Universal approach for unsupervised classification of univariate data

Maria El Abbassi¹, Jan Overbeck²,³,⁴, Oliver Braun²,³, Michel Calame²,³,⁴, Herre S.J. van der Zant¹ and Mickael L. Perrin²,∗

¹ Kavli Institute of Nanoscience, Delft University of Technology, Lorentzweg 1, 2628 CJ Delft, The Netherlands
² Empa, Swiss Federal Laboratories for Materials Science and Technology, Überlandstrasse 129, CH-8600 Dübendorf, Switzerland.
³ Department of Physics, University of Basel, Klingelbergstrasse 82, CH-4056 Basel, Switzerland.
⁴ Swiss Nanoscience Institute, University of Basel, Klingelbergstrasse 82, CH-4056 Basel, Switzerland.
∗ email: Mickael.Perrin@empa.ch.

Abstract: Unsupervised machine learning, and in particular data clustering, is a powerful approach for the analysis of datasets and identification of characteristic features occurring throughout a dataset. It is gaining popularity across scientific disciplines and is particularly useful for applications without a priori knowledge of the data structure. Here, we introduce a universal approach for unsupervised data classification relevant for any dataset consisting of a series of univariate measurements. It is therefore ideally suited for a wide range of measurement types. Here, we apply it to the field of nanoelectronics and spectroscopy to identify meaningful structures in data sets, providing physically relevant information about the system under study. An important step in our approach is the guidelines for the estimation of the optimum number of clusters. In addition, we have performed an extensive benchmark of ML approaches found in literature for the classification of molecular break junction traces. We find that several feature space construction methods we have introduced and clustering algorithms yield accuracies up to 20% higher than methods reported so far, increasing the Fowlkes-Mallows index from 0.77 up to 0.91.

INTRODUCTION

Machine learning (ML) and artificial intelligence (AI) are among the most significant recent technological advancements, with currently billions of dollars being invested in this emerging
technology\(^1\). In a few years, complex problems which had been around for decades, such as image\(^2\)
and facial recognition\(^3,4\), speech\(^5,6\) and text\(^7,8\) understanding, have been addressed. Machine learning promises to be a game-changer for major industries like health care\(^9\), pharmaceuticals\(^10\), information technology\(^11\), automotive\(^12\) and other industries relying on big data\(^13\). Its underlying strength is the excellence at recognizing patterns, either by relying on previous experience (supervised ML), or without any \textit{a priori} knowledge of the system itself (unsupervised ML). In both cases, ML relies on large amounts of data, which, in the last two decades, have become increasingly available due to the fast rise of cheap consumer electronics and the internet of things.

The same trend is also observed for scientific research, including the field of nanoscience, where tremendous progress has been made in the data acquisition\(^14–16\) and public databases have become available containing, for instance, a vast number of material structures and properties\(^17,18\). Inspiring examples of the use of the predictive power of supervised machine learning have, for instance, been realized in quantum chemistry for the prediction of the quantum mechanical wave function of electrons\(^19\) and in nanoelectronics for the tuning of quantum dots\(^20\), the identification of 2D material samples\(^21\), and the classification of breaking traces in atomic contacts\(^22\). Unsupervised machine learning methods, on the other hand, are intended for the investigation of the underlying structure of datasets without any \textit{a priori} knowledge of the system. Such approaches are ideally suited for the analysis of large experimental datasets and can help to significantly reduce the issue of conformation bias in the data analysis\(^23\). However, only few studies involving data clustering in nanoelectronics applications have been reported to date\(^24–28\).

In the study by Lemmer \textit{et al.}\(^24\), the univariate measurement data (conductance versus electrode displacement) is treated as an \(M\)-dimensional vector and compared to a reference vector for the feature space construction, after which the Gustafson-Kessel (GK) algorithm\(^29\) is employed for classification. A variation of this method was applied by El Abbassi \textit{et al.}\(^27\) to current-voltage characteristics. In a more recent study\(^26\), the need for this reference vector was eliminated by creating a 28x28 image of the each measurement trace. However, the high number of dimensions resulting from this approach is problematic for many clustering algorithms, as the data becomes sparse for increasing dimensionality (curse of dimensionality\(^30\)), thereby restricting the available clustering algorithms. Finally, the study by Huang \textit{et al.}\(^28\) is based on a deep auto-encoder for feature extraction from the raw data itself. In all above-mentioned studies, only a single feature space construction method and clustering algorithm were investigated, without a systematic benchmark of their accuracy against a large number of datasets of known classes and with varying
partitions. This makes it difficult to compare the performance of one method to another. In addition, no guidelines are provided for the estimation of the number of clusters, a critical step in data partitioning.

Here, we provide a universal workflow for the classification of univariate data sets. Our three-step approach consist of 1. the feature space construction, 2. the clustering algorithm, and 3. the internal validation to define the optimum number of clusters (NoC). We provide an extensive benchmark of a wide range of feature space construction methods (28) as well as clustering algorithms (16) using 900 datasets of simulated data with a number of classes varying between 2 and 10. We show that with the proper combination of step 1 and 2, high Fowlkes-Mallows classification indices up to 0.91 are obtained, significantly higher than other methods reported in literature\textsuperscript{24–28}, that all score below 0.77 in our benchmark. We then apply our workflow to several distinctively different measurement type (break-junction conductance traces, current-voltage characteristics, and Raman spectra), yielding extracted clusters that are distinctively different and reflect the physical properties of the system under study. Importantly, our approach does not require any \textit{a priori} knowledge of the system under study and therefore drastically reduces the confirmation bias that may be present in the analysis of large scientific datasets.

**RESULTS**

A schematic of the workflow for the unsupervised classification of univariate measurements is depicted in Fig. 1, starting from a dataset consisting of \( N \) univariate and discrete functions \( f(x_i), i \in [1,N] \). Each of these measurement curves is converted into an \( M \)-dimensional feature vector, resulting in an \( M \)-dimensional feature space containing \( N \) data points. After this step, a clustering algorithm is applied. As the number of classes is not known \textit{a priori}, this clustering step is repeated for a range of cluster numbers (in this illustration for 2-4 clusters). Here, we define a class as the ground truth distribution of each dataset, and a cluster the result of a clustering algorithm. Then, in order to determine the most suited NoC and assess the quality of the partitioning of the data, up to 29 internal cluster validation indices (CVI) are employed. Each CVI provides a prediction for the NoC, after which the optimal NoC is chosen based on a histogram of the predictions obtained from all CVIs. These CVIs are also used to estimate the optimal feature space method and clustering algorithm.
FIG. 1: Concept of the universal approach for univariate data classification. Any dataset in which the data depends on a single variable (for instance current $I$ vs. bias voltage $V$, conductance $G$ vs. electrode displacement $d$, force $F$ vs. displacement $d$, intensity $Int$ vs. energy $E$, etc...) can be converted into a feature vector. The feature space spanning the entire dataset is then split into clusters (represented using different colors) using a clustering algorithm. Finally, cluster validation indices (CVI) are used to estimate the optimum number of clusters (NoC).

**Benchmarking of algorithm performance on simulated dataset**

In the following, a large variety of feature space construction methods and clustering algorithms will be investigated and their performance benchmarked against artificially created datasets with known classes. The simulated datasets are conductance-displacement traces - also known as breaking traces - as commonly measured using the mechanically controllable break junction (MCBJ) technique for measuring the conductance of a molecule. For a detailed description of the construction of the simulated (labeled) data, we refer to section 1 of the Supporting Information.

In short, we generated 900 datasets, each consisting of 2000 breaking traces with known labels, with a varying number of clusters between 2 and 10 (100 x 2 clusters ... 100 x 10 clusters). The classes of the datasets were generated based on an experimental dataset consisting of conductance vs. distance curves recorded on the OPE3 molecule. This is in contrast to previous studies
where the benchmark data was purely synthetic\textsuperscript{24,28}. To account for possibly large variations in cluster population which may occur experimentally, the distribution of clusters is logarithmically distributed with the most probable cluster having 10 times more traces than the least occurring one. For example, for 2 clusters the distribution is 9.09\% and 90.91\%, for 3 clusters 6.10\%, 33.35\% and 60.55\%, etc.

We applied a variety of feature space construction processes and clustering algorithms to each of these 900 datasets. We investigated vector-based feature space construction methods based on a reference vector as described in Lemmer et al.\textsuperscript{24}, feature extraction from the raw data itself\textsuperscript{28}, and conversion to images (two-dimensional histogram)\textsuperscript{26}. In the latter case, inspired by the MNIST datasets\textsuperscript{33}, measurements are converted into images of 28x28 pixels. This has the advantage that all inputs for the feature space construction method have the same size, independent on the number of data points in each measurement. The number of pixels can be chosen to fine-tune the resolution for the feature extraction, independently from the number of data points in the measurements. We find that 28x28 is a good compromise between resolution and computational cost, although for certain types of data, a denser grid may be beneficial. In the following, the three different approached will be referred to as Lemmer, raw, and 28x28. The high number of dimensions for the raw and 28x28 case is known to lead to the \textit{curse of dimensionality}\textsuperscript{30}; the data becomes highly sparse and causes severe problems for many common clustering algorithms. To avoid this limitation, we have investigated a range of dimensionality reduction techniques, such as principal component analysis\textsuperscript{34} (PCA), kernel-PCA\textsuperscript{34}, multi-dimensional scaling\textsuperscript{34} (MDS), deep autoencoders\textsuperscript{34} (AE), Sammon mapping\textsuperscript{35}, stochastic neighbor embedding\textsuperscript{36} (SNE), t-distributed SNE\textsuperscript{37} and uniform manifold approximation and projection\textsuperscript{38} (UMAP). For the last two methods, three distance measure approaches were used (Euclidean, Chebyshev and cosine, abbreviated as Eucl., Cheb. and cos., respectively), bringing the total number of feature space construction methods to 28. For all methods containing dimensionality reduction, we used a reduction down to 3 dimensions. A description of each method is presented in section 2 of the Supporting Information. In section 3 of the Supplementary Information, we show that by increasing the dimensions for t-SNE (cos.) from 3 to 7 only a marginal gain in Fowlkes-Mallows index can be achieved for the five selected algorithms.

After each of the 900 datasets was run through the 28 feature space construction methods, 16 clustering algorithms were tested, covering a large spectrum of classification methods such as
FIG. 2: Benchmarking of various feature spaces and clustering algorithm on simulated data.  

a) Overview of the accuracy (Fowlkes-Mallows index) for all combinations of the various feature space construction methods and clustering algorithms. For this analysis, a total dataset consisting of 900 datasets of 2000 traces each is used with 2-10 clusters. The rows and columns of the table have been sorted by increasing average FM-index, with the best combination of feature spaces and algorithm in the lower right corner. 

b) Extracted clusters from an example dataset using the best performing feature space method 28x28 + t-SNE(cos.) and clustering method (GAL).

Distance minimization methods (k-means, k-medoids), fuzzy methods (fuzzy C-mean\(^{39}\) (FCM) and GK\(^{29}\)), self-organizing maps\(^{40}\) (SOM), hierarchical methods\(^{41}\) with various distance measures, expectation-maximization methods (Gaussian mixed model\(^{42}\) (GMM)), graph-based agglomerative methods (graph degree linkage\(^{43}\) (GDL) and graph average linkage\(^{44}\) (GAL)), spectral methods (Shi and Malik\(^{45}\) (S&M) and Jordan and Weiss\(^{46}\) (J&W)) and density-based methods (OPTICS\(^{47}\)). A description of each method can be found in section 4 of the Supporting Information. We note that we restricted ourselves to algorithms in which the number of clusters can be...
explicitly defined as input parameter. This step is needed further on to calculate the data partitioning for 2 to 9 clusters and determine the optimum number of clusters using clustering validation indices. This restriction excludes algorithms such as DBSCAN (density-based spatial clustering of applications with noise), hierarchical DBSCAN (HDBSCAN), and affinity propagation. We also note that many different image classification algorithms are available that can be run directly on the 28x28 image before dimensionality reduction, such as Deep Adaptive image Clustering (DAC), Associative Deep Clustering (ADC) and Invariant Information Clustering (IIC). Most of these algorithms, however, are based on neural networks and are significantly more expensive in terms of computational cost, thus limiting their applicability. The execution speeds of the various feature space and clustering methods applied here is presented in section 5 of the Supporting Information.

The accuracy of the classification is evaluated using the Fowlkes-Mallows (FM) index; it is an external cluster validation index (CVI) which scales between 0 and 1, where 1 represents the case of clusters perfectly reproducing the original classes. The FM index is defined as $FM = \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}$, where $TP$ is the number of true positives, $FP$ is the number of false positives, and $FN$ is the number of false negatives. The mean Fowlkes-Mallows indices for all combinations of feature space and clustering approach based on all 900 datasets are shown in Fig. 2b. We note that the NoC used for clustering is chosen to be the same number as the number of classes provided in the simulated dataset. The table is sorted by increasing average FM index per column and row, respectively, with the most accurate combination in the lower right corner. In this extensive benchmark, the least accurate algorithm is raw + SNE combined with FCM with a FM index of 0.47, while the most accurate one is the 28x28 + t-SNE(cos.) feature space, combined with the GAL algorithm. Based on our benchmark, this combination has a FM index of 0.91 and outperforms by at least 20% previously used methods to classify similar datasets in literature, which all show FM indices below 0.77.

The table also shows that both 28x28 + t-SNE and 28x28 + UMAP perform similarly well and provide a significant improvement in accuracy with respect to the other feature space methods investigated. In the following, we will therefore focus on these two feature space methods using the cosine distance measure.

In terms of the clustering algorithm, the table shows that the GAL algorithm yields the highest accuracy. This observation follows a previous study demonstrating that GAL outperforms many state-of-the-arts algorithms for image clustering and object matching. However, as this bench-
mark is performed on synthetic data, the performance of the algorithms may be different than on actual data. Therefore, we select the five best performing algorithms, namely GK, the most accurate of the spectral methods (J&W), GMM, the most accurate graph-based method (GAL), and OPTICS. In the remainder of this paper, we will restrict our use to these five methods.

From the fact that the row-to-row variation of FM indices, i.e., between feature space methods, is larger than the difference between columns (clustering methods), we conclude that the role of the feature space is more important than that of the algorithm. This can be rationalized, as a better feature space method will produce distinctively separated clusters, making it easier for the algorithm to find these clusters.

Finally, to ensure that the benchmark is not biased by the use of a logarithmically distributed cluster population, we produced the same table as shown in Fig. 2c but on datasets containing equal-size clusters (see section 6 of the Supporting Information). This benchmark yields very similar results in terms of best performing feature spaces and clustering algorithms, with overall slightly lower Fowlkes-Mallows indices.

**Application to MCBJ dataset**

We now apply our workflow to an experimental dataset of unknown classes and illustrate the different steps in Fig. 3. The starting point is an MCBJ dataset consisting of 10'000 traces recorded on the OPE3 molecule (see Fig. 3a for the 2D conductance-displacement histogram), to which we apply the two selected feature space methods + t-SNE (cos.) and + UMAP (cos.). Subsequently, these feature spaces are classified using the five selected clustering methods for a NoC ranging from 2 to 8. This gives a total of 5x2x7 = 70 different clustering distributions. For each of them, we calculate a wide range of internal cluster validation indices (CVI). These internal clustering validation indices are used to assess the compactness and separation of the clusters. Each index is calculated for a varying NoC, from which the optimum NoC can be estimated by different means (minimum, maximum, elbow, etc...). Here, we choose 29 CVI, including the well-known Silhouette index, Dunn and Davies-Bouldin index, that only require a maximization/minimization of the index. As such, the index can be used to compare different clustering methods, feature space, and NoCs, and determine the optimum combination. A complete list of all the indices and their implementation can be found in section 7 of the Supplementary Information.
FIG. 3: **Application of the workflow to measured MCBJ data.** a) Experimental 2D conductance-displacement histogram based on 10’000 MCBJ traces. b) Determination of the most suited clustering algorithm and number of clusters. c) Feature space constructed from the data of a) using the 28x28 + UMAP (cos.) method and clustered using the J&W algorithm for 5 clusters. d) 1D conductance and 2D conductance-displacement histogram for the cluster assignment in c).

The left panel of Fig. 3b presents the calculated values of the Davies-Bouldin index as a matrix, with as columns the NoC and as rows all combinations of feature space and the clustering algorithm. From this matrix, the maximum/minimum value of the index is obtained to determine the optimum NoC and method as determined by this particular CVI. We note that the use of CVIs to estimate the NoC is not straightforward as each of them has implicit assumptions, in particular on the distribution of the clusters. For this reason, we only consider NoC estimations that are unambiguous, in other words, a well-defined peak or dip in the cluster validation index. This means that we calculate the CVIs for 2 to 8 clusters, but we only take the CVI into account if the optimum NoC lies between 3 to 7 clusters. This procedure is repeated for all 29 CVIs and a 2D histogram is constructed (Fig. 3b right panel). Finally, this allows us to directly access the overall best feature space (28x28 + UMAP), algorithm (GAL) and NoC (5). The resulting feature space, with the individual breaking traces colored by cluster assignment, is plotted in Fig. 3c.

The resulting clusters are visualized as 2D conductance displacement histograms built from the individual breaking traces (see Fig. 3d). The plots show that the resulting 2D histograms exhibit distinctively different breaking behaviors, which can serve as a starting point for the investigation.
of the connections between the clusters resulting from this purely statistical procedure and the physical processes at play.

Based on our knowledge of these junctions, one can speculate that Cluster 1 corresponds to gold junctions breaking directly to below the noise floor, Cluster 2 to tunneling traces with some hints of molecular signatures, Cluster 3 to a fully stretched OPE3 molecule, Cluster 4 to tunneling traces without any molecular presence, and Cluster 5 to a two step breaking process involving molecule-electode interactions. The exact attribution of the various clusters, however, requires a detailed understanding of the microscopic picture of the molecular junction, possibly supported by ab-initio calculations, and is beyond the scope of this article.

The approach we employ here provides a significant advantage over the clustering algorithm employed in our previous work on the same dataset. First, the number of clusters is not arbitrarily chosen but originates from a statistical basis. Second, we find two clusters that have not been observed previously, Cluster 1 and Cluster 4. In particular Cluster 1 is of interest as the breaking traces in that cluster are distinctively different from the other clusters, but with a small population of less than 4%. Nevertheless, the approach is able to extract it.

To illustrate the versatility of our approach for different measurements types, we now proceed with the classification of two more datasets: the first one consists of 67 current-voltage (I(V)) characteristics, while the second one has 4900 Raman spectra. For the IV classification, we note that the OPTICS algorithm was excluded as it fails using the default parameters due to the limited amount of measurements.

**Application to current-voltage characteristics**

To illustrate the versatility of our approach for different measurements types, we now proceed with the classification of two more datasets: the first one consists of 67 current-voltage (I(V)) characteristics, while the second one has 4900 Raman spectra. For the IV classification, we note that the OPTICS algorithm was excluded as it fails using the default parameters due to the limited amount of measurements.

Figure 4a presents a 2D current-voltage histogram of the 67 IVs recorded on a diH-AC molecule. All IVs have been normalized to focus on the shape of the curves, not on the absolute values in current. The same procedure is repeated as described previously and the best feature space and clustering algorithm is determined to be 28x28 + UMAP(cos.) and GAL, respectively.
FIG. 4: **Application of the method on IV characteristics.** a) Experimental 2D current-voltage histogram based on 67 IVs recorded on a diH-AC molecule for the entire dataset and the different clusters b) Determination of the most suited clustering algorithm and number of clusters. c) Feature space constructed using the 28x28 + UMAP (cos.) method and clustered using the GAL algorithm for 5 clusters.

for an optimal number of clusters of 5. The corresponding feature space is presented in Fig. 4b, colored according to the clusters produced by the GAL algorithm. The 2D current-voltage histograms of the five resulting clusters are shown in Fig. 4a. Cluster 1 shows perfectly linear IVs, while cluster 2 shows a pronounced negative differential conductance (NDC) feature, with first a linear slope around zero bias, a sharp peak around 30 mV, followed by a rapid decrease of the current for increasing bias voltage. Cluster 3 contains mostly IV’s with a gap around zero bias. Cluster 4 exhibits NDC as well, but with a more rounded peak compared to cluster 2, and a more gentle decrease in current. Cluster 5 shows close-to-linear IVs with some deviations from the perfect line.
Application to Raman spectra

As a final application, we investigate the classification of Raman spectra recorded on a well-studied reference system, namely graphene irradiated by helium ions. The effect of He-induced defects on the Raman spectrum of graphene is known from literature\textsuperscript{60,61}, but for our analysis we explicitly do not rely on any a-priori knowledge of the system, \textit{i.e.}, we do not need to know beforehand which Raman bands will be altered by the irradiation and by what spatial pattern of the graphene has been irradiated. Instead, we use our clustering approach to identify the different types of Raman spectra present in the sample from which we infer the spatial distribution of He-irradiation doses and their effect on the graphene spectrum. The sample under study consists of a free-standing graphene membrane (6 $\mu$m diameter), suspended over a silicon nitride frame coated with 5 nm/40 nm of Ti/Au. The membrane has been divided in four quadrants, each exposed with a different dose of helium ions. An illustration of the sample layout is presented in Fig. 5a. On this sample, a two-dimensional map containing 70x70 spectra was acquired using a confocal WITec alpha300 R Raman microscope with a 532 nm excitation laser. A description of the sample preparation and Raman measurements is provided in Section 8 of the Supporting Information.

The Raman spectra were fed to the 28x28 + UMAP (cos.) feature space construction method and split in 7 clusters using the GAL algorithm (see section 8 of the Supporting Information for more details). Figure 5b presents the partitioned feature space, containing several well-separated clusters. From this partitioning, we construct the two-dimensional map of the clusters to investigate their spatial distribution (see Fig. 5c). The plot shows that the extracted clusters match well the physical topology of the sample: the Clusters 1-4 are located on the suspended graphene membrane, reproducing the four quadrants. Clusters 5-7 form concentric rings located at the edge of the boundary between the SiN/Ti/Au support and the hole and on the support itself. Figure 5d shows the average spectrum obtained per cluster. Cluster 1 shows a flat background, with pronounced peaks at 1585 cm$^{-1}$ and 2670 cm$^{-1}$. For Clusters 2 to 4 (corresponding to increasing He-dose), a peak at 1340 cm$^{-1}$ appears with steadily increasing intensity while the intensity of the peak at 2670 cm$^{-1}$, on the other hand, decreases. Cluster 5, located at the edge of the support possess all three above-mentioned peaks, while for Clusters 6 and 7, a broad fluorescence background originating from the gold is present and all graphene-related peaks drastically decrease in prominence.
Interestingly, the four quadrants have only been identified as distinct clusters on the suspended part, but not on the substrate. This implies that the clustering algorithm identifies spectral changes upon irradiation as characteristic features for the freely suspended material, whereas the additional fluorescence background from the gold is a more characteristic attribute of the supported material than the variation between quadrants. Nevertheless, when inspecting Cluster 6 and 7, some substructure is still visible, and performing a clustering on that subset may reveal additional structure.

The three observed peaks correspond to the well-known D-, G- and 2D-peak, and follow the behavior expected for progressive damage to graphene by He-irradiation\textsuperscript{60,61}. We would like to stress that our approach allowed to extract the increase of the D-peak and the decrease of the 2D-peak when introducing defects in graphene, without any before-hand knowledge of the system: neither the type of Raman spectra under consideration, nor where on the sample the He-irradiation occurred.

![Figure 5: Application of the method on Raman spectra. a) Sample layout: suspended graphene membrane irradiated with four different He-ion doses. b) Partitioned feature space. c) Spatial map of the extracted clusters. d) Average Raman spectrum of each cluster.](image)
DISCUSSION

In the synthetic data, the t-SNE and UMAP algorithms score equally well in reducing each measurement from a 784 dimensional space (28x28) down to the 3 dimensional feature space. On the experimental datasets, however, UMAP tends to perform better. This difference emphasizes the need for labelled data which resembles as closely as possible the experimental data, as synthetic data may not capture all the experimental complexity. We note that UMAP has become the new state-of-the-art method for dimensionality reduction, surpassing t-SNE in several applications\textsuperscript{62,63}. While t-SNE reproduces well the local structure of the data, UMAP reproduces both the local and large-scale structure\textsuperscript{38}. Moreover, one could also investigate more advanced variants of UMAP\textsuperscript{64} that could lead to even higher FM indices. Along the same lines, the use of more sophisticated clustering algorithms involving convolutional neural networks that can directly be applied to the 28x28 image merit additional research as some of them have proven to be highly accurate on the MNIST and other databases\textsuperscript{53}, despite their high computational cost. We would like to stress that the large collection of internal clustering validation indices we employ here only provide an estimate of the optimum number of clusters. To the best of our knowledge, no CVI exists that performs well in all situations. In particular clusters of largely varying densities are challenging as well as clusters of arbitrary shape. Therefore, the CVI should be used mere as a guideline.

CONCLUSION

In conclusion, we have introduced an optimized three-step workflow for the classification of univariate measurement data. The first two steps (feature space construction and partition algorithm) are based on an extensive benchmarked of a wide range of novel and existing methods using 900 simulated datasets with known classes synthesized from experimental break junction traces. By doing so, we have identified specific combinations of feature space construction and partition algorithm yielding high accuracies, with Fowlkes-Mallows indices up to 0.91. The third step in our method provides guidelines for the estimation of the optimal number of clusters using a wide range of cluster validation indices. We show that our approach can readily be applied to various types of measurements such as MCBJ conductance-breaking traces, IV curves and Raman spectra, thereby splitting the dataset into statically relevant clusters reflecting the physical properties of the measured samples.
Materials and Methods

All codes used in this study are freely available online at https://github.com/MickaelPerrin74/DataClustering. This repository contains the datasets and codes used for the benchmark, as well as a graphical user interface for the analysis of custom datasets.

Acknowledgements

The authors would like to thank Dr. Ivan Shorubalko (Empa) for the help in developing the graphene membranes technology and for the He-FIB exposures (supported by Swiss National Science Foundation REquip 206021-133823). The authors would also like to thank Dr. Riccardo Frisenda for sharing with us the dataset recorded on OPE326.

Funding

M.P. acknowledges funding by the EMPAPOSTDOCS-II program which is financed by the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement number 754364. This work was in part supported by the FET open project QuIET (no. 767187).

Author contributions

M.P. performed the machine learning analysis, with input from all authors. J.O. and O.B. performed the Raman measurements. All authors discussed the data and wrote the manuscript. M.P. supervised the study.

1 Worldwide Spending on Artificial Intelligence Systems Will Be Nearly $98 Billion in 2023, According to New IDC Spending Guide. URL https://www.idc.com/getdoc.jsp?containerId=prUS45481219.

2 Schmidhuber, J. Deep Learning in neural networks: An overview. Neural Networks 61, 85–117 (2015).
3 Sun, Y., Wang, X. & Tang, X. Deep learning face representation from predicting 10,000 classes. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1891–1898 (IEEE Computer Society, 2014).

4 Liu, Z., Luo, P., Wang, X. & Tang, X. Deep learning face attributes in the wild. In 2015 IEEE International Conference on Computer Vision (ICCV), 3730–3738 (2015).

5 Mikolov, T., Karafit, M., Burget, L., Cernock, J. & Khudanpur, S. Recurrent neural network based language model. In Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, vol. 2, 1045–1048 (2010).

6 Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29, 82–97 (2012).

7 Zhang, X., Zhao, J. & LeCun, Y. Character-level convolutional networks for text classification. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, 649–657 (Curran Associates, Inc., 2015).

8 Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571, 95–98 (2019).

9 Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal 13, 8–17 (2015).

10 Vamathevan, J. et al. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery 18, 463–477 (2019).

11 Hutto, C. & Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media 216, 18 (2014).

12 Bojarski, M. et al. End-to-End Learning for Self-Driving Cars (2016). 1604.07316.

13 Chen, X. W. & Lin, X. Big data deep learning: Challenges and perspectives. IEEE Access 2, 514–525 (2014).

14 Graf, D. et al. Spatially resolved Raman spectroscopy of single- and few-layer graphene. Nano Letters 7, 238–242 (2007). 0607562.

15 El Abbassi, M. et al. Unravelling the conductance path through single-porphyrin junctions. Chemical Science 10, 8299–8305 (2019).

16 Brown, K. A., Britman, S., Maccaferri, N., Jariwala, D. & Celano, U. Machine Learning in
Nanoscience: Big Data at Small Scales. *Nano Letters* **20**, 2–10 (2020).

17 Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nature Materials* **12**, 191–201 (2013).

18 Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Computational Materials Science* **111**, 218–230 (2016). 1504.01163.

19 Schütt, K. T., Gastegger, M., Tkatchenko, A., Müller, K. R. & Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications* **10** (2019). 1906.10033.

20 Lennon, D. T. *et al.* Efficiently measuring a quantum device using machine learning. *npj Quantum Information* **5** (2019). 1810.10042.

21 Masubuchi, S. *et al.* Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials. *npj 2D Materials and Applications* **4**, 3 (2020).

22 Lauritzen, K. P. *et al.* Perspective: Theory of quantum transport in molecular junctions. *The Journal of Chemical Physics* **148**, 84111 (2018).

23 Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Medicine* **2**, e124 (2005).

24 Lemmer, M., Inkpen, M. S., Kornysheva, K., Long, N. J. & Albrecht, T. Unsupervised vector-based classification of single-molecule charge transport data. *Nature Communications* **7** (2016).

25 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. *Physical Review Letters* **120** (2018). 1705.06161.

26 Cabosart, D. *et al.* A reference-free clustering method for the analysis of molecular break-junction measurements. *Applied Physics Letters* **114** (2019).

27 El Abbassi, M. *et al.* Robust graphene-based molecular devices. *Nature Nanotechnology* **14**, 957–961 (2019).

28 Huang, F. *et al.* Automatic classification of single-molecule charge transport data with an unsupervised machine-learning algorithm. *Physical Chemistry Chemical Physics* (2019).

29 Gustafson, D. E. & Kessel, W. C. Fuzzy Clustering with a Fuzzy Covariance Matrix. In *Proceedings of the IEEE Conference on Decision and Control*, 761–766 (IEEE, 1978).

30 Bellman, R. *Dynamic programming* (Princeton University Press, 2010).

31 Xu, B. Q. & Tao, N. J. Measurement of Single-Molecule Resistance by Repeated Formation of Molec-
ular Junctions. *Science* **301**, 1221–1223 (2003).

32 Frisenda, R., Stefani, D. & van der Zant, H. S. J. Quantum Transport through a Single Conjugated Rigid Molecule, a Mechanical Break Junction Study. *Acc. Chem. Res.* **51**, 1359–1367 (2018).

33 MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges. URL http://yann.lecun.com/exdb/mnist/.

34 Van Der Maaten, L., Postma, E. & Van den Herik, J. Dimensionality reduction: a comparative review. *J Mach Learn Res* **10**, 66–71 (2009).

35 Sammon, J. W. A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers* **C-18**, 401–409 (1969).

36 Hinton, G. E. & Roweis, S. T. Stochastic neighbor embedding. In *Advances in neural information processing systems*, 857–864 (2003).

37 Van Der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

38 McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861 (2018).

39 Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms* (Kluwer Academic Publishers, USA, 1981).

40 Kohonen, T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **43**, 59–69 (1982).

41 Silla, C. N. & Freitas, A. A. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery volume* **22**, 31–72 (2011).

42 Williams, C. K. I. & Rasmussen, C. E. Gaussian processes for regression. In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, NIPS95, 514–520 (MIT Press, Cambridge, MA, USA, 1995).

43 Zhang, W., Wang, X., Zhao, D. & Tang, X. Graph degree linkage: Agglomerative clustering on a directed graph. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7572 LNCS, 428–441 (2012).

44 Zhang, W., Zhao, D. & Wang, X. Agglomerative clustering via maximum incremental path integral. *Pattern Recognition* **46**, 3056–3065 (2013).

45 Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 888–905 (2000).
Ng, A. Y., Jordan, M. I. & Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS01, 849–856 (MIT Press, Cambridge, MA, USA, 2001).

Ankerst, M., Breunig, M. M., Peter Kriegel, H. & Sander, J. Optics: Ordering points to identify the clustering structure. In *Proceedings ACM SIGMOD International Conference on Management of Data*, 49–60 (ACM Press, 1999).

Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231 (AAAI Press, 1996).

Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. *Advances in Knowledge Discovery and Data Mining* 160–172 (2013).

Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).

Chang, J., Wang, L., Meng, G., Xiang, S. & Pan, C. Deep Adaptive Image Clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, 5880–5888 (Institute of Electrical and Electronics Engineers Inc., 2017).

Haeusser, P., Plapp, J., Golkov, V., Aljalbout, E. & Cremers, D. Associative Deep Clustering: Training a Classification Network with No Labels. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11269 LNCS, 18–32 (Springer Verlag, 2019).

Ji, X., Henriques, J. F. & Vedaldi, A. Invariant Information Clustering for Unsupervised Image Classification and Segmentation (2018). 1807.06653.

Fowlkes, E. B. & Mallows, C. L. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* **78**, 553–569 (1983).

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M. & Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognition* **46**, 243–256 (2013).

Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. Nbclust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software* **61**, 1–36 (2014).

Hämäläinen, J., Jauhiainen, S. & Kärkkäinen, T. Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering. *Algorithms* **10**, 105 (2017).

Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. Nbclust: An r package for determining the relevant
number of clusters in a data set. *Journal of Statistical Software, Articles* **61**, 1–36 (2014).

59 Perrin, M. L. *et al.* Large negative differential conductance in single-molecule break junctions. *Nature Nanotechnology* **9**, 830–834 (2014).

60 Buchheim, J., Wyss, R. M., Shorubalko, I. & Park, H. G. Understanding the interaction between energetic ions and freestanding graphene towards practical 2D perforation. *Nanoscale* **8**, 8345–8354 (2016).

61 Shorubalko, I., Pillatsch, L. & Utke, I. Directwrite milling and deposition with noble gases. In *Helium Ion Microscopy*, 355–393 (Springer Verlag, 2016).

62 Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology* **37**, 38–47 (2019).

63 Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C. & Gravel, S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics* **15**, e1008432 (2019).

64 McConville, R., Santos-Rodriguez, R., Piechocki, R. J. & Craddock, I. N2D: (Not Too) Deep Clustering via Clustering the Local Manifold of an Autoencoded Embedding (2019). 1908.05968.