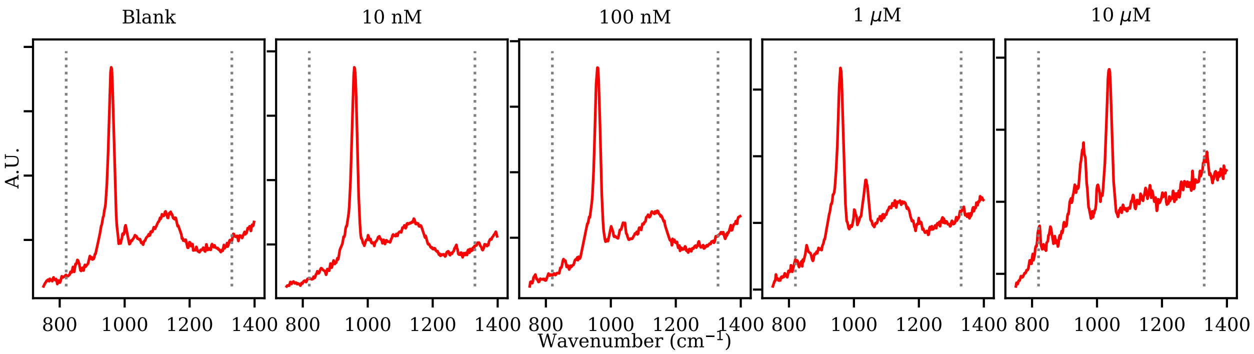

Fig. B.4 Example picric acid spectra at different concentration levels. We here select the top 1% of spectra with the highest fingerprint peak intensities from each SERS map and show the normalised averaged spectrum. We normalize it by subtracting the mean of the intensities and dividing by the standard deviation of the intensities. The fingerprint peaks at low concentrations are less visible compared to the peaks at high concentration level.

C Extra experimental details

C.1 Spectra-based model architecture

In this section, we explain the models that take the averaged spectra as the input. For the classical machine learning methods, including K Nearest Neighbors (KNN), Gradient Boosting, Random Forest, Support Vector Machine (SVM), and Decision Tree, we use the default implementations in scikit-learn `https://scikit-learn.org/stable/`. For the neural networks, we will explain the details below. All the implementations are publicly available.

- Xception: We use the same feature extractor as described by Li. et al[4]. We then append a single MLP layer on top of the feature extractor for either performing detection task or quantification task.
- Unified CNN: We use a similar feature extractor as described by Liu. et al [5]. We then append layers on top of the feature extractor following: `Linear-BatchNorm1d-Linear-Dropout-Linear`.
- ResNet: we use the same architecture as described by Ho. et al[6].

C.2 Augmentation procedure and hyperparameters

We explain the augmentation procedure in this section. We follow the augmentation scheme as described in Li. et al[4]. We use Gaussian noise with amplitude proportional to the standard deviation σ_w over a length K sliding window of the first difference $d_w = s_w - s_{w-1}$ of the intensities,

$$\mu_w = \frac{1}{K}\sum_{k=0}^{K-1} d_{w+k}, \qquad \sigma_w^2 = \frac{1}{K}\sum_{k=0}^{K-1}(d_{w+k}-\mu_w)^2, \qquad \hat{s}_w \sim \mathcal{N}(s_w, \kappa\cdot\sigma_w^2), \tag{C.1}$$

where s_w and $\hat{s}_w$ are the original and augmented spectra, and w is the wavenumber index. In our case, we set $K = 8$ and $\kappa \sim \mathcal{N}(0,1)$. For each SERS map measurement, we simulate 10 SERS maps for the 4-NBT dataset and 5 SERS maps for the DNP and the PA dataset. More detailed hyperparameters can be found in the repository.

C.3 Training and inference time

We show the number of parameters, training time, and evaluation time in Table C.1. While the ViT is a bit slower to train, the inference time is similar to other methods.

Table C.1 Comparison between different architectures. Models are trained for 300 epochs, and the evaluation is performed using Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz.

	ViT	Xception	ResNet	U-CNN
Model parameters	1,028,418	228,467	1,247,050	7,743,074
Training time (min)	11	7	7	6
Evaluation time (ms)	8	9	17	4

D Extra experimental results

D.1 Extra experimental results

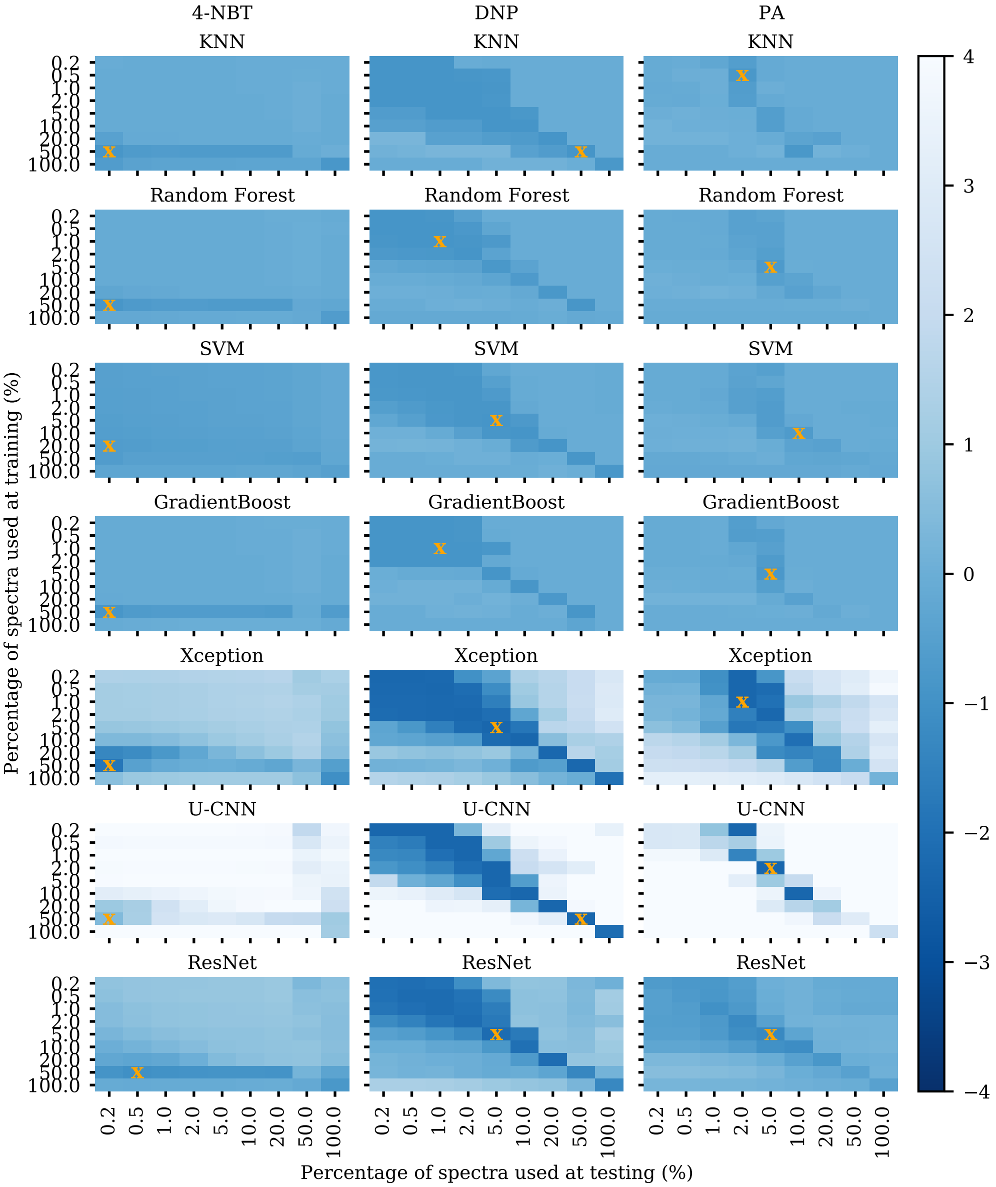

Fig. D.1 Classification loss (binary cross-entropy loss) on the validation datasets using different models. For better visualization, the colour bar indicates the log-transformed cross-entropy loss. The darker the colour is, the more accurate the model is. The lowest validation loss per heatmap is annotated with a x. Averaging the SERS map with a subset of spectra tends to be more accurate than averaging the entire SERS map.

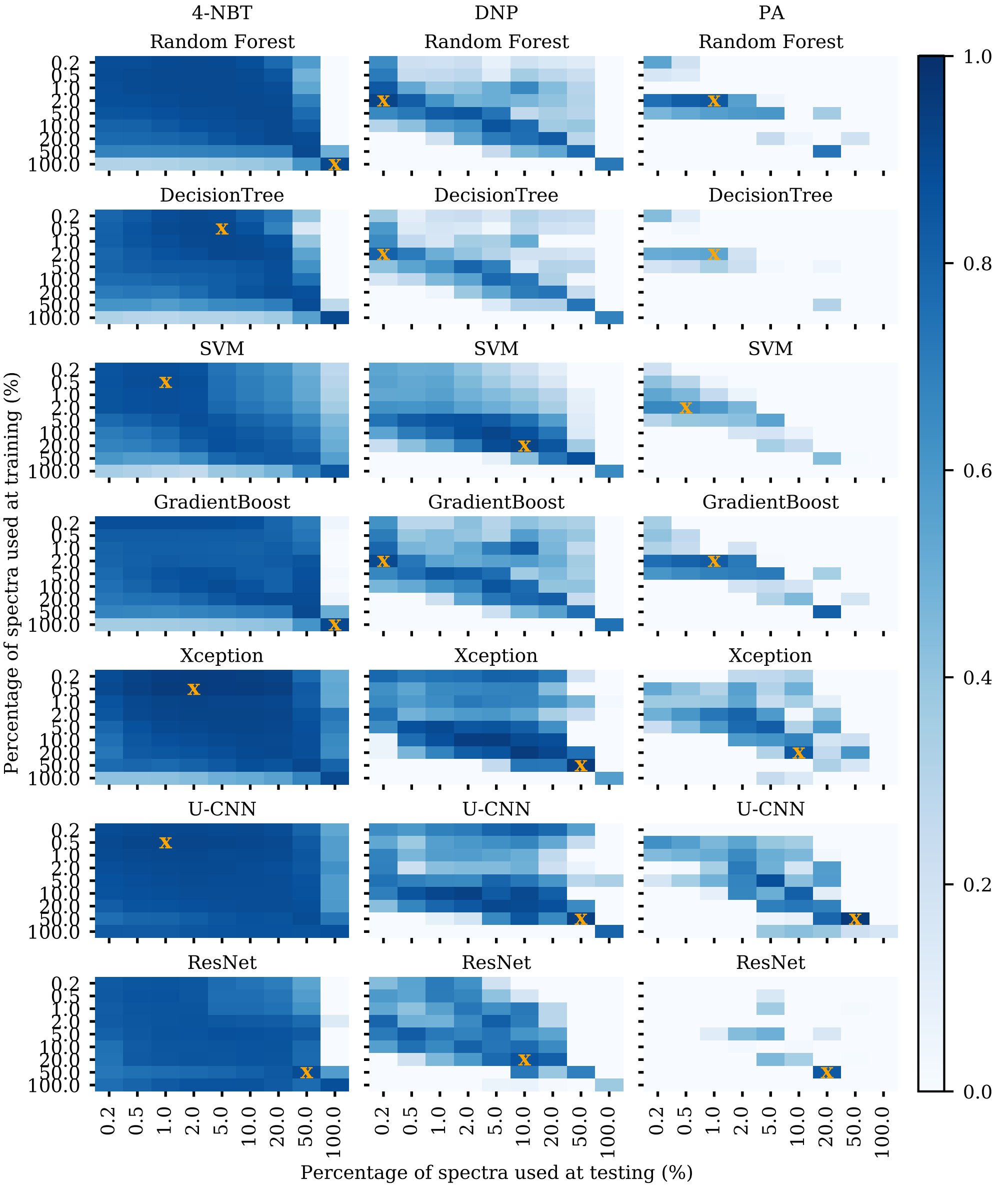

Fig. D.2 R^2 on the log-transformed concentrations (Eq. **??**) on the validation datasets using different models. The darker the colour is, the more accurate the predicted concentrations are. The highest R^2 per heatmap is annotated with a x. Averaging the SERS map with a subset of spectra gives a better prediction of the concentration than with all the spectra in the map. Due to the low signal in the Picric acid (PA) dataset, we achieve a lower validation R^2 compared to other datasets.

Table D.1 Detection and quantitation performance on multiple data sets.

	Global accuracy			R^2		
	4-NBT	DNP	PA	4-NBT	DNP	PA
KNN	0.77	0.87	0.80	0.80	0.60	0.55
Random forest	0.73	0.90	0.72	0.56	0.20	0.44
SVM	0.73	0.90	0.84	0.75	0.65	0.56
GradientBoost	0.70	0.87	0.72	0.77	0.51	0.65
Xception	0.73	0.93	0.92	0.86	0.85	0.36
U-CNN	0.73	0.90	0.88	0.81	0.81	**0.81**
ResNet	0.79	**1.00**	0.88	0.78	0.59	0.24
ViT	**1.00**	**1.00**	**0.96**	**0.94**	**0.92**	**0.81**

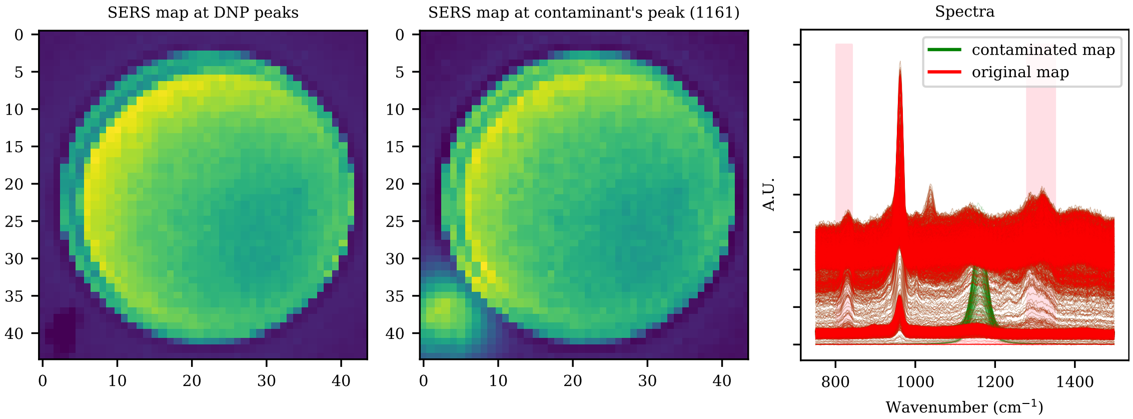

Fig. D.3 SERS measurements (DNP 100 nM) and contaminants. From left to right, we have the map that is calculated over the DNP peak regions (pink regions in the last subfigure), the map that is at the contaminant's peak, and the spectra (the pink regions denote the fingerprint region of DNP).

D.2 Detecting DNP in the presence of contaminants

A key use case for our method is detection of explosives in a scenario where strong contaminants are present[7]. We opt to further examine the performance of detecting DNP in the presence of contaminants, and since ResNet and ViT both have 100% accuracy for DNP in the setting with no contaminants, we focus our analysis on comparing these two methods. We have carried out experiments where the DNP measurements are contaminated with simulated contaminants at random wavenumbers, random intensities, and peak shapes (see Fig. D.3) following the procedure as explained in the Supplementary Information Section A.

We follow the procedure described in Section 2 from the manuscript to demonstrate the performance. During training, for each sample we add a random contaminant to 75% percent of training examples. Simulated contaminants with random intensity $\mathcal{N}(\mu_c, 10)$ are added to the SERS map, where μ_c =2843 is chosen as the 85-th intensity percentile within the DNP fingerprint region. We train the model for 300 epochs.Similar to the other experiments we perform leave-one-out crossvalidation, but evaluate the model's performance on 4000 examples created by combining the left-out test example with 4000 different contaminants. For the ResNet, we follow the established procedure and preprocess each sample to a single spectrum by averaging the top 5% of the spectra, which have the highest intensities across the DNP's fingerprint regions, as described in Section 2.3 in the manuscript. At test time, we also select the top 5% of the spectra per map (5% is the tuned hyperparameter, as shown in Fig.D.1 in the Supplementary Information). We include test results for three levels of contaminant amplitudes to demonstrate out of distribution performance.

The results are shown in Table D.2. No matter how strong the contaminants' signals are, the ViT achieves a detection accuracy $\geq$

Table D.2 Detection accuracy and balanced accuracy on samples that are created by mixing DNP measurements with random generated contaminants where the intensity of the contaminants follow $\mathcal{N}(\mu_c, 10)$.

Contaminant	Global accuracy		Balanced accuracy	
	ResNet	ViT	ResNet	ViT
None	**1.000**	**1.000**	**1.000**	**1.000**
$\mu_c = 1644$	0.964	**0.994**	0.967	**0.990**
$\mu_c = 2843$	0.962	**0.991**	0.957	**0.987**
$\mu_c = 4183$	0.957	**0.986**	0.936	**0.982**

Table D.3 Detection accuracy on samples that are created by mixing DNP measurements with random generated contaminants where the intensity of the contaminants follow $\mathcal{N}(\mu_c, 10)$. Each number shows the averaged accuracy of 4000 samples. We highlight the better performance (the difference between ResNet and ViT is larger than 0.02) within each contribution level. We see that ViT performs better than ResNet, especially when the signal of the contaminants is strong

Contaminant	Blank		1nM		10nM		100nM		1μM		10μM	
	ResNet	ViT	ResNet	ViT	ResNet	ViT	ResNet	ViT	ResNet	ViT	ResNet	ViT
μ_c=1644	0.983	0.985	1.000	1.000	0.981	0.993	1.000	0.992	0.820	**1.000**	1.000	1.000
μ_c=2843	0.960	**0.982**	1.000	0.999	0.982	0.987	1.000	0.989	0.832	**1.000**	1.000	1.000
μ_c=4183	0.915	**0.979**	1.000	0.996	0.978	0.976	1.000	0.980	0.846	**0.999**	1.000	1.000

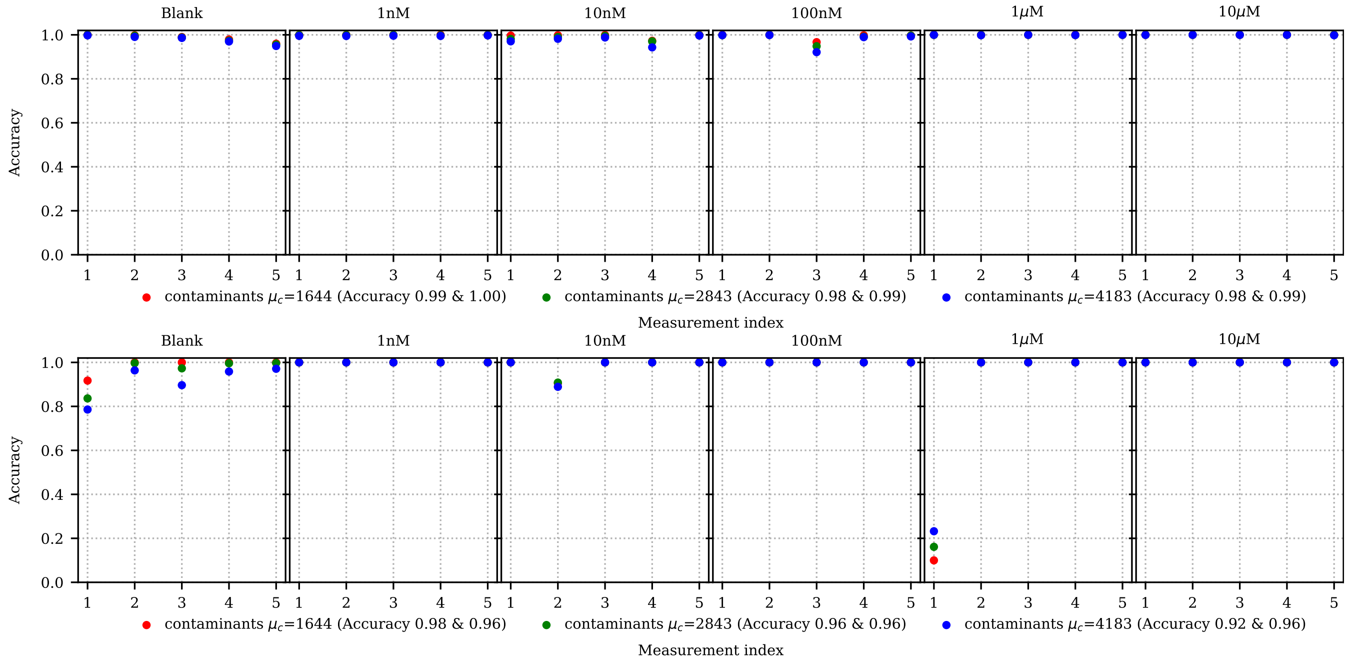

Fig. D.4 Detection performance using ViT (top) and ResNet (bottom). Each dot corresponds to the detection accuracy of 4000 samples where a SERS measurement is mixed with 4000 different contaminates. Different colour means the strength of the signal in the contaminants. A larger μ_c means that peaks of contaminants are with higher intensity. 4183 is the highest peak intensity within the DNP fingerprint regions in the measurements. In the legend, `Accuracy 0.99&1.00` means that ViT achieves on average 99% accuracy when DNP does not exist in the samples (Blank) and 100% accuracy when DNP exists in the samples (The rest of mixtures).

98%, whereas the ResNet is significantly less robust to the contamination. Detailing the results, Table D.3 and Fig. D.4 show that the ViT performs slightly worse on the samples that contain the 3rd measurement at 100 nM and contaminants. From Fig.5 in the manuscript, we see methods such as U-CNN, KNN, Random Forest, SVM, GradienBoost also have trouble of classifying this 3rd measurement (orange color coding) at concentration level of 100 nM.

D.3 Attention maps for DNP and Picric Acid datasets

We show the attention maps for datasets DNP and PA in this section. We observe a similar pattern here that ViT can capture the fingerprint characteristics better than the averaging scheme where we average a SERS map into a single spectrum.

Notes and references

1 M. D'Acunto, In situ surface-enhanced raman spectroscopy of cellular components: Theory and experimental results, Materials (Basel, Switzerland) 12 (9) (2019) 1564. doi:10.3390/ma12091564.
URL https://pubmed.ncbi.nlm.nih.gov/31086033

2 W. Demtroder, Laser Spectroscopy: Basic Concepts and Instrumentation, Springer Series in Chemical Physics, Springer Berlin Heidelberg, 2013.

3 S. Yang, X. Dai, B. B. Stogin, T.-S. Wong, Ultrasensitive surface-enhanced raman scattering detection in common fluids, Proceedings of the National Academy of Sciences 113 (2015) 268 – 273.

4 B. Li, M. N. Schmidt, T. S. Alstrøm, Raman spectrum matching with contrastive representation learning, Analyst (2022) –doi:

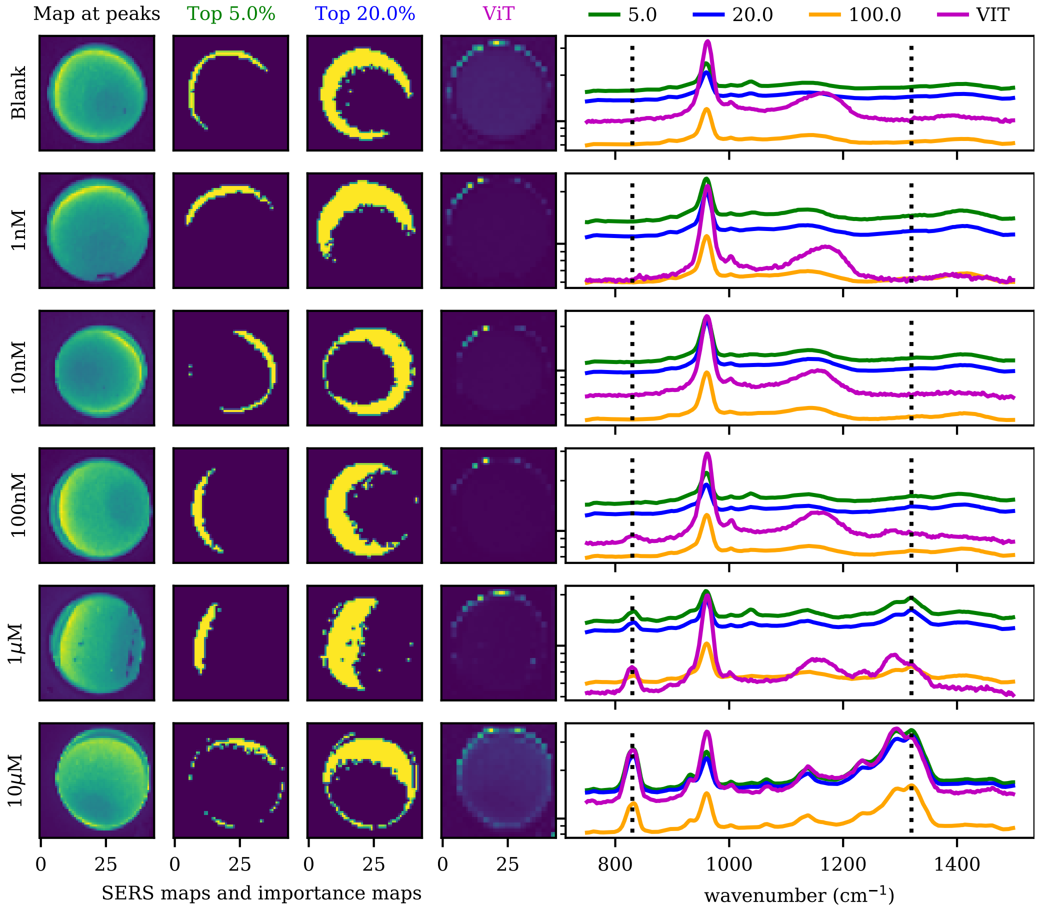

Fig. D.5 Example SERS maps and the selected spectra from dataset DNP. Each column from the left to right is the sum of the SERS maps from the peak locations, the locations of the top 5% and 20% spectra that are selected based on the peak intensities, attention map, and the corresponding spectra. Comparably, ViT can capture the fingerprint characteristics better than the peak-intensity selection criteria. The average over the entire map (100%) tends to alleviate the signal strength.

10.1039/D2AN00403H.
URL http://dx.doi.org/10.1039/D2AN00403H

5 J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, S. J. Gibson, Deep convolutional neural networks for raman spectrum recognition: a unified solution, Analyst 142 (2017) 4067–4074.

6 C.-S. Ho, N. Jean, C. Hogan, L. Blackmon, S. Jeffrey, M. Holodniy, N. Banaei, A. A. E. Saleh, S. Ermon, J. Dionne, Rapid identification of pathogenic bacteria using raman spectroscopy and deep learning, Nature Communications 10 (2019) 4927.

7 F. Lussier, V. Thibault, B. Charron, G. Q. Wallace, J.-F. Masson, Deep learning and artificial intelligence methods for raman and surface-enhanced raman scattering, TrAC Trends in Analytical Chemistry 124 (2020) 115796. doi:https://doi.org/10.1016/j.trac.2019.115796.
URL https://www.sciencedirect.com/science/article/pii/S0165993619305783

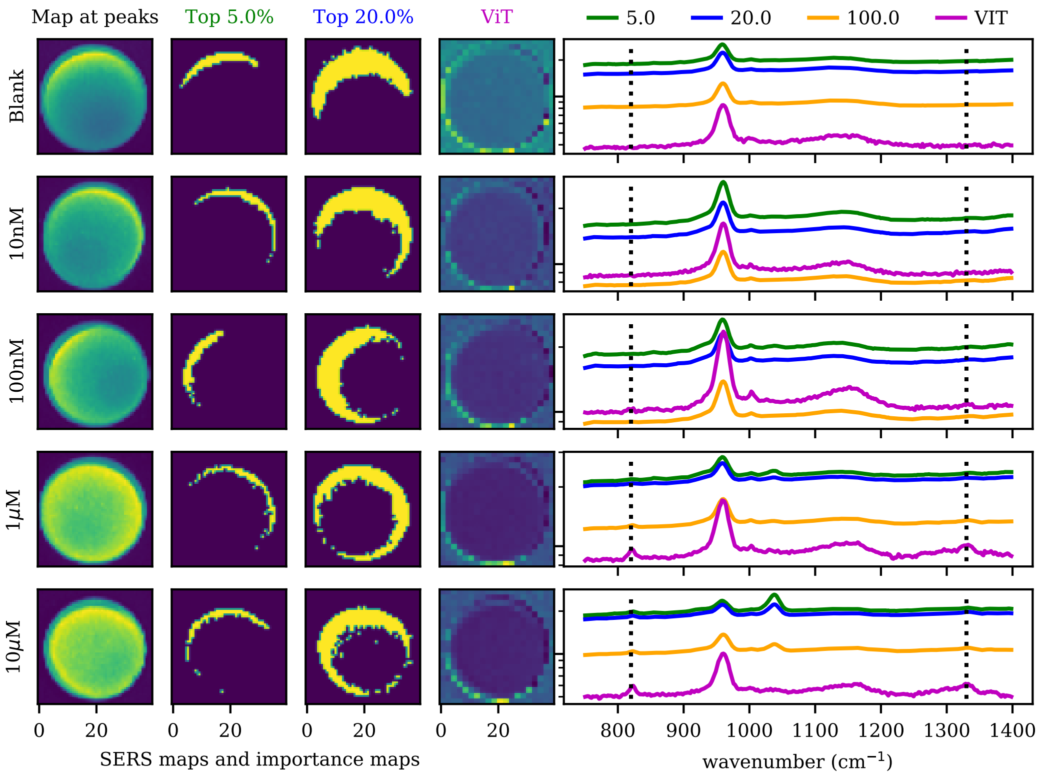

Fig. D.6 Example SERS maps and the selected spectra from dataset PA. Each column from the left to right is the sum of the SERS maps from the peak locations, the locations of the top 5% and 20% spectra that are selected based on the peak intensities, attention map, and the corresponding spectra. Comparably, ViT can capture the fingerprint characteristics better than the peak-intensity selection criteria. The average over the entire map (100%) tends to alleviate the signal strength.