Chemical space mapping for multicomponent gas mixtures

Airat Kotliar-Shapirov, Fedor S. Fedorov, Henni Ouerdane, Stanislav Evlashin, Albert G. Nasibulin, Keith J. Stevenson*

Skolkovo Institute of Science and Technology, 3 Nobel Street, 121205, Moscow, Russian Federation

KEYWORDS. Cyclic Voltammetry; Nitrogen Dioxide; Carbon Dioxide; Machine Learning; Gas Analytical System; Feature Importance

ABSTRACT. Analysis of gas mixtures using multivariable sensors or multisensor gas analytical systems requires adapted protocols of data processing. Notably, interference of analytes yields fingerprints that differ from a combination of the projected patterns of the single-components in the chemical space. Here, employing cyclic voltammetry only, we analyze the classification of single components, NO_2 and CO_2, and their mixtures based on the response of electrochemical sensors we specifically designed. The sensor sensitivity towards NO_2 is 15.7 nA/ppm, and the sensitivity towards CO_2 is 1.56 nA/ppm with the limits of detection of 1 ppm and 11 ppm, respectively. The overlap of the analytes' voltammetry profiles makes it difficult to extract their exact concentrations, so data-driven approaches were used, specifically principal component

analysis (PCA). After data classification, inverse PCA was used to determine the characteristics of the gas mixture response kinetics. Partial dependence analysis helps identify which measured potential contributes the most to the selective recognition of gases both in single-component gases and in mixtures. As an outcome, accurate gas mixture concentration profiles were obtained what allows deconvolution of the gas mixture in chemical space.

TEXT MANUSCRIPT

INTRODUCTION

Modern solutions for medical diagnostics, environmental monitoring, industrial safety, and other applications require new portable,[1] and highly sensitive and selective sensors with detection limits[2] of analyte concentrations down to a few ppm,[3,4] or even to the level of a few ppb.[5] Real, non-controlled environments may contain interfering species at a comparable concentration level.

A key-and-lock approach may be applied to selectively determine an analyte; it implies the development of receptor materials that can selectively bind the analyte of interest; this, however, might reduce the continuous use and reversibility of such sensors.[2,6] The other approach, first proposed by Persaud and Dodd,[7] requires the collection of several sensors into an array, which provides a vector signal as a "fingerprint" of an analyte when processed by classification tools mimicking the operation of the biological olfactory system.[8–10] In contrast, a good cross-sensitivity of a sensor is highly desired in this case. Such an approach is very well-adapted for many types of sensor systems enabling selective determination of analytes and preserving detection accuracy.[11–14] Moreover, current biological, chemical, electrochemical, i.e. other analytical studies are armed with machine learning tools to acquire or verify the information about the target analytes. These studies have stimulated a branching of a recent trend for the design of multivariable sensors where a single sensor on a single transducer provides multiple signals, e.g. by varying applied temperature or voltage, to be beneficial because it might give an advantage of the absence of uncorrelated drift and often simplified design.[2,15] In the latter approach, reconstruction of the chemical space[16] essentially relies on using machine learning protocols.

An option to distinguish between a few analytes, which falls to the multivariable sensor paradigm, presumes the application of spectroscopic methods, like impedance spectroscopy,[17] or, electrochemical methods like, voltammetry or chronoamperometry. Moreover, the advantages of electrochemical systems include high sensitivity and selectivity, a wide linear range, minimal space and power requirements, and low-cost instrumentation.[18,19] A primarily feature is a potential at which the reaction equilibrium is shifted towards reduction or oxidation of active species; given different electrochemical redox potentials of analytes, cyclic voltammetry (CV) enables their "spatial separation" manifested in current peaks reasonably close to these potentials. Still, the interfering compounds, i.e. having close or overlapping redox potentials, or components at low concentration might influence the selective distinction of analytes, e.g., due to a shift in potentials following the Nernst equation or overpotential losses associated with slow kinetics . In this case, use of machine learning tools can help to discriminate individual components from global overlapped, multiple peak voltammogram, and, thus, contribute to unveiling the kinetics of the occurring processes.[20–22]

Popular unsupervised learning or classification techniques like the principal component analysis (PCA) require a set of training tests to establish "chemical space", often done for particular single analytes "projected" on the PCA space. The use of such methods requires *a priory* knowledge of analytes to be further determined during the operation, which requires training of the sensor. The addition of extra analytes to the "chemical space" of the trained system will change the overall projection making reconstruction of the environment rather difficult. Moreover, the detection of composite mixtures implies an established relationship of the responses of the sensor to a sequence of the mixture with different ratios of analytes, i.e. corrected on the kinetics of interaction of all analytes in the mixture with the material surface. While the kinetics might be rather complicated, i.e. occurring processes might overlap and influence each other as shown for nitrogen oxides[23–25] and carbon dioxide reduction[26,27] one might lack the precision of identification of particular analyte, as in the case of an electrochemical sensor employing cyclic voltammetry for identification of inorganic and organic analytes in the mixtures.[28] Still, the use of the machine and deep learning in chemical identification can increase the precision as shown in some previous studies,[29,30] yet they were barely considered for mixtures of analytes. Thus, although chemometric methods have a long history of being applied to electrochemical data, the analysis of why the concrete model gives such prediction, especially for nonlinear methods (e.g. neural networks which boast better performances) are rarely reported. One of the approaches to evaluate the model's predictive power is to apply feature importance analysis, which for cyclic voltammetry correspond to the most "impactful" potentials on multiple peak voltammogram. More precisely, by assigning a score to a feature, one can select those features that have high scores, indicating the actual main contributions, which allows for proper interpretation and prediction.

Here we considered individual asphyxiating gas CO_2, toxic gas $NO_{2,}$ and their mixtures to test several approaches for recognition and to establish a "chemical space" in the frame of the developed multivariable sensor utilizing the cyclic voltammetry method. Both gases are present in indoor environments, and sometimes at concentration levels that may cause discomfort or even present a danger.[31] A concentration level of 2500 ppm of carbon dioxide indicates poor indoor air quality, and it starts becoming harmful from a level of 5000 ppm when people's cognitive functions are strongly affected.[32,33] Nitrogen dioxide, on the other hand, is well-known for causing severe adverse effects on the respiratory system from concentration levels of 5 ppm in indoor environments,[34,35] and even lower for young children and people who have a preexisting health condition such as asthma. It is therefore important to develop efficient solutions to identify these gases and measure accurately their concentrations when they are mixed in ambient air.

Here, we show that the profile of the mixture of analytes differs from those of individual gases and that nonlinear methods can provide much better gas concentration prediction accuracy for the gas mixtures compared to linear methods. We show CO_2 and NO_2 detection at low ppm concentrations and mapping their mixtures, including inverse PCA analysis, helps to unveil redox kinetics of a mixture of components at the sensor surface. Models that have exponential partial dependence plots provide better accuracies, and mapping the original features to that exponential space can substantially increase linear model performance. This approach increases the accuracy of existing electrochemical sensors and provides impetus in the design of new sensing paradigms with better analytical figures of merit.

MATERIALS AND METHODS

Setup. The schematic representation of the utilized setup is given in Figure 1a. The setup included several gas vessels *(1)* containing mixtures of CO_2 and NO_2, both 100 ppm, with N_2. Sources of N_2 and O_2 were mixed in a ratio to provide synthetic air (80/20). The vessels were connected to mass-flow controllers (Bronkhorst®) *(3)* via pressure reducer *(2)*. Programmable mass-flow controllers were set to flow gases in set proportions to achieve concentrations for pure gases of CO_2 and NO_2: 20 ppm, 40 ppm, and 80 ppm. Mixes were realized as NO_2 60 ppm + CO_2 20 ppm, NO_2 20 ppm + CO_2 60 ppm, NO_2 40 ppm + CO_2 20 ppm, NO_2 20 ppm + CO_2 40 ppm. The experiments were conducted under dynamic flow conditions at a flowrate of 50 sccm. All experiments had 10 sccm of pure oxygen added to the flow. Mixture flow and synthetic air flow (N_2 + O_2) were alternated every 8 minutes. During the gas flow, CV curves were constantly recorded. For a given parameter, one step of gas flow took exactly 12 CV cycles. The flow from mass-flow controllers was directed via pipelines to electrochemical cell *(5)* and then to exhaust

(6). Safety valves (4) were installed after the mass-flow controllers. The setup included potentiostat-galvanostat Elins P-40X (Elins, Russia) *(7)* connected to a PC *(8)* to conduct the measurements.

Electrochemical studies. The cross-section of the gas chamber can be seen in Figure 1b. All parts of the chamber were machine-cut from Teflon®. We used a 2-electrode cell configuration with one electrode to be Pt wire and the other was made of polypropylene porous film whose one side was sputtered with gold. The cell volume was ca. 25 cm^3 filled with electrolyte, 4 M sulfuric acid (RUSKHIM, LCC, Moscow, Russia). The cell was sealed using rubber O-rings. Cyclic voltammetry was performed in the potential range from -400 mV to 600 mV at a scan rate of 50 mVs^{-1}.

Electrode fabrication: 25µm thick, 12 mm diameter Celgard® polypropylene film was covered by 50 nm of gold by means of thermal evaporation. Further, it was platinized by cycling in 4 M H_2SO_4 in a voltage range from 400 mV to -600 mV at a scan rate of 50 mVs^{-1}. Another electrode was represented by 0.5 mm diameter 12 mm long platinum wire.

Scanning electron microscopy (SEM) Apreo Thermo Fisher Scientific was applied for analyzing the morphology of the platinized gold layer (Figure 1c). All images were recorded at a voltage of 5 keV and a current of 21 pA. The working distance was 4.4 mm. The through-lens detector (TLD) was used for collecting the signals.

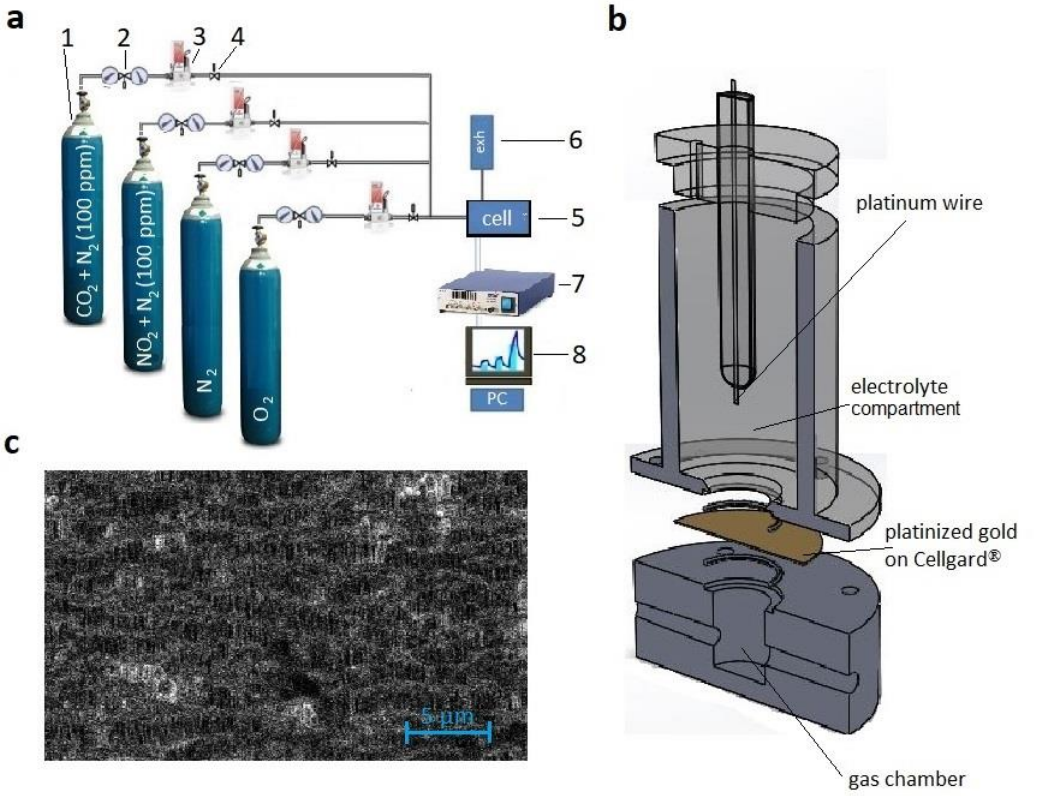


Figure 1. Gas mixing setup scheme and electrochemical cell scheme. (a) Gas mixing setup, *1* - gas cylinders with corresponding gases (NO$_2$ (100 ppm) + N$_2$, CO$_2$ (100 ppm) + N$_2$, O$_2$, and N$_2$); *2* reducer, *3* – mass flow meter, *4* – valve; *5* electrochemical gas cell, *6* - exhaust, *7* – potentiostat-galvanostat, *8* - PC, (b) cross-section of an electrochemical cell; (c) SEM image of Celgard® membrane with 50 nm gold, platinized by 200 cycles in 4 M H$_2$SO$_4$.

Model overview and applications in gas sensing. Linear regression (LR) is the simplest model that assumes a linear relation between features (current at certain voltages in our case) and predicted value (gas concentration). Usually, its performance is not very good, but its results are easily interpretable compared to a Deep neural network analysis. To explore the features selected by linear models, we applied L1 and L2-regularized linear regression (aka “ElasticNet”) for its simplicity and ability to reduce model complexity and prevent over-fitting.

Random forest (RF) is an approach built upon decision trees. At the core, decision tree models are nested if-else conditions. Their high tendency to overtraining is overcome by building a lot of decision trees on different random subsets of the data and averaging the prediction.

K-nearest Neighbours (KNN) is a type of non-parametric method used for classification and regression[36] where the function is only approximated locally and learning consists only of storing the training sample. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number of examples (K) closest to the query, then averages the labels. KNN has been used to predict air quality[37] and for gas identification.[38] In our approach we selected the number of neighbors n = 20, to assure including the data from other pulses too (one pulse length was 12 points) and used Euclidian distance as metric.

Support Vector Machine (SVM) classifier works by constructing a hyperplane or set of hyperplanes in a high- or infinite-dimensional space that separates the classes with maximum margin – thus allowing nonlinear boundaries. Its concepts can be generalized to become applicable to regression problems.[39,40] It has been successfully applied to the simultaneous voltammetric determination of morphine and noscapine,[41] explosives detection[42] and classification of green and black teas.[43] Its advantages over linear regression can be seen in the higher performance of SVR (see Table 2) and on the graph of mixture prediction

Deep neural network. A deep neural network (DNN) or deep learning models are known for their outstanding performance. Inspired by the biological neural networks, the DNN models with a single hidden layer containing a finite number of neurons are known for their ability to approximate almost any continuous functions[44] thus being very successful in approximating nonlinear relations. The number of papers devoted just to its application in chemical identification is overwhelming, among electrochemical sensors, there are examples of features based on square wave voltammetry,[29] cyclic voltammetry,[45] pulse voltammetry.[46,47]

Data analysis. For model training, 8 CV profile points out of each 12 CV gas pulse points were used considering sensor response time to be lower than 120 s (see Figure 6b).

Labeling: To avoid processing time-series points, the CV was labeled using an 8-cycle window inside each 12-cycle phase (2 cycles - 80 seconds for signal stabilization). Thus, every class is representing a stable signal for each concentration profile.

Table 1. Prediction models' hyperparameters

Model	Hyperparameters
ElasticNet	α = 0.005
Random Forest	Split criterion - mean squared error, max depth of the tree - 5, number of features to consider when looking for the best split – square root of 101, the minimum number of samples required to be at a leaf node - 6, the minimum number of samples required to split an internal node – 8; the number of trees in the forest = 10
K Nearest Neighbors	Number of neighbors -10; weight function used in prediction – Euclidian distance
Support Vector Regression	Radial basis function kernel; regularization parameter C = 1000
Deep Neural Network	Multilayer perceptron regressor with fully connected layers of size (101,200,100,50, 10,2); activation function – ReLu; solver - adam

RESULTS AND DISCUSSION

Sensor performance. In our studies, we have utilized a 2-electrode cell configuration with a Celgard® polypropylene membrane, 25 µm thick, with sputtered Au layer, whose one side is exposed to gas flow and the other side is exposed to the electrolyte, 4 M H_2SO_4 as detailed in Materials and Methods. The second electrode is Pt wire possessing a much smaller surface when compared to the other one, owing to the intentional asymmetry of the employed cell. Prior to the sensing tests, the cell is stabilized by cycling over 200 cycles from -600 mV till 400 mV at 50 mVs^{-1} that also favors platinization of the Au-sputtered membrane.[48] Use of 2-electrode cell configuration simplifies the design of the sensor and enables treating forward (or backward) CV scans only yet does not set any limitations to the use of a machine learning routine. In the 2-electrode configuration, we apply a potential that is to facilitate overvoltage difference associated with polarization effects ($E_{cat.} – E_{anod}$) and to overcome solution resistance:[49]

$$U = (E_{cat.} – E_{anod}) + iR, \qquad (1)$$

where i is the current and R is the solution resistance.

During the cycling, synthetic air (80% N_2 and 20% O_2) was used as a background atmosphere, with the pulsing of NO_2, and CO_2 at concentrations up to 80 ppm and their mixtures (see Materials and Methods). Representative cyclic voltammetry (CV) profiles for synthetic air and NO_2 (CO_2), 20, 40, and 80 ppm in the mixture with air are presented in Figure 2a,c. There are several current-potential peaks, being primarily located at -300 mV and -450 mV for NO_2 and CO_2, respectively. Note that CO_2–related peaks are less pronounced and more broadened over a wider potential range. The overall CV curves recorded in air, NO_2 (CO_2) analytes mixed with air (20, 40, and 80 ppm), and NO_2-CO_2 mixtures mixed with synthetic air are shown in Figure 2b. The presented data show the rather stable performance of the sensor over the course of the studies and analyte concentration-driven changes as profiled at the OY scale. Henceforth, only the forward scan of CV in the range [-100:-600] mV has been employed, similar to ref.[25], since it contains most of the information while the current profile at other potentials is not related to analyte redox reactions and could add noise to the system. Noting the different potentials associated with the detection of the analytes under study, current-potential profiles at -300, -400, and -500 mV were obtained for pure analytes and mixtures. Plots representing the gas profile response are depicted in Figure 2 with magnified slices of the current-potential profile at -400 mV are shown in Figure 2e,f. The first three (higher) peaks represent NO_2 injection (20, 40, and 80 ppm), the next three peaks (smaller) CO_2 injection (20, 40, and 80 ppm), next four peaks correspond to the mixtures of NO_2 and CO_2, accordingly (40 ppm + 20 ppm), (20 ppm + 40 ppm), (20 ppm + 60 ppm), (60 ppm + 20 ppm). One observes a notable current-potential response due to the appearance of analyte and, with a much more pronounced current, at -400 and -500 mV. Magnified pulses for the profile at -400 mV indicate feasibility detection of the lowest CO_2 concentration (Figure 2f).

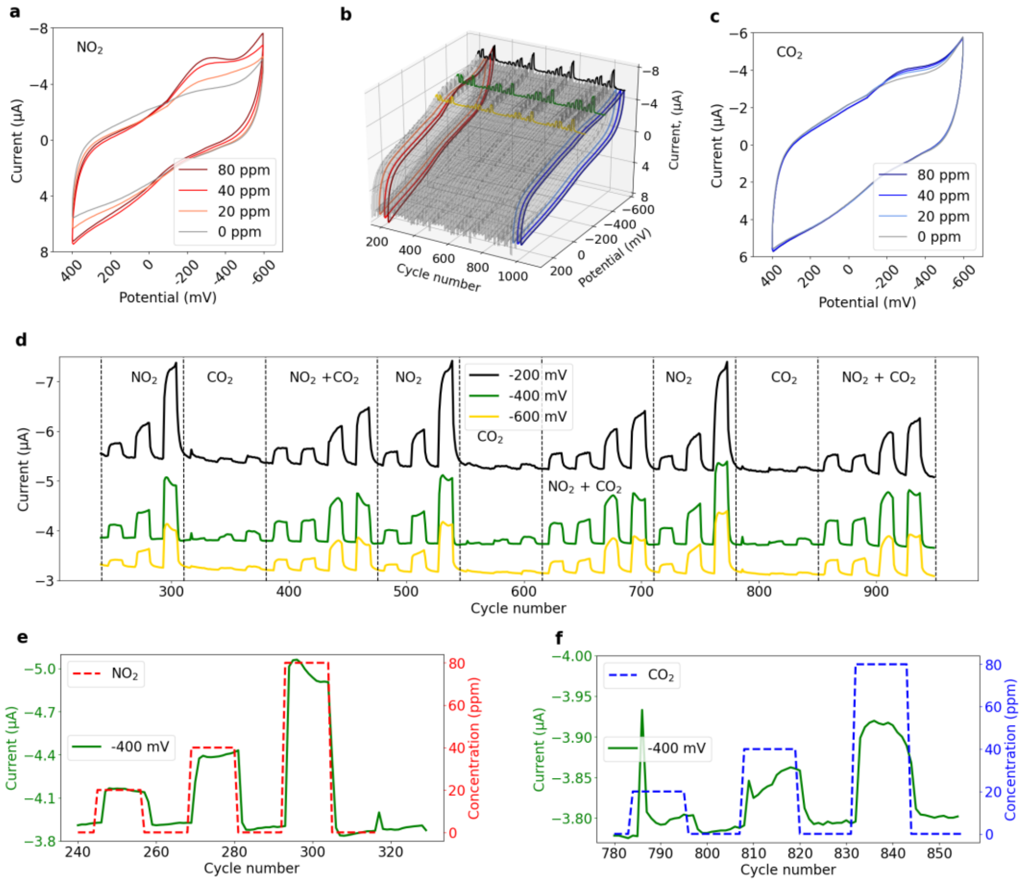


Figure 2. Cyclic voltammetry data for the sensor under exposure to NO_2, CO_2, and their mixture in the synthetic air. (a,c) CV profiles of different concentrations of applied pure NO_2 and CO_2, respectively (increasing concentrations are color-coded); (b) time evolution of CVs during gas pulses; (d) time evolution of sensor current during gas pulses at -200 mV (black line), -400 mV (green line) and -600 mV (yellow line) (e) sensor current (green) corresponding to pure NO_2 pulses (red dashed line); (f) sensor current (green) corresponding to pure CO_2 pulses (blue dashed line).

3D plots of NO_2 and CO_2 with subtracted synthetic air (Figure 3a,b) provide a good comparison of sensor sensitivity towards these analytes: while NO_2 surface can be approximated by plane (aka linear concentration dependence), the surface of CO_2 provides linear concentration dependence only at specific potentials (e.g., at - 400 mV). When considering a forward CV ramp for NO_2, CO_2, and their mixtures with subtracted synthetic air curve (Figure 3c-e) one notices at least two current peaks for CO_2 analyte where one peak (at -450 mV) shows a visible increase with the increase of concentration only. In the case of the NO_2 analyte, the detected current and modulations in current are much larger when compared with the CO_2 analyte. A background-subtracted CV ramp for the mixture, thus, should be much influenced by NO_2 presence and possesses a rather similar profile to a single NO_2 analyte that further requires machine learning tools for good differentiation. For the mixtures with higher CO_2 concentration, a plateau at ca. - 450 mV is observed.

Notably, the system showed significantly greater sensitivity to the NO_2 gas, 15.7 nA/ppm, than to the CO_2, 1.56nA/ppm calculated at -500 mV and further fitted by power dependence (Figure 3f,g), i.e.:

$$S_{NO2} = k_1[NO_2]^{n1}, n1 = 1.021 \pm 0.002, k1 = 0.015 \pm 0.001; \qquad (2)$$

$$S_{CO2} = k_2[CO_2]^{n2}, n2 = 0.883 \pm 0.005, k2 = 0.003 \pm 0.001. \qquad (3)$$

Such a nearly linear dependence on concentration, $R^2 = 0.985$ for NO_2 and $R^2 = 0.974$ for CO_2, corresponds to first-order kinetics which is inherent for many electrochemical processes, e.g. diffusion limited cases.[50] The limit of detection has been estimated as the concentration to be detected at the signal equal to three times the background noise of the CV at the exposure to synthetic air pulses divided by analyte sensitivity (slope of calibration curve). The limit of detection is ca. 1 ppm for NO_2 analyte and 11 ppm for CO_2 correspondingly.

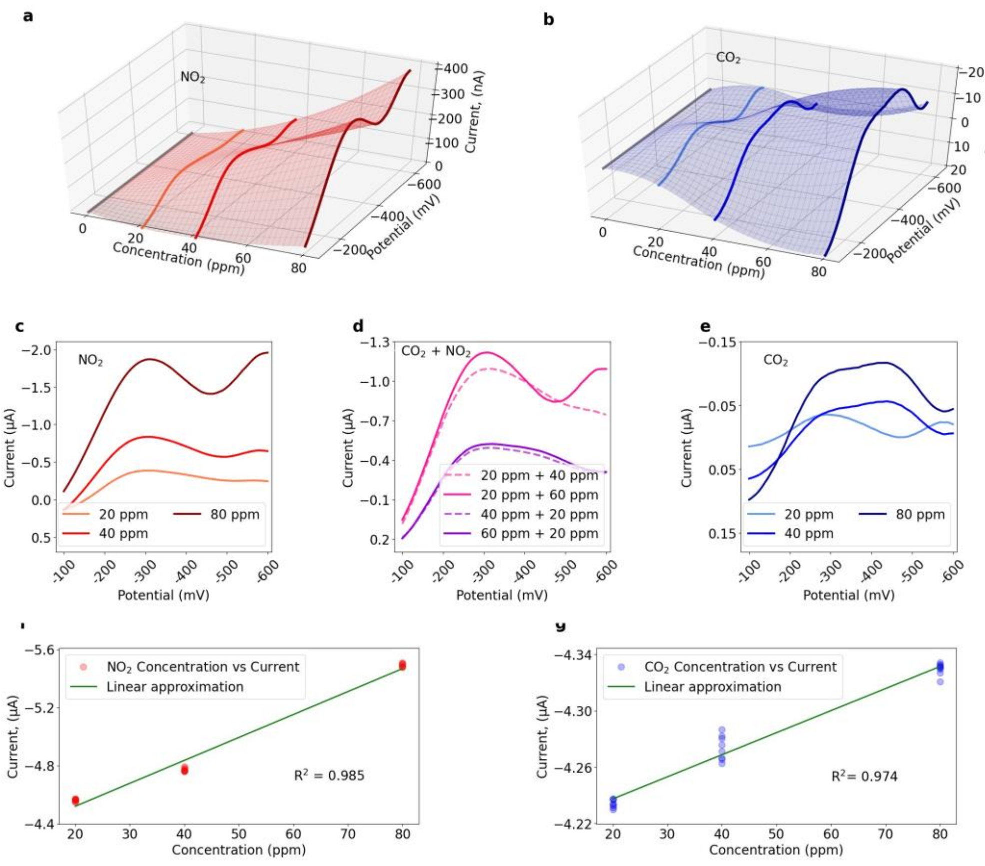


Figure 3. Current profile evolution. (a,b) 3D plots of current *versus* voltage *versus* concentration of NO_2/CO_2 respectfully; (c,d,e) current profiles for NO_2/CO_2 /mixtures respectfully with baseline (synthetic air) extracted; (f) current at voltage -500 mV *versus* concentration of NO_2, red dots represent experimental points, green line – linear approximation - slope 15.7 nA/ppm fits data with R^2 = 0.985; (g) current at voltage -500 mV *versus* concentration of CO_2, blue dots represent experimental points, green line – linear approximation - slope 1.56 nA/ppm fits data with R^2 = 0.974.

Gas recognition and classification. Recorded CV data in the range of potentials from - 100 mV to -600 mV for each pure gas/gas mixture pulses were analyzed by PCA, and from which, gas prediction ML models were constructed.

As a conventional approach to data visualization, PCA was applied and checked against the spanned scatter plot distributions for method limitations. PCA reveals the internal structure of the data in a way that best explains the variance in the data by a projection of the information carried by the original variables onto a smaller number of underlying ("latent") variables called principal components (PCs).[51] It should be noted that PCA only provides a visualization tool to check whether the samples group together in some regions, and cannot be fully considered as a pattern recognition method. Moreover, the best data of separation may not lie in the orthogonal basis of maximized variance.[45]

PCA, computed for NO_2/CO_2 analytes mixed with air, and their mixtures in air, with two components explains 98.9% and 0.7% of data variability (Figure 4). Change in PCA distribution represents the slope of increasing peak current during the same pulse (see Figure 2e,f). The relation between gas concentrations and PCA appears to be highly nonlinear, which was proven by mean gas concentration points and change in covariance distribution for each gas concentrations (blue line and grey shadow on Figure 4a,b) and self-intersection of gas mixture profile interpolations based on mean PCA values at NO_2 60 ppm (Figure 4c).

Visual analysis of the plot provides some expected trends: clusters corresponding to mixtures are located close to the compounds forming those mixtures. For example, mixtures of CO_2 (60, 40 ppm) and NO_2 (20 ppm) are clustered closer to pure NO_2 (20 ppm) and mixtures containing NO_2 (60 ppm) are clustered between pure NO_2 (40 ppm) and pure NO_2 (80 ppm). Thus, despite some overlapping regions for pure CO_2 concentrations, their fingerprints can be still distinguished due to the differentiated sensitivity exhibited by the electrode response. Therefore, by employing multivariate calibration methods resolution and individual quantification of gas compounds can be achieved from multicomponent mixtures. A primary hint is that position of clusters, projected at

2D PC components, expectedly changes when extra analytes are added. That means that the system has to be trained accounting for all the gases it is going to sense.

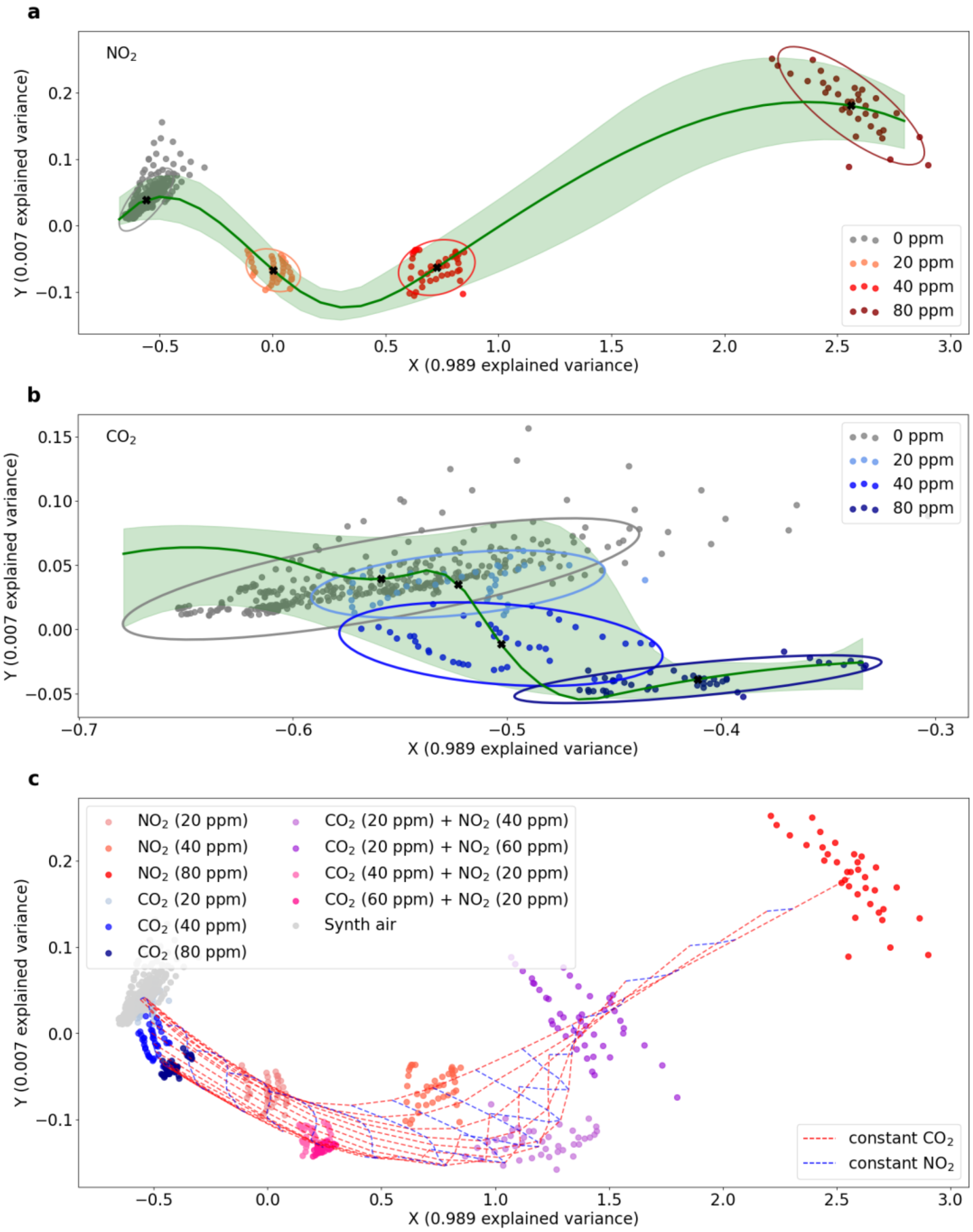


Figure 4. Principal component analysis of sensor data for pure gases and chemical space reconstruction for pure gases and mixtures. (a) PCA of NO_2 pulses. Scatter dots: grey – synthetic

air, orange – NO_2 (20 ppm), red – NO_2 (40 ppm), dark-red – NO_2 (80 ppm). Ellipses of the same colors represent covariance of distribution. green line – cubic interpolation of mean values of scatter dots, grey shadow – upper and lower bounds of mean graphs confidence; (b) PCA of CO_2 pulses. Scatter dots: grey – synthetic air, light-blue – CO_2 (20 ppm), blue – CO_2 (40 ppm), dark-blue – CO_2 (80 ppm). Ellipses of the same colors represent distribution covariance. Green line – cubic interpolation of mean values of scatter dots, grey shadow – upper and lower bounds of mean graphs confidence; (c) PCA of figures (a) and (b) combined together with mixture scatter added, color coding of gas scatter plots is given in the upper left part. Dashed lines represent equidistant concentration curves of NO_2 (red) and CO_2 (blue) calculated as an interpolation from available mean points of according scatter plots.

PCA provides a neat representation that two components from original data can be used to cluster the data with high accuracy. Although it tends to lose crucial information at higher concentrations limiting simultaneous gas detection at high concentrations. The shape of analyte distribution in the principal component space evidences a nonlinear relation between current and concentration. By projecting the interpolated points from the PCA (inverse PCA) curves back to the original space of reduction potentials we can provide artificial CVs for absent data points which allow plotting concentration curves with constant interfering gas (Figure 5a,b).

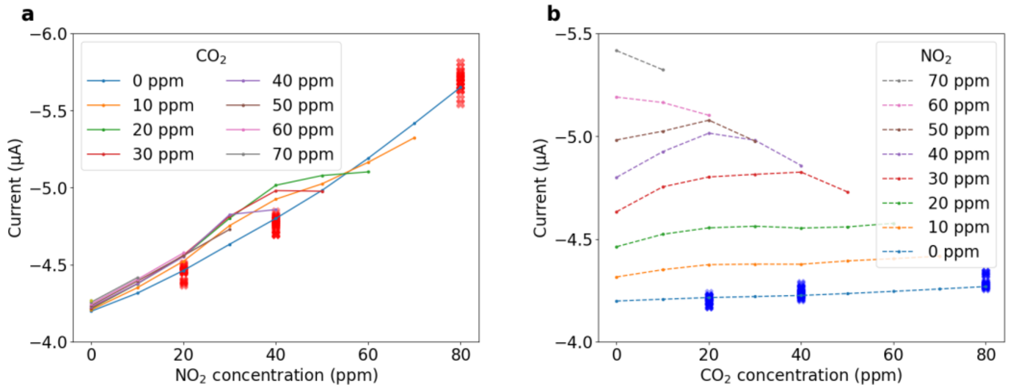


Figure 5. Reconstructed from PCA concentration curves of CO_2/NO_2 under constant interference of other gas NO_2/CO_2. (a) Current *versus* concentration dependence for NO_2 analyte with different constant concentrations of interfering CO_2 (lines) and experimental data points for pure NO_2 pulses (red crosses); (b) current *versus* concentration dependence for CO_2 analyte with different constant concentrations of interfering NO_2 (lines) and experimental data points for pure NO_2 pulses (blue crosses).

Under the assumption of no cross-reaction between NO_2 and CO_2 during electrochemical processes on the surface of the electrode, the currents were represented by $S = k_1C_1^{n1} + k_2C_2^{n2}$ where C_1 and C_2 are NO_2 and CO_2 concentrations, respectively. Optimizing the coefficients k_1, k_2 and powers n_1, n_2 to fit data obtained from inverse PCA we obtained $n1 = 1 \pm 0.003$, $n2= 0.300 \pm 0.095$, $k_1 = 0.016 \pm 0.001$, $k_2 = 0.015 \pm 0.004$ which indicate change in the kinetics of CO_2 related redox processes when it is mixed with NO_2 in air. For the single analytes, the dependence remains as represented by eqs. (2) and (3).

Gas concentration prediction. For gas concentration prediction the five most frequently reported methods were used: linear regression with L1 and L2 – norms regularization (aka ElasticNet), random forest, support vector machine regression, K-nearest neighbors, and deep learning (Fully connected neural network) model. Model performances measured as an absolute error in concentration prediction calculated via 5-fold cross-validation are presented in Table 2 and Figures 6-9. Optimal parameters for the first four were selected by grid search and can be found in Table 1 (see Materials and Methods) along with deep learning model layers configuration. Three different training samples: pure NO_2, pure CO_2, and mixtures were applied in each case. As a tool for visual feature importance comparison we plotted σ_{NO2}, $\sigma_{CO2,}$ and $\sigma_{mixtures}$ *versus* voltage on corresponding Figures 6-8 calculated as the standard deviation of current in each analyte class.

Table 2. Model prediction performance measured as average deviation error from actual NO_2, $CO_{2,}$ and mixes concentrations.

Model	Concentration prediction (mean ± std)		
	NO_2 (ppm)	CO_2 (ppm)	Mix (ppm)
ElasticNet	3.1 ± 0.5	20.3 ± 1.9	16.6 ± 1.0
RF	0.18 ± 0.15	6.5 ± 2.4	6.1 ± 2.1
KNN	0	5.4 ± 1.2	4.1 ± 1.0
DNN	0.7 ± 0.2	5.6 ± 1.3	4.9 ± 1.5
SVR	0.8 ± 0.3	5.6 ± 1.0	10.6 ± 1.2

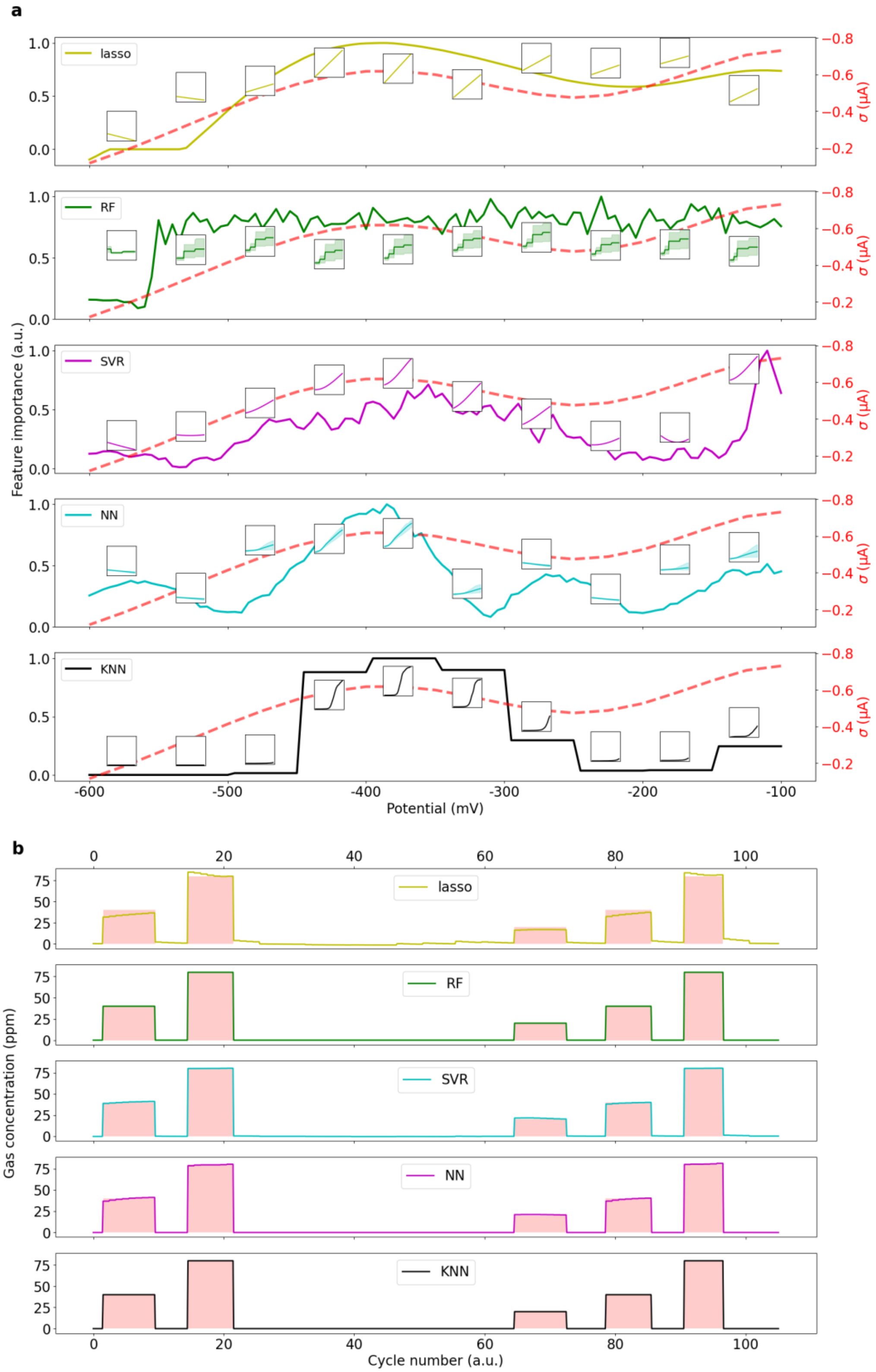

a
lasso
RF
SVR
NN
KNN
Feature importance (a.u.)
σ (μA)
1.0
0.5
0.0
−0.8
−0.6
−0.4
−0.2
-600
-500
-400
-300
-200
-100
Potential (mV)
b
lasso
RF
SVR
NN
KNN
Gas concentration (ppm)
75
50
25
0
0
20
40
60
80
100
Cycle number (a.u.)

Figure 6. Features selected by different models for NO_2 detection and their predictions on the test set. (a) NO_2 features (models are color-coded); red dashed transparent lines on the background are CVs for NO_2 pulses; (b) models' concentration prediction, color coding of lines is the same as on the left. Red transparent bars on the background represent true NO_2 concentrations. Feature importance is not defined for the KNN model; its prediction is represented by a black line.

As shown in Figure 6 and Table 1, almost all methods have perfect precision in predicting NO_2 concentrations. The only method that got an average error above 3 ppm is Lasso which still is comparable with the NO_2 limit of detection of 1 ppm. Almost all models' selected features are quite intuitive: high values at -310 mV are related to CV's peaks and represent higher sensitivity to the NO_2, lower values at the -440 mV represent the end of the peak. The overall shape matches the deviation of different gas current profiles (Figure 6a, red dashed line).

After yielding the models, their specific features were estimated via a partial dependence plot (PDP) to show the dependence between the target response and a set of 'target' features, marginalizing over the values of all other features (the 'complement' features).[52,53] Intuitively, interpretation of the partial dependence can be made as the expected target response as a function of the 'target' features. The number of features in our model is 100 (the number of potentials in the forward scan). To avoid overwhelming visualization, we plot the standard deviations of corresponding PDP which can be interpreted as a linear approximation of the model's sensitivity towards the selected potential current. All such PDP-STD plots are complemented with insets of actual PDPs calculated for 10 average regions of 50 mV width from -100 mV to -600 mV. (Figures 6a, 7a, 8a insets).

Random forest feature importance is almost uniform because the data can be easily separated throughout all provided voltage range forcing uniformity of possible bounds selected by trees.

Most interesting are the features selected by DNN: it selects the peak at - 310 mV and selects the potentials at -150 mV and -420 mV. SVR features can be seen as a transition between low-complexity and worse prediction capability linear model towards DNN.

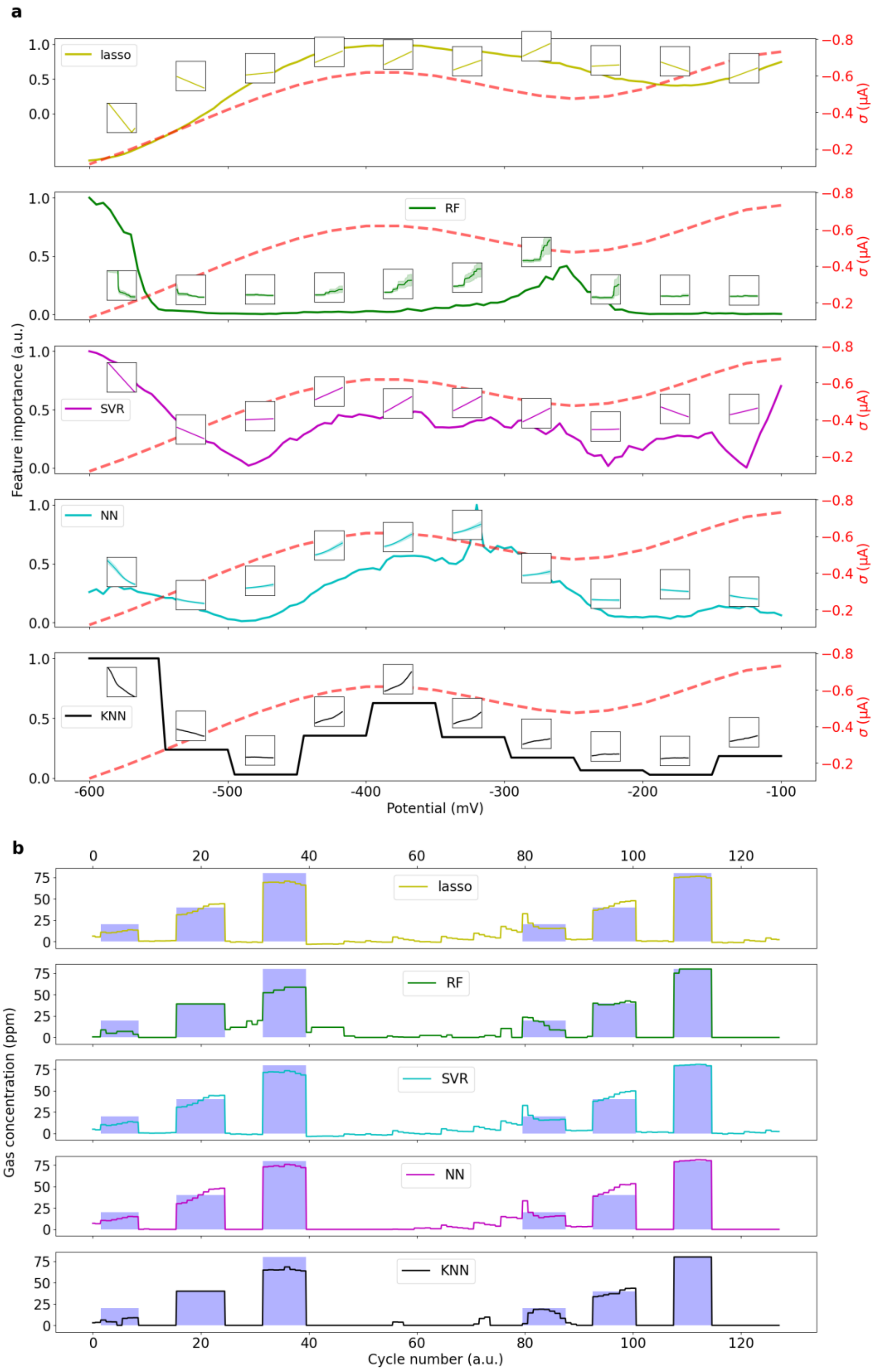

a
lasso
RF
SVR
NN
KNN
Feature importance (a.u.)
0.0
0.5
1.0
σ (μA)
−0.2
−0.4
−0.6
−0.8
-600
-500
-400
-300
-200
-100
Potential (mV)
b
lasso
RF
SVR
NN
KNN
Gas concentration (ppm)
0
25
50
75
0
20
40
60
80
100
120
Cycle number (a.u.)

Figure 7. Features selected by different models for CO_2 detection and their predictions on the test set. (a) CO_2 features (models are color-coded); blue dashed transparent lines on the background are CVs for CO_2 pulses; (b) models' concentration prediction, color coding of lines is the same as on the left. Blue transparent bars on the background represent true CO_2 concentrations. Features are not defined for the KNN model; its prediction is represented by a black line.

Since the sensitivity of the sensor towards CO_2 is comparably small, at a lower concentration, 10 ppm, due to a major overlap with synthetic air profile (Figure 4b), the prediction accuracy drops drastically (Table 2 and Figure 7b). Yet the explanation of selected features is mostly the same as in the NO_2 case - deviation of gas current profiles (Figure 3b, green rhombus dotted line). The difference in the sign for linear and all the rest models at lower voltages is due to the partial dependence method which allows only positive feature importance.

Random forest is the only model that focuses on potentials capturing the most dispersed current (-100 mV and -450 mV in Figure 3b). Notice the high success of KNN in predicting synthetic air as 0 ppm comparing to other methods (see Figure 7b cycles 40-80). The reason lays in method specificity – looking for neighbors - since the purge pulse was much longer than the gas pulse, which enlarged the training dataset by synthetic air examples. DNN and SVR are again learned to capture the most pronounced peak at -340 mV and its beginning -150 mV and ending -550 mV.

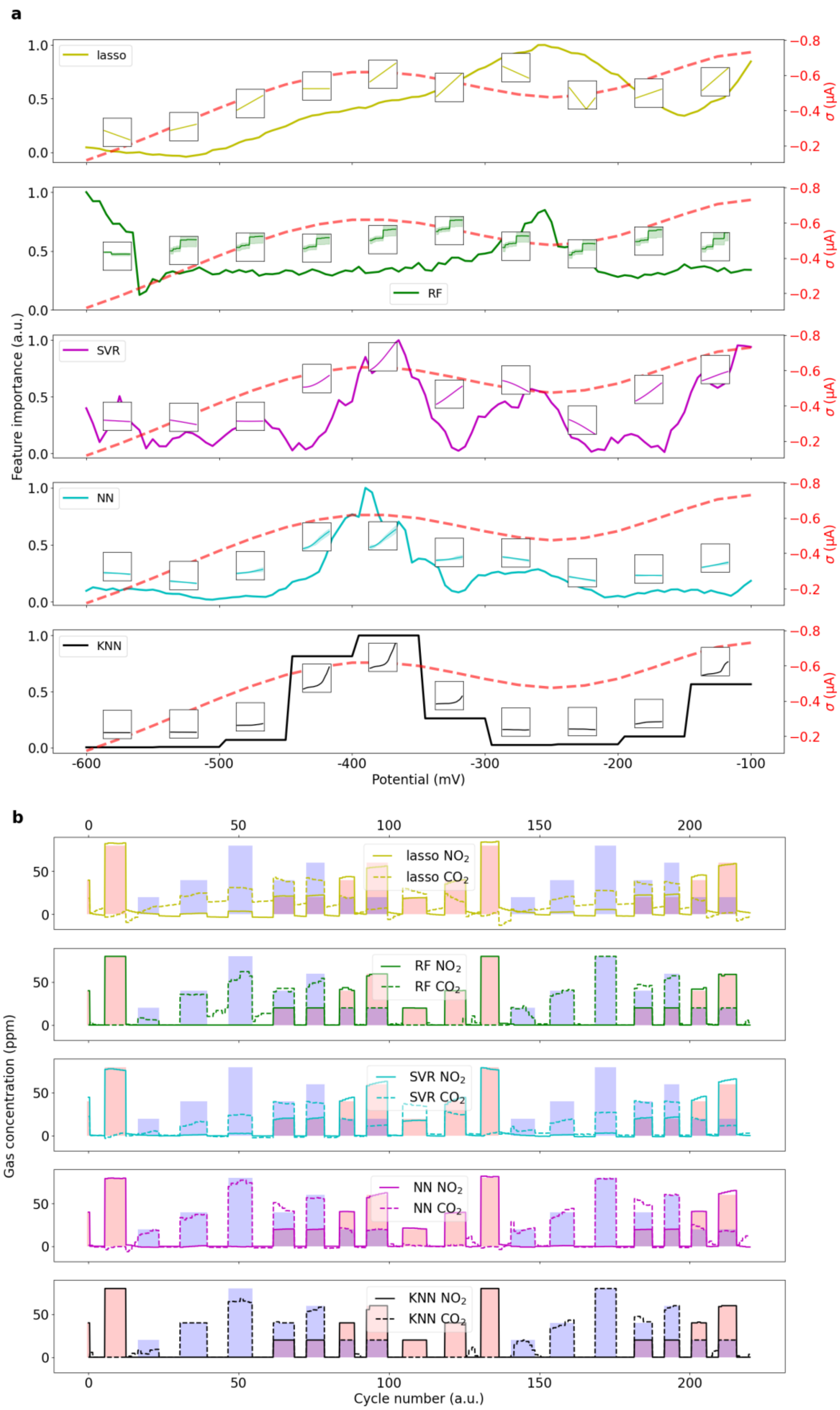

a
lasso
RF
SVR
NN
KNN
Feature importance (a.u.)
σ (μA)
1.0
0.5
0.0
−0.8
−0.6
−0.4
−0.2
-600
-500
-400
-300
-200
-100
Potential (mV)
b
lasso NO$_2$
lasso CO$_2$
RF NO$_2$
RF CO$_2$
SVR NO$_2$
SVR CO$_2$
NN NO$_2$
NN CO$_2$
KNN NO$_2$
KNN CO$_2$
Gas concentration (ppm)
50
0
0
50
100
150
200
Cycle number (a.u.)

Figure 8. Features selected by different models for gas mixtures detection and their predictions on the test set. (a) Mixtures features (models are color-coded); (b) models' concentration prediction, color coding of lines is the same as on the left. Red transparent bars on the background represent true NO_2 concentrations, blue transparent bars on the background represent true CO_2 concentrations. Thick lines correspond to the NO_2 prediction profile, dashed lines to CO_2 prediction profiles. Features are not defined for the KNN model; its prediction is represented by a black line.

As for mixture prediction, the features of Lasso still follow the current dispersion (Figure 8a dashed purple line) which no longer provides accurate information on gas concentrations since now mixtures of gases are involved and one should look for differences between two gases, not overall scatter – its low informativeness can be seen at the low success of linear method (Table 2 and Figure 8b thick yellow line) – it fails to discriminate between the analytes and has a major constant drift during interfering gas pulses. SVR is free from drift in prediction, yet it underestimates CO_2 concentrations during pure pulses and mixes (Figure 8b dashed blue line).

Features of SVR and DNN are in very close agreement so the difference in their precision lays in the methods themselves: DNN being much more suitable for nonlinear problems. The most pronounced peak corresponds to the potential at -300 mV with the most dispersed among gas profiles (Figure 3c). In addition to the fact that KNN's precision outperforms all other applied methods (see Table 2), a few important remarks must be made. Firstly, the number of presented individual analyte and mixed concentrations does not cover all "gas concentration space" between 20 to 80 ppm, rather it has a few clusters (Figure 3a,b) and the task tends to become closer to classification, rather than regression. This drawback can be overcome by using neighborhood components analysis or increasing the training concentrations. Secondly, KNN performance may be severely reduced by noisy or irrelevant features, thus its performance may drop significantly with an increased training set.

The analysis of features of specific models supports our assumption on the nonlinearity of sensor signal dependence. All models are selecting rather similar features, which can be explained as the potentials around redox reaction peaks for NO_2 with an expressed peak at -300 mV and with a small peak at -450 mV, and for CO_2 at -400 mV. The features for gas mixtures not only differ in the amplitude but are shifted in the potential range as well. This may indicate competition between CO_2 and NO_2 molecules for adsorption at specific active sites on the electrode surface.

After constructing training samples based on KNN partial dependence plot polynomials (Figure 8a, black line) and training linear regression on it we obtained much better performance (Figure 9) compared to the original prediction (Figure 8b first row). KNN proved to be the best-suited method here, however its advantage over deep neural networks in case of a larger number of analytes needs to be tested.

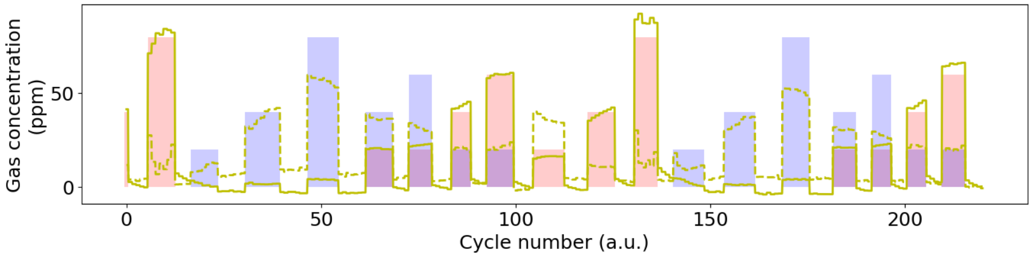


Figure 9. An elastic model trained on polynomial features (constructed from KNN pdp) predictions on the test set. Red transparent bars on the background represent true NO_2

concentrations, blue transparent bars on the background represent true CO_2 concentrations. Thick lines correspond to the NO_2 prediction profile, dashed lines to CO_2 prediction profiles.

We thus demonstrated how to make the linear methods prediction more accurate employing features estimated by means of PDP plots. Here, we considered two gases, CO_2 and NO_2, whose concentrations levels are typically different by orders of magnitude in an indoor environment; but considering that other chemical species, including volatile organic compounds are also present at different concentrations, our approach may provide a way to identify and deconvolute even more chemical species and their concentrations, for improved detection and monitoring protocols.

AUTHOR INFORMATION

Corresponding Author

Fedor S. Fedorov - Skolkovo Institute of Science and Technology, 3 Nobel Street, 121205, Moscow, Russian Federation; ORCID ID: 0000-0002-2283-0086 - link

E-mail: f.fedorov@skoltech.ru

Authors

Airat Kotliar-Shapirov - Skolkovo Institute of Science and Technology, 3 Nobel Street, 121205, Moscow, Russian Federation;

Henni Ouerdane - Skolkovo Institute of Science and Technology, 3 Nobel Street, 121205, Moscow, Russian Federation; ORCID ID: 0000-0002-1914-0244 – link

Stanislav Evlashin - Skolkovo Institute of Science and Technology, 3 Nobel Street, 121205, Moscow, Russian Federation; ORCID ID: 0000-0001-6565-3748 - link

Albert G. Nasibulin - Skolkovo Institute of Science and Technology, 3 Nobel Street, 121205, Moscow, Russian Federation; ORCID ID: 0000-0002-1684-3948 - link

Keith J. Stevenson - Skolkovo Institute of Science and Technology, 3 Nobel Street, 121205, Moscow, Russian Federation; ORCID ID: 0000-0002-1082-2871 - link.

Author Contributions

The manuscript was written through contributions of all authors.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

F.S.F. acknowledges the Russian Science Foundation for support by grant no. 19-72-00136.

REFERENCES

(1) Lewis, A.; Edwards, P. Validate Personal Air-Pollution Sensors. *Nature* **2016**, *535* (7610), 29–31.

(2) Potyrailo, R. A. Multivariable Sensors for Ubiquitous Monitoring of Gases in the Era of Internet of Things and Industrial Internet. *Chem. Rev.* **2016**, *116* (19), 11877–11923.

(3) van den Broek, J.; Abegg, S.; Pratsinis, S. E.; Güntner, A. T. Highly Selective Detection of Methanol over Ethanol by a Handheld Gas Sensor. *Nat. Commun.* **2019**, *10* (1), 4220.

(4) Fedorov, F.; Solomatin, M. A.; Uhlemann, M.; Oswald, S.; Kolosov, D. A.; Morozov, A.; Varezhnikov, A. S.; Ivanov, M. A.; Grebenko, A.; Sommer, M.; et al. Quasi-2D Co_3O_4 Nanoflakes as Efficient Gas Sensor versus Alcohol VOCs. *J. Mater. Chem. A* **2020**, *8*, 7214–7228.

(5) Lichtenstein, A.; Havivi, E.; Shacham, R.; Hahamy, E.; Leibovich, R.; Pevzner, A.; Krivitsky, V.; Davivi, G.; Presman, I.; Elnathan, R.; et al. Supersensitive Fingerprinting of Explosives by Chemically Modified Nanosensors Arrays. *Nat. Commun.* **2014**, *5*, 4195.

(6) Elghanian, R.; Storhoff, J. J.; Mucic, R. C.; Letsinger, R. L.; Mirkin, C. A. Selective Colorimetric Detection of Polynucleotides Based on the Distance-Dependent Optical Properties of Gold Nanoparticles. *Science* **1997**, *277* (5329), 1078–1081.

(7) Persaud, K.; Dodd, G. Analysis of Discrimination Mechanisms in the Mammalian Olfactory System Using a Model Nose. *Nature* **1982**, *299* (5881), 352–355.

(8) Buck, L.; Axel, R. A Novel Multigene Family May Encode Odorant Receptors: A Molecular Basis for Odor Recognition. *Cell* **1991**, *65* (1), 175–187.

(9) Wilson, C. D.; Serrano, G. O.; Koulakov, A. A.; Rinberg, D. A Primacy Code for Odor Identity. *Nat. Commun.* **2017**, *8* (1), 1477.

(10) Keller, A.; Gerkin, R. C.; Guan, Y.; Dhurandhar, A.; Turu, G.; Szalai, B.; Mainland, J. D.; Ihara, Y.; Yu, C. W.; Wolfinger, R.; et al. Predicting Human Olfactory Perception from Chemical Features of Odor Molecules. *Science.* **2017**, *355* (6327), 820–826.

(11) Albert, K. J.; Lewis, N. S.; Schauer, C. L.; Sotzing, G. A.; Stitzel, S. E.; Vaid, T. P.; Walt, D. R. Cross-Reactive Chemical Sensor Arrays. *Chem. Rev.* **2000**, *100* (7), 2595–2626.

(12) Lundstrom, I. Artificial Noses: Picture the Smell. *Nature* **2000**, *406* (6797), 682–683.

(13) Rakow, N. A.; Suslick, K. S. A Colorimetric Sensor Array for Odour Visualization. *Nature* **2000**, *406* (6797), 710–713.

(14) Fedorov, F.; Podgainov, D.; Varezhnikov, A.; Lashkov, A.; Gorshenkov, M.; Burmistrov, I.; Sommer, M.; Sysoev, V. The Potentiodynamic Bottom-up Growth of the Tin Oxide Nanostructured Layer for Gas-Analytical Multisensor Array Chips. *Sensors* **2017**, *17* (8), 1908.

(15) Hierlemann, A.; Gutierrez-Osuna, R. Higher-Order Chemical Sensing. *Chem. Rev.* **2008**, *108* (2), 563–613.

(16) Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L. How to Explore Chemical Space Using Algorithms and Automation. *Nat. Rev. Chem.* **2019**, *3* (2), 119–128.

(17) Fedorov, F. S.; Varezhnikov, A. S.; Kiselev, I.; Kolesnichenko, V. V.; Burmistrov, I. N.; Sommer, M.; Fuchs, D.; Kübel, C.; Gorokhovsky, A. V.; Sysoev, V. V. Potassium Polytitanate Gas-Sensor Study by Impedance Spectroscopy. *Anal. Chim. Acta* **2015**, *897*, 81–86.

(18) Wang, J. Portable Electrochemical Systems. *TrAC Trends Anal. Chem.* **2002**, *21* (4), 226–232.

(19) Wang, J. Real-Time Electrochemical Monitoring: Toward Green Analytical Chemistry. *Acc. Chem. Res.* **2002**, *35* (9), 811–816.

(20) Pravdová, V.; Pravda, M.; Guilbault, G. G. Role of Chemometrics for Electrochemical Sensors. *Anal. Lett.* **2002**, *35* (15), 2389–2419.

(21) Esteban, M.; Ariño, C.; Díaz-Cruz, J. M. Chemometrics for the Analysis of Voltammetric Data. *TrAC Trends Anal. Chem.* **2006**, *25* (1), 86–92.

(22) Ni, Y.; Wang, Y.; Kokot, S. Simultaneous Kinetic Spectrophotometric Analysis of Five Synthetic Food Colorants with the Aid of Chemometrics. *Talanta* **2009**, *78* (2), 432–441.

(23) Bergman, I. The Voltammetry of Some Oxidising and Reducing Toxic Gases Direct from the Gas Phase, at Gold and Platinum Metallised-Membrane Electrodes in Acid and Alkali. *J. Electroanal. Chem. Interfacial Electrochem.* **1983**, *157* (1), 59–73.

(24) Ye, S.; Kita, H. Adsorbed HNO2/NO Redox Couple at Pt(111) Single-Crystal Electrode. *J. Electroanal. Chem.* **1993**, *346* (1), 489–495.

(25) Roh, S.-W.; Stetter, J. R. Amperometric Sensing of NO_x with Cyclic Voltammetry. *J. Electrochem. Soc.* **2003**, *150* (11), H266.

(26) Evans, J.; Pletcher, D.; Warburton, P. R. G.; Gibbs, T. K. Amperometric Sensor for Carbon Dioxide: Design, Characteristics, and Performance. *Anal. Chem.* **1989**, *61* (6), 577–580.

(27) Pletcher, D. The Cathodic Reduction of Carbon Dioxide—What Can It Realistically Achieve? A Mini Review. *Electrochem. commun.* **2015**, *61*, 97–101.

(28) Knake, R.; Guchardi, R.; Hauser, P. C. Quantitative Analysis of Gas Mixtures by Voltammetric Sensing. *Anal. Chim. Acta* **2003**, *475* (1), 17–25.

(29) Dean, S. N.; Shriver-Lake, L. C.; Stenger, D. A.; Erickson, J. S.; Golden, J. P.; Trammell, S. A. Machine Learning Techniques for Chemical Identification Using Cyclic Square Wave Voltammetry. *Sensors* **2019**, *19*, 2392.

(30) Reichenbach, S. E.; Zini, C. A.; Nicolli, K. P.; Welke, J. E.; Cordero, C.; Tao, Q. Benchmarking Machine Learning Methods for Comprehensive Chemical Fingerprinting and Pattern Recognition. *J. Chromatogr. A* **2019**, *1595*, 158–167.

(31) Carbon Dioxide https://www.osha.gov/chemicaldata/.

(32) Satish, U.; Mendell, M. J.; Shekhar, K.; Hotchi, T.; Sullivan, D.; Streufert, S.; Fisk, W. J. Is CO_2 an Indoor Pollutant? Direct Effects of Low-to-Moderate CO_2 Concentrations on Human Decision-Making Performance. *Environ. Health Perspect.* **2012**, *120* (12), 1671–1677.

(33) Allen, J. G.; MacNaughton, P.; Satish, U.; Santanam, S.; Vallarino, J.; Spengler, J. D. Associations of Cognitive Function Scores with Carbon Dioxide, Ventilation, and Volatile Organic Compound Exposures in Office Workers: A Controlled Exposure Study of Green and Conventional Office Environments. *Environ. Health Perspect.* **2016**, *124* (6), 805–812.

(34) Int Panis, L.; Provost, E. B.; Cox, B.; Louwies, T.; Laeremans, M.; Standaert, A.; Dons, E.; Holmstock, L.; Nawrot, T.; De Boever, P. Short-Term Air Pollution Exposure Decreases Lung Function: A Repeated Measures Study in Healthy Adults. *Environ. Heal.* **2017**, *16* (1), 60.

(35) Garrett, M. H.; Hooper, M. A.; Hooper, B. M.; Abramson, M. J. Respiratory Symptoms in Children and Indoor Exposure to Nitrogen Dioxide and Gas Stoves. *Am. J. Respir. Crit. Care Med.* **1998**, *158* (3), 891–895.

(36) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46* (3), 175–185.

(37) Dragomir, E. G. Air Quality Index Prediction Using K-Nearest Neighbor Technique. *Ser. Mat. - Informatică - Fiz.* **2010**.

(38) Brahim-Belhaouari, S.; Hassan, M.; Walter, N.; Bermak, A. Advanced Statistical Metrics for Gas Identification System with Quantification Feedback. *IEEE Sens. J.* **2015**, *15* (3), 1705–1715.

(39) Awad, M.; Khanna, R. Support Vector Regression BT - Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers; Awad, M., Khanna, R., Eds.; Apress: Berkeley, CA, 2015; pp 67–80.

(40) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*; COLT '92; Association for Computing Machinery: New York, NY, USA, 1992; pp 144–152.

(41) Niazi, A.; Ghasemi, J.; Zendehdel, M. Simultaneous Voltammetric Determination of Morphine and Noscapine by Adsorptive Differential Pulse Stripping Method and Least-Squares Support Vector Machines. *Talanta* **2007**, *74* (2), 247–254.

(42) Asadpour-Zeynali, K.; Soheili-Azad, P. Simultaneous Polarographic Determination of 2-Nitrophenol and 4-Nitrophenol by Differential Pulse Polarography Method and Support Vector Regression. *Environ. Monit. Assess.* **2012**, *184* (2), 1089–1096.

(43) Liu, N.; Liang, Y.; Bin, J.; Zhang, Z.; Huang, J.; Shu, R.; Yang, K. Classification of Green and Black Teas by PCA and SVM Analysis of Cyclic Voltammetric Signals from Metallic Oxide-Modified Electrode. *Food Anal. Methods* **2014**, *7* (2), 472–480.

(44) Cybenko, G. Approximation by Superpositions of a Sigmoidal Function. *Math. Control. Signals Syst.* **1989**, *2* (4), 303–314.

(45) Cetó, X.; O' Mahony, A. M.; Wang, J.; Del Valle, M. Simultaneous Identification and Quantification of Nitro-Containing Explosives by Advanced Chemometric Data Treatment of Cyclic Voltammetry at Screen-Printed Electrodes. *Talanta* **2013**, *107*, 270–276.

(46) Schaumlöffel, L. de S.; Bolognese Fernandes, P. R.; Sartori Piatnicki, C. M.; Gutterres, M. A Chemometric-Assisted Voltammetric Method for Simultaneous Determination of Four Antioxidants in Biodiesel Samples. *Energy & Fuels* **2020**, *34* (1), 412–418.

(47) Schaumlöffel, L. de S.; Dambros, J. W. V.; Bolognese Fernandes, P. R.; Gutterres, M.; Piatnicki, C. M. S. Direct and Simultaneous Determination of Four Phenolic Antioxidants in Biodiesel Using Differential Pulse Voltammetry Assisted by Artificial Neural Networks and Variable Selection by Decision Trees. *Fuel* **2019**, *236*, 803–810.

(48) Tavakkoli, M.; Holmberg, N.; Kronberg, R.; Jiang, H.; Sainio, J.; Kauppinen, E. I.; Kallio, T.; Laasonen, K. Electrochemical Activation of Single-Walled Carbon Nanotubes with Pseudo-Atomic-Scale Platinum for the Hydrogen Evolution Reaction. *ACS Catal.* **2017**, *7* (5), 3121–3130.

(49) Roffia, S.; Conciallini, V.; Paradisi, C. The Interconversion of Electrical and Chemical Energy: The Electrolysis of Water and the Hydrogen Oxygen Fuel Cell. *J. Chem. Educ.* **1988**, *65* (3), 272–273.

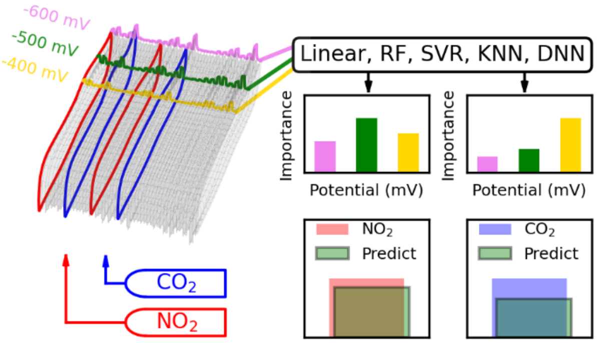
(50) Cottrell, F. G. Der Reststrom Bei Galvanischer Polarisation, Betrachtet Als Ein Diffusionsproblem. *Zeitschrift für Physikalische Chemie*. **1903**, *42U* (1), 385-431.

(51) Jolliffe, I. T. Principal Components in Regression Analysis; 1986.

(52) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, Second.; Springer-Verlag, 2009.

(53) Molnar, C. *Interpretable Machine Learning*; 2019.

TABLE OF CONTENT FIGURE


-600 mV
-500 mV
-400 mV
Linear, RF, SVR, KNN, DNN
Importance
Potential (mV)
Importance
Potential (mV)
NO_2
Predict
CO_2
Predict
CO_2
NO_2