Machine learning for composition analysis of ssDNA using chemical enhancement in SERS

PHUONG H. L. NGUYEN, BRANDON HONG, SHIMON RUBIN, AND YESHAIAHU FAINMAN

University of California San Diego, Electrical and Computer Engineering, La Jolla, CA 92161, USA
fainman@ece.ucsd.edu

Abstract: Surface-enhanced Raman spectroscopy (SERS) is an attractive method for biochemical sensing due to its potential for single molecule sensitivity and the prospect of DNA composition analysis. In this manuscript we leverage metal specific chemical enhancement effect to detect differences in SERS spectra of 200-base length single-stranded DNA (ssDNA) molecules adsorbed on gold or silver nanorod substrates, and then develop and train a linear regression as well as neural network models to predict the composition of ssDNA. Our results indicate that employing substrates of different metals that host a given adsorbed molecule leads to distinct SERS spectra, allowing to probe metal-molecule interactions under distinct chemical enhancement regimes. Leveraging this difference and combining spectra from different metals as an input for PCA (Principal Component Analysis) and NN (Neural Network) models, allows to significantly lower the detection errors compared to manual feature-choosing analysis as well as compared to the case where data from single metal is used. Furthermore, we show that NN model provides superior performance in the presence of complex noise and data dispersion factors that affect SERS signals collected from metal substrates fabricated on different days.

© 2020 Optical Society of America under the terms of the OSA Open Access Publishing Agreement

1. Introduction

Surface-enhanced Raman spectroscopy (SERS) discovered in the 1970s [1–3] provides an attractive method for bio-sensing applications [4–8], as it combines high degree of specificity inherent to Raman scattering with high scattering cross section mainly due to electromagnetic enhancement (EM) mechanism. These features turn SERS into an appealing method for DNA composition analysis in order to discriminate between DNA sequences according to the total number of bases of each type, with a variety of potential applications in genome evolution studies [9–11], cell sorting [12,13] and mutation detection [14], where the information of exact DNA sequence is not crucial. Furthermore, SERS admits several advantages over standard Raman spectroscopy such as overcoming of strong fluorescent background and requiring less excitation power, leading to a prominent increase of the signal-to-noise ratio (SNR) [2] and simplifying the complexity of optical spectrometers and detection systems necessary for biomedical analysis and biomedical sensing applications. Moreover, in contrast to fluorescent microscopy, SERS is a label-free technique which does not require complex preparation steps such as specific probe design, and results in simpler bio-assays [15,16]. Despite its large potential for a wide range of applications in bio-detection and sensing, especially due to the prospect of single molecule sensitivity [15,17,18], the results are known to be highly sensitive to preparation methods due to several physical mechanisms which affect adsorbed molecule orientation [19] and the chemical enhancement effect (CE) which stems mostly from the charge transfer mechanism between molecule and the metal [20,21]. The latter leads to discrepancies in the SERS spectra reported in the literature [22,23], and also gives rise to recent attempts that take advantage of the CE effect for nucleotide detection [24]. Given the complexity of the involved effects and numerous features of the corresponding SERS spectra, signal processing seems to be a highly relevant resource to account for spectral variability and heterogeneity. In particular, principal component analysis
(PCA), which employs a linear transformation to identify a smaller set of linearly uncorrelated variables referred as principal components (PCs), is one of the most commonly used techniques in analyzing Raman spectroscopic data. For instance, Raman spectral analysis using PCA has been used to interpret complex tumor signatures [25], or to assist with identification of the microstructure of a DNA helix [26]. Furthermore, pairing PCA method with a supervised machine learning (ML) algorithm is known to improve classification of Raman scattering results. For example, Satile et al. [27] combined PCA with a linear discriminant analysis algorithm to develop a reliable HIV bio-marker, work [28] reported PCA paired with Euclidean-distance classification to discriminate melanoma from normal skin cells, work [29] employed PCA paired with SVM model for superior diagnosis of prostate cancer, and work [30] used PCA paired with SVM for detection of drugs in human urine using dynamic SERS. However, despite the relative simplicity of the PCA method and the basic ML techniques, these have not been widely explored in the realm of DNA composition analysis especially for long (>30 bases) ssDNA molecules. The latter are particularly suited for overall composition analysis because DNA has complementary base pairing characteristic; the composition of one of the DNA strands uniquely determines the composition of the other strand and of the full DNA molecule.

In this manuscript, with Fig. 1 schematically describing its main concept, we employ gold and silver nanorod array substrates, fabricated using a straightforward single-step oblique-angled deposition (OAD) method [31] known to provide a prominent EM effect [32–34], to experimentally study SERS spectra of 200-base length ssDNA molecules adsorbed on these substrates. The need for using SERS is demonstrated in Appendix Section A2, where we include a comparison between normal Raman spectra (i.e. without metal substrate) and SERS spectra of same ssDNA molecules used in this work (see Fig. 8). Clearly SERS spectra admit much higher SNR values, and consequently are more efficient than standard Raman spectroscopy for detection applications. We experimentally show that adsorption of ssDNA molecules to gold and silver gives rise to a distinct CE effect, which manifests as characteristic spectral peak shifts and selective intensification of various vibrational modes, as also qualitatively supported by our numerical simulation results. More importantly, while several works exploited SERS for DNA sensing on gold [7] and on silver [8,16], in our work we employ both metals to demonstrate distinct CE effect associated with each of them, which can be used for enhanced specificity. Particularly, we employ PCA in order to enable identification of the corresponding orthogonal features present in the experimentally acquired ones, and then use these features (i.e. PCs) as a training set of our linear regression model. Furthermore, we also incorporate several data pre-processing methods, including Gaussian smoothing, normalization and multiplicative-scattering correction (MSC). The use of MSC here is especially important due to numerous noise generating factors such as multiplicative light scattering [35], and the fact that SERS is an extremely sensitive method. In principle, its spectra can contain very detailed information which allows to detect small differences from sample to sample, which in turn can strongly affect the visual assessment of the spectrum, by causing small arbitrary spectral shifts that may not contain information relevant to ssDNA composition (see Appendix section A7). We show that PCA multiple-feature linear regression greatly benefits from such noise elimination procedure, and consequentially demonstrates elevated sensitivity relative to the more basic one-feature peak-ratio regression model. Beside linear regression, we also utilize another ML - deep learning model namely neural network (NN). The NN concept was established in the 1980-90s and has since then been developed by numerous scientists and researchers [36,37]. Most NN organize their neurons into layers, and in layered NN the neurons in the input layer can accept numeric data points as their inputs. In particular, each neuron admits a weight, which upon multiplication with the input data yields neuron output and is transferred to the next layer [38]. We employ this NN model and its feature for training and testing data combined from both metals for superior performance.
Fig. 1. Schematic description of key experimental and data processing components: (a) Rough metal surface (gold or silver) formed by an array of nanorods of mean height $h$ and mean distance $\Lambda$, functionalized with Raman active ssDNA molecules comprised of adenine (pink) and of cytosine (blue) bases. (b) DNA composition analysis which employs PCA and linear regression (for gold and silver’s separate datasets) and neural network (for gold and silver’s combined datasets) to predict the percentage of adenine and cytosine bases in the ssDNA molecule. List in (c) presents the 200 bases long ssDNA training molecules (see for specific sequences and the random test sample in Appendix section A1 Table 5.)

2. Methods

2.1. Metal nanorod array substrate fabrication

The metal nanorod array structure is fabricated by using OAD technique with a Denton Discovery Sputter system. The average height of the nanorods, schematically described in Fig. 1(a), is $h = 200$ nm, and the average diameter of the rods is approximately $\Lambda = 50$ nm for gold and $\Lambda = 100$ nm for silver (see relevant SEM images in Fig. 9 in Appendix section A3). The substrate is tilted at an angle such that the zenithal deposition angle is at $\alpha = 75^\circ$. Following Barranco et. al. [31], we set the tilt angle of the nanorods to approximately 60° relative to the substrate normal (see also Appendix section A4 for detailed calculation).

2.2. ssDNA functionalization

The ssDNA solutions are prepared by diluting the DNA stock solution to 25 $\mu$M in 10 mM 4-(2-hydroxyethyl)-1-piperazine ethanesulfonic acid (HEPES), and then forming a 1:3 mixture with a 10 mM MgCl$_2$ solution. We then drop-cast this ssDNA solution onto the metal nanorod substrate and let it dry overnight. Before SERS measurements, all samples are rinsed with deionized water in order to remove excess of crystallized salt and unbound ssDNA molecules, and then blow-dried.

A table of the sequences for each ssDNA mixtures is listed in the Appendix section A1. The ssDNA concentration is chosen to provide sufficiently large surface concentration enabling us to measure SERS signal over a map of units on substrates uniformly. The salt ratio is used to neutralize the negative phosphate backbone and enable bonds between the DNA bases and the metal substrate [18].

2.3. SERS spectra acquisition and data processing

The SERS spectra are collected by using a Renishaw inVia Raman spectrometer with the following settings. Each spectrum is obtained by employing a 785 nm Raman excitation laser with 50
mW output power, acquisition time of 5 s and 1 accumulation per spectrum. The objective magnification is 50x with NA = 0.75. The grating type used is 1200 l/mm at 785 nm. The grating setting in the built-in spectrometer software is set to a static regime with acquired spectrum range extending between 600 cm\(^{-1}\) to 1700 cm\(^{-1}\). The resultant spectral resolution in our setup is approximately 1 cm\(^{-1}\). For presentation purposes, most of the SERS spectra presented in this work are cropped from 600 cm\(^{-1}\) to 1200 cm\(^{-1}\) only. Full SERS spectra can be found in Appendix section A7. The mapping setting of the spectrometer is used to acquire 100 measurements from a total substrate area of dimensions 50 \(\times\) 50 \(\mu\)m\(^2\); this area is divided into 10 \(\times\) 10 square units, where the dimension of each area unit is 5 \(\times\) 5 \(\mu\)m\(^2\) and each acquired spectrum measurement is taken from a different unit.

To analyze the CE effect, 100 SERS spectra are acquired from DNA bases without phosphate backbone adsorbed to gold and silver nanorod substrates. To demonstrate ssDNA composition analysis, 200 SERS measurements in total (divided into 2 maps, 100 measurements in each map) are made for each 200-base ssDNA composition (5 control sequences and a single test sequence).

The single-feature linear regression model is developed with MATLAB using `LinearModel.fit` algorithm, whereas the PCA multiple-feature linear regression model is developed in Python using ML libraries including `scikit-learn`, `numpy`, `scipy`, `pandas` and `matplotlib`, and the NN model is built in Tensorflow and Python using NN and ML libraries including `keras`, `scikit-learn`, `scipy` and `matplotlib`. In all models, we have five control sequences to be our training dataset with the following A and C compositions: 100% A - 0% C, 75% A - 25% C, 50% A - 50% C, 25% A - 75% C, 0% A - 100% C, and a single MATLAB generated random sequence with composition of 54% A - 46% C for testing (the percentage of A and C in the testing sequence is checked by MATLAB after generation, see Appendix section A1 for specific sequences used). In some models, i.e. the PCA linear regression and NN, a validation set which is randomly extracted from one third of the testing set is also used for model validation. For each control sequence, we have three samples of dataset made on different dates. Multiple samples are needed for training of some model to improve performance such as the PCA linear regression. Similarly, we have three different samples used for testing and calculation of errors in order to justify the robustness of the system. The list of pre-processing steps for the data consists of baseline subtraction and cosmic-ray removal which are performed by employing built-in algorithms of Renishaw WiRE 4 software, Gaussian smoothing (smoothing window = 5), signal normalization and MSC which are implemented in both Python and MATLAB, and are described in details in Appendix section A7.

### 3. Results and discussion

#### 3.1. CE effect of a single DNA base: comparison between gold and silver nanorod substrates

To probe the distinct CE effects introduced by gold and silver on the SERS spectra of DNA bases adsorbed to these metals, we first consider a numerical simulation using density functional theory (DFT) by employing Gaussian 09 [39] software on the Gordon supercomputer at the University of California, San Diego [40]. In particular, we consider a simplified model which considers only a single nucleotide without a phosphate backbone adsorbed to the corresponding metal with a fixed nitrogen atom, and examine the effect of modifying the type of the metal on the corresponding Raman spectra. The simulation result is presented in Appendix section A5, where Fig. 10(a,b) presents the simulated Raman spectra intensity (see [41] for formal definition of Raman intensity and activity) of adenine (A) and cytosine (C), respectively, by employing a B3LYP computational method and LANL2DZ computational basis function. Each base is bound to a tetrahedral nanoparticle comprised of 20 silver/gold atoms. In so doing, the model probes the CE effect for a given orientation by introducing a metal-nitrogen bond, and bypasses computationally expensive crystalline structures which typically require larger amount of metal.
atoms [42]. Importantly, the orientation dependent effects are eliminated because the models describe DNA bases bound to the different metals with the same orientation.

In particular, the simulated RBM peak of A unbound to metal is centered at 711 cm\(^{-1}\), and is shifted to 712 cm\(^{-1}\) when bound to silver and to 713 cm\(^{-1}\) when bound to gold. More significant shift is detected for the simulated RBM of C; 754 cm\(^{-1}\) without metal, 757 cm\(^{-1}\) when bound to silver and 769 cm\(^{-1}\) when bound to gold. Furthermore, the simulated spectra show different Raman intensity of RBM mode. For example, the intensity of RBM of A bound to gold is higher than that of A bound to silver. We attribute the differences in the features of these SERS spectra of molecules adsorbed to gold and silver due to the difference in Fermi energy levels of these metals, which in turn affects the charge transfer effect [4]. With this said, a model that relates the strength of dominant SERS spectral mode to the work function of relevant metal is beyond the scope of this work. Nevertheless, the distinct CE effects observed for the simple numerical cases suggest that distinct CE effects should be present in the experimental results considered below.

In the next step, we perform experimental SERS measurements of A and C (i.e. DNA bases without phosphate backbone) adsorbed to gold and silver nanorod array substrates, and to the same substrates but covered with a 2-nm thickness dielectric layer of Al\(_2\)O\(_3\) deposited via the atomic-layer deposition (ALD) method operated by the Beneq ALD system. Small thickness of this layer guarantees a prominent EM effect, but is expected to eliminate the CE effect by blocking the charge transfer between the adsorbed ssDNA molecule and the metal surface [24], and also allows quantitative measures for the CE factor of each one of the metals as described shortly below. Fig. 2(a,b) present arithmetic mean of 100 SERS spectra measurements results of A and C, respectively, whereas Table 1 presents the intensities of the corresponding RBM modes and their dependence on the CE effect. These results indicate that RBM peaks in SERS spectra are higher when EM and CE effects are both present compared to a case with just EM effect is present. We can estimate the relative strength of the CE effect on a certain vibrational mode by considering the so-called chemical enhancement factor (EF) [43], given by the following ratio,

\[
\Gamma^{(i,j)}_{CE} = \frac{I_{surf}(EM+CE)}{I_{surf}(EM)}; \quad i = Au, Ag; \quad j = A, C,
\]

where the indices \(i, j\) stand for the substrate metal and the DNA base adsorbed to that metal, respectively. Here, \(I_{surf}(EM+CE)\) and \(I_{surf}(EM)\) correspond to peak intensities of relevant vibrational mode in the SERS spectra with and without the CE effect, respectively; i.e. in our setup corresponds to the cases of metal nanorod substrate without and with the thin Al\(_2\)O\(_3\) layer. All SERS intensity spectra in our plots are normalized between zero and unity, both for uniform presentation and as a preparation step for ML model training. The peak intensities that serve the purpose of this experiment were extracted prior to normalization. Note that the enhancement factor defined in Eq. (1) is a proper figure of merit for the strength of the CE effect under the plausible assumption that the two cases with and without CE effect (i.e. without and with ALD of Al\(_2\)O\(_3\) dielectric layer) admit the same total number of optically excited molecules (see Appendix section A.6) and identical EM effect.

Table 1 below indicates that the chemical EF of adenine RBM adsorbed to gold nanorod substrate is approximately given by \(\Gamma^{(Au,A)}_{CE} \approx 23\), whereas in a case when it is adsorbed to silver nanorod substrate the EF of adenine RBM is given by \(\Gamma^{(Ag,A)}_{CE} \approx 2\). Similarly, the corresponding EFs of cytosine adsorbed to gold and silver nanorod substrates are both given by \(\Gamma^{(Au,C)}_{CE} = \Gamma^{(Ag,C)}_{CE} \approx 2\); both EFs correspond to the RBM. In both numerical simulation and experimental results, the CE effect appears to be stronger in DNA bases adsorbed to gold, compared to when they are adsorbed to silver. However, the calculation in our simulation only took into account the non-resonant charge transfer effect, which includes charge redistribution within the molecule or the metal structure itself at ground state. Therefore, a higher CE effect in our experiment is most likely due to the involvement of the resonant charge transfer between the...
Fig. 2. Experimental results presenting normalized SERS spectra of: a) adenine (A) and b) cytosine (C) bases bound to gold and to silver nanorod array substrate. The resulted SERS spectrum of DNA bases binding to gold nanorod substrate appears to be higher without the Al$_2$O$_3$ layer, similar trend is presented also in the SERS spectrum of ssDNA binding to silver nanorod substrate.

Table 1. RBM peak values of experimentally acquired SERS spectra of A and C molecules on Ag, Au nanorod substrates as well as Ag, Au covered with thin Al$_2$O$_3$ film, and the corresponding values of the chemical enhancement factors, $\Gamma_{CE}$, defined by Eq. (1). All spectra were obtained by averaging the 100 measurements results.

| RBM (A) | Ag  | Ag+Al$_2$O$_3$ | Au  | Au+Al$_2$O$_3$ |
|---------|-----|----------------|-----|---------------|
| Raman intensity (a.u.) | 811 ± 359 | 430 ± 158 | 1.67 × 10$^4$ ±7.16 × 10$^3$ | 734 ± 276 |
| $\Gamma_{CE}$ | 1.88 ± 1.08 | | 22.8 ± 12.9 |

| RBM (C) | Ag  | Ag+Al$_2$O$_3$ | Au  | Au+Al$_2$O$_3$ |
|---------|-----|----------------|-----|---------------|
| Raman intensity (a.u.) | 593 ± 346 | 270 ± 64.5 | 2.60 × 10$^3$ ±7.10 × 10$^2$ | 1.09 × 10$^3$ ±2.88 × 10$^2$ |
| $\Gamma_{CE}$ | 2.19 ± 1.38 | | 2.39 ± 0.906 |

metal and our DNA bases. Moreover, we also observe a high standard deviation in the EF values, which stems from fluctuation of SERS intensity signal due to low number of adsorbed molecules (surface concentration), leading to uneven distribution of hotspots in the nanorod substrate [44]. Specifically, the final concentration of the DNA solution we used in this experiment was about 6 $\mu$M, which converts to about 0.5 molecule/ cm$^2$ of metal, if all molecules have the same chance to bind. This number is not high, and could be much lower in reality because many molecules in the droplet might not come in contact with the metal surface.

Additionally, Fig. 2(a,b) presents a shift of the RBM peaks in SERS spectra of DNA bases adsorbed to metals, relative to the cases when the bases are measured in bulk or on top of an Al$_2$O$_3$ layer. For instance, A’s RBM peak shifts from 723 cm$^{-1}$ to 738 cm$^{-1}$ and to 740 cm$^{-1}$ respectively, when adsorbed to gold and silver, compared to when adsorbed to the Al$_2$O$_3$ layer or when measured in its bulk form.Fig. 2(a,b) results indicate that the CE effect is practically eliminated in case when the nanorod array substrate is covered with a thin Al$_2$O$_3$ layer, presumably due to a blockage of charge transfer between the DNA bases and the metal. This phenomenon is demonstrated by the resemblance between SERS spectra of ssDNA molecules adsorbed to the dielectric layer, and Raman spectra of bulk samples. More importantly, the different CE effects induce changes in SERS scattering wavelengths and intensities, resulting in distinguished spectral shifts and enhancement ratios, which account for the DNA prediction uncertainty for different types of metal.
The differences between experimental SERS spectra and numerical simulation can originate from the effect of different DNA bases binding sites to the metal substrate which result in different orientation of the molecule relative to the metal, which is known to depend on numerous experimental conditions and occasionally lead to controversial results [19]. Specifically, while in our simulation we consider the CE effect of a DNA base with a fixed binding site and orientation, in practice each nucleotide admits several binding possibilities, each giving rise to a different orientation and in principle leading to a different CE effect [42].

3.2. Composition detection of ssDNA using single-feature linear regression model

In this section, we analyze SERS spectra of ssDNA sequences and employ a simple linear regression model to probe their chemical compositions. Fig. 3 and Fig. 4 present SERS spectra of 200-base long ssDNA molecules bound to gold and silver nanorod array substrates, respectively. Interestingly, Fig. 3(a–e) and Fig. 4(a–e) both present two prominent peaks; \( p_A \) at \( \sim 725 \text{ cm}^{-1} \) which is associated with ssDNA sequences where A is present, and \( p_{A+C} \) at \( \sim 790 \text{ cm}^{-1} \) which is present in all cases including those where A or C are missing. Consequently, we employ the ratio between peaks’ intensities, \( p_A/p_{A+C} \), as a single feature in the regression model. The ratio values of each ssDNA composition undergo a lognormal distribution fit, and the natural logarithm of their medians are taken as control values which are used for building a linear regression model. Fig. 3(f–j) and Fig. 4(f–j) present the corresponding probability distribution functions (PDFs) of the ratio for each one of the control sequences, together with their median values (\( R \)) and the corresponding natural logarithm values (\( \ln(R) \)). The latter reflects a nonlinear relation between \( R \) and \( C_A \) caused by a nonlinearity of the CE effect [45] as a function of \( C_A \). Similar effect was also reported in work [16], for the case of adenine adsorbed to silver random islands. Fitting a linear regression model against these five control \( \ln(R) \) values yields the following linear regression function \( f(C_A) \) and the corresponding inverse function \( C_A = f(C_A) - b_i/a_i \).

\[
f(C_A) = a_i \cdot C_A + b_i; \quad C_A = \frac{f(C_A) - b_i}{a_i}.
\]

Here \( a_i \) and \( b_i (i = Au, Ag) \) are obtained by employing a built-in MATLAB curve fitting algorithm, given by

\[
a_{Au} = 0.0215; \quad b_{Au} = -1.68; \quad a_{Ag} = 0.0231; \quad b_{Ag} = -1.87,
\]

and which are employed in Fig. 3(m) and Fig. 4(m), respectively. Applying the linear regression model described by Eq. (2) and Eq. (3), to SERS spectra of the test sample described in Fig. 3(k–l) and Fig. 4(k–l) leads to the following prediction of the adenine concentration:

on gold: \( \bar{C}_A = 51.2\% \pm 7.32\% \);
on silver: \( \bar{C}_A = 54.8\% \pm 2.55\% \),

where the corresponding predicted cytosine concentration satisfies \( \bar{C}_C = 100\% - \bar{C}_A \) on each one of the substrates with the same detection errors. Detailed calculations are described in Appendix section A9. Since the ground-truth composition of the test ssDNA molecule contains 54\% A (and 46\% C), we conclude that in our setup silver nanorod array substrate appears to give a better prediction than gold substrate; silver substrate yields both smaller difference relative to the actual concentration value and lower detection error. The higher detection error in gold could stem from several factors, physically from the SERS enhancement effects or from the data analysis process. As we have seen from section 3.1 above, SERS enhancement of Au nanorod substrate is higher and fluctuating more than with Ag nanorod substrate. Moreover, the lognormal distribution
Fig. 3. Experimental results presenting SERS spectra of control and test ssDNA samples, measured on gold nanorod substrate, and the corresponding linear regression curve. (a-e) SERS spectra of ssDNA for the five control sequences: (a) 100% A - 0% C; (b) 75% A - 25% C; (c) 50% A - 50% C; (d) 25% A - 75% C; (e) 0% A - 100% C; and (f-j) the corresponding PDFs of the $p_A / p_A + p_C$ ratio accompanied by the median values $R$. (k,l) SERS spectrum of test sample, PDF of the $p_A / p_A + p_C$ ratio and the corresponding median $R$. (m) Linear regression curve based on $\ln(R)$ values of the five control samples, and position of the test point on the curve. Here, x-axis represents the percentage of adenine component. The spectra plotted here are arithmetic means of 200 measurements results. When used in data analysis, all spectra are pre-processed as described in the Methods section and treated separately as individuals.

goodness of fit was not as good for Au substrate compared to Ag, as described in details in Appendix section A8 (see Figs. 16, 17).

To test repeatability of the experiments we consider model performance on three samples with the same ssDNA test sequence performed on substrates fabricated on different days. Table 2 below summarizes results where the training was performed on set #1 and then applied to predict test sequence on other substrates (see also Table 7 in Appendix section A11 for more details). For a convenient metric for model’s sensitivity, we utilize the following two parameters: regression residual (RR) and standard deviation (SD). The RR value presents how much the predicted (mean) value deviates from the actual ground truth value, while the SD, which is given by the root mean square error (RMSE), indicates the spread of the predicted values around the predicted mean. As expected, the table indicates that on average SERS spectra acquired from samples fabricated on different days leads to higher error than performing training and test on the same day. We will find similar conclusion in the multiple-feature analysis below.

3.3. Composition detection of ssDNA using PCA with multiple-feature linear regression

Consider the experimentally acquired SERS spectra as being composed from a set of L features with $L=1021$, corresponding to the number of wavenumbers in one spectrum. In so doing, we notice that each spectrum in the training set shares some common features with others.
Fig. 4. Experimental results presenting SERS spectra of control and test ssDNA samples, measured on silver nanorod array substrate, and the corresponding linear regression curve. (a-e) SERS spectra for the five control ssDNA sequences: (a) 100% A - 0% C; (b) 75% A - 25% C; (c) 50% A - 50% C; (d) 25% A - 75% C; and (e) 0% A - 100% C; and (f-j) the corresponding PDFs of the \( p_A/p_{A+C} \) ratio accompanied by the median values \( R \). (k,l) SERS spectrum of the test sample, PDF of the \( p_A/p_{A+C} \) ratio and the corresponding median \( R \).

Table 2. Single-feature linear model regression results trained only on data set #1 presenting RR and SD of the predicted adenine concentration from three different test datasets, as well as the average values (Ave.). GT (ground truth) adenine concentration is 54%.

|                | Test set #1 | Test set #2 | Test set #3 |
|----------------|-------------|-------------|-------------|
|                | RR          | SD          | RR          | SD          | RR          | SD          |
| Au             | 2.80%       | 7.32%       | 0.40%       | 5.38%       | 2.20%       | 3.49%       |
| Ag             | 0.80%       | 2.55%       | 3.40%       | 6.65%       | 4.30%       | 5.85%       |
| Ave. RR        | 1.90% (Au); | 3.20% (Ag)  |             |             |             |             |
| Ave. SD        | 5.17% (Au); | 5.46% (Ag)  |             |             |             |             |
| GT             | 54.0%       |             |             |             |             |             |

By diagonalizing the covariance matrix formed by treating each spectra as a vector in an \( L \)-dimensional space, we are able to describe each spectrum by the weighted summation of a smaller subset of features also known as PCs. By determining this PC subset, as briefly shown below, we are able to extract additional features which was not possible in the single-feature model considered above. After Gaussian smoothing and MSC, we implement PCA transformation, which reduces the number of features to some number \( n \ll L \). First, we perform error analysis study as a function of \( n \) in order to determine the optimal number \( n \) specific to our data. The numbers of PCs used for training were chosen by looking at the Mean-square error (MSE)
curves, which were calculated from the predicted and the actual values of the validation ssDNA composition set. The two chosen \( n \) numbers for gold and silver substrates are \( n = 43 \) and \( n = 145 \), respectively. Detailed results and analysis are listed in Appendix section A10.

We then construct a linear regression model and train it with this set by employing built-in Python capabilities (see more details in Methods section) to predict the chemical composition of our test sample. Fig. 5 presents numerical simulation results of ground truth versus predicted adenine count for gold (Fig. 5(a,b)) and silver (Fig. 5(c,d)), leading to the following predicted \( \bar{C}_A \) values

\[
\begin{align*}
\text{on gold:} & \quad \bar{C}_A = 53.5\% \pm 2.17\% \\
\text{on silver:} & \quad \bar{C}_A = 54.3\% \pm 1.30\%.
\end{align*}
\]

These results indicate that silver nanorod array substrate admits a better detection sensitivity than the gold one; the corresponding RMSE between the predicted and the actual values are given by 1.36\% for silver substrate and 2.20\% for gold substrate, equivalent to approximately two and four bases in our 200-base length sequences. Importantly, our analysis indicates that the RR and SD values of \( C_A \) are reduced significantly when PCA and ML linear regression model is employed (Fig. 5) as opposed to just single-feature linear regression analysis (Fig. 3).

\[\text{Fig. 5. PCA linear regression lines and predicted values of 200 test measurements binding on gold (a, b) and silver (c, d) nanorods. The histogram plots (b, d) indicate the medians and RMSE values of the predicted } C_A.\]

For these datasets, the two most important pre-processing steps we use are Gaussian smoothing and MSC. PCA linear regression greatly benefits from pre-processing procedures that eliminate the multiplicative scattering noise because the whole spectrum is taken as its input. However, single-feature linear regression in some cases may not require such steps because the noise originates mostly from spectral regions which are not near the two RBM-related peaks (see Figs. 11, 12 in Appendices section). Therefore, if we include the MSC in the input data of the single-feature regression, the sensitivity may not improve as much as the case of PCA regression. The standard detection errors are summarized in Table 6 in Appendix section A7.
Table 3. PCA multiple-feature linear model regression results trained only on data set #1 presenting RR and SD of the predicted adenine concentration from three different test datasets, as well as the average values (Ave.). GT (ground truth) adenine concentration is 54%.

| Test set | RR    | SD  | RR    | SD  | RR    | SD  |
|----------|-------|-----|-------|-----|-------|-----|
|          |       |     |       |     |       |     |
| Au       | 0.50% | 2.17% | 6.10% | 1.24% | 9.20% | 3.82% |
| Ag       | 0.30% | 1.30% | 12.7% | 4.69% | 7.70% | 3.65% |
| Ave. RR  | 6.25% (Au); 7.10% (Ag) |
| Ave. SD  | 2.98% (Au); 3.55% (Ag) |
| GT       | 54.0% |

3.4. Combining Au and Ag models for composition detection of ssDNA

Although PCA linear regression model could already be powerful compared to the single-feature one, we are still limited by the number of features contained in a certain type of metal. Therefore, by developing a fusion model where the data acquired from different metals like gold and silver can be merged together as one input, we could increase the number of input features that could be beneficial for effective training. Instead of using the simple multivariate linear regression, the model we used to fit this combined dataset is a neural network (NN) model. Since according to Fig. 5, Ag model appears to have a better ssDNA composition detection results than Au, we would like to assign different weights to input data corresponding to different metals in order to optimally utilize the Au-Ag combined dataset for training. This task can be done in NN training, which is markedly different from our linear regression model described above, where all input spectra are treated equally (with the same weight).

Our NN model consists of five layers in total; one input and one output layer, and three hidden layers connecting between the input and output (see Fig. 6(a)). There are several criteria to determine the number of hidden layers and hidden neurons [38], which allow to bypass underfitting or overfitting of the model. In this work we chose to implement an NN network where the number of hidden neurons in one layer is 2/3 of the number of neurons in the one directly prior to that, which provides the lowest detection RMSE in our case. In particular, for each spectrum we have 1021 wavenumbers, representing 1021 features. When combining two metals together, our total number of features becomes $1021 \times 2 = 2042$, which also act as inputs of our NN model. The number of our model’s hidden neurons, according to the mentioned rule would be: $2042 \times 2/3 \approx 1361$ in the second hidden layer, and $1361 \times 2/3 \approx 907$ in the third hidden layer. We also used a nonlinear activation function activation, ReLU, for input and hidden layers, which allows back-propagation for error minimization while the model was trained and validated through several rounds, namely "epochs.” The termination criteria for training is when the mean-squared errors between the predicted values and ground-truth values of the validation set is at its lowest and unchanged within the next 30 epochs. Fig. 6, presents comparison between a linear regression for gold (b), linear regression for silver (c) and a NN model which combines gold and silver (d). Importantly, the result of combining the data of two metals appears to be superior to the result obtained by considering each one of the metal models separately.

Table 4 presents NN model performance in predicting the composition of the test sequence on substrates fabricated on different days with training done only on the first day data (referred below as Test set #1). The results over three test sets indicate smaller errors and more consistent prediction values compared to the linear multiple-feature PCA model. For a fair comparison with the PCA regression method, we also build additional NN models in which different numbers
Fig. 6. a) Schematic diagram of our NN model with 2024 input features (labeled as f1 to f2024) from Au and Ag substrates’ SERS spectra, and comparison between b) Au, c) Ag linear regression models and d) combined Au-Ag NN model. All three models are ML-based, and the result in c) shows the smallest RMSE value, which indicates a better detection performance for the combined model.

Table 4. NN model results trained only on data set #1 presenting RR and SD of the predicted adenine concentration from three different test datasets, as well as the average values (Ave.). GT (ground truth) adenine concentration is 54%.

|        | Test set #1 | Test set #2 | Test set #3 |
|--------|-------------|-------------|-------------|
| RR     | 0.10%       | 1.70%       | 1.70%       |
| SD     | 1.24%       | 1.22%       | 1.64%       |
| Au + Ag| Ave. RR     | Ave. SD     | GT (54.0%)  |
|        | 1.30%       | 1.45%       |             |

While in this work we employed ssDNA molecules with two DNA bases and showed that NN can reduce significantly the prediction errors relative to linear regression models, incorporating all four bases in the molecule should be in principle possible in future works as all four DNA bases admit different features in SERS spectra (see [42] for studies of bases adsorbed to silver) and thus well adapted to NN method.
4. Conclusions

In this work we experimentally and numerically studied CE effect of nucleotides adsorbed to gold and silver nanorod structure and employed PCA paired with ML algorithm as well as NN model to demonstrate a highly sensitive ssDNA composition detection method for 200-base long ssDNA sequences. In particular, we demonstrated that our method is superior to a single-feature linear regression model which takes into account only SERS intensity of two dominant peaks, and leads to lower RMSE prediction values for SERS spectra acquired on both gold and silver nanorod array substrates. Importantly, our results indicate that each metal possesses a distinct CE effect with the same molecule, and therefore it operates as an independent probe which provides additional information due to distinct metal-molecule interaction regime. In particular, our NN-based analysis indicates that the prediction error drops once input data from both metals is used as a training set compared to the case when only input data from only one of the metals is used. Furthermore, comparing performance of multi-feature PCA and NN on cases where training and test data belong to samples prepared on different days, indicates that NN provides superior performance compared to multi-feature PCA, indicating that it is useful to mitigate the effects of data dispersion which may emerge due to bio-degradation of the samples as a function of time or due to slight differences of the nanorod substrates fabricated on different days. The latter is particularly relevant as fabricating substrates with nano-scale size features, which are required for SERS sensing, with high levels of uniformity/reproducibility is still challenging [46]. In future studies it would be interesting to investigate the effect of these and other potential factors on the SERS signal and determining more effective NN modalities. Straightforward future directions that can be employed to extend our basic NN model include improvement of detection performance and of its explainability/ interpretability, as well as incorporation of non-linear models with data visualization in order to enhance ssDNA sensitivity. We hope that our work will stimulate future studies where different ML and deep learning models could be exploited to realize all-optical and SERS-based single-base detection sensitivity of long ssDNA and dsDNA molecules.
Appendices

A1. ssDNA sequences

| ssDNA composition  | ssDNA sequence                                      |
|-------------------|-----------------------------------------------------|
| 100% Adenine - 0% Cytosine | AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA AAA A
Similar to the single-feature model above, we test prediction performance of the multiple-feature model on test sequences deposited on samples fabricated on different days where the training was done on various days. However, in this case, due to the variation in the data components from one day to another, a model that has been trained on one dataset with a fixed number of PCs does not perform well on datasets of different dates (see Table 3 below). The prediction errors can be reduced once there are enough samples coming from different measurement dates. In order to effectively monitor the prediction improvement, we averaged out the RR and SD values of the above three test datasets. Fig. 19 indicates that increasing the number of datasets used in the training from one to two and to three, leads to enhanced average detection sensitivity. The average detection composition improved from $47.1 \pm 3.55\%$ to $54.7 \pm 3.77\%$ for Ag, and
from 47.8% ± 2.98% to 55.3% ± 2.63% for Au. Detailed predicted values for each set in different training rounds and their average values are listed in Appendices section A11 (see Tables 8, 9).

A3. SEM images of Au and Ag nanorod-array substrates

![SEM images of gold (left) and silver (right) nanorod array substrates.](image)

Fig. 9. SEM images of gold (left) and silver (right) nanorod array substrates.

A4. Oblique-angled deposition (OAD) tilted angle calculation

The tangent rule reported in [31] is given by

$$\tan \alpha = 2 \tan \beta$$

$$\beta = \arctan \frac{\tan \alpha}{2}.$$  \hspace{1cm} (6)

Inserting $\alpha = 75^\circ$ into the equation above leads to the nanorod tilt angle of value $\beta = 61.8^\circ$.

A5. Simulation of the DNA bases bound to metal models

Fig. 10(a,b) present the effect of different metals on the ring-breathing mode (RBM) of the adsorbed A and C bases; peak #1 and peak #2 stand for RBM modes of A and C, respectively.

![DFT Simulation results presenting SERS spectra of: (a) adenine (A) and (b) cytosine (C), each bound to silver (Ag), gold (Au) and unbound to metal (normal). CE effect leads to spectral shifts and intensity changes of several peaks for a given DNA base orientation as a function of metal type.](image)
A6. Calculation of CE effect factor $\Gamma_{CE}$

From the average SERS EF equation $EF = (I_{surf}/N_{surf})/(I_{bulk}/N_{bulk})$, where $I_{surf}$, $N_{surf}$ are SERS intensity and number of measured molecules, and $I_{bulk}$, $N_{bulk}$ are normal Raman intensity and number of measured molecules in bulk respectively, we have the EF calculations for system with and without CE to be:

$$EF_{EM+CE} = \frac{I_{surf}(EM+CE)/N_{surf}(EM+CE)}{I_{bulk}/N_{bulk}}$$  \hspace{2cm} (7a)

$$EF_{EM} = \frac{I_{surf}(EM)/N_{surf}(EM)}{I_{bulk}/N_{bulk}}$$  \hspace{2cm} (7b)

$$\Gamma_{CE} = \frac{EF_{EM+CE}}{EF_{EM}} = \frac{I_{surf}(EM+CE)/N_{surf}(EM+CE)}{I_{surf}(EM)/N_{surf}(EM)}$$  \hspace{2cm} (7c)

Here, $I_{surf}$, $N_{surf}$ are SERS intensity and amount of DNA molecules measured on the surface, and $I_{bulk}$, $N_{bulk}$ numerical simulation results of ground tr are Raman bulk intensity and amount of DNA molecule measured in the bulk. Because the amount of DNA solution drop cast on the substrate is similar, we can assume the number of molecules that got excited from the incoming laser is the same regardless of the CE effect, $N_{surf}(EM+CE) = N_{surf}(EM)$. Therefore, Eq. (7c) becomes

$$\Gamma_{CE} = \frac{I_{surf}(EM+CE)}{I_{surf}(EM)}$$  \hspace{2cm} (8)

A7. Signal processing of SERS measurements

After being collected, the SERS spectra are put through Gaussian smoothing, Normalization and Multiplicative Scatter Correction (MSC) before putting into calibration curve. For single-feature linear regression, the data only undergoes Gaussian smoothing and normalization. For multiple-feature linear regression, the data undergoes all the pre-processing steps: Gaussian smoothing, normalization and MSC. The reason has been explained in the text above. To normalize data in between 0 and 1, each spectrum is scaled using the following equation:

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)}$$  \hspace{2cm} (9)

Where $X$ represent the SERS intensities in one spectrum. Afterwards, the MSC can be either performed or not on the dataset. In the MSC processing, we assume that the measured spectrum was scaled and added with some white noise, which we need to remove to make the data closer to the ideal data. In other words, the light scattering or change in path length for each spectrum in this case is estimated relative to that of an ideal spectrum:

$$X_{c} = b \times R_{c} + \epsilon$$  \hspace{2cm} (10)

$$P_{E}(\epsilon; b, \nu) = \frac{1}{\sqrt{2\pi\sigma_{E}^{2}}} \exp \left(-\frac{\epsilon^{2}}{2\sigma_{E}^{2}}\right)$$  \hspace{2cm} (11)

Here we assume that the noise is Gaussian white noise: $E \sim G(0, \nu I)$. Because we have $\epsilon = X_{c} - b \times R_{c}$, the variable $X_{c}$ also has a Gaussian PDF:

$$P_{X_{c} \mid R}(x_{c} | r; b, \nu) = \frac{1}{\sqrt{2\pi
u I}} \exp \left(-\frac{1}{2\nu} \|X_{c} - bR_{c}\|^{2}\right)$$  \hspace{2cm} (12)
The scattering factor \( b \) can be calculated using the Maximum Likelihood estimation (MLE) function:

\[
    b = \text{argmax}_b \ln P_{X_c|R_c}(X_c|R_c|b, \nu) \\
    = \text{argmax}_b \left( -\frac{1}{2} \ln 2\pi\nu I - \frac{1}{2\nu} ||X_c - bR_c||^2 \right) \\
    = \text{argmin}_b ||X_c - bR_c||^2
\]  

To find value of \( b \) such that \( ||X_c - bR_c||^2 \) is minimum, we need to find the zero value for the first derivative:

\[
    \frac{d}{db} ||X_c - bR_c||^2 = 0 \\
    \frac{d}{db} \sum_i (x_i - br_i)^2 = 0 \\
    \sum_i \frac{d}{db} (x_i^2 - 2x_i br_i + br_i^2) = 0 \\
    \sum_i (-2x_i r_i + 2br_i^2) = 0 \\
    b = \frac{\sum X_c R_c}{\sum R_c^2}
\]  

We input data that has and has not been pre-processed with MSC into our two linear regression methods, and compared the detection errors obtained from the two methods. The results are shown in Fig. 13, 14, 15.

Fig. 11. SERS spectra of different ssDNA compositions binding on gold nanorod substrate before and after MSC processing
Fig. 12. SERS spectra of different ssDNA compositions binding on silver nanorod substrate before and after MSC processing

Table 6. Detection SD of ssDNA composition (in percentage), bound to gold and silver nanorods, before and after MSC pre-processing of data.

| Linear Regression methods | Gold detection SD (%) | Silver detection SD (%) |
|---------------------------|-----------------------|------------------------|
| Before MSC                | After MSC             | Before MSC             | After MSC             |
| Single-feature            | 7.15                  | 7.32                   | 2.37                  | 2.55                   |
| Multiple-feature          | 3.88                  | 2.20                   | 2.82                  | 1.36                   |
A7.1. MSC processing on multiple-feature PCA regression

Fig. 13. PCA regression before and after MSC processing of ssDNA binding on gold nanorod substrate

Fig. 14. PCA regression before and after MSC processing of ssDNA binding on silver nanorod substrate.

A7.2. MSC processing on single-feature linear regression
A8. Probability functions with lognormal distribution fit

For the single-feature linear regression calculation, the data was fit under Lognormal Distribution because the ratios of peak cannot be negative, and this distribution was proven before to have high goodness of fit [16]. In this paper, the PDFs and CDFs are plotted again just to confirm the fit still holds. The plots were created from the \textit{histfit} function in MATLAB. If the empirical CDF is similar to the theoretical CDF, the chosen distribution is a good fit for our data. From the plots, we can see that the fit was not as good for gold compared to silver. As the percentage of Adenine increase, the histogram becomes more skewed and it was unable to find a good fit for the data. However, lognormal distribution was the best choice compared to other type of distribution. The lognormal fit was good for other mixtures, and exceptionally good for those measured on silver.
Fig. 16. Probability density function and cumulative distribution function of the ratio $A/(A+C)$ of different DNA mixtures binding on gold.
Fig. 17. Probability density function and cumulative distribution function of the ratio $A/(A+C)$ of different DNA mixtures binding on silver.
A9. Single-feature linear regression model and DNA composition calculation

The single-feature Linear Regression was performed in MATLAB. We used a built-in MATLAB function called \textit{LinearModel}, which takes in the five training datapoints (lognormal of ratio A/(A+C) for the five DNA compositions) and give out a line that fit with the five datapoints. The testing data is then plugged into this linear equation for the prediction of DNA composition. The quantification error for this case is calculated from the RMSE of the fitting model. Consider the general form of the regression line and of the DNA composition $C_A$ as:

$$f(C_A) = \ln(R) = a_1 + a_2 C_A$$

$$C_A = A_1 \% \pm A_2 \%$$ (15)

Where $R$ is the ratio of the two peaks mentioned in the main text, $A_1$ is the central value of adenine percentage, and $A_2$ is the prediction uncertainty. With the standard deviation of the peak ratios $\text{std}(R)$ included, the percentage of adenine becomes:

$$C_{A+} = \frac{\ln(R \pm \text{std}(R)) - a_1}{a_2}$$ (16a)

$$A_1 = \frac{\ln(R) - a_1}{a_2}$$ (16b)

$$A_2 = A_1 - C_{A-} = A_1 - \frac{\ln(R - \text{std}(R)) - a_1}{a_2} = \frac{\ln(R) - \ln(R - \text{std}(R))}{a_2}$$ (16c)

For example, the lines derived from the \textit{LinearModel} function for gold and silver without MSC processing are:

$$f_{Au}(x) = 0.0220x - 1.72$$ (17a)

$$f_{Ag}(x) = 0.0248x - 2.03$$ (17b)

Therefore, from the lognormal of the ratio obtained from the testing sample, we can estimate the adenine percentage and the prediction uncertainty.

The PCA linear regression was performed in Python notebook. A built-in function \textit{PCA} from \textit{sklearn} library was used to transform and reduce the dimension number of the input matrix. Afterwards, another built-in function \textit{LinearRegression} also from \textit{sklearn} was used to take in the PCA-processed data for training and prediction of the testing sample.

A10. Determination of the PC number for training and testing data in PCA linear regression model

Fig. 18 indicates that for both gold (Fig. 18(a)) and silver (Fig. 18(b)), the variance is a rapidly converging function towards unity; at 10 PCs about 97.5\% of the data is explained whereas above 30 PCs approximately 100\% of the data is explained for ssDNA binding.

From the red dotted lines in Fig. 18, we conclude that the optimal PC numbers which correspond to the lowest amount of quantification errors for gold and silver substrates, are $n = 43$ and $n = 145$, respectively.
Fig. 18. Explained variance and MSE of prediction plots from gold (a) and silver (b) nanorod array substrates. The explained variance plot (blue) shows the number of PCs required to express 100% of the data variance (where explained variance = 1). The MSE plot (red) points out the number of PCs needed for the detection error to be lowest.

A11. Predicted $C_A$ values for different test datasets using different ML models and training sets

We have 200 spectra in test sets #1 and #2, and 400 spectra in test set #3. The average RR and SD values were calculated based on the value collected from each test set and the number of spectra in each set. In particular, we will have:

$$\text{Average RR} = \frac{2 \times |\text{test}\#1 - GT| + 2 \times |\text{test}\#2 - GT| + 4 \times |\text{test}\#3 - GT|}{8}$$

$$\text{Average SD} = \sqrt{\frac{2 \times (\text{RMSE}\#1)^2 + 2 \times (\text{RMSE}\#2)^2 + 4 \times (\text{RMSE}\#3)^2}{8}}$$

$$GT = 54\%$$

A11.1. Single-feature linear regression

| Test set #1 | Test set #2 | Test set #3 | Ave. RR | Ave. SD |
|------------|------------|------------|---------|---------|
| Au         | 51.2\% ± 7.32\% | 53.6\% ± 5.38\% | 56.2\% ± 3.49\% | 1.90\% | 5.17\% |
| Ag         | 54.8\% ± 2.55\% | 57.4\% ± 6.65\% | 58.3\% ± 5.85\% | 3.20\% | 5.46\% |

A11.2. PCA multiple-feature linear regression

| Train set | Test set #1 | Test set #2 | Test set #3 | Ave. RR | Ave. SD |
|-----------|------------|------------|------------|---------|---------|
| #1        | 53.5\% ± 2.17\% | 47.9\% ± 1.24\% | 44.8\% ± 3.82\% | 6.25\% | 2.98\% |
| (#1 + #2) | 52.6\% ± 2.13\% | 51.9\% ± 1.30\% | 49.6\% ± 3.76\% | 3.08\% | 2.94\% |
| (#1 + #3) | 52.9\% ± 4.45\% | 57.6\% ± 1.51\% | 55.3\% ± 1.67\% | 1.83\% | 2.63\% |
| (#1 + #2 + #3) | 52.3\% ± 4.51\% | 57.2\% ± 1.50\% | 55.8\% ± 1.59\% | 2.13\% | 2.63\% |
Table 9. Predicted values of $C_A$ on silver for different test sets using different number of datasets for training PCA multiple-feature linear regression.

|                   | Test set #1 | Test set #2 | Test set #3 | Ave. RR | Ave. SD |
|-------------------|-------------|-------------|-------------|---------|---------|
| Train set #1      | 54.3% ± 1.30% | 41.3% ± 4.69% | 46.3% ± 3.65% | 7.10%   | 3.55%   |
| Train set (#1 + #2)| 48.9% ± 0.98% | 56.5% ± 3.44% | 53.9% ± 4.97% | 1.95%   | 3.94%   |
| Train set (#1 + #3)| 48.9% ± 3.31% | 54.8% ± 3.15% | 56.5% ± 3.61% | 2.73%   | 3.43%   |
| Train set (#1 + #2 + #3) | 50.4% ± 3.46% | 54.5% ± 3.53% | 56.9% ± 4.02% | 2.48%   | 3.77%   |

Fig. 19. Plots of predicted average adenine composition values when model was trained on different number of datasets. The results were averaged from three test sets described above. As more datasets got trained, a higher detection sensitivity was obtained for the model.

A11.3. Neural Network models

- For (Au + Ag) fusion

Table 10. Predicted values of $C_A$ by NN for different test sets using different number of datasets for training.

|                   | Test set #1 | Test set #2 | Test set #3 | Ave. RR | Ave. SD |
|-------------------|-------------|-------------|-------------|---------|---------|
| Train set #1      | 54.1% ± 1.23% | 52.3% ± 1.22% | 52.3% ± 1.64% | 1.30%   | 1.45%   |
| Train set (#1 + #2)| 50.8% ± 1.99% | 54.3% ± 0.79% | 54.2% ± 1.66% | 0.980%  | 1.59%   |
| Train set (#1 + #3)| 51.4% ± 1.92% | 53.5% ± 1.04% | 53.5% ± 1.24% | 1.03%   | 1.40%   |
| Train set (#1 + #2 + #3) | 52.8% ± 1.54% | 51.3% ± 0.81% | 53.3% ± 1.10% | 1.33%   | 1.17%   |
Fig. 20. RR and SD of the NN fusion regression model under training with different number of sample sets. The results were averaged from three test sets described above.

- For Au only

**Table 11.** NN model predicting the percentage of adenine base in ssDNA molecule adsorbed to gold nanorod substrate, for different number of datasets used for training.

|                  | Test set #1 | Test set #2 | Test set #3 | Ave. RR | Ave. SD |
|------------------|-------------|-------------|-------------|---------|---------|
| Train set #1     | 52.6% ± 2.03% | 60.4% ± 2.14% | 61.8% ± 3.10% | 5.85%   | 2.64%   |
| Train set (#1 + #2) | 51.0% ± 1.80% | 53.6% ± 0.80% | 54.2% ± 2.21% | 0.95%   | 1.85%   |
| Train set (#1 + #3) | 50.0% ± 1.39% | 50.3% ± 0.91% | 52.9% ± 2.79% | 2.48%   | 2.14%   |
| Train set (#1 + #2 + #3) | 49.4% ± 1.55% | 53.8% ± 0.80% | 54.2% ± 2.41% | 1.30%   | 1.91%   |

- For Ag only

**Table 12.** NN model predicting the percentage of adenine base in ssDNA molecule adsorbed to silver nanorod substrate, for different number of datasets used for training.

|                  | Test set #1 | Test set #2 | Test set #3 | Ave. RR | Ave. SD |
|------------------|-------------|-------------|-------------|---------|---------|
| Train set #1     | 54.7% ± 1.11% | 48.5% ± 1.59% | 47.6% ± 1.63% | 4.75%   | 1.51%   |
| Train set (#1 + #2) | 53.6% ± 1.51% | 53.9% ± 1.43% | 53.1% ± 1.56% | 0.575%  | 1.52%   |
| Train set (#1 + #3) | 52.7% ± 1.55% | 54.0% ± 1.86% | 53.0% ± 2.01% | 0.825%  | 1.87%   |
| Train set (#1 + #2 + #3) | 52.6% ± 1.00% | 53.8% ± 1.77% | 53.8% ± 2.19% | 0.500%  | 1.85%   |
Fig. 21. RR and SD of the NN single-metal regression models for gold and silver under training with different number of sample sets. The results were averaged from three test sets described above.

Fig. 22. Comparison of the detection RR and SD values of different ML models; single-feature linear regression, PCA multiple-feature linear regression, Neural Network single-metal and and Neural network multi-metal. Results from training on one dataset (left), and on multiple datasets (right)

**Funding**

Defense Advanced Research Projects Agency (DSO’s NLM and NAC Programs); National Science Foundation (CBET-1704085, CCF-1640227, DMR-1707641, ECCS-180789, ECCS-190184); Semiconductor Research Corporation; Army Research Office; Office of Naval Research; Cymer.

**Acknowledgments**

We would like to thank Dr. Lindsay Freeman for her guidance on building biochemical model simulations and DNA solution preparation techniques, which have been used in this paper. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.
Disclosures

The authors declare no conflicts of interest.

References

1. M. Fleischmann, P. Hendra, and A. McQuillan, “Raman spectra of pyridine adsorbed at a silver electrode,” Chem. Phys. Lett. 26(2), 163–166 (1974).
2. D. L. Jeanmaire and R. P. Van Duyne, “Surface raman spectrotelectrochemistry: Part I. Heterocyclic, aromatic, and aliphatic amines adsorbed on the anodized silver electrode,” J. Electroanal. Chem. Interfacial Electrochem. 84(1), 1–20 (1977).
3. M. G. Albrecht and J. A. Creighton, “Anomally intense raman spectra of pyridine at a silver electrode,” J. Am. Chem. Soc. 99(15), 5215–5217 (1977).
4. J. R. Lombardi and R. L. Birke, “A Unified Approach to Surface-Enhanced Raman Spectroscopy,” J. Phys. Chem. C 112(14), 5605–5617 (2008).
5. B. Sharma, R. R. Frontiera, A.-I. Henry, E. Ringe, and R. P. Van Duyne, “SERS: Materials, applications, and the future,” Mater. Today 15(1-2), 16–25 (2012).
6. S. Arsalani, T. Ghodselahi, T. Neishaboornejad, and O. Baffa, “DNA Detection Based on Localized Surface Plasmon Resonance Spectroscopy of Ag@Au Biocomposite Nanoparticles,” Plasmonics 14(6), 1419–1426 (2019).
7. W. H. Kim, J. U. Lee, S. Song, S. Kim, Y. J. Choi, and S. I. Sim, “A label-free, ultra-highly sensitive and multiplexed SERS nanoplasmic biosensor for mRNA detection using a head-flocked gold nanopillar,” Analyst 144(5), 1768–1776 (2019).
8. G. Braun, S. J. Lee, M. Dante, T.-Q. Nguyen, M. Moskovits, and N. Reich, “Surface-Enhanced Raman Spectroscopy for DNA Detection by Nanoparticle Assembly onto Smooth Metal Films,” J. Am. Chem. Soc. 129(20), 6378–6379 (2007).
9. F. Rodríguez-Trelles, R. Tarrío, and F. J. Ayala, “Fluctuating Mutation Bias and the Evolution of Base Composition in Drosophila,” J. Mol. Evol. 50(1), 1–10 (2000).
10. R. Tarrío, F. Rodríguez-Trelles, and F. J. Ayala, “Shared Nucleotide Composition Biases Among Species and Their Impact on Phylogenetic Reconstructions of the Drosophilidae,” Mol. Biol. Evol. 18(8), 1464–1473 (2001).
11. X. Tian, J. E. Strassmann, and D. C. Queller, “Genome Nucleotide Composition Shapes Variation in Simple Sequence Repeats,” Mol. Biol. Evol. 28(2), 899–909 (2011).
12. G. A. Kwong, C. G. Radu, K. Hwang, C. J. Shu, C. Ma, R. C. Koya, B. Comin-Anduix, S. R. Hadrup, R. C. Bailey, O. N. Witte, T. N. Schumacher, A. Ribas, and J. R. Heath, “Modular Nucleic Acid Assembled pMHC Microarrays for Multiplexed Sorting of Antigen-Specific T Cells,” J. Am. Chem. Soc. 131(28), 9695–9703 (2009).
13. S. N. Dahotre, Y. M. Chang, A. Wieland, S. R. Stammen, and G. A. Kwong, “Individually addressable and dynamic DNA gates for multiplexed cell sorting,” Proc. Natl. Acad. Sci. U. S. A. 115(17), 4357–4362 (2018).
14. R. Cotton, “Current methods of mutation detection,” Mutat. Res. Mol. Mech. Mutagen. 285(1), 125–144 (1993).
15. L.-J. Xu, Z.-C. Lei, J. Li, C. Zong, C. J. Yang, and B. Ren, “Label-Free Surface-Enhanced Raman Spectroscopy Detection of DNA with Single-Basis Sensitivity,” J. Am. Chem. Soc. 137(15), 5149–5154 (2015).
16. L. M. Freeman, L. Pang, and Y. Fainman, “Self-reference and random sampling approach for label-free identification of DNA composition using plasmonic nanomaterials,” Sci. Rep. 8(1), 7398 (2018).
17. K. Kneipp, Y. Wang, H. Kneipp, L. T. Perelman, I. Itzkan, R. R. Dasari, and M. S. Feld, “Single Molecule Detection Using Surface-Enhanced Raman Scattering (SERS),” Phys. Rev. Lett. 78(9), 1667–1670 (1997).
18. E. Papadopoulou and S. E. J. Bell, “Label-Free Detection of Single-Base Mismatches in DNA by Surface-Enhanced Raman Spectroscopy,” Angew. Chem. Int. Ed. 50(39), 9058–9061 (2011).
19. S. G. Harron, “The Controversial Orientation of Adenine on Gold and Silver,” ChemPhysChem 19, 1003–1015 (2018).
20. M. Moskovits, “Surface-enhanced spectroscopy,” Rev. Mod. Phys. 57(3), 783–826 (1985).
21. A. Campion, J. E. Ivanecy, C. M. Child, and M. Foster, “On the Mechanism of Chemical Enhancement in Surface-Enhanced Raman Scattering,” J. Am. Chem. Soc. 117(47), 11807–11808 (1995).
22. W. E. Doering and S. Nie, “Single-Molecule and Single-Nanoparticle SERS: Examining the Roles of Surface Active Sites and Chemical Enhancement,” J. Phys. Chem. B 106(2), 311–317 (2002).
23. J.-P. Su, Y.-T. Lee, S.-Y. Lu, and J. S. Lin, “Chemical mechanism of surface-enhanced raman scattering spectrum of pyridine adsorbed on Ag cluster: Ab initio molecular dynamics approach,” J. Comput. Chem. 34(32), 2806–2815 (2013).
24. L. M. Freeman, L. Pang, and Y. Fainman, “Maximizing the Electromagnetic and Chemical Resonances of Surface-Enhanced Raman Scattering for Nucleic Acids,” ACS Nano 8(8), 8383–8391 (2014).
25. J. S. Harder, Q. Matthews, M. Isabell, A. G. Brolo, J. J. Lum, and A. Jirasek, “A Raman Spectroscopic Study of Cell Response to Clinical Doses of Ionizing Radiation,” Appl. Spectrosc. 69(2), 193–204 (2015).
26. D. Bharanidharan and N. Gautham, “Principal component analysis of DNA oligonucleotide structural data,” Biochim. Biophys. Res. Commun. 340(4), 1229–1237 (2006).
27. L. Sitole, F. Steffens, and D. Meyer, “Raman Spectroscopy-based Metaboomics of HIV-infected Sera Detects Amino Acid and Glutathione Changes,” Curr. Metabolomics 3(1), 65–75 (2015).
28. B. Bodanese, F. L. Silveira, R. A. Zângaro, M. T. T. Pacheco, C. A. Pasqualucci, and L. Silveira, “Discrimination of Basal Cell Carcinoma and Melanoma from Normal Skin Biopsies in Vitro Through Raman Spectroscopy and Principal Component Analysis,” Photomed. Laser Surg. 30(7), 381–387 (2012).
29. S. Li, Y. Zhang, J. Xu, L. Li, Q. Zeng, L. Lin, Z. Guo, Z. Liu, H. Xiong, and S. Liu, “Noninvasive prostate cancer screening based on serum surface-enhanced Raman spectroscopy and support vector machine,” Appl. Phys. Lett. 105(9), 091104 (2014).
30. R. Dong, S. Weng, L. Yang, and J. Liu, “Detection and Direct Readout of Drugs in Human Urine Using Dynamic Surface-Enhanced Raman Spectroscopy and Support Vector Machines,” Anal. Chem. 87(5), 2937–2944 (2015).
31. A. Barranco, A. Borras, A. R. Gonzalez-Elipe, and A. Palmero, “Perspectives on oblique angle deposition of thin films: From fundamentals to devices,” Prog. Mater. Sci. 76, 59–153 (2016).
32. S. Shanmukh, L. Jones, J. Driskell, Y. Zhao, R. Dluhy, and R. A. Tripp, “Rapid and Sensitive Detection of Respiratory Virus Molecular Signatures Using a Silver Nanorod Array SERS Substrate,” Nano Lett. 6(11), 2630–2636 (2006).
33. S. B. Chuney, S. Shanmukh, R. A. Dluhy, and Y.-P. Zhao, “Aligned silver nanorod arrays produce high sensitivity surface-enhanced Raman spectroscopy substrates,” Appl. Phys. Lett. 87(3), 031908 (2005).
34. R. Gao, Y. Zhang, F. Zhang, S. Guo, Y. Wang, L. Chen, and J. Yang, “SERS polarization-dependent effects for an ordered 3d plasmonic tilted silver nanorod array,” Nanoscale 10(17), 8106–8114 (2018).
35. T. Isaksson and T. Naes, “The Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy,” Appl. Spectrosc. 42(7), 1273–1284 (1988).
36. R. Hecht-Nielsen, “Theory of the backpropagation neural network,” Neural Networks 1, 593–605 vol.1 (1989).
37. J. Bell, Machine Learning: Hands-On for Developers and Technical Professionals (Wiley, 2014).
38. J. Heaton, Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks, Artificial Intelligence for Humans (Createspace Independent Publishing Platform, 2015).
39. M. J. Frisch, et al., “Gaussian 09, Revision D.01,” (2009).
40. J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lefika, G. D. Peterson, R. Roskies, J. Scott, and N. Wilkins-Diehr, “XSEDE: Accelerating Scientific Discovery,” Comput. Sci. Eng. 16(5), 62–74 (2014).
41. P. L. Polavarapu, “Ab initio vibrational Raman and Raman optical activity spectra,” J. Phys. Chem. 94(21), 8106–8112 (1990).
42. L. M. Freeman, A. Smolyaninov, L. Pang, and Y. Fainman, “Simulated Raman correlation spectroscopy for quantifying nucleic acid-silver composites,” Sci. Rep. 6(1), 23535 (2016).
43. W. Cai, B. Ren, X. Li, C. She, F. Liu, X. Cai, and Z. Tian, “Investigation of surface-enhanced Raman scattering from platinum electrodes using a confocal Raman microscope: dependence of surface roughening pretreatment,” Surf. Sci. 406(1-3), 9–22 (1998).
44. D. P. dos Santos, M. L. A. Temperini, and A. G. Brolo, “Intensity Fluctuations in Single-Molecule Surface-Enhanced Raman Scattering,” Acc. Chem. Res. 52(2), 456–464 (2019).
45. A. B. Myers, “Resonance Raman Intensities and Charge-Transfer Reorganization Energies,” Chem. Rev. 96(3), 911–926 (1996).
46. R. Pilot, R. Signorini, C. Durante, L. Oriani, M. Bhamidipati, and L. Fabris, “A review on surface-enhanced raman scattering,” Biosensors 9(2), 57 (2019).