Electronic Supplementary Material (ESI) for Materials Chemistry Frontiers.

Detection of a SARS-CoV-2 sequence with genosensors using data analysis based on information visualization and machine learning techniques

Juliana Coatrini Soares[1], Andrey Coatrini Soares[2], Valquiria Cruz Rodrigues[1], Pedro Ramon Almeida Oiticica[1], Paulo Augusto Raymundo-Pereira[1], José Luiz Bott-Neto[1], Lorenzo Antonio Buscaglia[1], Lucas Daniel Chiba de Castro[1], Lucas C. Ribas[1,3], Leonardo Scabini[1], Laís C. Brazaca[4,5], Daniel S. Correa[2], Luiz Henrique C. Mattoso[2], Maria Cristina Ferreira de Oliveira[3], André Carlos Ponce Leon Ferreira de Carvalho[3], Emanuel Carrilho[4,5], Odemir Bruno[1], Matias Eliseo Melendez[6], Osvaldo Novais Oliveira Jr[1*].

[1] São Carlos Institute of Physics (IFSC), University of São Paulo (USP), 13566-590 São Carlos, SP, Brazil.
[2] Nanotechnology National Laboratory for Agriculture (LNNA), Embrapa Instrumentação, 13560-970 São Carlos, SP, Brazil.
[3] Institute of Mathematics and Computer Science (ICMC), University of São Paulo (USP), 13566-590 São Carlos, SP, Brazil.
[4] São Carlos Institute of Chemistry (IQSC), University of São Paulo (USP), 13566-590 São Carlos, SP, Brazil.
[5] National Institute of Science and Technology in Bioanalytics – INCTBio, 13083-970 Campinas, SP, Brazil
[6] Pelé Little Prince Research Institute, Little Prince College, Little Prince Complex Curitiba, 80250-060 Curitiba, PR, Brazil

1. Experimental Procedure

1.1 SEM images

Scanning electron microscopy (SEM) images were obtained with a Zeiss-LEO-440 electron microscope equipped with a detector 7060 (Oxford Instruments) and operating at 15 kV. Prior to the analysis, the genosensing units were affixed onto aluminum (Al) stubs and covered with a 3 nm Au layer deposited by sputter coating using a BAL-TEC MED 020 coating system for improved electrical contact and imaging. In order to collect a reliable and representative dataset, 10 images were acquired in duplicate from different regions of each sample. The total magnifications

used in the SEM images were 10,000× and 1,000×, where the latter was chosen to coincide with the upper limit of optical microscopes. This analysis was carried out to verify the plausibility of using optical microscopes in the future for the same purpose. Typical SEM images of the genosensing units are shown in Figure S4.

1.2 Mechanism behind SARS-CoV-2 Detection

The mechanism behind the detection of the SARS-CoV-2 sequence was elucidated using polarization-modulated infrared reflection absorption spectroscopy (PM-IRRAS), which also served to verify the film architecture of the genosensors. The measurements were performed with a spectrophotometer PMI 550 (KSV Instruments), with the Au electrode spectrum as a reference. The incident angle of the incoming IR beam was 81º, and the spectral resolution was 8 cm^{-1}. The PM-IRRAS signal was obtained from *s* and *p* reflectivity components through Eq. 1, where R_p and R_s are the parallel and perpendicular components to the plane of incidence of the IR light, respectively

$$\frac{\Delta R}{R} = \frac{R_p - R_s}{R_p + R_s} \qquad (1)$$

1.3. Fabrication of the Plasmonic Substrates

The fabrication of the plasmonic substrates used in the optical genosensor device was monitored with UV-VIS spectroscopy, with the color change and peak in the spectrum in Figure S1 indicating the formation of nanoparticles after the thermal annealing process. The gold nanoparticles were produced according to methods in the

literature.[1,2] The AFM images were obtained using a Tip made of silicon with rectangular geometry, radius of 7nm (10 nm max), and 42 N/m spring constant (model OTESPAW from Bruker). The frequency of the tip during the measurement was approximately 345.5 kHz and the image was acquired using the software Nasoscope in the Soft TappingMode method. The AFM image presented in the paper was acquired from the Height sensor on a scan area of 3 x 3 μm^2 with sampling resolution is given by 512 samples/line and 512 lines. The sampling resolution represents the pixel size of the acquired image and is given by 3μm/512 = 5.85 nm (aspect ratio 1:1). This is smaller than the tip radius, then the lateral resolution should be limited by the tip radius and shape. It is worth mentioning that the most suitable method to characterize gold nanoparticles is transmission electron microscopy or scanning electron microscopy with a field emission gun. AFM was only used here owing to the ready availability of the instrument and because we took the view that the nanoparticle characterization was not crucial for the analysis of the genosensing results.

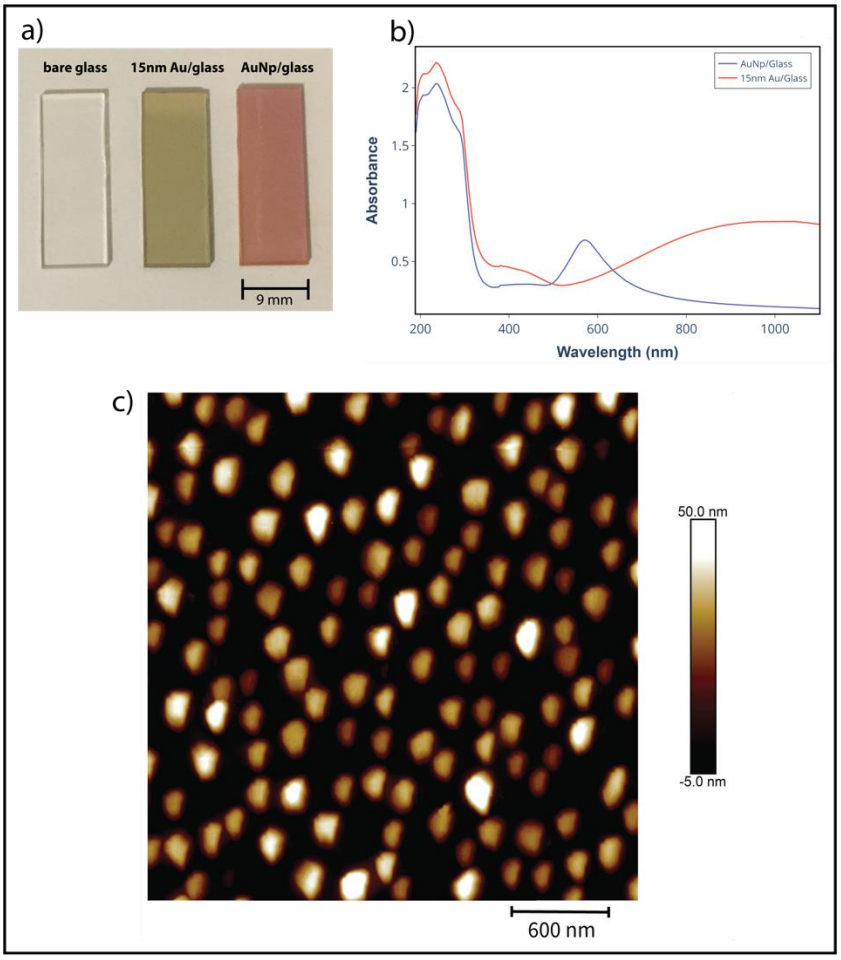


Figure S1. a) Photograph of glass substrates with 25 × 9 mm dimensions. From left to right: bare glass, 15-nm thick gold film evaporated onto the bare glass, and plasmonic substrate formed after thermal annealing. b) UV-Vis spectra of the substrates with evaporated 15-nm thick gold film (red trace) and the plasmonic substrate (blue trace). c) AFM image of the plasmonic substrate.

2. Results and Discussion

The results from the detection of the SARS-CoV-2 positive sequence at various concentrations using impedance spectroscopy with the homemade impedance spectrometer are depicted in the IDMAP plot of Figure S2 (right), shown with a photo of the instrument (left). The data points corresponding to the spectra obtained after the genosensor was exposed to the highest concentrations are located further away from

PBS (marked as cpDNA in the figure).

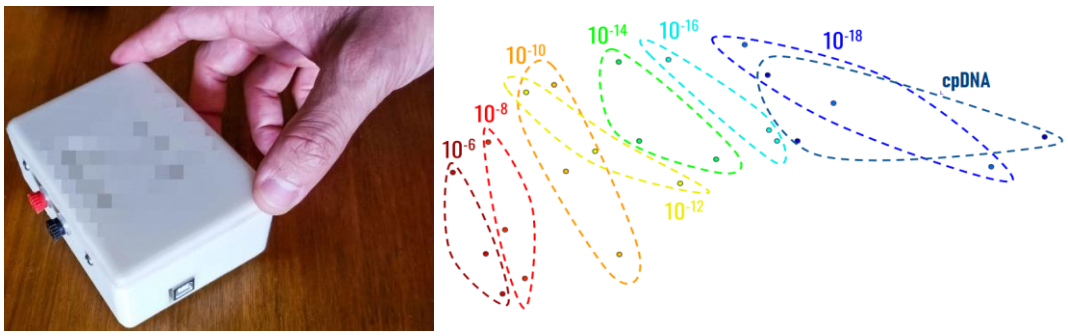


Figure S2. On the left, the low-cost portable impedance spectrometer. On the right, the IDMAP projection of the capacitance spectra for cpDNA (Probe) and various ssDNA SARS-CoV-2 concentrations measured with the homemade impedance spectrometer.

Optical setup for LSPR genosensor detection.

The light coupling and optical alignments were performed using fiber optics and a collimation lens, as illustrated in Figure S3. The collimated beam at the sample has 4 mm diameter and approximately 150 µW radiation power, distributed along with the spectral range of the radiation source. The spectral resolution of the USB4000 spectrometer is limited to 0.20 nm in the region containing the LSPR peak absorption of the plasmonic substrates (approximately 570 nm). A support with a y-stage was used to collect the spectrum from different points on the sample film, and then the difference was measured for each point before and after the interaction with the analyte molecules.

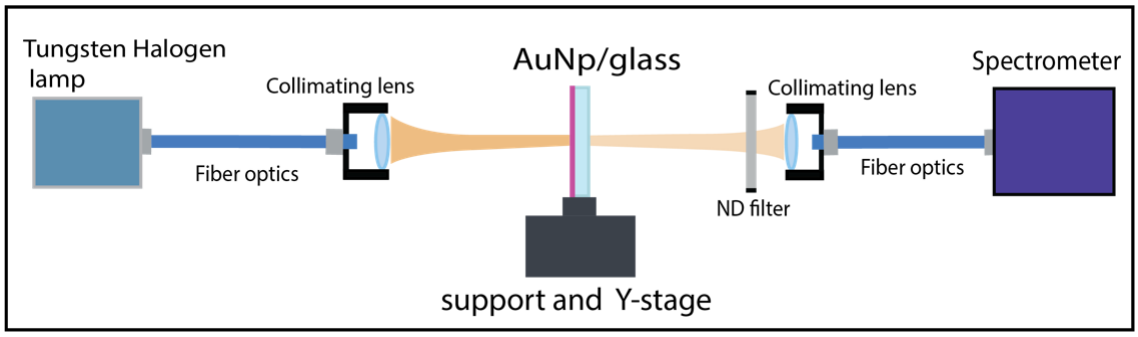


Figure S3. Transmission/absorption setup used in optical SARS-COV-2 detection tests by measuring the wavelength shift of the LSPR spectrum of the plasmonic genosensor.

Typical SEM images (1,000×) of genosensors are shown in Figure S4.

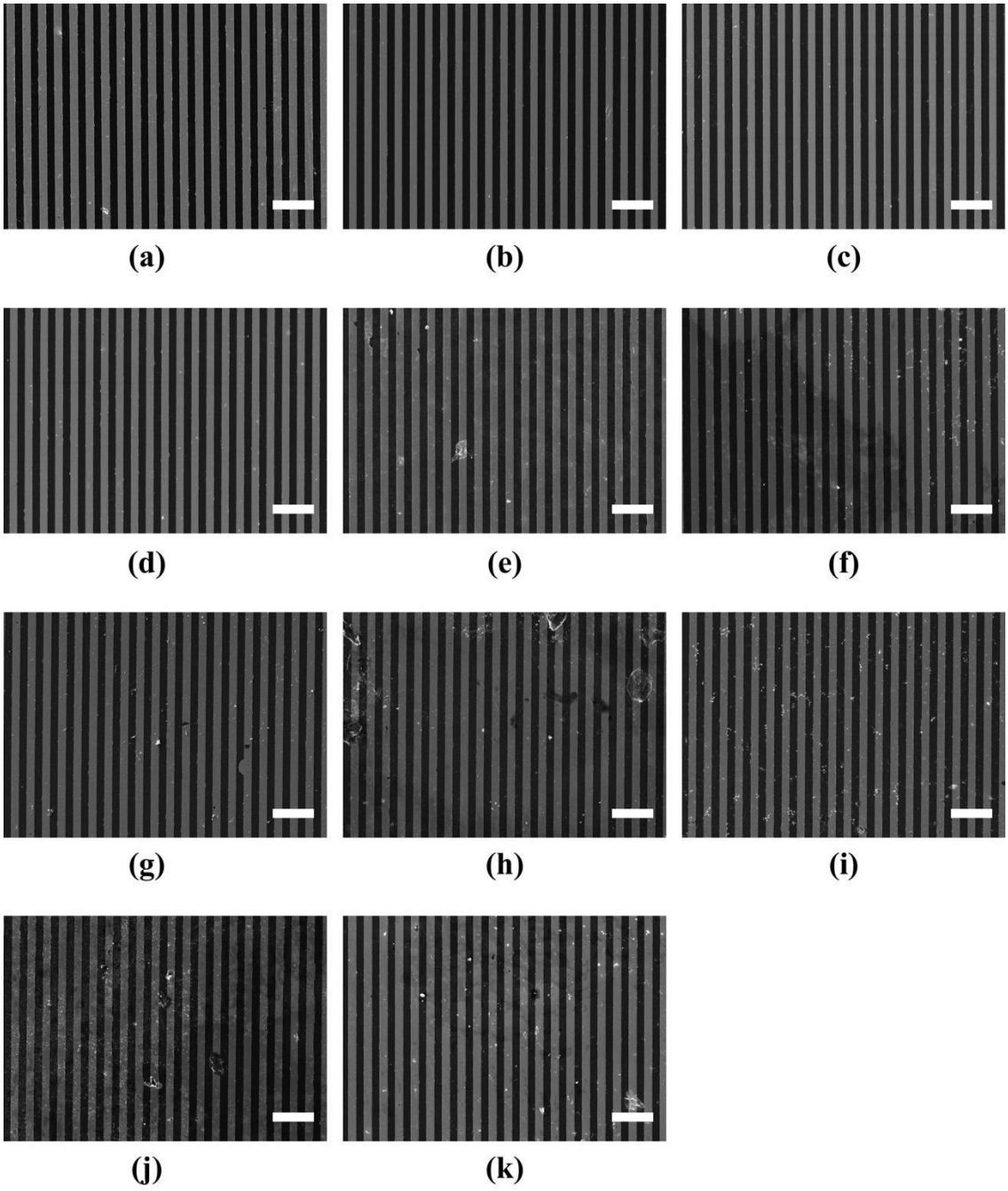


Figure S4. SEM images (1,000×) of the (a) blank control; genosensing units exposed to (b) negative control; (c) HPV16 interferent; (d) PCA3 interferent, and to different concentrations ($molL^{-1}$) of the ssDNA SARS-CoV-2 positive control: (e) 10^{-18} $molL^{-1}$; (f) 10^{-16} $molL^{-1}$; (g) 10^{-14} $molL^{-1}$; (h) 10^{-12} $molL^{-1}$; (i) 10^{-10} $molL^{-1}$; (j) 10^{-8} $molL^{-1}$; (k) 10^{-6} $molL^{-1}$. Scale bar: 50µm.

Figure S5 shows the IDMAP plot of a complete set of experiments with impedance spectroscopy, including the control ones. The most important feature is that the different concentrations of positive sequences for SARS-CoV-2 are clustered separately from negative sequences and other biomarkers and buffers. Also significant is that distinction among the positive sequences is not as high if ethanolamine is used in the genosensors, as indicated in the data cluster 4 in the figure.

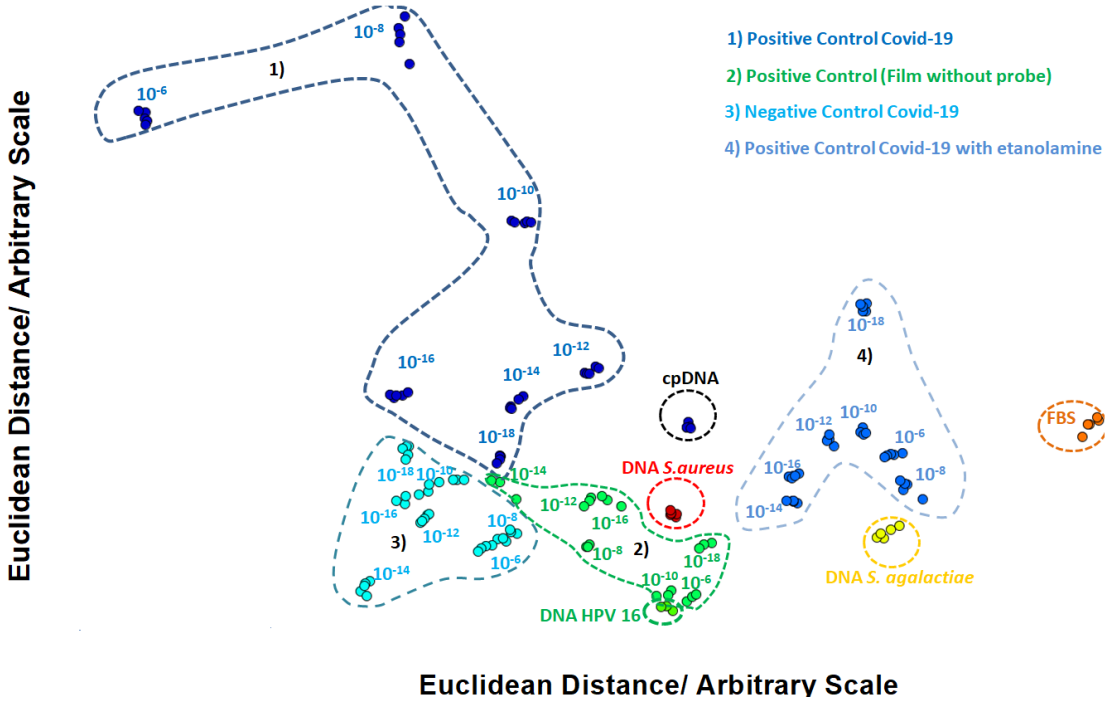


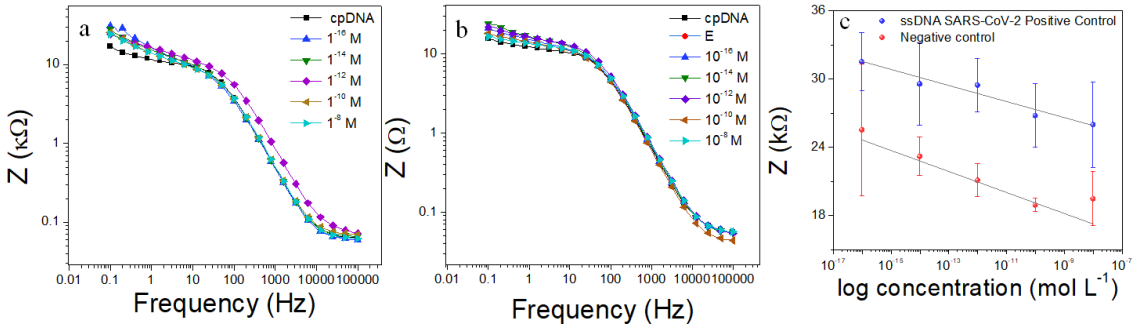
Figure S5. IDMAP projection of the capacitance spectra for various ssDNA SARS-CoV-2 concentrations using genosensors made with a 11-MUA monolayer coated with a layer of DNA sequence. In the projection, the data points from the control experiments (FBS, DNA *S. Agalactiae*, cpDNA, DNA S. aureus, groups 2 and 3) are separated from those of the genosensor exposed to the SARS-CoV-2 DNA concentrations (groups 1 and 4), showing high specificity.

Figures S6a and S6b show the electrochemical impedance spectra with the impedance increasing with the concentration of the complementary and non-complementary sequences mainly at low frequencies (between 0.1 and 100Hz), which is the region governed by changes in the electrical double layer. The changes at 1Hz are

shown in the analytical curve in Figure S6c, which is linear with the logarithm of ssDNA concentration between 1.0×10^{-16} and 1.0×10^{-8} $molL^{-1}$ with equation S1

$$(Z - Z_0) = a + b\log C \quad \text{(S1)}$$

where C is ssDNA concentration. As mentioned in the main text, the distinguishing ability was not as efficient as observed with electrical impedance spectroscopy, according to the IDMAP plots in Figure S6d.



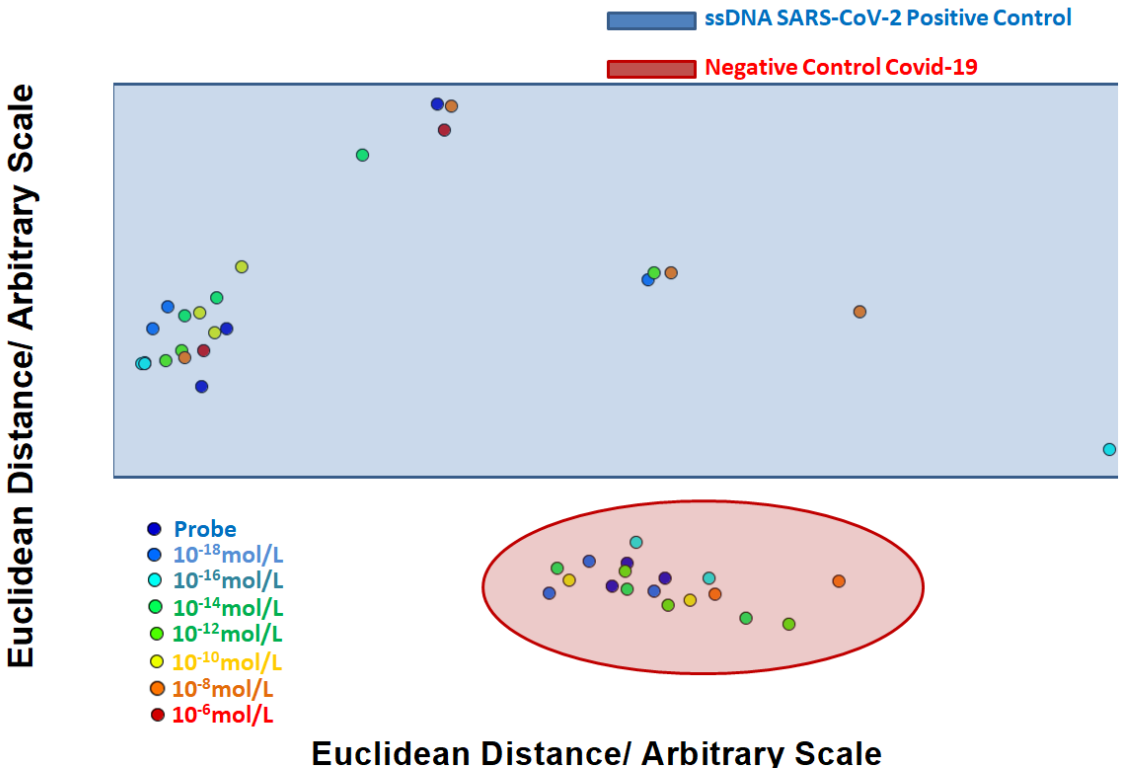


Figure S6. Impedance spectra for genosensors made with Au/SAM-MUA/cpDNA to detect a complementary in (**a**) and non-complementary sequence in (**b**) the concentration range between 1×10^{-18} and 1×10^{-6} $molL^{-1}$. Conditions: 5 $mmolL^{-1}$ $[Fe(CN)6]^{3-/4-}$ in PBS/$MgCl_2$ (1 $mmolL^{-1}$). **(c)** Analytical curves obtained from the electrochemical impedance spectra at 0.1 Hz for the complementary ssDNA sequence (in triplicate). **(d)** IDMAP projection of the capacitance spectra for various ssDNA SARS-CoV-2 concentrations using genosensors made with 11-MUA SAM and coated with an active layer. The silhouette coefficient for electrochemical measurements was -0.18

The results from the selectivity tests with LSPR measurements are shown in the IDMAP plot of Figure S7, where the positive and negative sequences can be distinguished. The measurements with pure PBS/$MgCl_2$ (blank) are similar to those of negative sequences, as expected. In the plot, each point corresponds to an LSPR spectrum normalized by the reference obtained from each genosensor device before the detection test. No feature selection procedure was adopted. Distinction ability can be improved in the future by optimizing the analysis.

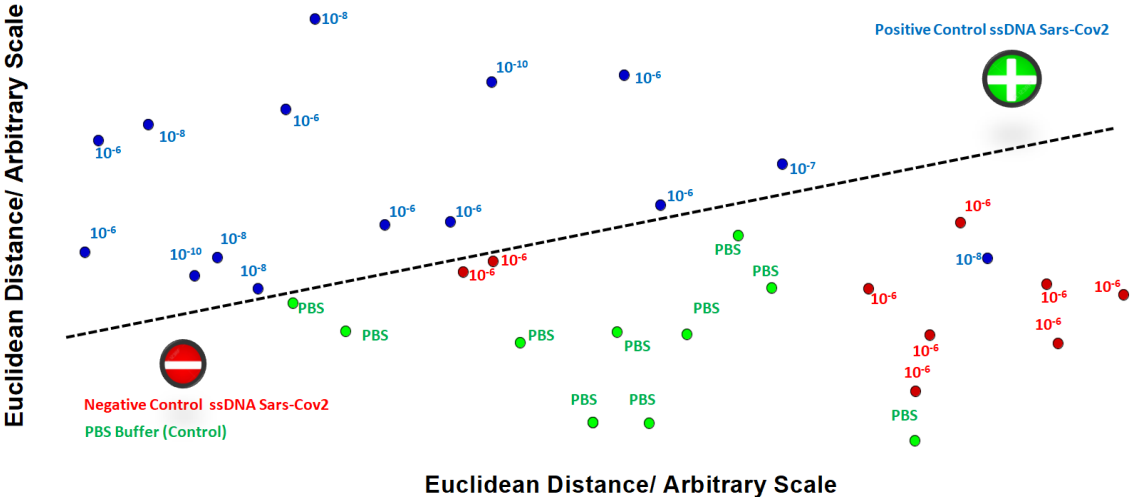


Figure S7. Information visualization plot obtained for the optical detection tests. The dashed line separates the detection tests with ssDNA SARS-CoV-2 from the blank and non-complementary ssDNA test.

Confirming the Mechanism Behind SARS-CoV-2 Detection

Figure S8 shows the PM-IRRAS spectra for the genosensor and after its exposure to different concentrations of ssDNA SARS-CoV-2. Figure S9A shows the PM-IRRAS spectra of genosensors exposed to a positive and a negative sequence for SARS-CoV-2. Three regions are affected by different types of interactions. The CH_2 band shifts to shorter wavelengths after interaction between negative control and Probe (1418□1384 cm^{-1}),[3,4] probably owing to a non-specific interaction that increases the oscillation energy of the C-H dipole. The same applies to the cytosine band from 1545 to 1525 cm^{-1}.[5] Changes were observed in the orientation of the guanine/COOH group at 1735 cm^{-1},[4,5] also due to non-specific interactions. In addition to the region of nitrogenous bases, interactions between the active layer and positive/negative controls affect the bands at (2844/2837 cm^{-1})[3,4] and 2920 cm^{-1}. According to Figure S9B, there are two types of interaction leading to: 1) displacement of the CH_3 band to shorter

wavenumbers, indicating a possible non-specific interaction between the negative control and the probe; 2) Increase of the band area of CH_2 and CH_3 groups after interaction between the positive control and the probe. Nevertheless, it should be noted that the PM-IRRAS spectra indicate that the genosensors were able to distinguish positive and negative samples, despite the non-specific adsorption observed for the negative sequences.

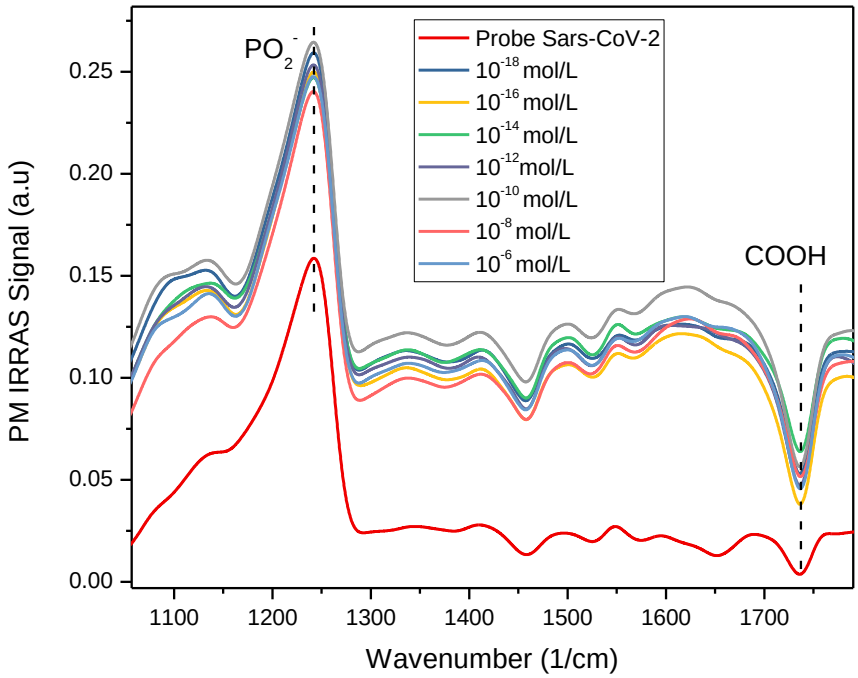
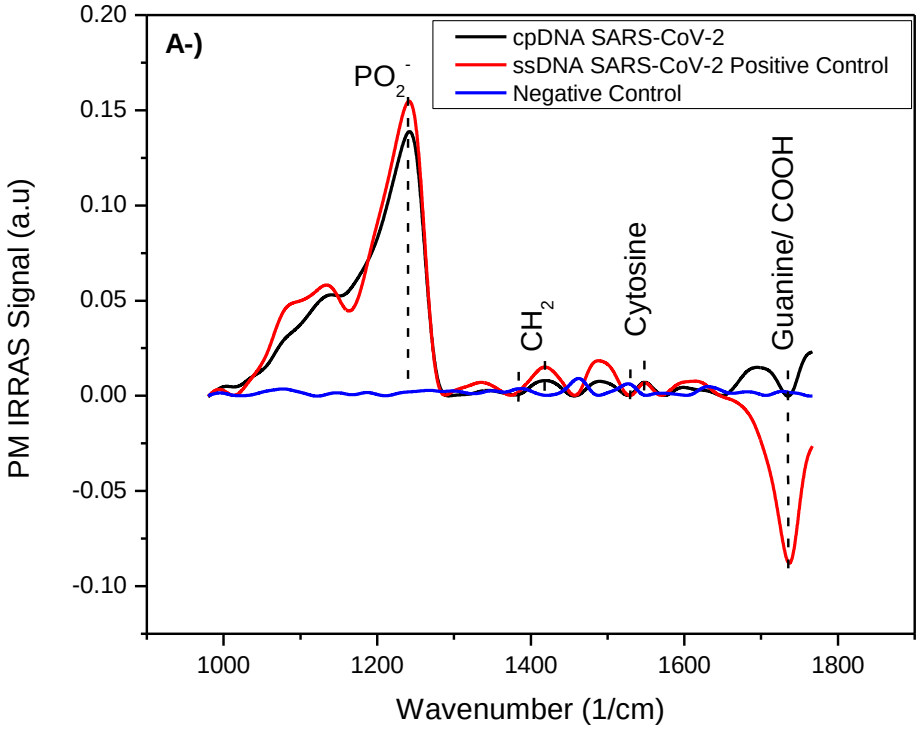


Figure S8. PM-IRRAS spectra for the 11-MUA/cpDNA genosensor before and after exposure to different concentrations of ssDNA SARS-CoV-2 positive sequences (complementary). Also shown in red trace is the spectrum for the genosensor exposed to PBS.


A-)
cpDNA SARS-CoV-2
ssDNA SARS-CoV-2 Positive Control
Negative Control
PO2-
CH2
Cytosine
Guanine/ COOH
PM IRRAS Signal (a.u)
Wavenumber (1/cm)
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
1000
1200
1400
1600
1800

.

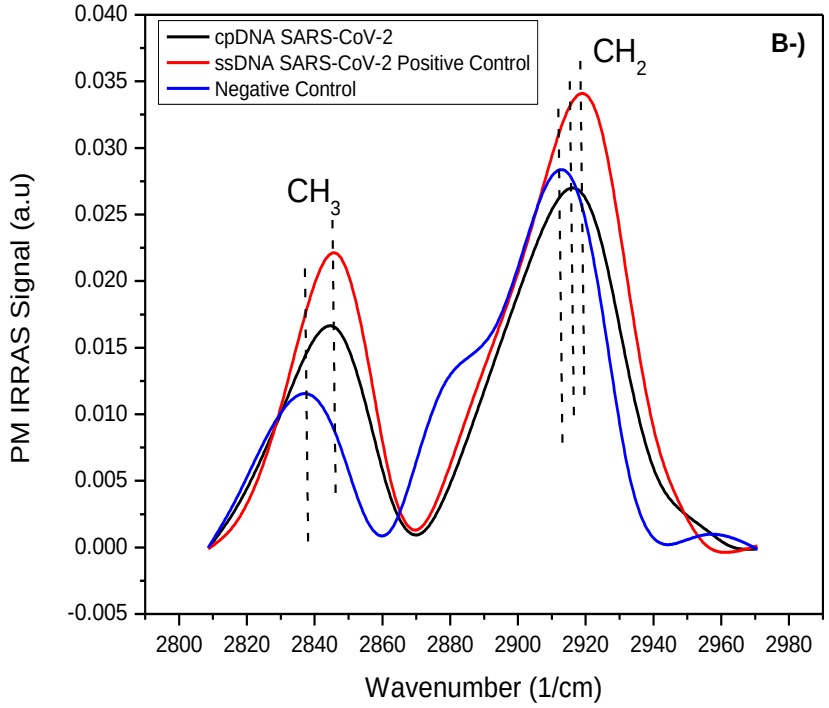

B-)
cpDNA SARS-CoV-2
ssDNA SARS-CoV-2 Positive Control
Negative Control
CH2
CH3
PM IRRAS Signal (a.u)
Wavenumber (1/cm)
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
-0.005
2800
2820
2840
2860
2880
2900
2920
2940
2960
2980

Figure S9. Normalized PM-IRRAS spectra of 11-MUA films functionalized with cpDNA (black line), cpDNA/ssDNA SARS-CoV-2 negative (blue line) and cpDNA/ssDNA Sars-CoV-2 positive (red line) to study the adsorption process within (a) 997-1800 cm^{-1} and (b) 2810-2970 cm^{-1} region.

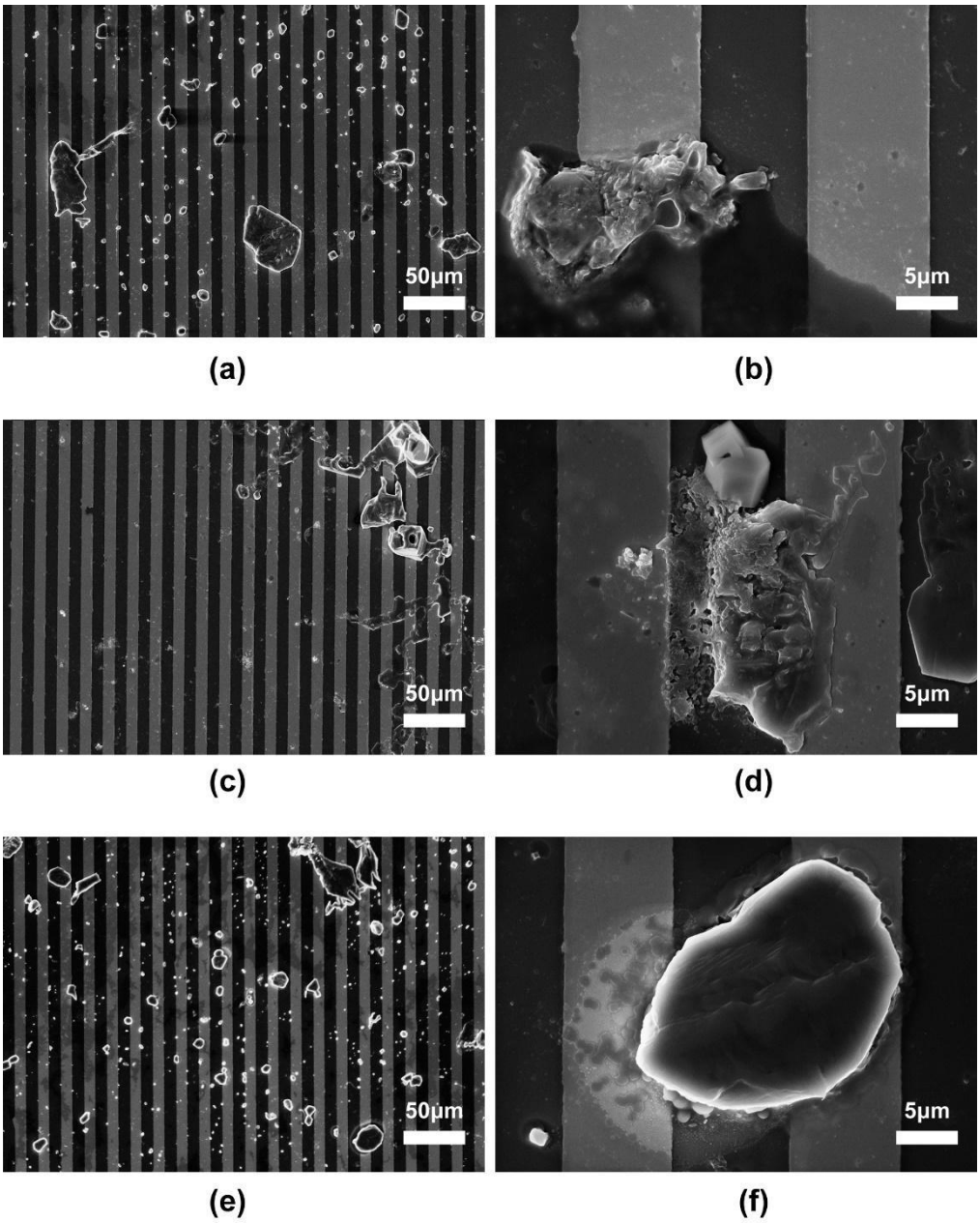


Figure S10. A set of representative SEM images of the genosensing units exposed to different concentrations ($molL^{-1}$) of the ssDNA SARS-CoV-2 positive control: (a - b) 10^{-16} $molL^{-1}$; (c - d) 10^{-10} $molL^{-1}$; and (e - f) 10^{-6} $molL^{-1}$. Here it is possible to observe the presence of defects and artifacts in two different magnifications: (a; c and e) 1,000×; and (b; d and f) 10,000×.

Table S1. Characteristics of the balanced datasets used in the classifications.

Experiment	Total Examples	Examples per Class
multiclass (1,000×)	200	20
binary (1,000×)	200	50
multiclass (10,000×)	150	10
binary (10,000×)	150	10

Table S2: Accuracy in Binary and Multiclass (9 classes) classifications for the SEM images with 10,000×.

Methods	Binary		Multiclass	
	LDA	SVM	LDA	SVM
AHP	90.90 (7.05)	95.50 (4.41)	48.00 (6.25)	58.49 (5.13)
CLBP	95.65 (3.93)	96.95 (3.55)	63.51 (4.91)	61.94 (4.56)
CNTD	95.00 (5.73)	93.95 (5.19)	49.50 (6.24)	56.51 (4.55)
GLDM	88.25 (9.98)	94.85 (4.68)	32.45 (6.04)	54.21 (4.85)
LCFNN	95.20 (5.17)	94.95 (5.25)	45.77 (5.64)	47.40 (5.02)
Fourier	66.00 (13.71)	76.45 (10.52)	32.55 (5.97)	35.29 (5.09)

Machine Learning pipeline

Here we give more details about the machine learning experiment performed in this work. Figure S11 illustrates the pipeline employed for the detection of SARS–CoV-2 based on machine learning. Figure 11(a) shows the steps to train a classifier from a set of training images and obtain a prediction model for detection. In this case, we get feature vectors with image analysis techniques, and then we use these feature vectors to train the classifiers. Once a prediction model is obtained, new images can be classified as indicated in Figure S11(b). The original images we employed on this work are available for download at https://github.com/scabini/covid_genosensors_ML.

Validation

We adopted a 10-fold cross-validation scheme to divide the dataset for training and testing purposes to validate the machine learning framework. This scheme separates the dataset into 10 balanced subsets. Then, one fold is used for testing, and the remaining folds for training the model. This process is repeated using all folds for testing, and the accuracy is computed as the average of the 10 folds. The accuracy is the percentage of correctly predicted samples. Next, we report the final accuracy as the average of 100 random trials. In other words, the 10 folds are chosen randomly 100 times, and for each trial, we test the model and obtain an accuracy.

Feature extraction with image analysis techniques

The feature extraction based on image analysis techniques consists of extracting significant features from the images to compose a feature vector (i.e., a descriptor). This vector represents the image in the training and classification process. The features are real values that quantify essential information in the images related to micro and macro texture patterns. To obtain discriminative features, we test several image analysis techniques that explore different manners to describe the SEM images' patterns.

GLDM is a statistical-based approach that computes statistics from the gray-level co-occurrence matrices. AHP and CLBP are structural-based approaches that explore primitive arrangements in the images. We also employ Fourier descriptors to analyze the textures in the frequency domain. On the other hand, the Fractal and CNDT describe

the texture images using mathematical models representing and computing the image complexity to compose the feature vector. More recently, learning-based methods have been used to learn the image features with promising results. This work uses the LCFNN method that learns the features from a complex network framework using a randomized neural network. Besides, we tested other deep and complex neural network architectures to extract features. They are DenseNet201, InceptionResNetV2 and MobileNet. All features obtained with each of these methods are available for download at https://github.com/scabini/covid_genosensors_ML.

Classifiers

Linear Discriminant Analysis (LDA)

The Linear Discriminant Analysi is a well-known method based on a linear combination of features used for dimensionality reduction, regression, and classification. This method applies a transformation in feature space to obtain uncorrelated features based on linear combinations. Such transformations aim to estimate a feature projection in which the variance between classes is larger than the intra-class variance. For a classification task, let us consider a random variable X that belongs to one class i = 1, ..., C, and a density function fi(x). The goal is to employ a discriminant rule to separate the feature space in C (number of classes) different regions where each region corresponds to a class. Thus, a new sample x is predicted to be of class k if x is in the region representing class k. That is, for each class i, the density function fi(x) is computed and the highest value k = arg maxi=1,...,C{fi(x)} is predicted as class of x.

The function fi(x) is defined by

$$f_i(x) = -\frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i) \quad \text{(S2)}$$

where Σ is the covariance matrix and μ_i is the average of the class i.

Support Vector Machine (SVM)

SVM is one of the most popular supervised learning techniques for classification and regression. The method aims to find a set of maximum-margin hyperplanes in high-dimensional space to separate the classes from a set of training data. Thus, this method learns the decision frontier using a set of support/training vectors. In an SVM with a linear kernel, the decision frontier is given by w.xi+b=0, where w and b are parameters and the class yi for the vector xi is predicted by

$$y_i = \begin{cases} 1, & \text{if } w.x_i + b > 0 \\ -1, & \text{if } w.x_i + b < 0 \end{cases} \quad \text{(S3)}$$

In the SVM training process, the parameters w and b that represent the decision frontier are estimated from the training set. These parameters are calculated to maximize the decision frontier, which is equal to minimize the objective function

$$\min_w = \frac{|w|^2}{w} \quad \text{(S4)}$$

Since the objective function is quadratic and the parameters are linear, the Lagrangian multipliers solve the minimization. For nonlinear problems, the method uses kernel trick in the vectors xi to transform the nonlinear space to a linearly separable space.

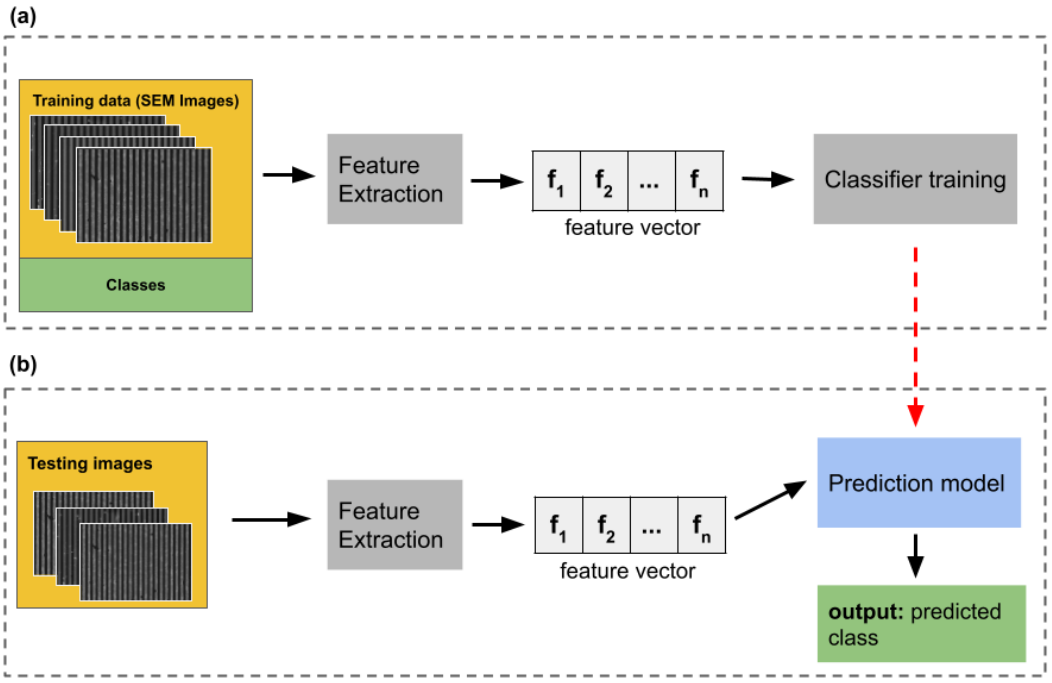


Figure S11. Pipeline of the machine learning experiment.

References

1 A. Vaskevich and I. Rubinstein, in *Nanoplasmonic Sensors*, ed. A. Dmitriev, Springer New York, New York, NY, 2012, pp. 333–368.

2 S. Badilescu, D. Raju, S. Bathini and M. Packirisamy, Gold Nano-Island Platforms for Localized Surface Plasmon Resonance Sensing: A Short Review, *Molecules*, 2020, **25**, 4661.

3 N. B. Colthup, L. H. Daly and S. E. Wiberley, *Introduction to infrared and Raman spectroscopy*, Academic Press, Boston, 3rd ed., 1990.

4 A. Więckowski, C. Korzeniewski and B. Braunschweig, Eds., *Vibrational spectroscopy at electrified interfaces*, Wiley, Hoboken, New Jersey, 2013.

5 M. L. S. Mello and B. C. Vidal, Changes in the Infrared Microspectroscopic Characteristics of DNA Caused by Cationic Elements, Different Base Richness and Single-Stranded Form, *PLoS ONE*, 2012, **7**, e43169.