Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

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] and Tommy S. Alstrøm[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Rapidly and accurately detecting and quantifying the concentrations of nitroaromatic explosives is critical for public health and security. Among existing approaches, explosives detection with Surface-enhanced Raman Spectroscopy (SERS) has received considerable attention due to its high sensitivity. Typically, a preprocessed single spectrum that is the average of the entire or a selected subset of a SERS map is used to train various machine learning models for detection and quantification. Designing an appropriate averaging and preprocessing procedure for SERS maps across different concentrations is time-consuming and computationally costly, and the averaging of spectra may lead to the loss of crucial spectral information. We propose an attention-based vision transformer neural network for nitroaromatic explosives detection and quantification that takes the raw SERS maps as input without any preprocessing. We produce two novel SERS datasets, 2,4-dinitrophenols (DNP), picric acid (PA), and one benchmark SERS dataset, 4-nitrobenzenethiol (4-NBT), which have repeated measurements down to concentrations of 1 nM to illustrate the detection limit. We experimentally show that our approach outperforms or is on par with the existing methods in terms of detection and concentration prediction accuracy. With the produced attention maps, we can further identify the regions with the higher signal-to-noise ratio in the SERS maps. Based on our findings, the molecule of interest detection and concentration prediction using the raw SERS maps is a promising alternative to existing approaches.

1 Introduction

With increasing attention to terrorist and chemical warfare attacks, efficient detection of nitroaromatic explosives can save lives[1–3]. Nitroaromatic compounds such as 2,4-dinitrophenol (DNP) and picric acid have been extensively used for preparing military-industrial materials[2,4]. The US Department of Homeland Security lists them as chemicals of interest. Commercially, these compounds are used in manufacturing dyes, pesticides, and wood preserves[5,6]. However, due to their high toxicity, contamination of food, water, or soil may bring severe physical or even irreversible damage to human eyes, skin, kidneys, liver, and heart muscle[7–12]. Therefore, developing a rapid and accurate method to detect nitroaromatic explosives is a pressing public health and safety need.

Many different methods have been used to detect and identify nitroaromatic explosives, including chromatography[13], mass spectroscopy[14], ion mobility spectroscopy[15], and electrochemical methods[16]. Despite the success of these approaches, they usually require complicated operational steps and well-trained technicians. Together with the expensive cost, it may limit their usage for the speedy detection of explosives[17]. Therefore, a cost-efficient, rapid, simple, and sensitive method is crucial for explosives detection and quantitation.

Surface Enhanced Raman Spectroscopy (SERS) has received great attention as a highly sensitive, reusable, and fast sensing platform for various tasks[18,19] such as identifying chemical hazards[20,21], bacteria[22,23], medical diagnosis[24,25], and food quality control[26]. By adsorbing analytes onto, e.g., a gold or silver nanostructure surface[18,27], SERS can enhance Raman scattering by a factor up to $10^{10} \sim 10^{11}$ [27].

Explosives detection with SERS is usually carried out by measuring the SERS spectra across multiple spatial locations to form a SERS map. Preprocessing methods, including smoothing[24], baseline subtraction, and normalization[18], are typically applied to the spectra to remove unwanted variations, such as instrumental artefacts[28]. In some cases, signal variation is further reduced

[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

by mapping onto a lower dimensional representation using techniques such as principal component analysis (PCA)[24,25,29]. Next, the spectra are typically averaged across the entire or a suitably selected subset of the SERS map[25] to yield a single noise-reduced spectrum. Finally, the processed spectrum is used as the input in algorithms such as support vector machines (SVM)[30], logistic models[31], k nearest neighbours (KNN)[32], neural networks[22], and others[18] for detection, classification, or quantitation.

Averaging SERS maps properly and using proper preprocessing steps are essential for these methods to work well in practice[18,33]. However, designing an averaging scheme that can effectively extract the fingerprint characteristics from a SERS map is time-consuming and complicated, especially in real-world scenarios where contaminants are unavoidable[18]. Averaging a SERS map into a single spectrum may also inadvertently remove useful information, and selecting appropriate preprocessing usually requires extensive domain knowledge[28]. To alleviate these problems, we investigate how to perform explosive detection using raw SERS maps without preprocessing.

The modality of a SERS map is similar to an image in the computer vision domain, with the key difference that each pixel in a SERS map represents a SERS spectrum. Deep neural networks have achieved great success in computer vision tasks such as image classification, object detection, and reconstruction[34–37]. Inspired by this, we investigate how recent advances in image analysis can transfer to the analysis of SERS maps.

In this paper, we propose a deep neural network model based on a vision-transformer (ViT)[36,38] architecture. The ViT model achieves state-of-the-art results for image processing tasks such as classification[36], object detection[39], and image segmentation[40]. It is based on a spatial attention mechanism, which furthermore provides valuable model interpretation. We train a ViT using the raw SERS maps as input, and our method requires no preprocessing or spectral averaging.

To demonstrate its use for detecting and quantifying explosives, we produce two novel datasets consisting of SERS maps of two nitroaromatic explosives, DNP and picric acid. The datasets consist of repeated measurements at several concentration levels ranging from 1 nM to 10 μM and blank measurements. These datasets are made publicly available to serve as future benchmarks. To demonstrate the generality of our approach, we additionally test our ViT model on an existing 4-NBT dataset. We achieve better results for both detection and quantitation than the previous state-of-the-art in all our experiments, indicating that the ViT is more efficient in extracting information from noisy measurements. Besides, our reported performance on multiple repeated measurements and experiments are more reliable and unbiased than the existing approaches that demonstrate detection limit with a single chip experiment[41].

As a key benefit, the ViT uses data to learn which parts of the SERS map are most important, and it produces attention maps that can be used to interpret fingerprint characteristics. At low concentrations, in particular, we find that the ViT focuses on the edges of the SERS substrate, where it has been cut from a larger wafer. Further analysis indicates that the signal-to-noise ratio is exceptionally high in this region. The result indicates that the preferred analyte binding areas are located at the edge of the SERS substrate.

Compared with recent approaches that use convolutional neural networks, such as[42], our method has several advantages, 1) we do not need any preprocessing, 2) we do not split a single SERS map into both training and testing datasets which potentially can confound results (information leakage), and 3) our method is trained end-to-end without any auxiliary tasks.

Our main contributions are:

- An optimized attention-based vision transformer that achieves state-of-the-art results in the detection and quantitation of nitroaromatic explosives using raw SERS maps.
- Two novel publicly available SERS datasets with repeated measurements of nitroaromatic explosives at low concentrations.
- A method for producing interpretable attention maps for locating important spatial regions in the SERS maps, which in our experiments points to the observation that the SERS signal is strongest at the edges of the substrate.

2 Materials and methods

2.1 Chemicals and SERS maps measurements

The molecule set as a benchmark is 4-Nitrothiophenol (4-NBT), technical grade (80% of purity) and purchased from Sigma-Aldrich. 4-NBT was solubilized in 50 ml absolute Ethanol (EtOH) ($\geq$ 99.8%) (VWR Chemicals BDH, VWR International, Radnor, Pennsylvania, USA) to prepare 10 mM stock solution. The stock solution was diluted in absolute EtOH to prepare standards of 1 μM, 100 nM, 10 nM, 1 nM and 100 pM concentrations.

The first nitroaromatic explosive used is 2,4-Dinitrophenol (DNP). The molecule was purchased from Sigma-Aldrich ($\geq$ 98% purity) and moistened with water. DNP was solubilized in Methanol (MeOH) (VWR Chemicals BDH, VWR International, Radnor, Pennsylvania, USA) to prepare a 100 μM stock solution. The stock solution was diluted in Milli-Q water to prepare standards of 10 μM, 1μM, 100 nM, 10 nM, and 1 nM concentrations.

The second nitroaromatic explosive is 2,4,6-Trinitrophenol or picric acid, provided by the CBRN Defence and Security Division of the Swedish Defence Research Agency FOI. Picric acid was solubilized in ultrapure water (18.2 MΩcm) from a Milli-Q purification device (Milli-Q® IQ 7000 Purification System, Merck, Darmstadt, Germany) to make a 44 μM stock solution. The stock solution was then diluted in Milli-Q water to prepare standards of 10 μM, 1 μM, 100 nM, 10 nM, and 1 nM concentrations.

2.1.1 SERS substrates fabrication

The SERS substrates used for the analysis were silver (Ag) covered silicon (Si) nanopillar (NP) structures fabricated using a two-step fabrication process: (i) maskless reactive ion etching (RIE) of Si and (ii) electron e-beam evaporation of Ag[43,44]. In brief, an RIE process with a SF_6 and O_2 gas mixture flow was performed for 5:20 minutes in ICP Metal Etcher (PRO ICP, SPTS Technologies Ltd., Newport, UK). Two polished 6-inch black Si wafers were etched to produce vertically standing Si NP structures with

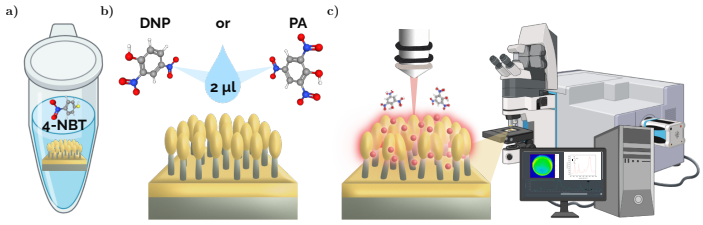

Fig. 1 Protocols and experimental procedures performed for SERS analysis. (a) 2 h incubation of SERS substrates in 4-NBT solutions of different concentrations. (b) 2 µl droplet deposition of DNP and picric acid solutions at different concentrations. (c) SERS chip performance evaluation and map acquisition. (This figure is created with BioRender.com.)

the following features: NP density 20 NP/μm^2, height 640 nm, width 50 nm. The second step of the fabrication process consists of metal deposition of 200 nm of Ag, using a Temescal FC-2000 e-beam evaporator from Ferrotec (Tokyo, Japan) (Fig. A1.a). The resulting SERS substrates are Ag-capped Si NPs (Fig. A1.b). Finally, the processed Si wafers were diced into 3 × 3 mm chips from the backside using a laser micromachining tool (D-09126, 3D-Micromac AG, Chemnitz, Germany) and stored in a desiccator (Fig. A1.c).

2.1.2 SERS measurements

The 3 × 3 mm^2 SERS chips were then exposed to 4-NBT, DNP and picric acid solutions of varying concentrations (5 chips per concentration). In the case of 4-NBT, the SERS substrates were incubated for 2 h in 4-NBT solutions, washed with EtOH to remove excess (unbonded) 4-NBT and then left to dry (Fig. 1.a). In the case of DNP and picric acid, 2 μl analyte droplets were deposited on substrates from the same wafer and left to dry (Fig. 1.b). Each analyte was prepared and used on different days to avoid cross-contamination on the SERS substrates. The SERS chip maps were acquired using a DXRxi Raman Microscope (Thermo Scientific, Waltham, MA, USA) (Fig. 1.c). All SERS maps were acquired using a 780 nm laser excitation wavelength, a 10x objective, optical focus on the chip surface, a single acquisition per point, an exposure time of 0.05 s, and a step size of 50 μm. 4-NBT SERS maps were acquired using a laser power of 2 mW, while in the case of DNP and picric acid, the laser power was set to 5 mW. The SERS maps were produced by mapping the entire 3 × 3 mm^2 chip area (between 1800 and 2000 spectra per chip). See Fig. 2 for an example of the SERS spectra.

2.2 Methods

We propose a vision transformer (ViT) neural network architecture tailored for low concentration detection and quantitation from SERS maps. After outling the major components of the architecture, we explain each of the components in detail in the following sections. A schematic overview of the model is given in Fig. 3.

The input to the system is a SERS map, which is first split into small patches which cover multiple spectra. Each patch is then mapped to a feature vector using a shared neural network and augmented with a learned spatial embedding vector. Each patch

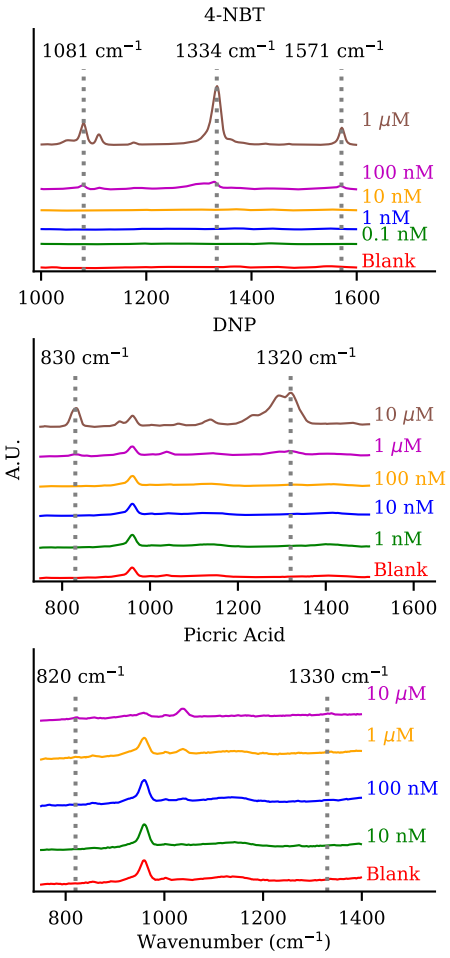

Fig. 2 Example spectra at different concentration levels. We here select the top 1% of spectra with the highest fingerprint peak intensities from each SERS map and show the averaged spectrum. The fingerprint peaks are more obvious at higher concentrations. The prominent fingerprint peaks for picric acid (820 cm^{-1} and 1330 cm^{-1}) are not visible to the naked eye even at the highest concentration of 10μM. The normalised SERS spectra are shown in the Supporting Information B.2

is thus represented as a single vector, referred to as a token, which includes information about the spectra and the spatial location of the patch. From this point, the data is considered as a unordered set of tokens. In addition to the patch tokens, a special token called the class embedding (CLS) token is included, which serves the purpose of representing the entire map. The initial class embedding is taken to be a learnable parameter. Next, the tokens are updated through a series of self attention steps in order to contextualize the tokens and capture their dependencies. Each token is updated using an attention weighted sum of all the other tokens, where the weights are computed by a measure of compatibility between the tokens. After a number of such self attention updates, the value of the CLS token is taken as the final representation for the whole map. Finally, we attach a prediction head in the form of a neural network classifier or regression model to give the final output for detection or quantitation. The entire model is trained supervised end-to-end, on a labelled dataset.

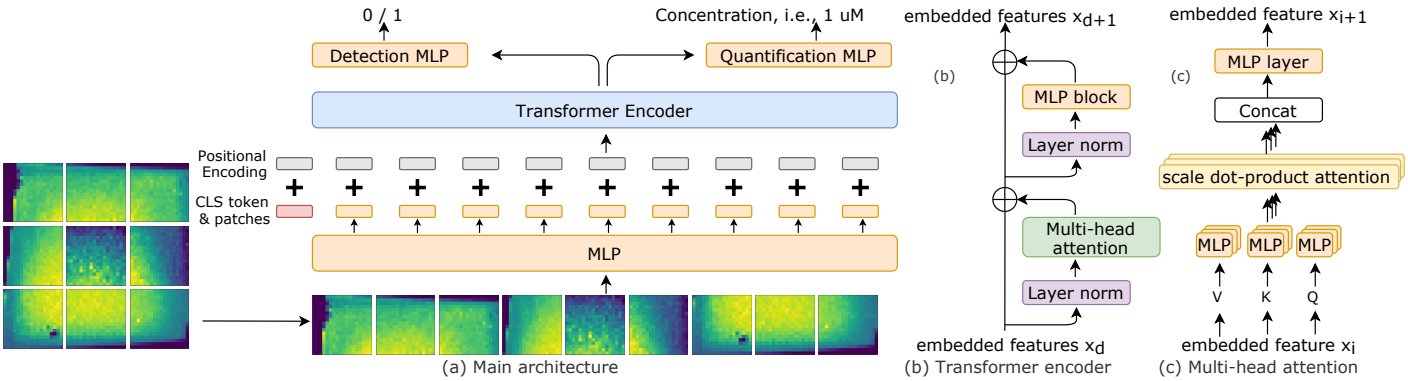

Fig. 3 Our proposed approach. A vision transformer neural network architecture for detecting or quantifying the concentration of the molecule of interest using SERS maps. Given a SERS map as the input, we split it into patches and apply a shared MLP layer on each patch to create a feature vector. (The illustration shows a large patch size here, whereas, in practice, we use patches of size 2×2 pixels.) We then append a learned CLS token at the beginning of the feature vector and add positional encoding on each feature. The obtained features are used as the input for a transformer encoder, as shown in (b) and (c), to produce the attention map that indicates the importance of each patch for decision-making. The attention map is then used as the input for an MLP layer for performing either a detection task or a concentration prediction task.

2.2.1 Patch embeddings

Given a SERS map $X \in \mathbb{R}^{N_x \times N_y \times N_w}$, where N_x, N_y are the width and height of the SERS map and N_w is the number of wavenumbers, we reshape the SERS map into a set of flattened square 2D patches $x_m \in \mathbb{R}^{N_p \times (m^2 \cdot N_w)}$ where m is the patch width and height, and $N_p = \frac{N_x N_y}{m^2}$ is the number of patches. A shared multilayer perceptron (MLP) layer is then applied to each patch to create feature embeddings.

Following Dosovitskiy *et al.*[36], Devlin *et al.*[45], we prepend a learnable class embedding (CLS token) to the feature vector, which serves to aggregate the learned representation from the entire input[45]. As there are no convolutional operations to preserve the spatial information, we add learnable positional encodings to the input embeddings to capture the relative and/or absolute spatial position of a patch[46] and provide the sum as the input to the transformer encoder.

2.2.2 Transformer encoder

The transformer encoder is a stack of D transformer layers, where each layer consists of a multi-head self-attention layer followed by an MLP block (Fig. 3.b). Each transformer layer takes a set of feature embeddings as the input and outputs a new set of features with the same dimensionality. We also employ residual connections[35] and layer normalization[47]. Depending on the orders of layer normalization, residual connection, multi-head attention, and MLP blocks, there are different variants of transformer architectures in the literature. In this study, we use layer norms prior to both the multi-head attention and MLP since this architecture has been demonstrated to be more efficient and produce more stable gradients, especially at the beginning of the network training[48,49].

The attention head is composed of three vectors (query, key, and value) computed from the embedded feature of each token by an MLP. For each token, the attention to every other token is computed as the scale dot-product similarity between its query and their keys, and the output of the attention head for each patch is the sum of the values weighted by the attention. We use H attention heads in parallel as shown in Fig. 3.c as it is more beneficial to extract information from different representation spaces[46] simultaneously. We concatenate the multi-head attention outputs and project them into the final output using an MLP. We then use the output from the multi-head attention block as the input for another layer normalization and MLP block as shown in Fig. 3.b.

2.2.3 Prediction head

After applying the transformation layer D times (Fig. 3.b), the updated CLS token is taken as the attention map that shows an aggregation of the representations over all the patches. This attention map then serves as the input for a prediction head that consists of a single linear layer. For the detection task, we output the probability of a SERS map containing the explosive molecule. For the quantification task, we directly produce the predicted concentration.

2.3 Comparison baselines

We compare our approach with established deep neural networks as well as classical machine learning methods that take an averaged spectrum from a SERS map as the input. Spectral averaging combined with a suitable supervised machine learning method is an often used practical approach[18,26,30,50–53] and we consider this to be the current state of the art. Different preprocessing methods exist for calculating the averaged spectrum from a SERS map, such as averaging over the entire map or averaging over a subset of spectra per map[33]. To set the baselines in the best light, we assume access to information about the Raman peak regions of the analyte, and following[33] we choose the spatial region based on the signal intensity summed over the peak regions. We rank the spectra in the SERS map according to

$$P_{x,y} = \sum_{w \in \mathscr{P}} X_{x,y,w}, \tag{1}$$

where $\mathscr{P}$ denotes the set of wavenumbers that are identified as peak locations, and we then select the top α percentage of spectra according to $P_{x,y}$. Fig. 4 shows an example of selecting the top 5%

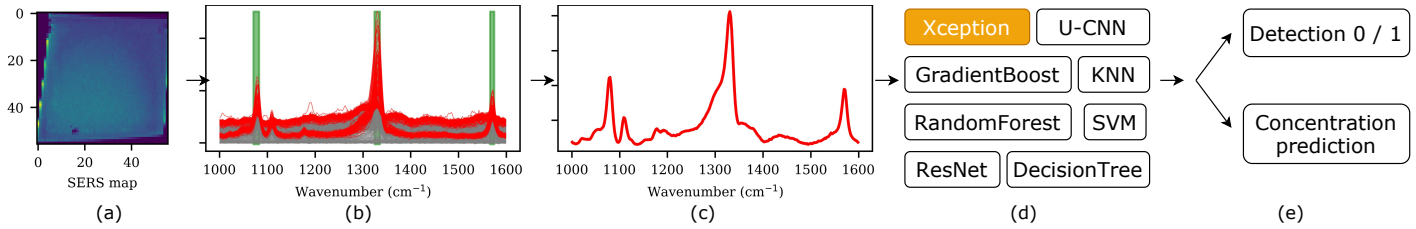

Fig. 4 Baseline approaches. Explosive detection and quantitation using the average spectrum per SERS map. For example, (a) shows a 4-NBT SERS map with $56 \times 56 = 3136$ spectra. We first rank the extracted spectra from the SERS map in (a) based on the summed intensities over the peak regions (green) and use the ranking to select a subset, here chosen as the top 5% = 157 spectra (red) in (b). The average of the selected spectra, as shown in (c), is used as the input in a supervised machine learning model in (d) (e.g., an Xception deep neural network), which predicts the detection and concentration results.

spectra according to this criterion. Results for other selection criteria (such as spectral mean or standard deviation) can be found in Supporting Information A.

We average the selected spectra to produce a single spectrum per SERS map and repeat this process for all the SERS maps. As different choices of α can influence the averaged spectrum, we optimize the averaging scheme by performing a grid search to select both the percentage of spectra used for training, α_{train}, and for performance evaluation at test time, α_{eval}.

Using the averaged spectrum from a SERS map for tasks such as detection and concentration can be used with many different supervised machine learning methods. In this paper, we choose eight different models that have achieved success in many different domains: K Nearest Neighbors (KNN), Gradient Boosting[54], Random Forest[55], Support Vector Machine (SVM)[55], Decision Tree[56], Xception[32,34,57], U-CNN[58], and ResNet[22,35]. Detailed information about each method is shown in Supporting Information C. We denote these as "spectra-based models" and train them in a similar fashion as ViT as documented below.

2.4 Model training

We train the detection model using the binary cross-entropy objective with an L_2 regularization term as shown in Eq. 2a, where θ denotes the parameters in the neural network. We train the quantification model using mean-squared-error (Eq. 2b) between the predicted concentrations $\hat{c}_i$ and the true log-concentration c_i as this can emphasize the prediction of the lower concentrations.

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1-y_i)\log(1-p_i) + \lambda_1 ||\theta||^2, \tag{2a}$$

$$L = \frac{1}{N}\sum_{i=1}^{N} ||c_i - \hat{c}_i||_2^2 + \lambda_2 ||\theta||^2. \tag{2b}$$

2.4.1 Data preparation and augmentation

As the number of SERS maps is limited, we perform leave-one-out cross-validation to demonstrate our model performance. In each experiment, we leave one SERS map out as the test data. We then select one map per concentration from the rest of the data and use those as a validation dataset to optimize model hyperparameters. The rest of the data are used as the training data. There is no overlap between training, validation, and test data. We repeat this process N times where N equals the number of SERS maps per dataset.

Since the performance of a deep neural network is highly influenced by the size of the dataset[32,58], it is common to augment the data, for example, by adding slightly modified copies of the existing data. We propose to augment the training and validation dataset following[32]. We augment each spectrum in the SERS map by adding noise simulated to mimic realistic spectra variation with the highest variance near the peaks. See Supporting Information C for more details.

2.4.2 Ensembling

Rather than using a single model, we use an ensemble of models to boost the detection and quantification performance[57,59–61]. Following[62], we train an ensemble of five models with the same training data but different random initialization. We then use the averaged predictions from the ensemble to make decisions for the detection and quantification tasks.

2.4.3 Evaluation criteria

We evaluate the detection accuracy using Eq. 3a and quantification accuracy using Eq. 3b as these are commonly used metrics for demonstrating the classification accuracy and goodness-of-fit[42]. For optimizing the preprocessing procedure in the baseline approaches, we evaluate the validation binary cross-entropy loss (Eq. 2a with $\lambda_1 = 0$) for the detection task and R^2 (Eq. 3b) for the quantification task across multiple $\{\alpha_{\text{train}}, \alpha_{\text{eval}}\}$. We then adopt the combination of $\{\alpha_{\text{train}}, \alpha_{\text{eval}}\}$ that achieves the lowest validation loss in the detection task and the highest R^2 in the quantification task for evaluating the test performance on each dataset and each model.

$$\text{Global accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}. \tag{3a}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(c_i - \hat{c}_i\right)^2}{\sum_{i=1}^{N}\left(c_i - \frac{\sum_{i=1}^{N} c_i}{N}\right)^2}. \tag{3b}$$

2.4.4 Experimental setup

For the ViT model, we use $D = 2$ stacked transformer encoders and $H = 3$ attention heads. We choose to use a patch size of two-by-two pixels ($m = 2$) to reduce the computational cost but also to get fine-grained attention maps. For the spectra-based deep neural network models, the architectures are explained in Supporting Information B. For the grid search, we include α-values

0.2%, 0.5%, 1.0%, 2%, 5%, 10%, 20%, 50% and 100%, where the latter corresponds to averaging the entire SERS map. We train all the models in our study with the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9. Following[36], we use a linear learning rate warmup up to 50 epochs and cosine learning decay until the learning rate reaches 10^{-5}. The learning rates are tuned experimentally based on the validation performance. All the weight matrices W are initialised by the Xavier initialization[63] such that each element in the weight matrices $\mathbb{R}^{n_{\text{in}},n_{\text{out}}}$ are drawn from Gaussian distribution $\mathcal{N}(0, \frac{2}{n_{\text{in}}+n_{\text{out}}})$. The bias vectors are initialized by zeros. We also apply gradient clipping at global norm one as we found that it can help to stabilize the training[36,49].

3 Results and discussion

We report and compare the performance between the model that takes a raw SERS map as the input and the models that take the averaged spectrum as the input in this section. We quantitatively show the detection and concentration prediction performance over the three datasets: 4-NBT, DNP, and picric acid.

3.1 Detection performance

We first look at the influence of the choices of $\{\alpha_{\text{train}}, \alpha_{\text{eval}}\}$ on the detection performance in the baseline approaches in Fig.D.1 (Supporting Information D). We observe that the optimal $\{\alpha_{\text{train}}, \alpha_{\text{eval}}\}$ tends to be different for different datasets and models. For example, we benefit more by using a bigger α_{train} and a smaller α_{eval} for the 4-NBT dataset and benefit more by using similar α_{train} and α_{eval} (diagonal) for the DNP and picric acid dataset. Compared to averaging the entire SERS map into a single spectrum, we perform better by averaging a subset of spectra. These observations indicate that selecting a generally appropriate preprocessing procedure for analyzing SERS maps in the traditional spectra-based approaches is difficult and usually requires domain knowledge.

We then report the test performance using the baseline approaches by selecting the most optimal combination of $\{\alpha_{\text{train}}, \alpha_{\text{eval}}\}$ and our proposed approaches in Fig. 5. Each marker in Fig. 5 represents the predicted probability of the correct class for a single SERS map measurement. We count predictions as correct if the probability of the correct class is higher than 50% (50% is a commonly used threshold for classification/detection tasks in the literature[22,40]). Different colours represent the measurement indices within a single concentration level, allowing direct comparison across methods. The same colour across different concentrations corresponds to different measurements. The averaged predicted probability per concentration level is shown as the red line. The spectra-based models tend to make more and the same mistakes at lower concentrations. For example, all the spectra-based models make a wrong classification for the same (orange) measurement at a concentration of 0.1 nM. For all the datasets, ViT performs better or is on par with the methods that take the spectra as the input without the need to search for the most appropriate way of preprocessing the SERS maps. The slightly worse performance on the picric acid may be due to the low signal strength as shown in Fig. 2.

3.2 Quantification performance

In addition to detection, it is beneficial to quantify the concentration of the explosives, as different strategies may be required given the concentration of the explosives. Therefore, we next show the quantification experiment results.

We follow the same procedure as for the detection experiment: we first choose α_{train} and α_{eval} based on the validation performance as shown in Fig.D.2 (Supporting Information D). We observe a similar trend with best results near the diagonal, corresponding to approximately the same percentage of spectra used for training and testing. Again, averaging over a subset of spectra is better than averaging the entire SERS map, but the optimal selection of α_{train} and α_{eval} depends strongly on the data and to a slightly less degree on the model. We take the optimal combination of α_{train} and α_{eval} for each model and apply them to the test dataset to report the quantification test performance.

Fig. 6 shows the quantification performance across multiple datasets and models. Each marker represents the predicted concentration for a single SERS map measurement, with the true concentration on the x-axis. Following Fig. 5, we use the same colour coding here such that each colour represents a single measurement within one concentration level and the same colour over concentrations corresponds to different measurements. The red line shows the averaged predicted concentration with a 95% confidence interval $\frac{1.95\sigma(\hat{c})}{\sqrt{5}}$. The grey diagonal line indicates the optimal prediction.

All the models give very accurate predictions at the highest concentration level. However, classical spectra-based models, such as Random Forest, Decision Trees, SVM, and Gradient Boost, can hardly differentiate between the measurements at low concentrations due to the weaker signal. Additionally, they tend to make the same mistake as in the detection experiment (Fig. 5). For example, all the spectra-based models predict a higher concentration value for one of the measurements at 0.1 nM (orange) for 4-NBT, which is also predicted wrong by all the spectra-based models in the detection task. Our model ViT outperforms or is on par with the methods that take the spectra as the input. We can better differentiate the concentrations until the second lowest concentration level as the 95% confidence interval are non-overlapping until the second lowest concentration level.

3.3 Attention maps

We next provide an analysis of why and how ViT performs better than the baseline approaches. Specifically, we show and compare the selected spectra from the baseline approaches and the attention map from our proposed method. Fig. 7 shows an example of using the 4-NBT dataset. The attention maps for other datasets are described in Supporting Information D.

We take a single SERS map from each concentration and show the sum of the SERS maps at the fingerprint peak locations in the first column in Fig. 7. To understand what kind of spectra are contributing more to learning the models, we show the selected spectra locations based on the peak intensities from the spectra-

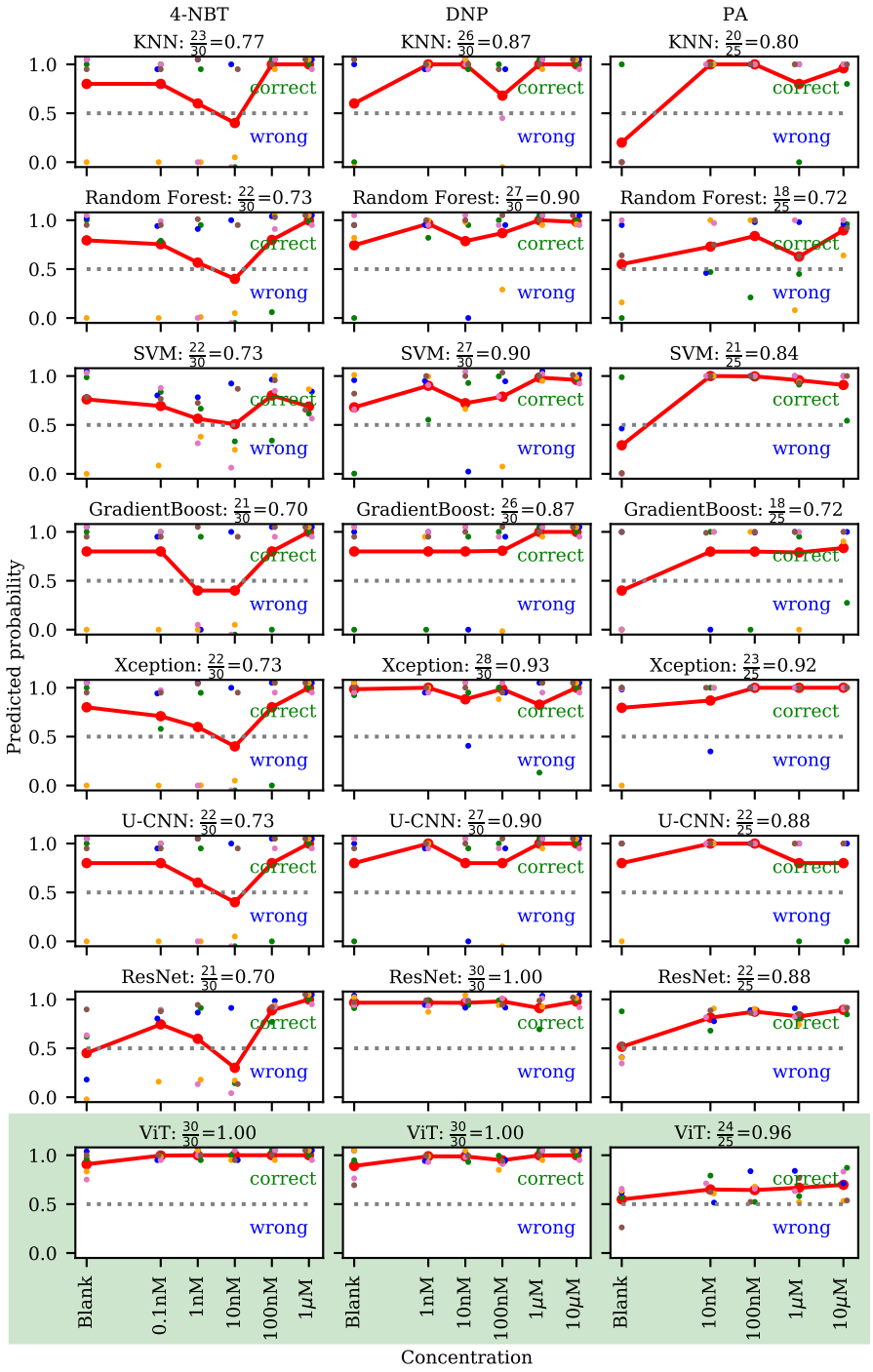

Fig. 5 Detection performance on multiple datasets. Each marker represents a single measurement, and the colour represents the measurement index within a single concentration level. The same colour across concentrations corresponds to different measurements. We achieve better detection accuracy using ViT on all the datasets compared to the spectra-based methods.

based models and the attention maps from the ViT. The attention map is an aggregation of the learned representations over all the patches and thus can be used to infer the spectra that contribute more to decision-making [46].

We choose to select the top 5% and 20% of spectra based on the peak intensities and annotate their locations in a map with the same shape as the SERS map. For example, the yellow dots in the third column in Fig. 7 means the corresponding spectra are used to calculate the averaged spectrum as shown in the fifth column in Fig. 7. We show the averaged spectrum for when α is 5%, 20%, and 100% in the last column. For the ViT model, we select the locations whose attention weights are higher than the 99%-th quantile of the attention map and average the spectra within those locations. Note that we learn the ViT model with the raw SERS maps and only average the spectra here for visual comparison purposes.

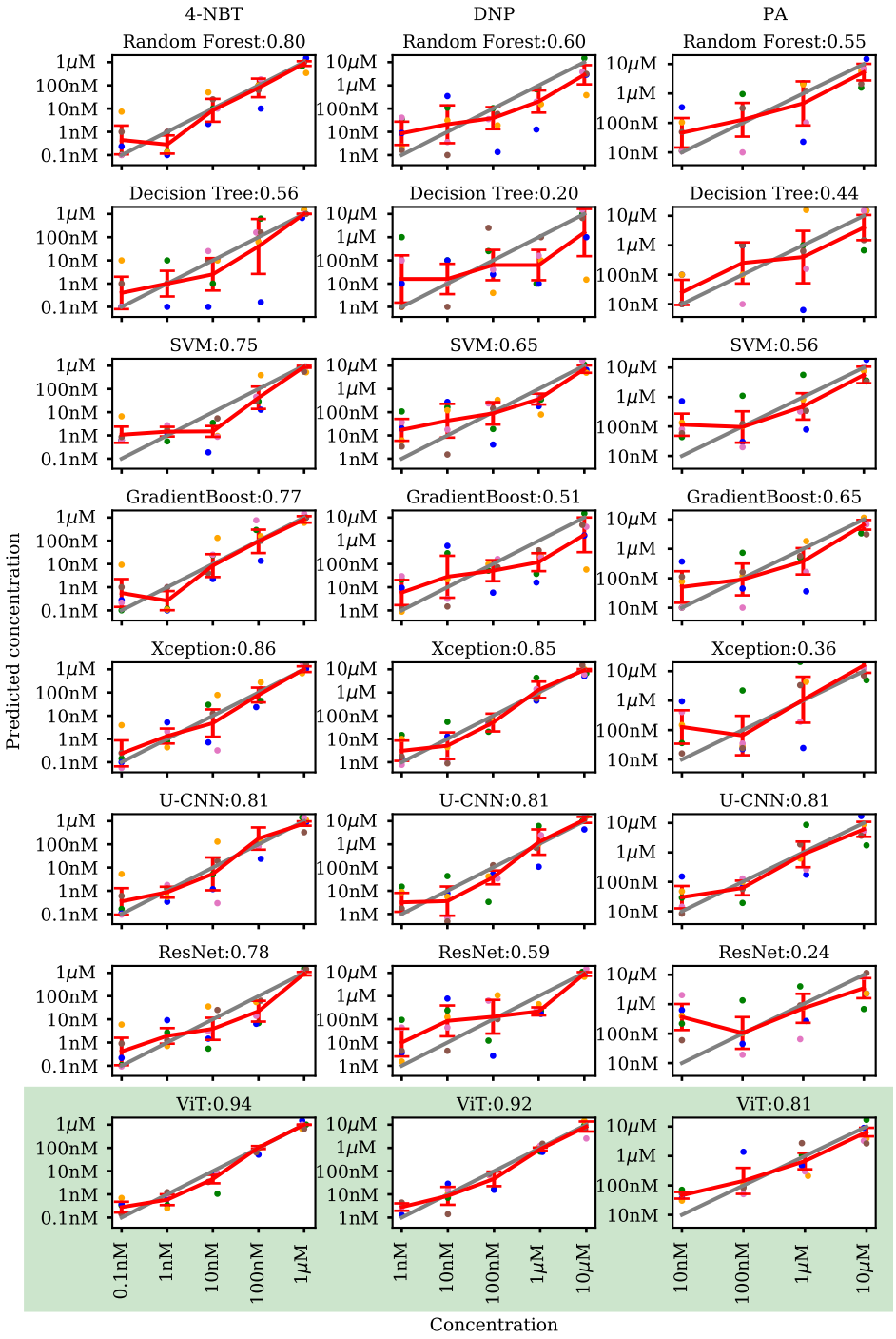

Fig. 6 Quantification performance. Predicted concentrations together with 95% confidence interval over five runs ($1.96\frac{\sigma(\hat{c})}{\sqrt{5}}$). Each marker represents a single measurement, and the colour represents the measurement index within a single concentration level. We achieve better performance using ViT on all the datasets compared to the spectra-based methods

As the signal is strong (for the concentration of 1 μM), we can observe clear peaks at the fingerprint regions no matter how we average the spectra. However, when we inspect lower concentrations such as 1 nM, spectra-based methods select more spectra at the centre of the SERS maps, whereas ViT focuses more on the edges maps. We can only observe the fingerprint peaks at wavenumber 1081 cm^{-1} and 1571 cm^{-1} when we look at the spectrum from ViT. We also observe a similar pattern for DNP and picric acid (the results can be found in Supporting Information D) that the attention maps can capture the fingerprint peak better. Therefore, it shows that the ViT is better at utilizing the fingerprint characteristics of the corresponding molecule compared to the methods that select the spectra based on the peak intensities.

The above results indicate that the preferred analyte binding

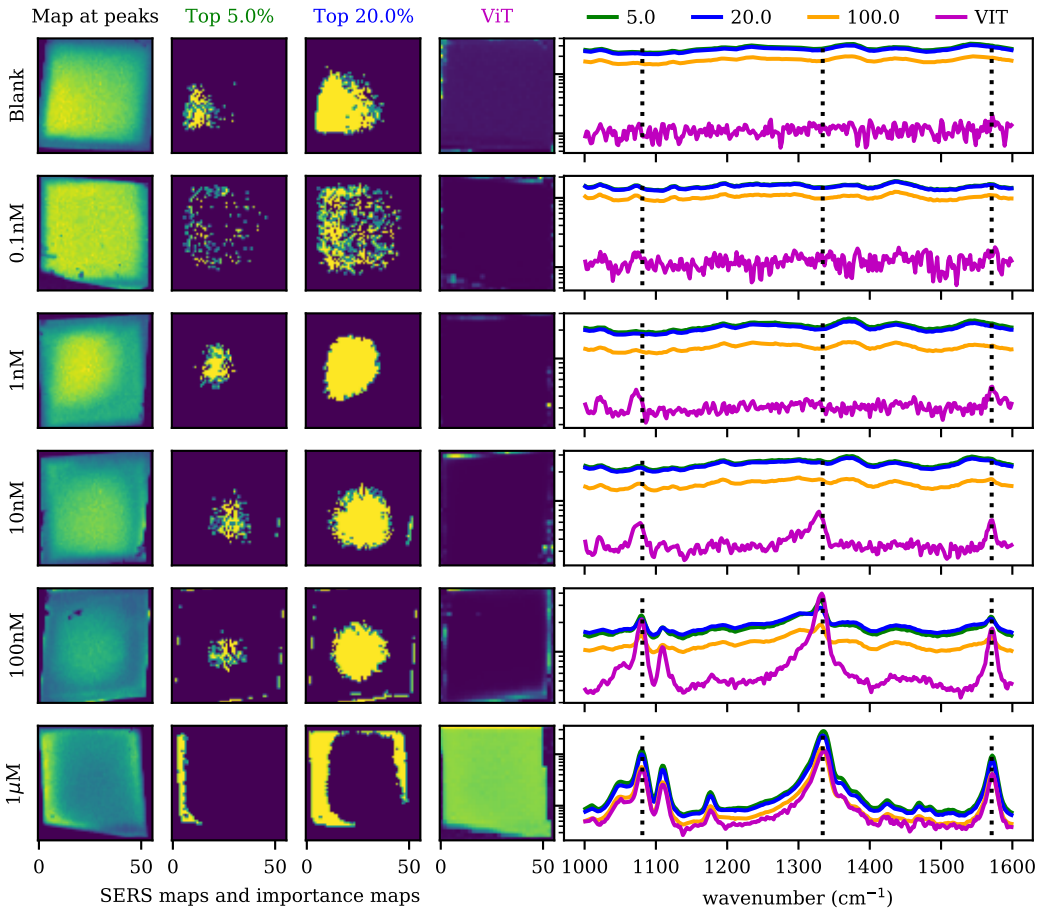

Fig. 7 Example SERS maps and the selected spectra from dataset 4-NBT. Each column from left to right is the sum of the SERS maps from the peak locations, the locations of the top 5% and 20% spectra that are selected based on the peak intensities, attention map, and the corresponding spectra. Comparably, ViT can capture the fingerprint characteristics better than the peak-intensity selection criteria. Averaging over the entire map (100%) tends to alleviate the signal strength, especially at low concentrations.

areas are located at the edge of the SERS substrate. The drying of the analyte is usually denoted as the *coffee ring* effect[64]. Despite this effect and the inhomogeneous distribution of the SERS signal, our proposed approach is robust and accurate compared to the approaches where spectra are extracted based on statistical criteria.

4 Conclusion

In this paper, we proposed to use a vision transformer-based neural network for nitroaromatic detection and quantification with Surface-enhanced Raman Spectra maps. Given a raw SERS map as the input, we trained the vision transformer to learn an attention map that showed which regions in the SERS map contained the best representative fingerprint characteristics. We then performed detection and quantification of the explosives based on the attention maps. To demonstrate the benefit of our approaches, we produced two novel nitroaromatic explosive datasets consisting of DNP and picric acid and one benchmark dataset, 4-NBT. We make these datasets publicly available for the benefit of future benchmarking. We empirically demonstrated that our proposed approach is more accurate and efficient in identifying and quantifying explosive compounds, especially at lower concentrations. Our method uses raw SERS maps and exempts us from designing preprocessing procedures, which usually require extensive domain knowledge. Our model applies to all planar SERS substrates to detect SERS-active analytes in the gas/liquid phase as we have no requirements on the SERS substrates that are being used. We experimentally observed that we only need around 0.4 seconds to perform explosive detection and quantification for each SERS map on a computer equipped with a medium-grade CPU. The computational software will be released upon publication.

By taking a closer look at the attention maps, we discovered that the signal-to-noise ratio is higher on the edges of the SERS substrate. The observation indicates that in the low-concentration regime, analyte binding is most efficient at the boundary of the chip. This could be attributed to the intrinsic nanopillar SERS substrate properties utilized in this study, which is likely related to the nanopillar leaning effect. Importantly, the results show that the ViT model can be used to extract quantitative SERS data from a SERS map displaying inhomogeneous analyte binding patterns, which is highly relevant for researchers working on real-life SERS applications that utilize different types of SERS substrates.

Data and code availability

We implement all the models in Python 3.7.9 using PyTorch Lightning 1.5.1. For the reviewing process, we have provided a private link `https://figshare.com/s/0596e7a36420911d28c3` that contains the experiments (model checkpoints for the ViT models), SERS maps (4-NBT, DNP, and picric acid), and the software, which can be used to reproduce the results in the paper. Detailed instruction is explained in the software description. We will open-source this information upon paper publication.

Conflicts of interest

There are no conflicts to declare

Acknowledgements

The authors thank for financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 883390 (H2020-SU-SECU-2019 SERSing Project). The authors thank the NVIDIA Corporation for the donation of GPUs used for this research.

Notes and references

1 M. H. Wong, J. P. Giraldo, S.-Y. Kwak, V. B. Koman, R. Sinclair, T. T. S. Lew, G. Bisker, P. Liu and M. S. Strano, *Nature Materials*, 2017, **16**, 264–272.

2 X. Chen, C. Sun, Y. Liu, L. Yu, K. Zhang, A. M. Asiri, H. M. Marwani, H. Tan, Y. Ai, X. Wang and S. Wang, *Chemical Engineering Journal*, 2020, **379**, 122360.

3 K. C. To, S. Ben-Jaber and I. P. Parkin, *ACS Nano*, 2020, **14**, 10804–10833.

4 U. S. Epa, *Provisional Peer-Reviewed Toxicity Values for Dinitrophenol, 2,4*, 2007.

5 *Toxicological Profile for Dinitrophenols. Public Health Service, U.S. Department of Health and Human Services*, 1995.

6 M. A. Marletta, *Hepatology*, 1985, **5**, 165–165.

7 X.-G. Li, Y. Liao, M.-R. Huang, V. Strong and R. B. Kaner, *Chem. Sci.*, 2013, **4**, 1970–1978.

8 B. Gogoi, N. Paul, D. Chowdhury and N. S. Sarma, *J. Mater. Chem. C*, 2015, **3**, 11081–11089.

9 A. H. Malik, S. Hussain, A. Kalita and P. K. Iyer, *ACS Applied Materials & Interfaces*, 2015, **7**, 26968–26976.

10 S. Sanda, S. Parshamoni, S. Biswas and S. Konar, *Chem. Commun.*, 2015, **51**, 6576–6579.

11 S. S. Nagarkar, B. Joarder, A. K. Chaudhari, S. Mukherjee and S. K. Ghosh, *Angewandte Chemie International Edition*, 2013, **52**, 2881–2885.

12 J. Grundlingh, P. I. Dargan, M. El-Zanfaly and D. M. Wood, *J. Med. Toxicol.*, 2011, **7**, 205–212.

13 L. Barron and E. Gilchrist, *Analytica Chimica Acta*, 2014, **806**, 27–54.

14 R. Mu, H. Shi, Y. Yuan, A. Karnjanapiboonwong, J. G. Burken and Y. Ma, *Analytical Chemistry*, 2012, **84**, 3427–3432.

15 J. S. Caygill, F. Davis and S. P. J. Higson, *Talanta*, 2012, **88**, 14–29.

16 X. Chen, X. Cheng and J. J. Gooding, *Analytical Chemistry*, 2012, **84**, 8557–8563.

17 M. Rong, L. Lin, X. Song, T. Zhao, Y. Zhong, J. Yan, Y. Wang and X. Chen, *Anal. Chem.*, 2015, **87**, 1288–1296.

18 F. Lussier, V. Thibault, B. Charron, G. Q. Wallace and J.-F. Masson, *TrAC Trends in Analytical Chemistry*, 2020, **124**, 115796.

19 J. Langer, D. Jimenez de Aberasturi, J. Aizpurua, R. A. Alvarez-Puebla, B. Auguié, J. J. Baumberg, G. C. Bazan, S. E. J. Bell, A. Boisen, A. G. Brolo, J. Choo, D. Cialla-May, V. Deckert, L. Fabris, K. Faulds, F. J. García de Abajo, R. Goodacre, D. Graham, A. J. Haes, C. L. Haynes, C. Huck, T. Itoh, M. Käll, J. Kneipp, N. A. Kotov, H. Kuang, E. C. Le Ru, H. K. Lee, J.-F. Li, X. Y. Ling, S. A. Maier, T. Mayerhöfer, M. Moskovits, K. Murakoshi, J.-M. Nam, S. Nie, Y. Ozaki, I. Pastoriza-Santos, J. Perez-Juste, J. Popp, A. Pucci, S. Reich, B. Ren, G. C. Schatz, T. Shegai, S. Schlücker, L.-L. Tay, K. G. Thomas, Z.-Q. Tian, R. P. Van Duyne, T. Vo-Dinh, Y. Wang, K. A. Willets, C. Xu, H. Xu, Y. Xu, Y. S. Yamamoto, B. Zhao and L. M. Liz-Marzán, *ACS Nano*, 2020, **14**, 28–117.

20 T. K. Naqvi, A. Bajpai, M. S. S. Bharati, M. M. Kulkarni, A. M. Siddiqui, V. R. Soma and P. K. Dwivedi, *Journal of Hazardous Materials*, 2021, **407**, 124353.

21 D. Lin, R. Dong, P. Li, S. Li, M. Ge, Y. Zhang, L. Yang and W. Xu, *Talanta*, 2020, **218**, 121157.

22 C.-S. Ho, N. Jean, C. Hogan, L. Blackmon, S. Jeffrey, M. Holodniy, N. Banaei, A. A. E. Saleh, S. Ermon and J. Dionne, *Nature Communications*, 2019, **10**, 4927.

23 J.-y. Lim, J.-s. Nam, S.-e. Yang, H. Shin, Y.-h. Jang, G.-U. Bae, T. Kang, K.-i. Lim and Y. Choi, *Analytical Chemistry*, 2015, **87**, 11652–11659.

24 H. Dies, J. Raveendran, C. Escobedo and A. Docoslis, *Sensors and Actuators B: Chemical*, 2018, **257**, 382–388.

25 K. Sivashanmugan, K. Squire, A. Tan, Y. Zhao, J. A. Kraai, G. L. Rorrer and A. X. Wang, *ACS Sensors*, 2019, **4**, 1109–1117.

26 Q. Bao, H. Zhao, S. Han, C. Zhang and W. Hasi, *Anal. Methods*, 2020, **12**, 3025–3031.

27 E. J. Blackie, E. C. Le Ru and P. G. Etchegoin, *Journal of the American Chemical Society*, 2009, **131**, 14466–14472.

28 J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet and L. M. Buydens, *TrAC Trends in Analytical Chemistry*, 2013, **50**, 96–106.

29 H. Shin, H. Jeong, J. Park, S. Hong and Y. Choi, *ACS Sensors*, 2018, **3**, 2637–2643.

30 R. Dong, S. Weng, L. Yang and J. Liu, *Analytical Chemistry*, 2015, **87**, 2937–2944.

31 J. Wang, R. Ding, F. Cao, J. Li, H. Dong, T. Shi, L. Xing and J. Liu, *Chemical Engineering Journal*, 2022, **442**, 136064.

32 B. Li, M. N. Schmidt and T. S. Alstrøm, *Analyst*, 2022, –.

33 J. Yang, M. Palla, F. G. Bosco, T. Rindzevicius, T. S. Alstrøm, M. S. Schmidt, A. Boisen, J. Ju and Q. Lin, *ACS Nano*, 2013, **7**, 5350–5359.

34 C. Szegedy, S. Ioffe, V. Vanhoucke and A. A. Alemi, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 4278–4284.

35 K. He, X. Zhang, S. Ren and J. Sun, *CoRR*, 2015, **abs/1512.03385**, year.

36 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, *ICLR*, 2021.

37 R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017, pp. 618–626.

38 J.-B. Cordonnier, A. Loukas and M. Jaggi, *arXiv*, 2019.

39 Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, 2021, pp. 9992–10002.

40 R. Ranftl, A. Bochkovskiy and V. Koltun, 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, 2021, pp. 12159–12168.

41 A. Hakonen, F. Wang, P. O. Andersson, H. Wingfors, T. Rindzevicius, M. S. Schmidt, V. R. Soma, S. Xu, Y. Li, A. Boisen and H. Wu, *ACS Sensors*, 2017, **2**, 198–202.

42 W. J. Thrift and R. Ragan, *Analytical Chemistry*, 2019, **91**, 13337–13342.

43 M. S. Schmidt, J. Hübner and A. Boisen, *Advanced Materials*, 2011, **24**, OP11–OP18.

44 K. Wu, T. Rindzevicius, M. S. Schmidt, K. B. Mogensen, A. Hakonen and A. Boisen, *Journal of Physical Chemistry Part C*, 2015, **119**, 2053–2062.

45 J. Devlin, M. Chang, K. Lee and K. Toutanova, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

46 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.

47 L. J. Ba, J. R. Kiros and G. E. Hinton, *CoRR*, 2016, **abs/1607.06450**, year.

48 R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang and T. Liu, Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, 2020, pp. 10524–10533.

49 A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit and L. Beyer, *CoRR*, 2021, **abs/2106.10270**, year.

50 P. G. Etchegoin, M. Meyer, E. Blackie and E. C. Le Ru, *Analytical Chemistry*, 2007, **79**, 8411–8415.

51 Z. Zhang, D. Li, X. Wang, Y. Wang, J. Lin, S. Jiang, Z. Wu, Y. He, X. Gao, Z. Zhu, Y. Xiao, Z. Qu and Y. Li, *Chemical Engineering Journal*, 2022, **438**, 135589.

52 R. Beeram, D. Banerjee, L. M. Narlagiri and V. R. Soma, *Anal. Methods*, 2022, **14**, 1788–1796.

53 J.-L. Li, D.-W. Sun, H. Pu and D. Jayas, *Food Chem.*, 2017, **218**, 543–552.

54 J. H. Friedman, *The Annals of Statistics*, 2001, **29**, 1189 – 1232.

55 T. K. Ho, Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, pp. 278–282 vol.1.

56 D. H. Moore II, *Cytometry*, 1987, **8**, 534–535.

57 C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 1–9.

58 J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon and S. J. Gibson, *Analyst*, 2017, **142**, 4067–4074.

59 T. G. Dietterich, Multiple Classifier Systems, First International Workshop, MCS 2000, 2000, pp. 1–15.

60 B. Lakshminarayanan, A. Pritzel and C. Blundell, Advances in Neural Information Processing Systems 30, 2017, pp. 6402–6413.

61 J. Q. Candela, C. E. Rasmussen, F. H. Sinz, O. Bousquet and B. Schölkopf, Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW, 2005, pp. 1–27.

62 S. Lee, S. Purushwalkam, M. Cogswell, D. J. Crandall and D. Batra, *CoRR*, 2015, **abs/1511.06314**, year.

63 X. Glorot and Y. Bengio, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010, 2010, pp. 249–256.

64 P. Šimáková, E. Kočišová and M. Procházka, *Journal of Nanomaterials*, 2021, **2021**, 4009352.