ACS NANOSCIENCE Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

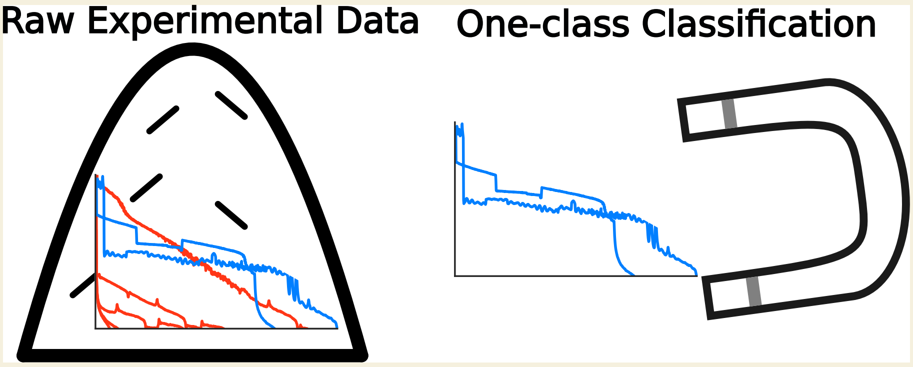
Open Access

Making the Most of Nothing: One-Class Classification for Single-Molecule Transport Studies

William Bro-Jørgensen, Joseph M. Hamill, Gréta Mezei, Brent Lawson, Umar Rashid, András Halbritter,* Maria Kamenetska,* Veerabhadrarao Kaliginedi,* and Gemma C. Solomon*

Cite This: *ACS Nanosci. Au* 2024, 4, 250−262 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



ABSTRACT: Single-molecule experiments offer a unique means to probe molecular properties of individual molecules—yet they rest upon the successful control of background noise and irrelevant signals. In single-molecule transport studies, large amounts of data that probe a wide range of physical and chemical behaviors are often generated. However, due to the stochasticity of these experiments, a substantial fraction of the data may consist of blank traces where no molecular signal is evident. One-class (OC) classification is a machine learning technique to identify a specific class in a data set that potentially consists of a wide variety of classes. Here, we examine the utility of two different types of OC classification models on four diverse data sets from three different laboratories. Two of these data sets were measured at cryogenic temperatures and two at room temperature. By training the models solely on traces from a blank experiment, we demonstrate the efficacy of OC classification as a powerful and reliable method for filtering out blank traces from a molecular experiment in all four data sets. On a labeled 4,4′-bipyridine data set measured at 4.2 K, we achieve an accuracy of 96.9 ± 0.3 and an area under the receiver operating characteristic curve of 99.5 ± 0.3 as validated over a fivefold cross-validation. Given the wide range of physical and chemical properties that can be probed in single-molecule experiments, the successful application of OC classification to filter out blank traces is a major step forward in our ability to understand and manipulate molecular properties.

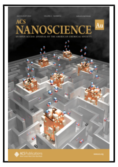
KEYWORDS: *machine learning, single-molecule junctions, one-class modeling, molecular electronics, Gaussian mixture model, support vector machine*

■ INTRODUCTION

Chemists have long been able to determine the chemical properties of a single molecule by making bulk measurements of the compound of interest, a remarkable feat that we often take for granted. In common techniques such as ultraviolet−visible (UV−vis) spectroscopy, infrared spectroscopy, or nuclear magnetic resonance spectroscopy, determining the chemical properties is largely possible because blank measurements can be used to subtract the background contributions from the matrix or solvent.[1] Subsequently, we assume that the remaining signal originates from the molecule of interest and interpret the results accordingly. While there can be a degree of variability between the contributions of the individual molecules in the sample, each experiment measures the average signal from numerous molecules. This averaged signal is the result that is reported for a compound. This workflow is depicted in the top row of Figure 1.

Single-molecule measurements provide a new perspective on molecular properties. By definition, the idea of blank measurements seems unnecessary as we directly measure the properties of the single molecule of interest. The molecule-to-

Received: May 1, 2024
Revised: May 25, 2024
Accepted: May 29, 2024
Published: June 6, 2024



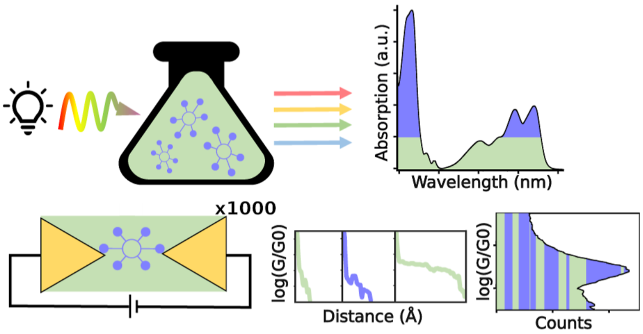


Figure 1. Comparison of UV−vis and single-molecule transport experiments. The top row depicts the standard UV−vis experiment (left), in which light passes through the solution of analyte (purple) and solvent (green), producing a spectrum that shows the wavelength at which the solution absorbs light. The bottom row displays a typical break-junction experiment (left), which produces thousands of samples with varying amounts of tunneling/contaminant signal (green) and analyte signal (purple). To understand the average trend of the molecule, normally all samples are compiled into a histogram (right).

molecule variability, unavoidably averaged in bulk measurements, becomes evident in successive single-molecule measurements. To determine the average behavior of a molecule, we average some numbers of single-molecule measurements. For instance, it is common to compile thousands of measured samples into 1D- and 2D histograms that provide information about the average properties of the measured molecule, such as conductance or indicate subpopulations. This workflow is depicted in the bottom row of Figure 1.

In the course of making repeated single-molecule measurements, sometimes the matrix or solution is measured instead of the molecule of interest. Consequently, the signal from these blank or background samples can impact the averaged observables, potentially obscuring rare events or altering the accuracy of results. Although each individual sample that makes up the 1D- and 2D-histograms is available, there is no straightforward method to remove the signal originating blank samples, as is possible in absorption spectroscopy. This inability to remove the background is partly because we do not know the exact number of blank traces in the molecular data nor their exact form.

To minimize the impact of background signal on the accuracy of results, a number of different methods to filter blank traces have been developed, including clustering techniques,[2−5] principal component analysis method,[6−9] and custom filtering algorithms.[10−36] The widespread use of filtering blank by several different group traces underlines the need for a structured and robust approach to this problem.

Alternatively, methods have been developed that circumvent the need to remove blank traces entirely. For instance, Bamberger et al.[37] have developed a method to extract traces with molecular features that were initially invisible in the combined raw data. They also developed a method to extract the main plateau of each trace,[38] which they later used to explain large variations in the reported conductance of different stilbene derivatives.

Reliable distinction between blank traces and traces with a molecular plateau enables new types of studies. For example, Fu et al.[39] correlate the existence of a molecular junction with experimentally measured features of each junction. Thus, they can investigate the conditions influencing junction formation.

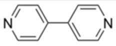
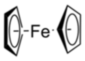
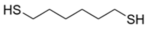
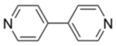
Single-molecule transport experiments are gaining importance as a tool to monitor chemical reactions at the single-molecule level and in the chemical and biological sciences in general.[40−42] For example, in cases where the analyte has a low junction formation probability, the overabundance of blank traces can obscure the molecular signal from the analyte.

In addition, removing unwanted signal from blanks and other traces could also find utility in single-molecule transport experiments used for driving and catalyzing electric field-induced chemical reactions.[43−45] These experiments have opened up new avenues for understanding and controlling chemical processes, enabling unprecedented precision in chemical transformations. Consequently, efficient removal of unwanted signal from, for example, starting material could facilitate the automation potential of such experiments.

In this paper, we show that the inclusion of blank experiments in a calibration step in single-molecule transport studies can improve the description of those experiments. We present the use of OC classification methods as a method to filter blank traces by training only on data from a blank experiment. This approach will allow us to perform the same background-subtracting step as in spectroscopy. Instead of using the blank measurement as a simple scalar to subtract, the OC model identifies the characteristic patterns of blank traces, allowing us to filter them out. This allows us to adaptively handle the inherent variability in single-molecule transport experiments. Thus, our methodology reinterprets the use of blank measurements, offering a dynamic and more precise tool for the removal of blank traces.

OC classification methods differ fundamentally from other commonly used methods to analyze single-molecule data sets

Table 1. Summary of Data Sets

Data set	Molecule	# molecular samples	# blank samples	Setup	Solvent	Featurization	Temperature[b]	Labels
4K-bpy		3129[a]	1863[a]	MCBJ	No	128 bins 1D-histogram	4.2 K	Yes
Fc		7916	6167	STM-BJ	No	100 bins 1D-histograms	5 K	No
C6-thiol		12010	23465	MCBJ	Yes	64 × 64 2D-histograms	RT	No
RT-bpy		9787	20129	MCBJ	No	32 × 32 2D-histograms	RT	No

[a]This data set contains samples from a molecular experiment, each subsequently labeled as either 'molecular' or 'blank'. [b]RT = Room temperature.

https://doi.org/10.1021/acsnanoscienceau.4c00015

such as supervised machine learning (ML) or clustering. Supervised ML, in particular, which is used to classify between two or more classes, relies on a training set containing at least one example of each class. However, the generation of such a training set in single-molecule transport studies can be challenging due to the nontrivial task of determining the class to which a trace belongs. On the other hand, no training set is needed when using clustering. Instead, practitioners choose a clustering algorithm, a similarity metric, and oftentimes, how many classes are expected. Subsequently, the algorithm groups similar samples together. In this way, no specific characteristics are assumed about any of the classes.

In contrast with these two methods, OC classification exploits the only class we can consistently generate large amounts of samples of blanks. Running a blank experiment in single-molecule transport studies is commonly performed before any measurements on the molecule of interest to ensure that the experimental setup is functioning correctly.

■ METHODS

Our exploration of OC classification involves analyzing four distinct data sets from three separate laboratories, with two collected at cryogenic temperatures and the remaining two at room temperature. Three of the data sets was collected using mechanically controlled break junction and the fourth with scanning tunneling microscopy break junction. An overview of these varied experimental conditions is presented in Table 1. In spite of these varied conditions, the OC classification method exhibits robust performance as we show in the following.

First, we analyze a data set of 4,4′-bipyridine that was measured at 4.2 K.[7] It is the only data set in this paper that is fully labeled (by inspection), allowing us to quantify the performance of the tested models. We refer to this data set as **4K-bpy**.

Next, we analyze a data set of ferrocene measured at 5 K.[32] With no labels, we rely on a qualitative evaluation of the performance of our models. We compare 1D- and 2D-histograms of the traces classified to be blanks and not-blanks with traces from the blank experiment. We also analyze the traces classified as not-blanks to investigate the nature of the outliers. We refer to this data set as **Fc**.

The final data set we analyze is of 4,4′-bipyridine measured at room temperature. We refer to this data set as **RT-bpy**. In addition to the qualitative examination, we perform a comparison of 2D-correlation histograms of the raw and the filtered **RT-bpy** data set. Notably, we observe a shift from a slightly positive correlation to a negative correlation in certain parts of the 2D-correlation histograms highlighting the impact of blank traces.

In the Supporting Information, we have included an analysis of a fourth data set which is 1,6-hexanedithiol measured at room temperature. Despite the increased amount of noise in this data set compared to the cryogenic data sets, we are still able to filter out a considerable number of blank traces from a molecular experiment. We refer to this data set as **C6-thiol**.

The first type of model used to filter blank traces is the one-class support vector machine (OC-SVM). It was originally proposed by Schölkopf et al.[46] as a tool for novelty detection.

Consider some training data $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathcal{X}$ where n is the number of training samples. To obtain a more complex, nonlinear decision boundary, the OC-SVM uses a transformation function referred to as the feature map, Φ: $\mathcal{X} \rightarrow F$, that maps each of the training samples into a new higher-dimensional space. The OC-SVM will only need the inner product between the samples in this higher-dimensional space, but this inner product might be computationally expensive to compute directly. Fortunately, we can make use of the so-called 'kernel trick':[47]

We define a kernel function that satisfies the following equation

$$k(\mathbf{x}, \mathbf{x}') = (\Phi(\mathbf{x})\cdot\Phi(\mathbf{x}')) \tag{1}$$

With an appropriately defined kernel function, we circumvent the need to calculate the inner product in the higher-dimensional space and instead get it directly from k. One popular choice of k is the radial basis function (RBF) kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \,\|\mathbf{x} - \mathbf{x}'\|^2) \tag{2}$$

Here, $\mathbf{x}$ and $\mathbf{x}'$ are input vectors, and γ is one of the hyperparameters that can be tuned for the model to obtain a better fit. Intuitively, γ defines how far the influence of a single-training example reaches. The larger the γ is, the closer the other examples must be to be affected. The kernel function is used to lift the data into a higher-dimensional vector space where it is potentially easier to separate our classes.

In the following, we will also make use of the linear kernel which is defined as

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\cdot\mathbf{x}' \tag{3}$$

The goal of the OC-SVM is to draw a boundary around our data. We want to do this in a way that contains the maximum number of points, while keeping the boundary as small as possible. To achieve that, the following constrained optimization problem has to be solved

$$\min_{w,\varepsilon,\rho} \|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\varepsilon_i - \rho \tag{4}$$

subject to

$$(w\cdot\Phi(\mathbf{x}_i)) \geq \rho - \varepsilon_i,\ \varepsilon_i \geq 0 \tag{5}$$

Here, w is the components of the hyperplane separating the positive from the negative samples, ε_i are slack variables that control how tightly our decision boundary encloses our data, n is the number of samples, and ρ is an offset. The parameter ν is similar to the regularization parameter C in classical SVMs and helps to control the trade-off between the volume of the sphere and the fraction of outliers.

It can be shown that solving eq 4 leads to a decision function of the form

$$f(\mathbf{x}) = \mathrm{sgn}(\sum_{i}^{n}\alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho) \tag{6}$$

where α_i are the parameters solved for.

In OC-SVMs, ν set the upper bound on the fraction of outliers and a lower bound on the fraction of training examples used for modeling in the OC-SVM. We are certain that there are no molecular traces in our training data, so we set ν = 0.01.

The hyperparameter γ controls the width of the Gaussian of the RBF kernel (see eq 2) and its optimal value depends entirely on the data set. By default, γ is calculated with the following equation

$$\gamma = \frac{1}{n \times \sigma^2} \tag{7}$$

where n is the number of features and σ^2 is the variance of the input samples. For the two data sets measured at cryogenic temperatures, we calculate γ using eq 7. For the room-temperature data sets, we set γ = 0.0009 and γ = 0.004 for **C6-thiol** and **RT-bpy**, respectively.

In an ideal scenario, the choice of an optimal γ would be based on the performance of the model on a test data set, with both 'normal' and 'abnormal' classes. However, due to the lack of labels, we manually calibrated γ by gauging the appearance of molecular peaks in the 1D- and 2D-histograms. To prevent false positives, we set the value of γ such that there was no appearance of a peak at the molecular conductance in the class labeled as blanks. This conservative approach prioritizes classifying blank traces as molecular. We reasoned that this bias presents a lesser risk to subsequent analyses than misidentifying molecular traces as blanks. We note that the optimal γ setting might vary based on any specific subsequent analysis at hand. In scenarios where minimizing false negatives is crucial—avoiding the misclassification of tunneling traces as molecular, for instance—a different γ adjustment may be warranted. Hence, our

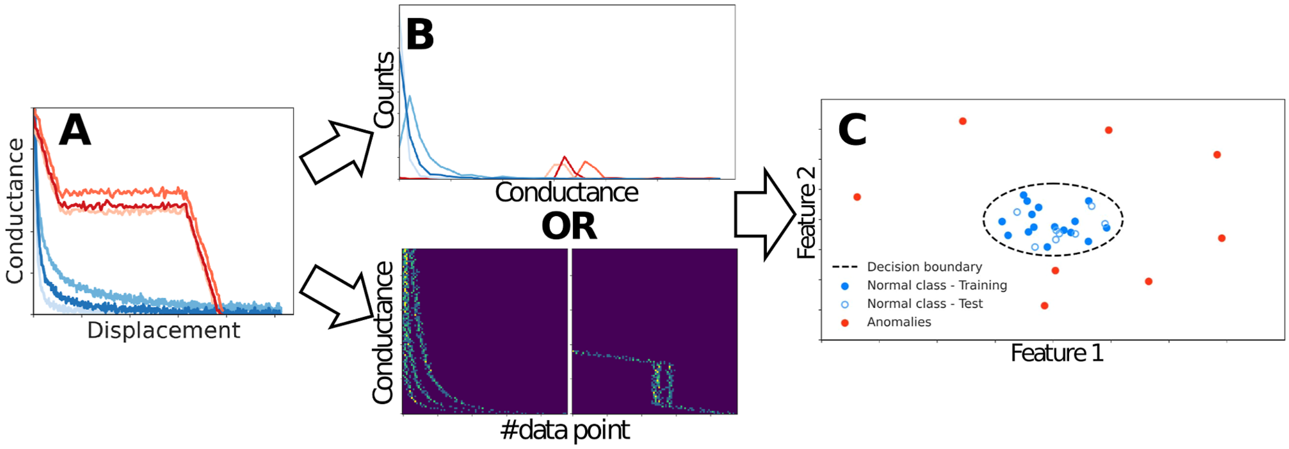


Figure 2. Schematic illustration of the OC classification workflow. (A) Simulated traces showing the idealized difference between molecular (red) and blank (blue) traces. (B) Each individual trace is either converted to a 1D- or 2D-histogram. (C) A model is trained solely on blank traces from a blank experiment (blue, solid circles). After training, the model classifies all traces from an experiment with a molecule. Samples that fall within its decision boundary are labeled blank (blue, open circles), and samples that fall outside are labeled not-blank (red, solid circles).

chosen setting for γ reflects an appropriate decision only under these assumptions rather than a universally optimal one.

The second type of model used to filter blank traces is a Gaussian mixture model (GMM). These models consists of a weighted sum of m Gaussian densities of the form

$$p(\mathbf{x}|w_i, \mu_i, \Sigma_i) = \sum_{i=1}^{n} w_i g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{8}$$

Here, n is the number of samples, $\mathbf{x}$ is the D-dimensional input vector, w_i are the mixture weights, $g(\mathbf{x}|\boldsymbol{\mu}_i, \Sigma_i)$ are the Gaussian densities, μ_i are the mean vectors of each Gaussian, and Σ_i are the covariance matrices. The mixture weights satisfy the constraint $\sum_{i=1}^{n} w_i = 1$.

Each of the Gaussians are a D-variate Gaussian of the form

$$g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right) \tag{9}$$

For a more thorough introduction to GMMs, we refer to Reynolds.[48]

For the GMMs, we change two hyperparameters: the number of Gaussian densities used to fit the data (components) and the threshold at which the data is classified as lying outside the normal class. The number of components are noted in the text whenever we mention a GMM model. For example, A GMM with three components is called GMM-3. The threshold for all GMM models is set to 0.99.

All other parameters are kept at their default values for both types of models. We use the OC-SVM from scikit-learn[49] and GMM from scikit-lego.[50]

Our modeling process, including the data input and preprocessing, varies due to the diverse nature of the data sets we analyze. These differences in data are not only due to random noise but rather reflections of unique molecular properties and varied experimental circumstances. This necessitates the application of different featurization strategies to optimally represent the characteristics of a specific data set. A detailed summary of the featurization strategies used for each data set can be found in Table 1. This tailored approach allows our OC classification model to distinguish effectively between molecular and blank traces across a wide spectrum of single-molecule transport experiments.

For all data sets except **C6-thiol**, we remove any data point above $10^{-0.5}$ G_0 and below 10^{-5} G_0. For **C6-thiol**, we remove any data point above 10^{-1} G_0. We refer to this as 'thresholding'.

For **4K-bpy**, we discard any trace that have no data points after thresholding. The remaining traces are converted into 1D-histograms with 128 bins, serving as input for our models.

For **C6-thiol**, **RT-bpy**, and **Fc**, we discard any trace that has fewer than 32 or more than 6000 data points after thresholding. Very short (<32 data points) and very long traces (>6000 data points) are likely to be samples where the conditions for measuring a single molecule were not met.

We convert the traces of **C6-thiol** into 2D-histograms with 64 × 64 bins and the traces of **RT-bpy** into 2D-histograms with 32 × 32 bins. These 2D-histograms are used as input for our models. For **Fc**, we convert each trace into a 1D-histogram with 100 bins as input for our models.

We use the number of data points in each trace as a proxy for the length of the molecular junction. Any motivated reader will be able to estimate the molecular junction lengths from the raw data that we have shared and from the original publications, where present.

In what follows, we distinguish between molecular traces and traces that are not-blanks. OC models attempt to establish a *normal* class such that they can distinguish normal and *anomalous*/*novel* samples. While the normal class is well-defined, the same cannot be said of the anomalous class. For example, if the normal class is chairs, then the anomalous class would be anything not a chair. Therefore, a sample classified as an *anomaly* does not guarantee that the sample is a molecular trace; instead, it could be a trace from an impurity not present during the blank experiment. In this paper, we refer to samples obtained from the blank experiments as "blank samples" and samples identified as the normal class by the OC model as "blanks"

All scripts used to analyze the data in this article are available at https://github.com/chem-william/one_class_smbj. The data sets can be downloaded here: https://erda.ku.dk/archives/5df033bfa19fd24b50c7c88300ea7640/published-archive.html.

■ RESULTS

The OC classification workflow is depicted schematically in Figure 2. First, the raw traces (Figure 2A) are converted to a 1D- or a 2D-histogram (Figure 2B). These two representations are common in single-molecule transport studies and show good performance in this study. It should be noted that there are other representation options. For example, raw traces could directly be used as input or each trace could be characterized by a series of descriptors such as length, mean conductance, variance, etc.

The final step of our process involves fitting a model (Figure 2C) using exclusively the blank samples, which in our case, are

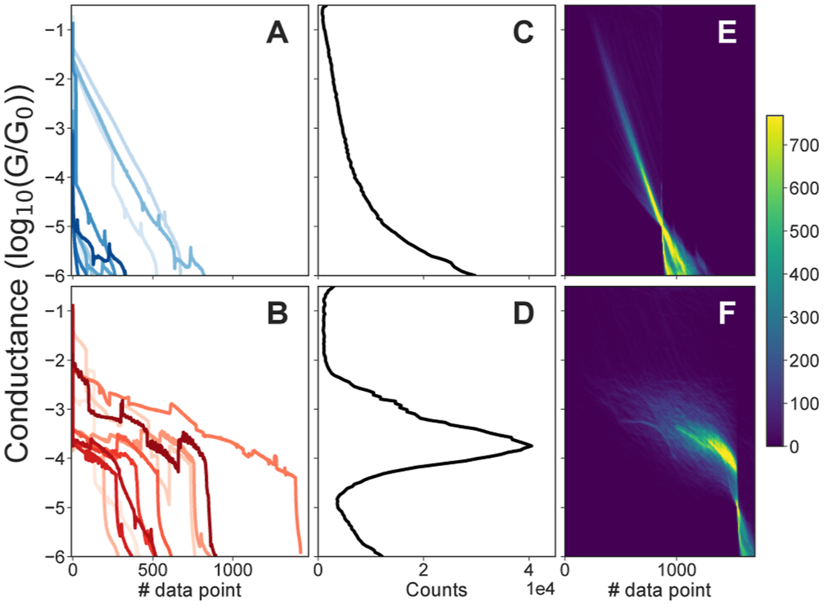


Figure 3. Visualization of the **4K-bpy** data set. (A) Example traces labeled as blanks. (B) Example traces labeled as molecular. (C) 1D-histogram of all blank traces. (D) 1D-histogram of all molecular traces. (E) 2D-histogram of all blank traces. (F) 2D-histogram of all molecular traces.

samples from a blank experiment (represented by blue, solid circles). The model fits a decision boundary (black, dashed line) that is neither too strict nor too loose, thus avoiding both over- and underfitting. All samples from a molecular experiment that belong to the normal class fall within the decision boundary (blue, hollow circles), while samples that fall outside are classified as anomalies (red, solid circles).

When we show 2D-histograms in the following, it is only the first data set, **4K-bpy**, that has had its traces aligned at $10^{-5}G_0$. Alternative alignments for each data set, relative to those presented here in the main article, can be found in the Supporting Information. In particular, Figures S1 and S2 in the SI illustrate 2D-histograms for the **4K-bpy** data set but without trace alignment. In contrast, Figures S3−S8 showcase 2D-histograms for the remaining data sets, where all traces have been aligned at $10^{-5}G_0$.

Results: Cryogenic-Temperature Experiments

We start by applying OC classification models to two data sets obtained at cryogenic temperatures. Conducting experiments at such low temperatures reduces thermal noise, enhancing the contrast between 'blank traces' and 'molecular traces', thereby facilitating easier data analysis and interpretation. The first data set, **4K-bpy**, is also the only data set that has been fully labeled, allowing us to quantify the performance of our models.

4K-bpy. In Figure 3, we present the **4K-bpy** data set. The experimental setup for gathering the data has been explained in a previous publication.[7] After preprocessing, it consists of 3219 blank traces and 1863 molecular traces. In Figure 3A,B, we show 10 blank and molecular traces, respectively. In Figure 3C,D, we show 1D-histograms of all traces labeled as blank and molecular traces, respectively. Finally, in Figure 3E,F, we show 2D-histograms of all traces labeled as blank and molecular traces, respectively.

As all traces have been labeled, we can quantify the performance of any given model. In Table 2, we report the mean accuracy and mean area under the receiver operating characteristic curve (AUROC) along with the standard deviation over a fivefold cross-validation. In the Supporting Information (Figure S9), we plot the receiver operating characteristic curve. Previously, we have shown why it is important to report more than just the accuracy.[51]

Table 2. Performance on fivefold Cross-Validation[a]

model	accuracy (%)	AUROC
OC-SVM (linear)	37.9 ± 2	67.4 ± 2
OC-SVM (RBF)	**96.9±0.3**	**99.5±0.3**
GMM-1	95.6 ± 0.7	99.0 ± 0.2
GMM-3	95.0 ± 0.8	97.9 ± 0.4

[a] Errors are the SD.

Except for the OC-SVM with a linear kernel, all models do an excellent job at distinguishing between blank and molecular traces despite having only been trained on blank traces. The best-performing model is the OC-SVM with an RBF kernel with an accuracy of 96.9 ± 0.3%. This model is slightly better than the GMM with one component, which has an accuracy of 95.6 ± 0.7%. Increasing the number of components in the GMM seems to slightly reduce the performance.

While quantifying the performance of models is important, it is also informative to gauge their performance qualitatively. In the left column of Figure 4, we show 1D-histograms of the **4K-bpy** data set after each trace has been classified. In the right column, we show 2D-histograms of all the traces that have been classified as blank traces. Each row corresponds to a new model: an OC-SVM with a linear kernel is used in Figure 4A and B; an OC-SVM with an RBF kernel in Figure 4C,D; a GMM with one component in Figure 4E,F; and a GMM with three components in Figure 4G,H.

Clearly the performance of the models qualitatively matches what we expected from the accuracy and AUROC reported in Table 2. The OC-SVM with a linear kernel performs very poorly as it barely classifies any traces as blank. The performance could potentially be improved by choosing a different feature set.

As is also clear from Table 2, the OC-SVM with an RBF kernel is the best model, although the two GMM models perform similarly. However, a slight bump is visible in the 1D-

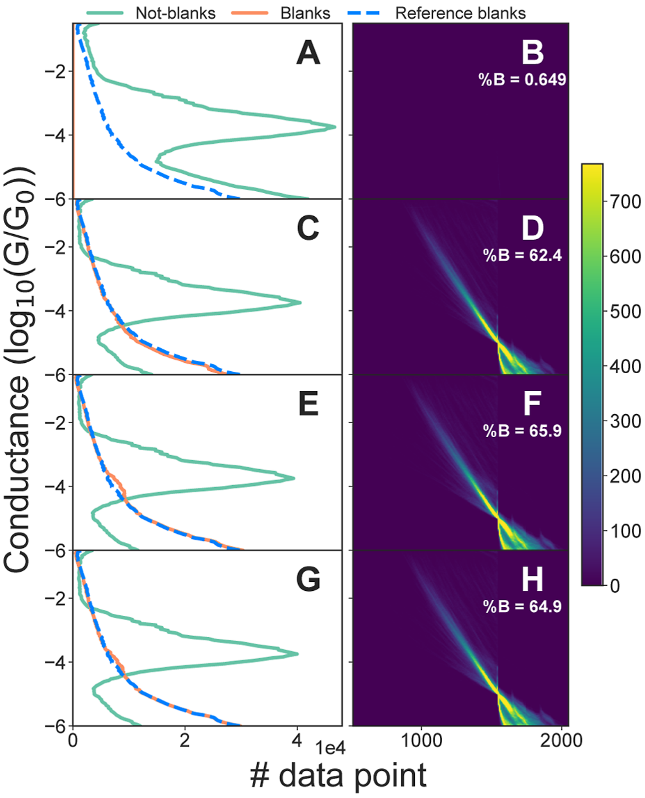


Figure 4. Left column shows 1D-histograms of classified molecular (green, solid line), classified blank (orange, solid line), and the reference blank traces (blue, dashed line) for the **4K-bpy** data set. The right column shows 2D-histograms of the classified blank traces. We show the following models: (A,B) OC-SVM with a linear kernel, (C,D) OC-SVM with an RBF kernel, (E,F) GMM with one component, and (G,H) GMM with three components. The %B denotes the percentage of traces out of the full data set that have been labeled as blank traces.

histograms from the two GMM models at the molecular peak, suggesting that some molecular traces may have been mislabeled as blank traces (see Figure 4E,G).

In the Supporting Information, we also test five different clustering algorithms that have been reported in the literature.[52−55] The best-performing clustering algorithm is a GMM achieving an accuracy of 92.1% with 1D-histograms as input. Except for the OC-SVM with a linear kernel, all OC models perform better than the best clustering algorithm. At the end of the manuscript, we discuss why such a comparison, while informative, can also be problematic.

Through an initial analysis using the **4K-bpy** data set, we have demonstrated the qualitative and quantitative efficacy of the examined OC classification models, which exhibit exceptional performance on high-quality experimental data. As we explore data sets without labeled tunneling and molecular traces, it is reassuring to note that our intuition of what constitutes a good model translates into robust quantitative metrics. This alignment bolsters our confidence in applying these models to the unlabeled data sets.

Fc. In Figure 5, we present the **Fc** data set. This experiment measures ferrocene, a very short molecule compared with 4,4′-bipyridine measured in the **4K-bpy** data set. Therefore, even traces that have a molecular plateau are expected to be short. The experimental setup for gathering the data has been explained in a previous publication.[32] After preprocessing, the data set consists of 7916 blank traces and 6167 traces from an experiment with molecules. In Figure 5A,B, we show 10 individual traces from a blank experiment and from an experiment with a molecule added, respectively. In Figure 5C,D, we show 1D-histograms of all traces from the blank and molecular experiments, respectively. Finally, in Figure 5E,F, we show 2D-histograms of all the traces from the blank and the molecular experiments, respectively.

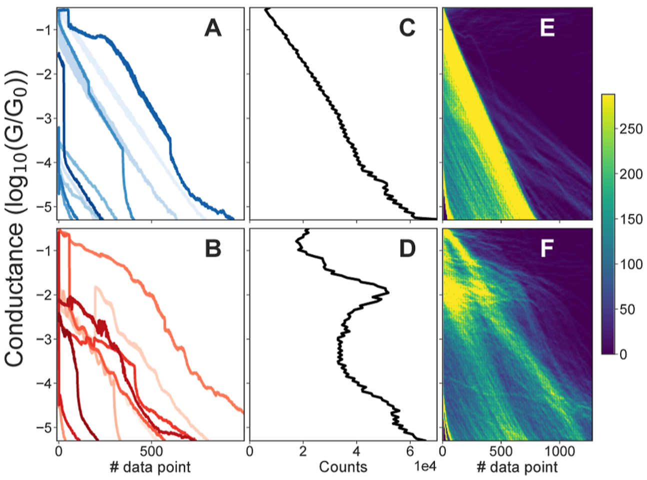


Figure 5. Visualization of the **Fc** data set. (A) Example traces from a blank experiment. (B) Example traces from an experiment with the molecule added. (C) 1D-histogram of all traces from the blank experiment. (D) 1D-histogram of all traces from the experiment with the molecule added. (E) 2D-histogram of all traces from the blank experiment. (F) 2D-histogram of all traces from the experiment with the molecule added.

In Figure 6, we show the results of filtering blank traces from the **Fc** data set. In the left column of Figure 6, we show 1D-histograms of the **Fc** data set after each trace has been classified. In the right column, we show 2D-histograms of all the traces that have been classified as blank traces. Each row corresponds to a new model: an OC-SVM with a linear kernel is used in Figure 6A,B; an OC-SVM with an RBF kernel in Figure 6C,D; a GMM with one component in Figure 6E,F; and a GMM with three components in Figure 6G,H.

Again, the OC-SVM with a linear kernel performs poorly compared to the other models. The OC-SVM with an RBF kernel and the two GMMs perform similarly well. The GMMs classify slightly more traces as blanks. In the 1D-histograms of the GMMs, there seems to be a slight hint of a peak at the molecular conductance around $10^{-2}G_0$. In the 2D-histograms, there also seems to be a slightly higher number of data points at $10^{-2}G_0$. Both details indicate that the GMMs might classify some molecular traces as blank traces.

For the **4K-bpy** data set, we are able to ascertain that almost all traces are correctly classified because the entire data set is labeled. We do not have the same certainty with the traces from the **Fc** data set and subsequent data sets. Therefore, we further analyze the traces labeled as not-blank.

For the next steps, we use all the traces that an OC-SVM model with an RBF kernel classified as not-blank. Qualitatively, approximately three different types of traces appear to be present. In Figure 7, we show these three classes (Figure 7A−F) alongside the traces from the blank experiment (Figure 7G,H). In the Supporting Information, we show all the traces labeled as not-blank in a single 2D-histogram (Figure S11).

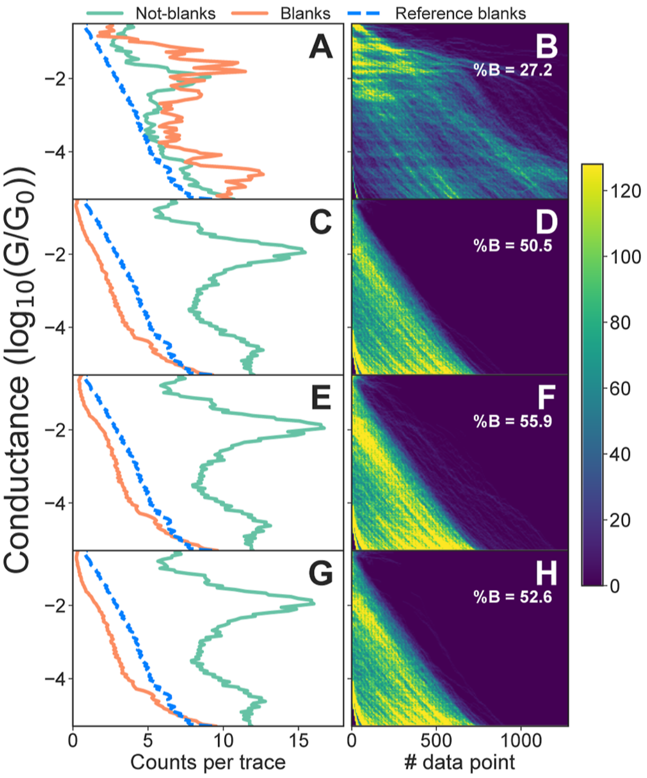


Figure 6. Left column shows 1D-histograms of classified molecular (green, solid line), classified blank (orange, solid line), and the reference blank traces (blue, dashed line) for the **Fc** data set. The right column shows 2D-histograms of the classified blank traces. We show the following models: (A,B) OC-SVM with a linear kernel, (C,D) OC-SVM with an RBF kernel, (E,F) GMM with one component, and (G,H) GMM with three components. The %B denotes the percentage of traces out of the full data set that have been labeled as blank traces.

There, we also explain how we extracted the three different classes.

We show individual traces from each class in the top row of Figure 7, and we plot a 2D-histogram of each class in the bottom row. In Figure 7A,B, we show examples of the class predominantly consisting of molecular traces. In Figure 7C,D, we show examples of traces that could be unusual molecular traces or traces from contaminants. In Figure 7E,F, we show examples of traces that are very similar to prototypical blank traces. For comparison, we have plotted the traces from the blank experiment alongside the rest in Figure 7G,H.

As expected, a subset of the traces identified as 'not-blank' displays a molecular signal, as shown in Figure 7A,B. These traces display distinct features indicative of the presence of molecules. On the other hand, the traces in Figure 7C,D are likely not molecular traces, but they differ from the blank traces shown in Figure 7G,H. These traces were omitted in the original publication as they are indicative of a junction that did not have a gold point contact before rupture.[32]

The traces depicted in Figure 7E,F highlight the effectiveness of OC classification for single-molecule transport studies. By comparing these traces with the blank traces from the molecular experiment in Figure 7E,F, it is evident that these two classes are not the same. Importantly, our model has been trained solely on the examples shown in Figure 7G,H.

Our model only filtered out blank traces that resemble those from the blank experiment. Consequently, any remaining traces are highly likely to have originated from different physical processes than the ones leading to the blank traces. In this particular instance, the introduction of a molecule to the experiment causes some of the blank-like traces to exhibit the noisy and curved behavior seen in Figure 7E,F rather than the clean signal of the traces shown in Figure 7G,H. The exact physical origin of this difference is unclear based on this experiment alone.

The utility of OC classification models in single-molecule transport studies is further underlined with the **Fc** data set. Notably, these models excel even with very short molecules and exhibit remarkable specificity, as evident in Figure 7E,F, where traces similar to tunneling traces from blank experiments are not classified as such. However, this data set also underscores the importance of a representative training set. As depicted in Figure 7C,D, the introduction of a new type of defect or contaminant in the molecular experiment may lead to its correct classification as 'not-blank' even though the practitioner might still have wanted it filtered out.

Furthermore, we use five clustering techniques documented in the literature and show the results in the Supporting

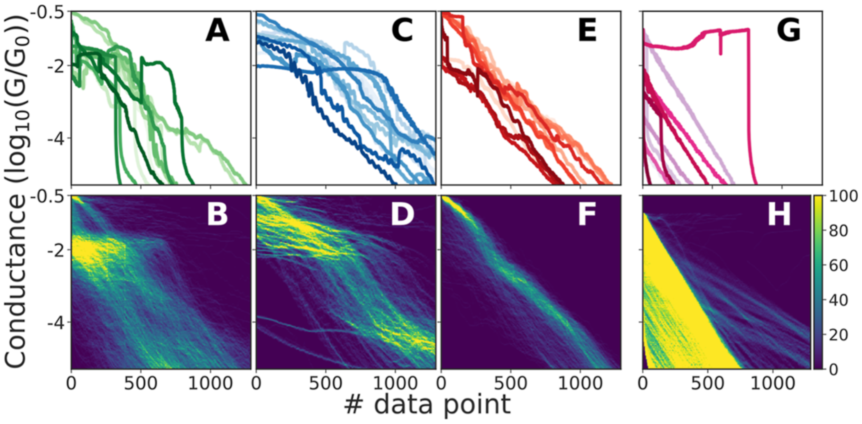


Figure 7. Analysis of the traces classified as not blank from the **Fc** data set. (A,B) Example traces that exhibit molecular features. (C,D) Example traces that does not have a 1 G_0 plateau. (E,F) Example traces that have no clear molecular plateau. (G,H) Example traces from the blank experiment. In the top row, we show 10 individual traces for each class. and in the bottom row, we show 2D-histograms of all traces from each class. We used the OC-SVM with an RBF kernel to classify blank traces.

Information. We then evaluate the impact of filtering out tunneling traces by comparing the clustering outcomes before and after filtering the tunneling traces.

Results: Room-Temperature Experiments

Compared with experiments conducted at cryogenic temperatures, room-temperature experiments exhibit more noise, both in each individual sample and in the overall phase space the experiment samples. Even so, room-temperature experiments are more common. Therefore, in the following, we investigate how OC models perform on a room-temperature data set. In the Supporting Information, we have included an analysis of a fourth data set (the **C6-thiol** data set) that is also measured at room-temperature but is from a different lab and of a different molecule.

RT-bpy. In Figure 8, we present the **RT-bpy** data set. The experimental setup for gathering the data is explained in the Supporting Information and in previous publications.[7,9] After preprocessing, it consists of 9787 blank traces and 20,129 traces from an experiment with a molecule. In Figure 8A,B, we show 10 traces from a blank experiment and an experiment with an added molecule, respectively. In Figure 8C,D, we show 1D-histograms of all traces from the blank and molecular experiment, respectively. Finally, in Figure 8E,F, we show 2D-histograms of all traces from the blank and molecular experiment, respectively. Please note the increased variance of the samples compared with the samples from the **4K-bpy** data set or the **Fc** data set.

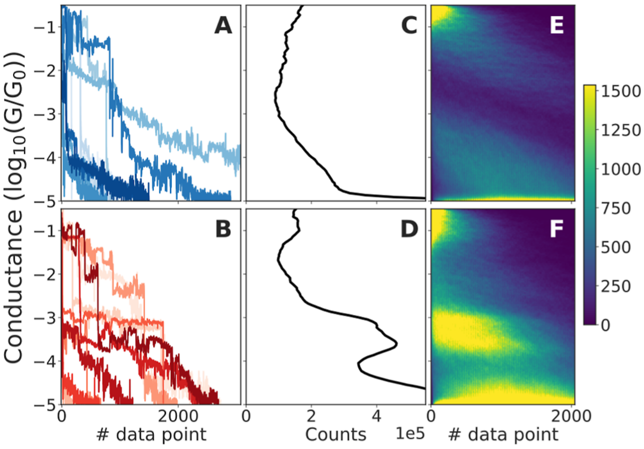


Figure 8. Visualization of the **RT-bpy** data set. (A) Example traces labeled as blank. (B) Example traces labeled as molecular. (C) 1D-histogram of all blank traces. (D) 1D-histogram of all molecular traces. (E) 2D-histogram of all blank traces. (F) 2D-histogram of all molecular traces.

In Figure 9, we show the results of filtering blank traces from the **RT-bpy** data set. In the left column of Figure 9, we show 1D-histograms of the **RT-bpy** data set after each trace has been classified. In the right column, we show 2D-histograms of all the traces that have been classified as blank traces. Each row corresponds to a new model: an OC-SVM with a linear kernel is used in Figure 9A,B; an OC-SVM with an RBF kernel in Figure 9C,D; a GMM with three components in Figure 9E,F; a GMM with nine components in Figure 9G,H; a GMM with 12 components in Figure 9I,J; and a GMM with 20 components in Figure 9K,L. We test more models and with a higher amount of components due to the higher complexity of the room-temperature data set compared with the data sets measured at cryogenic temperatures.

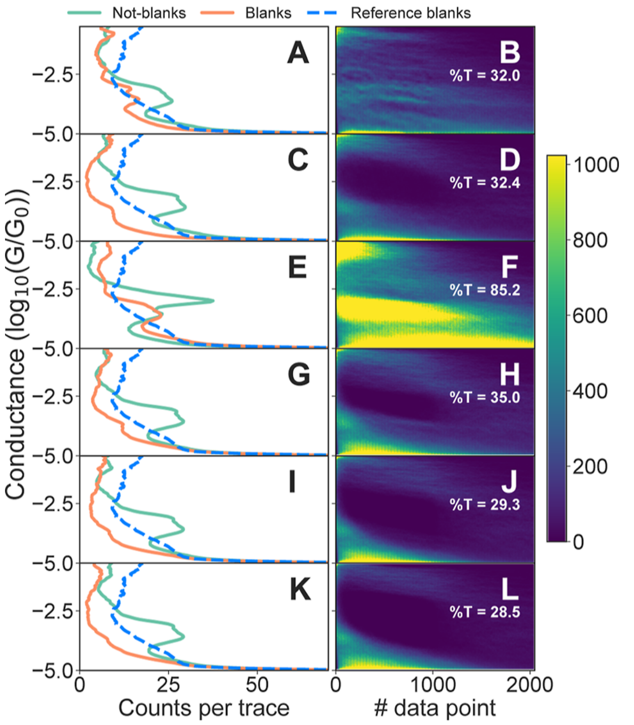


Figure 9. Left column shows 1D-histograms of classified molecular (green, solid line), classified blank (orange, solid line), and the reference blank traces (blue, dashed line) for the **RT-bpy** data set. The right column shows 2D-histograms of the classified blank traces. We show the following models: (A,B) OC-SVM with a linear kernel, (C,D) OC-SVM with an RBF kernel, (E,F) GMM with three components (G,H) GMM with nine components, (I,J) GMM with 12 components, (K,L) and GMM with 20 components. The 1D-histograms of each class have been divided by the amount of samples in their respective class. The %B denote the percentage of traces out of the full data set that have been labeled as blank traces.

We can see from Figure 9 that the OC-SVM with a linear kernel and the GMM with three components perform the poorest. The OC-SVM with an RBF kernel appears to perform at the same level as the GMM with nine components.

The GMMs with 12 and 20 components (Figure 9I−L, respectively) seem to perform the best based on the 1D-histogram of the classified blank traces, where there is almost no peak at the same location as the main molecular peak of the molecular traces. The GMM with 20 components classifies fewer traces as blanks than the rest of the models. It is difficult to determine whether this is due to fewer misclassifications or if there are truly fewer blank traces present.

To illustrate the impact that blank traces can have on downstream analysis tasks, we compare 2D correlation histograms of the filtered and unfiltered data set.[56] In Figure 10A,B, we show 2D correlation histograms of the **RT-bpy** data set before and after filtering, respectively.[56−58] We use the GMM with 20 components for filtering blank traces. In the Supporting Information, we show 2D correlation histograms for the **C6-thiol** and the **Fc** data sets.

It is clear from Figure 10B that filtering blank traces substantially changes the correlations. The most drastic change is seen in the blue box. In this region, parts of the traces are initially positively correlated but become negatively correlated after blank traces are filtered out.

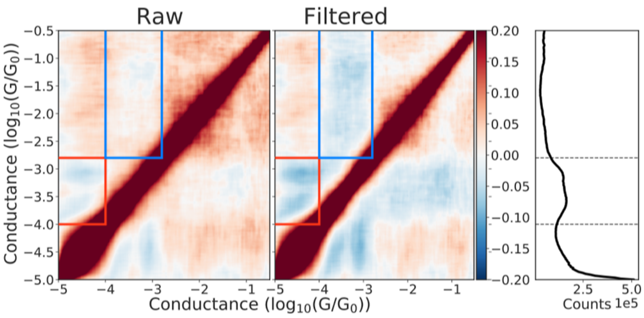


Figure 10. Correlation analysis of the **4K-bpy** data set before and after filtering for blank traces. (A) Correlation plot for the full data set. (B) Correlation plot for the data set after blank traces have been removed. (C) 1D-histogram of the data set after filtering for blank traces. The dotted, black lines are a guide for the eye. We use a GMM with 20 components to filter out blank traces. Note that the scales do not go from −1 to 1.

The majority of filtered traces exhibit little to no signal in the range of $10^{-0.5}G_0$ to $10^{-3}G_0$ (dark blue area), as shown in Figure 9K,L. Due to the nearly uniform values in this range, it leads to a positive correlation that obscures the underlying negative correlation present after filtering. In the Supporting Information, we have plotted a 2D-correlation histogram of only the traces classified as blanks.

A smaller change is seen in the red box. In this region, the negative correlation that is already present becomes more pronounced after blank traces have been removed. Such negative correlations have been reported elsewhere.[56]

Our findings on the **RT-bpy** data set echoes the findings from the **C6-thiol** data set that is presented in the Supporting Information. Despite the increased complexity of room-temperature data sets compared with cryogenic temperature data, the performance of our models remains robust. Notably, correlation analysis underscores that filtering blank traces can substantially influence subsequent analysis. As a result, the interpretations derived from filtered data sets can be qualitatively different from those based on the raw data set.

As with the previous, unlabeled data sets, we compare the use of five clustering techniques before and after filtering tunneling traces. These results are shown in the Supporting Information.

■ DISCUSSION

In recent years, clustering techniques have become increasingly popular for analyzing single-molecule transport data.[2−4,28,52,54,55,59−64] As mentioned in the introduction, these are sometimes used to filter out blank traces. However, filtering blank traces is often a byproduct of separating the molecular traces into different classes. Moreover, the end-user needs to decide which class (or classes) constitutes the blank class. In contrast, OC classification specifically targets the removal of blank traces from a molecular experiment. The defining characteristics of these blank traces have been explicitly learned from the training data obtained from a blank experiment. Therefore, we can have greater confidence that very few traces with molecular features will be included in the blank class.

While both techniques can be utilized to filter out blank traces, making a fair comparison between them is challenging. One of the major obstacles is the lack of labeled data sets. On a per-sample basis, it is difficult to determine the underlying event that led to a given measured sample. Therefore, any manual labeling of a data set may not accurately reflect the true labels of the data set. Even if we had a fully labeled data set, those labels would only represent one possible partitioning of the data set. There exists a plethora of clustering algorithms and similarity metrics, and any given combination of these will give a result that might not correspond to the fully labeled data set.

OC classification and clustering techniques address distinct problems that may overlap in certain situations. Neither technique can fully replace the other. By effectively filtering out blank traces from a data set, subsequent analyses will benefit from the reduced noise and confounding signals. We envision that both techniques can be integrated into data analysis pipelines to improve results and uncover new insights.

The fundamental assumption of using OC classification to remove blank traces from a molecular experiment is that the blank traces from a blank experiment are the same as the blank traces from a molecular experiment. There are valid reasons to believe that this assumption may not hold in practice. For instance, the chemical environment undergoes substantial changes with the addition of the molecule of interest. However, as demonstrated with the **Fc** data set, there still appears to be a large fraction of traces that are highly similar to the traces from the blank experiment. This result suggests that the assumption of different blank traces only partially breaks down and that OC classification remains applicable. The analysis of the **Fc** data set also demonstrates that even if the assumption of similar blank traces completely breaks down, the OC model would still classify the different blank traces as abnormal.

The introduction of solvents into the experiment could potentially enhance the differentiation of blank traces as solvent molecules can alter the evolution of the junction. Additionally, trace contaminants that may be present in the solvent could also affect the process. These factors can both lead to blank traces that have features distinct from those of typical, exponentially decaying tunneling traces.

One advantage of OC classification is the ability to model the blank class using a variety of blank traces. This allows for a diverse classification of traces that a clustering algorithm may potentially split into multiple classes. Another advantage is that an algorithm trained on a specific training set ensures that if we know the characteristics of that training set, we know what traces will be classified as blank traces.

In the Supporting Information, we compare the length distributions of traces from blank experiments with the traces classified as blank and not-blank for the **Fc**, **C6-thiol**, and **RT-bpy** data sets. For the well-performing models, samples that are classified as blanks tend to be shorter, while the traces classified as not-blanks tend to be longer. However, there is considerable overlap between the two distributions. This result matches our intuition: the conductance measured for traces with no molecule should fall off relatively quickly. It also illustrates that the models discriminate between traces based on more than just their length.

In the Supporting Information, we additionally provide histograms that display summary statistics such as the mean, median, standard deviation, and slope for each data set. This further illustrates the overarching traits of the traces that have been filtered out. To complement these statistics, we also include representative traces from each data set that have been identified as 'blanks' by the, respective, models.

Practical Considerations for Using OC Classification Models

It is advisible to initially experiment with a variety of OC models. We demonstrated that the OC-SVM consistently performs well, although the GMM also shows promising results. Furthermore, while we have not explored them in this paper, a wealth of other models exists.[65−67] This initial exploration allows for a more informed selection of the model that best suits the specific needs of the study.

Calibration of the chosen model with a known molecule is a crucial step and, if feasible, labeling some traces from the molecular experiment can provide a quantifiable calibration. Furthermore, maintaining consistency in experimental conditions between blank and molecular experiments is important. This ensures that any observed differences can be attributed to the molecule under study, rather than variations in the experimental setup.

Incorporating OC classification models into an analysis pipeline that also includes clustering can yield more nuanced insights. Once the data has been cleaned of tunneling traces using the OC classification model, a clustering algorithm may be able to discern subtle details more effectively. It is essential to adhere to all general standards and principles related to ML and model calibration, as outlined in the previous literature.[51,68] This will ensures the robustness and reproducibility of the findings.

CONCLUSION

In this work, we have demonstrated the use of OC classification methods to filter blank traces on four diverse data sets from three different laboratories. Using OC classification provides a robust and reliable approach to effectively remove blank traces from molecular experiments. As highlighted in the introduction, various laboratories employ different techniques for blank trace filtering. By utilizing OC classification, we achieve a more principled quantification of the traces that are filtered out while demonstrating excellent performance.

Using the labeled **4K-bpy** data set, we show that an OC-SVM with an RBF kernel achieves an accuracy of 96.9 ± 0.3% and an AUROC of 99.5 ± 0.3 by training solely on the blank traces. This accuracy surpasses the previously reported accuracy on the same data set using a supervised algorithm. We validate the excellent performance by visualizing 2D-histograms of the classified blank traces and comparing 1D-histograms of the classified blank traces with the reference blank traces.

On the more challenging **Fc** data set, which has also been measured at cryogenic temperatures, we also see good performance. An OC-SVM with an RBF kernel classifies approximately 50% of the traces as blanks. Additionally, we explore the traces classified as 'not-blanks' to gain a deeper understanding of the behavior of our model. It becomes apparent that the model can distinguish between subtly different traces. Despite the traces exhibiting the characteristic linear drop-off in conductance of prototypical blank traces, they are labeled as 'not-blanks'. This classification is likely due to their noisier and more curved profile compared with the traces from the blank experiment.

Furthermore, our approach opens up possibilities for studying the differences between blank traces from a molecular experiment and those from a blank experiment. By filtering out traces that are identical to the ones from the blank experiment, we are left with only the blank traces that are different. This provides an opportunity for further in-depth exploration and analysis of these distinct blank traces.

The last two data sets, both measured at room-temperature, exhibited higher levels of noise compared with the data sets measured at cryogenic temperatures. Despite the increased noise, the tested models still demonstrate good performance. Depending on the chosen model, between 15 and 40% of traces are identified and removed as blank traces from the **C6-thiol** data set without any indication of misclassification of molecular traces.

For the **RT-bpy** data set, the percentage of removed blank traces ranges from 28 to 35%. Moreover, we demonstrate that the removal of blank traces has a substantial impact on the 2D-correlation histograms calculated for the filtered data set, as compared with the raw data set.

Overall, our work builds upon the impressive capabilities of chemists to determine the chemical properties of single molecules in single-molecule transport experiments, while addressing inherent limitations in this approach. By providing a more precise and detailed understanding of molecular behavior, our aim is to contribute to the ongoing quest to comprehend the fundamental properties of matter at the single-molecule level.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsnanoscienceau.4c00015.

Experimental setups for **C6-thiol** and **RT-bpy**; analysis of C6-thiol data set; different alignments of 2D-histograms for all data sets; performance of models on the **4K-bpy** data set; clustering on **4K-bpy**; not-blank traces from **Fc**; clustering results on unlabeled data sets; 2D-correlation plots; correlation analysis of blanks from **RT-bpy**; length distributions; summary statistics; and visualization of traces classified as blanks (PDF)

AUTHOR INFORMATION

Corresponding Authors

András Halbritter − *Department of Physics, Institute of Physics, Budapest University of Technology and Economics, Budapest H-1111, Hungary; ELKH-BME Condensed Matter Research Group, Budapest H-1111, Hungary;* orcid.org/0000-0003-4837-9745; Email: halbritter.andras@ttk.bme.hu

Maria Kamenetska − *Department of Physics, Chemistry and Division of Material Science and Engineering, Boston University, Boston, Massachusetts 02215, United States;* orcid.org/0000-0002-0390-035X; Email: mkamenet@bu.edu

Veerabhadrarao Kaliginedi − *Department of Inorganic and Physical Chemistry, Indian Institute of Science, Bangalore 560012, India;* orcid.org/0000-0002-4361-741X; Email: vkaliginedi@iisc.ac.in

Gemma C. Solomon − *Department of Chemistry and Nano-Science Center, University of Copenhagen, Copenhagen Ø DK-2100, Denmark; NNF Quantum Computing Programme, Niels Bohr Institute, University of Copenhagen, Copenhagen N DK-2200, Denmark;* orcid.org/0000-0002-2018-1529; Email: gsolomon@chem.ku.dk

Authors

William Bro-Jørgensen − *Department of Chemistry and Nano-Science Center, University of Copenhagen, Copenhagen Ø DK-2100, Denmark;* orcid.org/0000-0001-8171-6374

Joseph M. Hamill − *Department of Chemistry and Nano-Science Center, University of Copenhagen, Copenhagen Ø DK-2100, Denmark;* orcid.org/0000-0002-9024-4636

Gréta Mezei − *Department of Physics, Institute of Physics, Budapest University of Technology and Economics, Budapest H-1111, Hungary; ELKH-BME Condensed Matter Research Group, Budapest H-1111, Hungary;* orcid.org/0009-0002-5763-6168

Brent Lawson − *Department of Physics, Chemistry and Division of Material Science and Engineering, Boston University, Boston, Massachusetts 02215, United States*

Umar Rashid − *Department of Inorganic and Physical Chemistry, Indian Institute of Science, Bangalore 560012, India;* orcid.org/0000-0001-6583-061X

Complete contact information is available at: https://pubs.acs.org/10.1021/acsnanoscienceau.4c00015

Author Contributions

CRediT: **William Bro-Jørgensen** conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualization, writing-original draft, writing-review & editing; **Joseph Martin Hamill** data curation, formal analysis, investigation, methodology, supervision, validation, writing-review & editing; **Gréta Mezei** data curation, resources, validation, writing-review & editing; **Brent Lawson** data curation, resources, validation, writing-review & editing; **Umar Rashid** data curation, methodology, resources, validation, writing-review & editing; **András Halbritter** data curation, funding acquisition, methodology, resources, supervision, validation, writing-review & editing; **Maria Kamenetska** data curation, funding acquisition, methodology, resources, supervision, validation, writing-review & editing; **Veerabhadrarao Kaliginedi** data curation, funding acquisition, methodology, resources, supervision, validation, writing-review & editing; **Gemma C. Solomon** conceptualization, data curation, funding acquisition, methodology, project administration, resources, supervision, validation, visualization, writing-review & editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Jesper Dahl Jensen for supplying a UV−vis spectrum used for the figure in the introduction. This research used resources of the Center for Functional Nanomaterials (CFN), which is a U.S. Department of Energy Office of Science User Facility, at Brookhaven National Laboratory under Contract no. DE-SC0012704. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 865870). V.K. and U.R. acknowledge the funding support from the Indian Institute of Science, DST-INSPIRE (DST/INSPIRE/04/2018/002983), SERB-Core research grant (CRG/2020/002302).

■ REFERENCES

(1) Harris, D. C. *Quantitative Chemical Analysis*, 8th ed.; Freeman Custom Publishing, 2010, p 491.

(2) Reznikova, K.; Hsu, C.; Schosser, W. M.; Gallego, A.; Beltako, K.; Pauly, F.; van der Zant, H. S. J.; Mayor, M. Substitution Pattern Controlled Quantum Interference in [2.2]Paracyclophane-Based Single-Molecule Junctions. *J. Am. Chem. Soc.* **2021**, *143*, 13944−13951.

(3) Schosser, W. M.; Hsu, C.; Zwick, P.; Beltako, K.; Dulić, D.; Mayor, M.; van der Zant, H. S. J.; Pauly, F. Mechanical conductance tunability of a porphyrin−cyclophane single-molecule junction. *Nanoscale* **2022**, *14*, 984−992.

(4) Hurtado-Gallego, J.; Davidson, R.; Grace, I. M.; Rincón-García, L.; Batsanov, A. S.; Bryce, M. R.; Lambert, C. J.; Agraït, N. Quantum interference dependence on molecular configurations for cross-conjugated systems in single-molecule junctions. *Mol. Syst. Des. Eng.* **2022**, *7*, 1287−1293.

(5) Hsu, C.; Schosser, W. M.; Zwick, P.; Dulić, D.; Mayor, M.; Pauly, F.; van der Zant, H. S. J. Mechanical compression in cofacial porphyrin cyclophane pincers. *Chem. Sci.* **2022**, *13*, 8017−8024.

(6) Hamill, J. M.; Zhao, X. T.; Mészáros, G.; Bryce, M. R.; Arenz, M. Fast Data Sorting with Modified Principal Component Analysis to Distinguish Unique Single Molecular Break Junction Trajectories. *Phys. Rev. Lett.* **2018**, *120*, 016601.

(7) Magyarkuti, A.; Balogh, N.; Balogh, Z.; Venkataraman, L.; Halbritter, A. Unsupervised feature recognition in single-molecule break junction data. *Nanoscale* **2020**, *12*, 8355−8363.

(8) Magyarkuti, A.; Balogh, Z.; Mezei, G.; Halbritter, A. Structural Memory Effects in Gold−4,4′-Bipyridine−Gold Single-Molecule Nanowires. *J. Phys. Chem. Lett.* **2021**, *12*, 1759−1764.

(9) Balogh, Z.; Mezei, G.; Tenk, N.; Magyarkuti, A.; Halbritter, A. Configuration-Specific Insight into Single-Molecule Conductance and Noise Data Revealed by the Principal Component Projection Method. *J. Phys. Chem. Lett.* **2023**, *14*, 5109−5118.

(10) Jang, S.-Y.; Reddy, P.; Majumdar, A.; Segalman, R. A. Interpretation of Stochastic Events in Single Molecule Conductance Measurements. *Nano Lett.* **2006**, *6*, 2362−2367.

(11) Frei, M.; Aradhya, S. V.; Koentopp, M.; Hybertsen, M. S.; Venkataraman, L. Mechanics and Chemistry: Single Molecule Bond Rupture Forces Correlate with Molecular Backbone Structure. *Nano Lett.* **2011**, *11*, 1518−1523.

(12) González, M. T.; Leary, E.; García, R.; Verma, P.; Herranz, M. A. ´.; Rubio-Bollinger, G.; Martín, N.; Agraït, N. Break-Junction Experiments on Acetyl-Protected Conjugated Dithiols under Different Environmental Conditions. *J. Phys. Chem. C* **2011**, *115*, 17973−17978.

(13) González, M. T.; Díaz, A.; Leary, E.; García, R.; Herranz, M. A. ´.; Rubio-Bollinger, G.; Martín, N.; Agraït, N. Stability of Single- and Few-Molecule Junctions of Conjugated Diamines. *J. Am. Chem. Soc.* **2013**, *135*, 5420−5426.

(14) Parker, C. R.; Leary, E.; Frisenda, R.; Wei, Z.; Jennum, K. S.; Glibstrup, E.; Abrahamsen, P. B.; Santella, M.; Christensen, M. A.; Della Pia, E. A.; et al. A Comprehensive Study of Extended Tetrathiafulvalene Cruciform Molecules for Molecular Electronics: Synthesis and Electrical Transport Measurements. *J. Am. Chem. Soc.* **2014**, *136*, 16497−16507.

(15) Moreno-García, P.; La Rosa, A.; Kolivoška, V.; Bermejo, D.; Hong, W.; Yoshida, K.; Baghernejad, M.; Filippone, S.; Broekmann, P.; Wandlowski, T.; et al. Charge Transport in C60-Based Dumbbell-type Molecules: Mechanically Induced Switching between Two Distinct Conductance States. *J. Am. Chem. Soc.* **2015**, *137*, 2318−2327.

(16) Inkpen, M. S.; Lemmer, M.; Fitzpatrick, N.; Milan, D. C.; Nichols, R. J.; Long, N. J.; Albrecht, T. New Insights into Single-Molecule Junctions Using a Robust, Unsupervised Approach to Data Collection and Analysis. *J. Am. Chem. Soc.* **2015**, *137*, 9971−9981.

(17) Lemmer, M.; Inkpen, M. S.; Kornysheva, K.; Long, N. J.; Albrecht, T. Unsupervised vector-based classification of single-molecule charge transport data. *Nat. Commun.* **2016**, *7*, 12922.

(18) Frisenda, R.; Janssen, V. A. E. C.; Grozema, F. C.; van der Zant, H. S. J.; Renaud, N. Mechanically controlled quantum interference in individual π-stacked dimers. *Nat. Chem.* **2016**, *8*, 1099−1104.

(19) Aragonès, A. C.; Haworth, N. L.; Darwish, N.; Ciampi, S.; Bloomfield, N. J.; Wallace, G. G.; Diez-Perez, I.; Coote, M. L. Electrostatic catalysis of a Diels−Alder reaction. *Nature* **2016**, *531*, 88−91.

(20) Chang, W.-C.; Chang, C.-W.; Sigrist, M.; Hua, S.-A.; Liu, T.-J.; Lee, G.-H.; Jin, B.-Y.; Chen, C.-h.; Peng, S.-M. Nonhelical heterometallic $[Mo_2M(npo)_4(NCS)_2]$ string complexes (M = Fe, Co, Ni) with high single-molecule conductance. *Chem. Commun.* **2017**, *53*, 8886−8889.

(21) Leary, E.; Zotti, L. A.; Miguel, D.; Márquez, I. R.; Palomino-Ruiz, L.; Cuerva, J. M.; Rubio-Bollinger, G.; González, M. T.; Agrait, N. The Role of Oligomeric Gold−Thiolate Units in Single-Molecule Junctions of Thiol-Anchored Molecules. *J. Phys. Chem. C* **2018**, *122*, 3211−3218.

(22) Leary, E.; Limburg, B.; Alanazy, A.; Sangtarash, S.; Grace, I.; Swada, K.; Esdaile, L. J.; Noori, M.; González, M. T.; Rubio-Bollinger, G.; et al. Bias-Driven Conductance Increase with Length in Porphyrin Tapes. *J. Am. Chem. Soc.* **2018**, *140*, 12877−12883.

(23) Leary, E.; Roche, C.; Jiang, H.-W.; Grace, I.; González, M. T.; Rubio-Bollinger, G.; Romero-Muñiz, C.; Xiong, Y.; Al-Galiby, Q.; Noori, M.; et al. Detecting Mechanochemical Atropisomerization within an STM Break Junction. *J. Am. Chem. Soc.* **2018**, *140*, 710−718.

(24) Aragonès, A. C.; Darwish, N.; Ciampi, S.; Jiang, L.; Roesch, R.; Ruiz, E.; Nijhuis, C. A.; Díez-Pérez, I. Control over Near-Ballistic Electron Transport through Formation of Parallel Pathways in a Single-Molecule Wire. *J. Am. Chem. Soc.* **2019**, *141*, 240−250.

(25) Tamaki, T.; Minode, K.; Numai, Y.; Ohto, T.; Yamada, R.; Masai, H.; Tada, H.; Terao, J. Mechanical switching of current−voltage characteristics in spiropyran single-molecule junctions. *Nanoscale* **2020**, *12*, 7527−7531.

(26) Dulić, D.; Rates, A.; Castro, E.; Labra-Muñoz, J.; Aravena, D.; Etcheverry-Berrios, A.; Riba-López, D.; Ruiz, E.; Aliaga-Alcalde, N.; Soler, M.; et al. Single-Molecule Transport of Fullerene-Based Curcuminoids. *J. Phys. Chem. C* **2020**, *124*, 2698−2704.

(27) Quintans, C.; Andrienko, D.; Domke, K. F.; Aravena, D.; Koo, S.; Díez-Pérez, I.; Aragonès, A. C. Tuning Single-Molecule Conductance by Controlled Electric Field-Induced trans-to-cis Isomerisation. *Appl. Sci.* **2021**, *11*, 3317.

(28) Xu, W.; Leary, E.; Sangtarash, S.; Jirasek, M.; González, M. T.; Christensen, K. E.; Abellán Vicente, L.; Agraït, N.; Higgins, S. J.; Nichols, R. J.; et al. A Peierls Transition in Long Polymethine Molecular Wires: Evolution of Molecular Geometry and Single-Molecule Conductance. *J. Am. Chem. Soc.* **2021**, *143*, 20472−20481.

(29) González, M. T.; Ismael, A. K.; García-Iglesias, M.; Leary, E.; Rubio-Bollinger, G.; Grace, I.; González-Rodríguez, D.; Torres, T.; Lambert, C. J.; Agraït, N. Interference Controls Conductance in Phthalocyanine Molecular Junctions. *J. Phys. Chem. C* **2021**, *125*, 15035−15043.

(30) Yao, X.; Vonesch, M.; Combes, M.; Weiss, J.; Sun, X.; Lacroix, J.-C. Single-Molecule Junctions with Highly Improved Stability. *Nano Lett.* **2021**, *21*, 6540−6548.

(31) Harashima, T.; Fujii, S.; Jono, Y.; Terakawa, T.; Kurita, N.; Kaneko, S.; Kiguchi, M.; Nishino, T. Single-molecule junction spontaneously restored by DNA zipper. *Nat. Commun.* **2021**, *12*, 5762.

(32) Lawson, B.; Zahl, P.; Hybertsen, M. S.; Kamenetska, M. Formation and Evolution of Metallocene Single-Molecule Circuits with Direct Gold-π Links. *J. Am. Chem. Soc.* **2022**, *144*, 6504−6515.

(33) Chen, Z.-Z.; Wu, S.-D.; Lin, J.-L.; Chen, L.-C.; Cao, J.-J.; Shao, X.; Lambert, C. J.; Zhang, H.-L. Modulating Quantum Interference Between Destructive and Constructive States in Double N-Substituted Single Molecule Junctions. *Adv. Electron. Mater.* **2023**, *9*, 2201024.

(34) Alangari, M.; Demir, B.; Gultakti, C. A.; Oren, E. E.; Hihath, J. Mapping DNA Conformations Using Single-Molecule Conductance Measurements. *Biomolecules* **2023**, *13*, 129.

(35) Ortuño, A. M.; Reiné, P.; Álvarez de Cienfuegos, L.; Márquez, I. R.; Dednam, W.; Lombardi, E. B.; Palacios, J. J.; Leary, E.; Longhi, G.; Mujica, V.; et al. Chiral Single-Molecule Potentiometers Based on Stapled ortho- Oligo(phenylene)ethynylenes. *Angew. Chem., Int. Ed.* **2023**, *62*, No. e202218640.

(36) Zhang, A.; Zhuang, X.; Liu, J.; Huang, J.; Lin, L.; Tang, Y.; Zhao, S.; Li, R.; Wang, B.; Fang, B.; et al. Catalytic cycle of formate dehydrogenase captured by single-molecule conductance. *Nature Catalysis* **2023**, *6*, 266−275.

(37) Bamberger, N. D.; Dyer, D.; Parida, K. N.; McGrath, D. V.; Monti, O. L. A. Grid-Based Correlation Analysis to Identify Rare Quantum Transport Behaviors. *J. Phys. Chem. C* **2021**, *125*, 18297−18307.

(38) Bamberger, N. D.; Ivie, J. A.; Parida, K. N.; McGrath, D. V.; Monti, O. L. A. Unsupervised Segmentation-Based Machine Learning as an Advanced Analysis Tool for Single Molecule Break Junction Data. *J. Phys. Chem. C* **2020**, *124*, 18302−18315.

(39) Fu, T.; Frommer, K.; Nuckolls, C.; Venkataraman, L. Single-Molecule Junction Formation in Break-Junction Measurements. *J. Phys. Chem. Lett.* **2021**, *12*, 10802−10807.

(40) Xie, X.; Li, P.; Xu, Y.; Zhou, L.; Yan, Y.; Xie, L.; Jia, C.; Guo, X. Single-Molecule Junction: A Reliable Platform for Monitoring Molecular Physical and Chemical Processes. *ACS Nano* **2022**, *16*, 3476−3505.

(41) Harashima, T.; Egami, Y.; Homma, K.; Jono, Y.; Kaneko, S.; Fujii, S.; Ono, T.; Nishino, T. Unique Electrical Signature of Phosphate for Specific Single-Molecule Detection of Peptide Phosphorylation. *J. Am. Chem. Soc.* **2022**, *144*, 17449−17456.

(42) Dief, E. M.; Low, P. J.; Díez-Pérez, I.; Darwish, N. Advances in single-molecule junctions as tools for chemical and biochemical analysis. *Nat. Chem.* **2023**, *15*, 600−614.

(43) Wang, X.; Zhang, B.; Fowler, B.; Venkataraman, L.; Rovis, T. Alkane Solvent-Derived Acylation Reaction Driven by Electric Fields. *J. Am. Chem. Soc.* **2023**, *145*, 11903−11906.

(44) Tang, C.; Stuyver, T.; Lu, T.; Liu, J.; Ye, Y.; Gao, T.; Lin, L.; Zheng, J.; Liu, W.; Shi, J.; et al. Voltage-driven control of single-molecule keto-enol equilibrium in a two-terminal junction system. *Nat. Commun.* **2023**, *14*, 3657.

(45) Kareem, S.; Vali, S. R.; Reddy, B. V. S. Electric-Field-Induced Organic Transformations. *Eur. J. Org. Chem.* **2023**, *26*, No. e202300103.

(46) Schölkopf, B.; Platt, J. C.; Shawe-Taylor, J.; Smola, A. J.; Williamson, R. C. Estimating the Support of a High-Dimensional Distribution. *Neural Comput.* **2001**, *13*, 1443−1471.

(47) Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*, 4th ed.; Theodoridis, S., Koutroumbas, K., Eds.; Academic Press: Boston, 2009; pp 151−260.

(48) Reynolds, D. A. Gaussian mixture models. *Ency. Biom.* **2009**, *741*, 659−663.

(49) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(50) Warmerdam, V. D.; Brouns, M.; et al. *koaning/scikit-lego*, Revision 0.6.11; Zenodo, 2022 .

(51) Bro-Jørgensen, W.; Hamill, J. M.; Bro, R.; Solomon, G. C. Trusting our machines: validating machine learning models for single-molecule transport experiments. *Chem. Soc. Rev.* **2022**, *51*, 6875−6892.

(52) El Abbassi, M.; Overbeck, J.; Braun, O.; Calame, M.; van der Zant, H. S. J.; Perrin, M. L. Benchmark and application of unsupervised classification approaches for univariate data. *Commun. Phys.* **2021**, *4*, 50.

(53) van Veen, F. H.; Ornago, L.; van der Zant, H. S. J.; El Abbassi, M. Benchmark Study of Alkane Molecular Chains. *J. Phys. Chem. C* **2022**, *126*, 8801−8806.

(54) Cabosart, D.; El Abbassi, M.; Stefani, D.; Frisenda, R.; Calame, M.; van der Zant, H. S. J.; Perrin, M. L. A reference-free clustering method for the analysis of molecular break-junction measurements. *Appl. Phys. Lett.* **2019**, *114*, 143102.

(55) Lin, L.; Tang, C.; Dong, G.; Chen, Z.; Pan, Z.; Liu, J.; Yang, Y.; Shi, J.; Ji, R.; Hong, W. Spectral Clustering to Analyze the Hidden Events in Single-Molecule Break Junctions. *J. Phys. Chem. C* **2021**, *125*, 3623−3630.

(56) Makk, P.; Tomaszewski, D.; Martinek, J.; Balogh, Z.; Csonka, S.; Wawrzyniak, M.; Frei, M.; Venkataraman, L.; Halbritter, A. Correlation Analysis of Atomic and Single-Molecule Junction Conductance. *ACS Nano* **2012**, *6*, 3411−3423.

(57) Halbritter, A.; Makk, P.; Mackowiak, S.; Csonka, S.; Wawrzyniak, M.; Martinek, J. Regular Atomic Narrowing of Ni, Fe, and V Nanowires Resolved by Two-Dimensional Correlation Analysis. *Phys. Rev. Lett.* **2010**, *105*, 266805.

(58) Chen, H.; Brasiliense, V.; Mo, J.; Zhang, L.; Jiao, Y.; Chen, Z.; Jones, L. O.; He, G.; Guo, Q.-H.; Chen, X.-Y.; et al. Single-Molecule Charge Transport through Positively Charged Electrostatic Anchors. *J. Am. Chem. Soc.* **2021**, *143*, 2886−2895.

(59) Liu, B.; Murayama, S.; Komoto, Y.; Tsutsui, M.; Taniguchi, M. Dissecting Time-Evolved Conductance Behavior of Single Molecule Junctions by Nonparametric Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 6567−6572.

(60) Vladyka, A.; Albrecht, T. Unsupervised classification of single-molecule data with autoencoders and transfer learning. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 035013.

(61) Huang, F.; Li, R.; Wang, G.; Zheng, J.; Tang, Y.; Liu, J.; Yang, Y.; Yao, Y.; Shi, J.; Hong, W. Automatic classification of single-molecule charge transport data with an unsupervised machine-learning algorithm. *Phys. Chem. Chem. Phys.* **2020**, *22*, 1674−1681.

(62) Hamill, J. M.; Weaver, C.; Albrecht, T. Multivariate Approach to Single-Molecule Thermopower and Electrical Conductance Measurements. *J. Phys. Chem. C* **2021**, *125*, 26256−26262.

(63) Li, J.; Zhuang, Z.; Shen, P.; Song, S.; Tang, B. Z.; Zhao, Z. Achieving Multiple Quantum-Interfered States via Through-Space and Through-Bond Synergistic Effect in Foldamer-Based Single-Molecule Junctions. *J. Am. Chem. Soc.* **2022**, *144*, 8073−8083.

(64) Yuan, S.; Qian, Q.; Zhou, Y.; Zhao, S.; Lin, L.; Duan, P.; Xu, X.; Shi, J.; Xu, W.; Feng, A.; et al. Tracking Confined Reaction Based on Host−Guest Interaction Using Single-Molecule Conductance Measurement. *Small* **2022**, *18*, 2104554.

(65) Pimentel, M. A.; Clifton, D. A.; Clifton, L.; Tarassenko, L. A review of novelty detection. *Signal Process* **2014**, *99*, 215−249.

(66) Seliya, N.; Abdollah Zadeh, A.; Khoshgoftaar, T. M. A literature review on one-class classification and its potential applications in big data. *J. Big Data* **2021**, *8*, 122.

(67) Perera, P.; Oza, P.; Patel, V. M. One-Class Classification: A Survey, 2021. arXiv:2101.03064. https://doi.org/10.48550/arXiv.2101.03064 (accessed 11 May 2024).

(68) Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best practices in machine learning for chemistry. *Nat. Chem.* **2021**, *13*, 505−508.