Benchmark and application of unsupervised classification approaches for univariate data

El Abbassi, Maria; Overbeck, Jan; Braun, Oliver; Calame, Michel; van der Zant, Herre S.J.; Perrin, Mickael L.

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

ARTICLE

Benchmark and application of unsupervised classification approaches for univariate data

Maria El Abbassi [1], Jan Overbeck [2,3,4], Oliver Braun [2,3], Michel Calame [2,3,4], Herre S. J. van der Zant [1] & Mickael L. Perrin [2✉]

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 an approach for unsupervised data classification of any dataset consisting of a series of univariate measurements. It is therefore ideally suited for a wide range of measurement types. We apply it to the field of nanoelectronics and spectroscopy to identify meaningful structures in data sets. We also provide guidelines for the estimation of the optimum number of clusters. In addition, we have performed an extensive benchmark of novel and existing machine learning approaches and observe significant performance differences. Careful selection of the feature space construction method and clustering algorithms for a specific measurement type can therefore greatly improve classification accuracies.

[1] Kavli Institute of Nanoscience, Delft University of Technology, Delft, The Netherlands. [2] Empa, Swiss Federal Laboratories for Materials Science and Technology, Dübendorf, Switzerland. [3] Department of Physics, University of Basel, Basel, Switzerland. [4] Swiss Nanoscience Institute, University of Basel, Basel, Switzerland. ✉email: Mickael.Perrin@empa.ch

Machine learning (ML) and artificial intelligence 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. ML 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 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 ML 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 ML methods, on the other hand, are intended for the investigation of the underlying structure of datasets without any 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].

Several studies involving data clustering in nanoelectronics applications have been reported to date[24–31]. In the study by Lemmer 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[32] is employed for classification. A variation of this method was applied by El Abbassi et al.[28] to current–voltage characteristics. In a more recent study[27], the need for this reference vector was eliminated by creating a 28×28 image of 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[33]), thereby restricting the available clustering algorithms. Several approaches have been proposed to reduce the number of dimensions, such as deep auto-encoder for feature extraction from the raw data itself[29], or the use of the approximately linear sections of the breaking traces[31]. Characteristic of the previous studies, however, is the fact that the clustering is performed on a feature space constructed from the individual breaking traces, an approach that can become computationally prohibitive in case large datasets are acquired. An appealing alternative has been introduced by Wu et al.[25], in which the clustering algorithms is run on the 2D conductance-displacement histogram.

In all the 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, few studies[25,31] provide guidelines for the estimation of the number of clusters (NoC), a critical step in data partitioning.

Here, we provide a workflow for the classification of univariate data sets. Our three-step approach consists of: (1) the feature space construction, (2) the clustering algorithm, and (3) the internal validation to define the optimum NoC. In the first part of the article, we benchmark a wide range of 28 feature space construction methods as well as 16 clustering algorithms using 900 datasets of simulated breaking traces with a number of classes varying between 2 and 10. In this benchmark, we identify the top five best performing clustering algorithms and top two feature spaces. We then apply our workflow to several distinctively different measurement types (break-junction conductance traces, current–voltage characteristics, and Raman spectra), yielding extracted clusters that are distinctively different. Importantly, our approach does not require any a priori knowledge of the system under study and therefore reduces the confirmation bias that may be present in the analysis of large scientific datasets. The attribution of the various clusters to the physical phenomena dictating their behavior, however, requires a detailed understanding of the microscopic picture of the system under study and is beyond the scope of this article.

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 measurement curve is converted into an M-dimensional feature vector, resulting in a feature space containing $M \times N$ data points. After this step, a clustering algorithm is applied. As the number of classes is not known 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 (CVIs) are employed. Each CVI provides a prediction for the NoC, after which the optimal NoC is estimated based on a histogram of the predictions obtained from all CVIs. These CVIs are also used to determine the optimal feature space method and clustering algorithm.

Benchmarking of algorithm performance on simulated mechanically controllable break-junction (MCBJ) datasets. In the following, a large variety of feature space construction methods and clustering algorithms are investigated and their performance is benchmarked against artificially created datasets with known classes. The aim of a benchmark is to rank the various algorithms according to their performance for a given set of parameters. Here, all algorithms were executed using their default parameters, both in the benchmark, as well as when applied to experimental datasets. The simulated datasets are conductance-displacement traces—also known as breaking traces—as commonly measured using the MCBJ technique and scanning tunneling microscope (STM) for measuring the conductance of a molecule[34]. For a detailed description of the construction of the simulated (labeled) data, we refer to Supplementary Method 1.

In short, we generated 900 datasets, each consisting of 2000 breaking traces with known labels, with a varying number of classes between 2 and 10 (100×2 classes ... 100×10 classes). The traces were generated based on an experimental dataset consisting of conductance vs. distance curves recorded on OPE3 molecules[35,36]. This is in contrast to previous studies where the benchmark data was purely synthetic[24,29]. To account for possibly large variations in cluster population which may occur experimentally, the distribution of classes is logarithmically
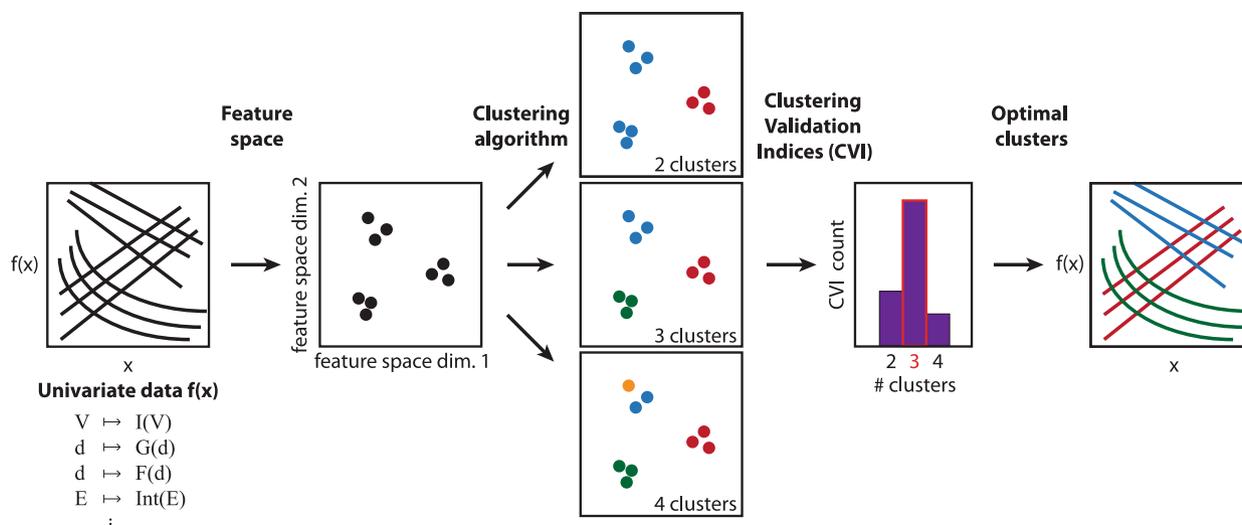
Fig. 1 Concept of our approach for univariate data classification. Any dataset in which the data depends on a single variable (for instance current *I* vs. bias voltage *V*, conductance *G* vs. electrode displacement *d*, force *F* vs. displacement *d*, intensity *Int* vs. energy *E*, etc.) can be converted 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 (CVIs) are used to estimate the optimal number of clusters (NoC).

distributed with the most probable class having 10 times more traces than the least occurring one. For example, for two classes the distribution is 9.09% and 90.91%, for three classes the distribution is 6.10%, 33.35%, 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.[24], feature extraction from the raw data itself[29], and conversion to images (two-dimensional histogram)[27]. In the latter case, inspired by the MNIST datasets[37], measurements are converted into images of 28 × 28 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. Here, we would like to stress that 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. In Supplementary Note 1, we show that 28 × 28, inspired by the MNIST database, is a good compromise between accuracy and computational cost. This choice implies that the distinction between features occurring below the bin size (0.25 orders of magnitude in conductance and 0.1 nm in distance) is limited as it relies only on the counts within the bin itself. To illustrate this, for fixed acquisition rate, a slanted plateau can be separated from a horizontal plateau as both would yield different counts in a particular bin. For a distinction between more elaborate shapes a denser grid would be beneficial. However, the use of more bins comes at higher computational costs and may lead to high-dimensional sparse data, which in turn is challenging to cluster, even after dimensionality reduction.

In the following, the three different approaches will be referred to as 'Lemmer', 'raw', and '28 × 28'. The high number of dimensions for the raw and 28 × 28 case is known to lead to the curse of dimensionality[33]; 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[38] (PCA), kernel-PCA[38], multi-dimensional scaling[38] (MDS), deep autoencoders[38] (AE), Sammon mapping[39], stochastic neighbor embedding[40] (SNE), t-distributed SNE[41] and uniform manifold approximation and projection[42] (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 Supplementary Method 2. In Supplementary Note 2, we show that by increasing the dimensions for t-SNE (cos.) from 3 to 7 only a marginal gain in Fowlkes–Mallows (FM) 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 distance minimization methods (k-means, k-medoids), fuzzy methods (fuzzy C-mean[43] (FCM) and GK[32]), self-organizing maps[44] (SOM), hierarchical methods[45] with various distance measures, expectation-maximization methods (Gaussian mixed model[46] (GMM)), graph-based agglomerative methods (graph degree linkage[47] (GDL) and graph average linkage[48] (GAL)), spectral methods (Shi and Malik[49] (S&M) and Jordan and Weiss[50] (J&W)) and density-based methods (Ordering Points To Identify the Clustering Structure (OPTICS[51])). A description of each method can be found in Supplementary Method 3. We note that we restricted ourselves to algorithms in which the NoC can be explicitly defined as input parameter. This step is needed further on to calculate the data partitioning for 2–9 clusters and determine the optimum NoC using clustering validation indices. This restriction excludes algorithms such as density-based spatial clustering of applications with noise[52] (DBSCAN), hierarchical DBSCAN (HDBSCAN[53]), and affinity propagation[54]. We also note that many different image classification algorithms are available that can be run directly on the 28 × 28 image before dimensionality reduction, such as deep adaptive image clustering[55] (DAC), associative deep clustering[56] (ADC), and invariant information clustering[57] (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 Supplementary Note 3.
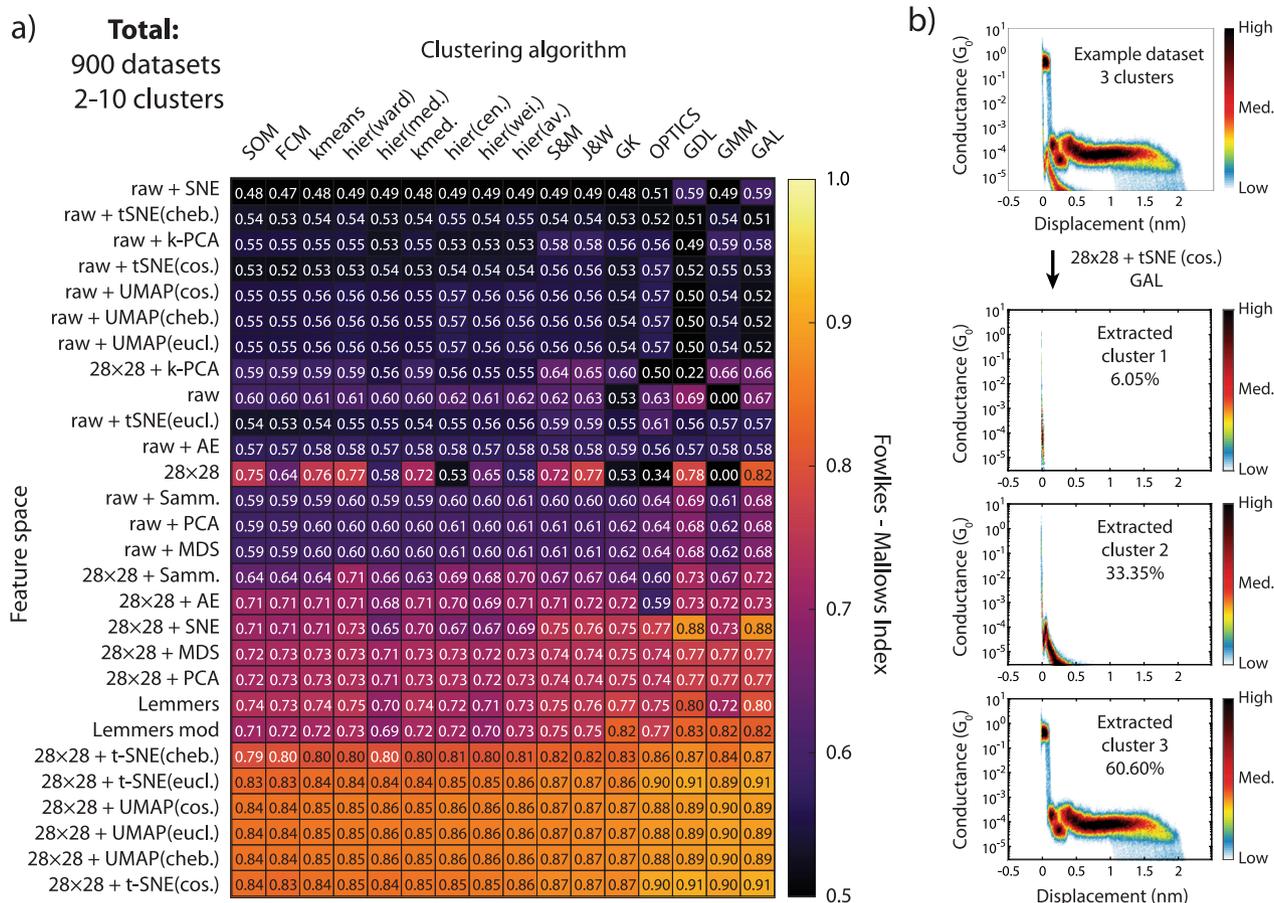
a) Total: 900 datasets 2-10 clusters

Clustering algorithm

Feature space

Feature space	SOM	FCM	kmeans	hier(ward)	hier(med.)	kmed.	hier(cen.)	hier(wei.)	hier(av.)	S&M	J&W	GK	OPTICS	GDL	GMM	GAL
raw + SNE	0.48	0.47	0.48	0.49	0.49	0.48	0.49	0.49	0.49	0.49	0.49	0.48	0.51	0.59	0.49	0.59
raw + tSNE(cheb.)	0.54	0.53	0.54	0.54	0.53	0.54	0.55	0.54	0.55	0.54	0.54	0.53	0.52	0.51	0.54	0.51
raw + k-PCA	0.55	0.55	0.55	0.55	0.53	0.55	0.53	0.53	0.53	0.58	0.58	0.56	0.56	0.49	0.59	0.58
raw + tSNE(cos.)	0.53	0.52	0.53	0.53	0.54	0.53	0.54	0.54	0.54	0.56	0.56	0.53	0.57	0.52	0.55	0.52
raw + UMAP(cos.)	0.55	0.55	0.56	0.56	0.56	0.55	0.57	0.56	0.56	0.56	0.56	0.54	0.57	0.50	0.54	0.52
raw + UMAP(cheb.)	0.55	0.55	0.56	0.56	0.56	0.55	0.57	0.56	0.56	0.56	0.56	0.54	0.57	0.50	0.54	0.52
raw + UMAP(eucl.)	0.55	0.55	0.56	0.56	0.56	0.55	0.57	0.56	0.56	0.56	0.56	0.54	0.57	0.50	0.54	0.52
28×28 + k-PCA	0.59	0.59	0.59	0.59	0.56	0.59	0.56	0.55	0.55	0.64	0.65	0.60	0.50	0.22	0.66	0.66
raw	0.60	0.60	0.61	0.61	0.60	0.60	0.62	0.61	0.62	0.62	0.63	0.53	0.63	0.69	0.00	0.67
raw + tSNE(eucl.)	0.54	0.53	0.54	0.55	0.55	0.54	0.55	0.56	0.56	0.59	0.59	0.55	0.61	0.56	0.57	0.57
raw + AE	0.57	0.57	0.58	0.58	0.57	0.58	0.58	0.57	0.58	0.58	0.58	0.59	0.56	0.57	0.58	0.58
28×28	0.75	0.64	0.76	0.77	0.58	0.72	0.53	0.65	0.58	0.72	0.77	0.53	0.34	0.78	0.00	0.82
raw + Samm.	0.59	0.59	0.59	0.60	0.59	0.59	0.60	0.60	0.61	0.60	0.60	0.60	0.64	0.69	0.61	0.68
raw + PCA	0.59	0.59	0.60	0.60	0.60	0.60	0.61	0.60	0.61	0.61	0.61	0.62	0.64	0.68	0.62	0.68
raw + MDS	0.59	0.59	0.60	0.60	0.60	0.60	0.61	0.60	0.61	0.61	0.61	0.62	0.64	0.68	0.62	0.68
28×28 + Samm.	0.64	0.64	0.64	0.71	0.66	0.63	0.69	0.68	0.70	0.67	0.67	0.64	0.60	0.73	0.67	0.72
28×28 + AE	0.71	0.71	0.71	0.71	0.68	0.71	0.70	0.69	0.71	0.71	0.72	0.72	0.59	0.73	0.72	0.73
28×28 + SNE	0.71	0.71	0.71	0.73	0.65	0.70	0.67	0.67	0.69	0.75	0.76	0.75	0.77	0.88	0.73	0.88
28×28 + MDS	0.72	0.73	0.73	0.73	0.71	0.73	0.73	0.72	0.73	0.74	0.74	0.75	0.74	0.77	0.77	0.77
28×28 + PCA	0.72	0.73	0.73	0.73	0.71	0.73	0.73	0.72	0.73	0.74	0.74	0.75	0.74	0.77	0.77	0.77
Lemmers	0.74	0.73	0.74	0.75	0.70	0.74	0.72	0.71	0.73	0.75	0.76	0.77	0.75	0.80	0.72	0.80
Lemmers mod	0.71	0.72	0.72	0.73	0.69	0.72	0.72	0.70	0.73	0.75	0.75	0.82	0.77	0.83	0.82	0.82
28×28 + t-SNE(cheb.)	0.79	0.80	0.80	0.80	0.80	0.80	0.81	0.80	0.81	0.82	0.82	0.83	0.86	0.87	0.84	0.87
28×28 + t-SNE(eucl.)	0.83	0.83	0.84	0.84	0.84	0.84	0.85	0.85	0.86	0.87	0.87	0.86	0.90	0.91	0.89	0.91
28×28 + UMAP(cos.)	0.84	0.84	0.85	0.85	0.86	0.85	0.86	0.86	0.86	0.87	0.87	0.87	0.88	0.89	0.90	0.89
28×28 + UMAP(eucl.)	0.84	0.84	0.85	0.85	0.86	0.85	0.86	0.86	0.86	0.87	0.87	0.87	0.88	0.89	0.90	0.89
28×28 + UMAP(cheb.)	0.84	0.84	0.85	0.85	0.86	0.85	0.86	0.86	0.86	0.87	0.87	0.87	0.88	0.89	0.90	0.89
28×28 + t-SNE(cos.)	0.84	0.83	0.84	0.85	0.84	0.85	0.85	0.85	0.86	0.87	0.87	0.87	0.90	0.91	0.90	0.91

Fowlkes - Mallows Index (colour scale 0.5 – 1.0)

b) Example dataset, 3 clusters — 2D conductance–displacement histogram (Conductance (G_0) vs Displacement (nm)).

28×28 + tSNE (cos.) GAL

- Extracted cluster 1, 6.05%
- Extracted cluster 2, 33.35%
- Extracted cluster 3, 60.60%

Fig. 2 Benchmarking of various feature spaces and clustering algorithm on simulated mechanically controllable break-junction data. a Overview of the accuracy, expressed as Folwkes–Mallows (FM) index, for all combinations of the various feature space construction methods and clustering algorithms. For this analysis, the average FM is shown based on 900 datasets of 2000 traces each, with 2–10 classes. The rows and columns of the heatmap have been sorted by increasing average FM-index, with the best combination of feature spaces and algorithm in the lower right corner. **b** 2D conductance-displacement histogram for an example dataset, including the 2D conductance-displacement histograms obtained by clustering using the best performing feature space method 28×28 + t-distributed stochastic neighbor embedding (t-SNE) using a cosine distance (cos.) and the graph average linkage (GAL) clustering method.

The accuracy of the classification is evaluated using the FM index[58]; it is an external cluster validation index (CVI) which varies 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 FM indices for all combinations of feature space and clustering approach based on all 900 datasets are shown in Fig. 2a, presented as heatmap. Figure 2b presents an example dataset that has been clustered using t-SNE (cos.) and GAL. 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 heatmap 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 28×28 + t-SNE (cos.) feature space, combined with the GAL algorithm. Based on the benchmark performed on this dataset, this optimal combination feature space and clustering algorithm exhibits a FM index of 0.91 and outperforms previously used methods to classify similar datasets in literature[24,27,29].

The heatmap also shows that both 28×28 + t-SNE and 28×28 + 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 heatmap 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[47].

To ensure that the benchmark is not biased by the use of a logarithmically distributed class population, we produced the same heatmap as shown in Fig. 2a but on datasets containing equal-size classes (see Supplementary Note 4). This benchmark yields very similar results in terms of best performing feature spaces and clustering algorithms. Finally, to account for different noises that may be present during experiments, we generated three additional datasets (see Supplementary Note 4 for details). One dataset had an increased amount of noise, while the two others contained heteroscedastic noise, either scaling with conductance or with displacement. The best performing feature spaces and clustering algorithms remain largely unaffected.

From the fact that the row-to-row variation of FM indices, i.e., between feature space methods, is larger than the difference
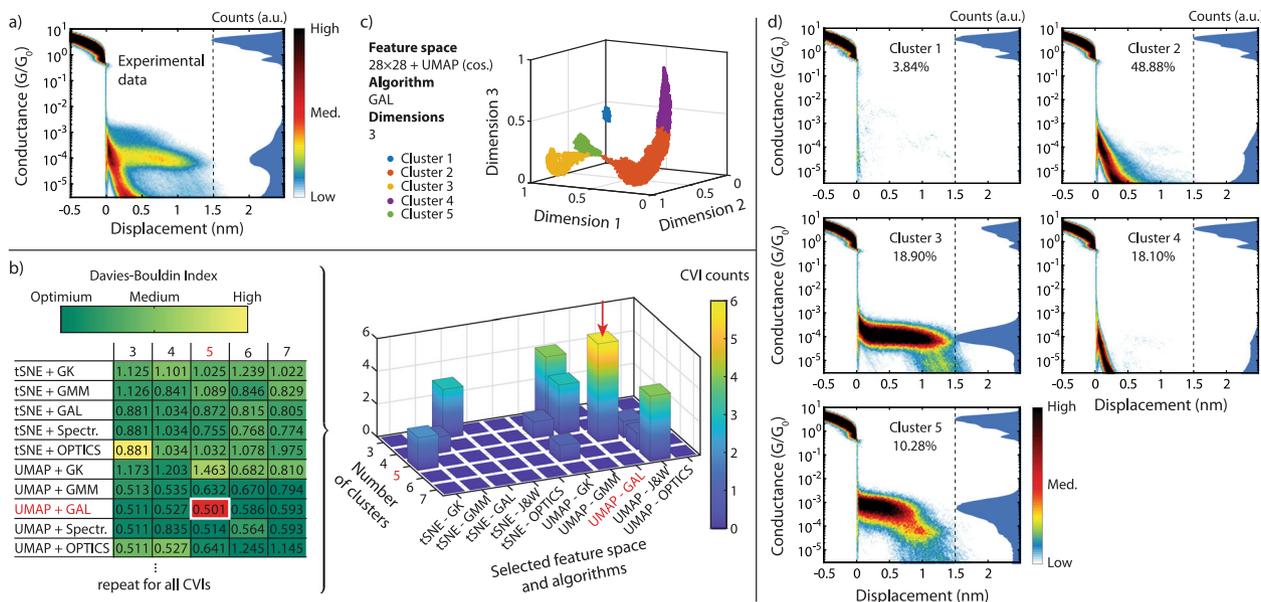
Fig. 3 Application of the workflow to measured mechanically controllable break-junction data. a Experimental 2D conductance-displacement histogram based on 10,000 breaking traces. The blue area represents the corresponding 1D conductance histogram. **b** Determination of the most suited feature space, clustering algorithm, and the optimal number of clusters using cluster validation indices (CVI). The feature space considered are $28 \times 28 + $ t-distributed stochastic neighbor embedding (t-SNE) and $28 \times 28 + $ uniform manifold approximation and projection (UMAP), both using a cosine (cos.) distance metric. The clustering algorithms considered are Gustafson–Kessel (GK), Gaussian mixed model (GMM), graph average linkage (GAL), spectral clustering following Jordan and Weiss (J&W), and Ordering Points To Identify the Clustering Structure (OPTICS). The heat map represents the Davies–Bouldin CVI, requiring a minimization of its value (red/white highlighted box). The histogram counts the occurrence of the feature space and clustering algorithm combinations, and of the optimal number of clusters, as predicted by the various CVIs. **c** Feature space constructed from the data of (**a**) using $28 \times 28 + $ UMAP (cos.) and the GAL clustering method for five clusters. **d** 1D conductance and 2D conductance-displacement histogram for the cluster assignment in (**c**).

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. However, as this benchmark 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 for further studies in the remainder of this paper.

Application to an experimental 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[35] (see Fig. 3a for the 2D conductance-displacement histogram), to which we apply the two selected feature space methods $28 \times 28 + $ t-SNE (cos.) and $28 \times 28 + $ 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 $5 \times 2 \times 7 = 70$ different clustering distributions. For each of them, we calculate internal CVIs[59–62]. 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 CVIs, 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 Supplementary Method 4.

The heat map shown in 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 CVI. This means that we calculate the CVIs for 2–8 clusters, but we only take the CVI into account if the optimum NoC lies between 3 and 7 clusters. This procedure is repeated for all 29 CVIs and a 2D histogram is constructed (Fig. 3b). Finally, this allows us to directly access the overall best feature space (28×28 + UMAP), algorithm (GAL), and NoC (5). As a verification of the robustness of the CVI prediction, we have performed the same analysis including, in addition, two poorly performing feature spaces (raw + t-SNE (cos.) and raw + UMAP (cos.)) and the same five clustering algorithms. Shown in Supplementary Note 5, the analysis shows that the combination of 28×28 + UMAP, GAL and 5 clusters again comes out as optimal. 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, and 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
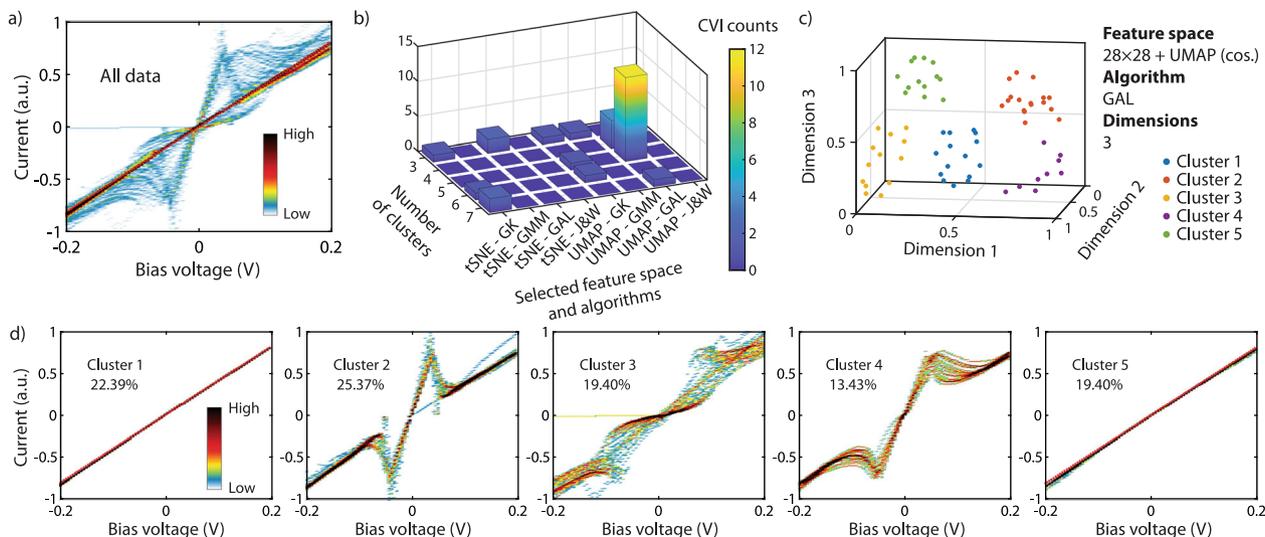
Fig. 4 Application of the method on current–voltage characteristics. a Experimental 2D current–voltage histogram based on 67 current–voltage characteristics recorded on a dihydroanthracene molecule[36,63,64]. **b** Determination of the most suited feature space, clustering algorithm and the optimal number of clusters using cluster validation indices (CVI). The feature space considered are 28×28 + t-distributed stochastic neighbor embedding (t-SNE) and 28×28 + uniform manifold approximation and projection (UMAP), both using a cosine (cos.) distance metric. The clustering algorithms considered are Gustafson–Kessel (GK), Gaussian mixed model (GMM), graph average linkage (GAL), spectral clustering following Jordan and Weiss (J&W), and Ordering Points To Identify the Clustering Structure (OPTICS). **c** Feature space constructed using the 28×28 + UMAP (cos.) and clustered using the GAL algorithm for five clusters. **d** 2D current–voltage histogram of the data shown in (**a**), clustered according to the partitioning shown in (**c**).

presence, and Cluster 5 to a two step breaking process involving molecule–electrode 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. Even though the CVIs show that five is the ideal NoC, this result should be taken with a grain of salt. 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 CVIs should be used merely as a guideline, and, as reference, we show the resulting cluster for 3–7 clusters in Supplementary Note 6.

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 (IV) characteristics, while the second one contains 4900 Raman spectra. For the current–voltage characteristics 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. Figure 4 a presents a 2D current–voltage histogram of 67 current–voltage characteristics recorded on a dihydroanthracene molecule[36,63,64]. The IVs have been normalized to focus on the shape of the curves, not on the absolute values in current. The same clustering procedure is repeated as described previously and the best feature space and clustering algorithm is determined to be 28×28 + UMAP(cos.) and GAL, respectively for an optimal NoC of 5 (Fig. 4b). The corresponding feature space is presented in Fig. 4c, colored according to the clusters produced by the GAL algorithm. The 2D current–voltage histograms of the five resulting clusters are shown in Fig. 4d. 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 IVs 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[36]. As Raman spectra are less stochastic than MCBJ measurements, we have performed a separate benchmark (see Supplementary Note 7) to rank the different algorithms. We find that, similar as for the MCBJs, the 28×28 + t-SNE and 28×28 + UMAP feature spaces perform the best. For the clustering algorithms, we find that most of them perform similarly well.

The Raman spectra are recorded on a well-studied reference system, namely a graphene membrane that has been divided into four quadrants, each exposed with a different dose of helium ions. The effect of He-induced defects on the Raman spectrum of graphene is known from literature[65,66], but for our analysis we explicitly do not rely on any a priori knowledge of the system, 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 μm diameter), suspended over a silicon nitride frame coated with Ti/Au (5 nm/40 nm). An illustration of the sample layout is presented in Fig. 5a. On this sample, a two-dimensional map containing 70×70 spectra was acquired using a confocal Raman microscope (WITec alpha300 R) with a 532 nm excitation laser. A description of the sample preparation and Raman measurements is provided in Supplementary Note 8.

The Raman spectra were fed to the 28×28 + UMAP (cos.) feature space construction method and split into 7 clusters using the GAL algorithm (see Supplementary Note 8 for more details). Figure 5b presents the partitioned feature space, containing several well-separated clusters. From this partitioning, we
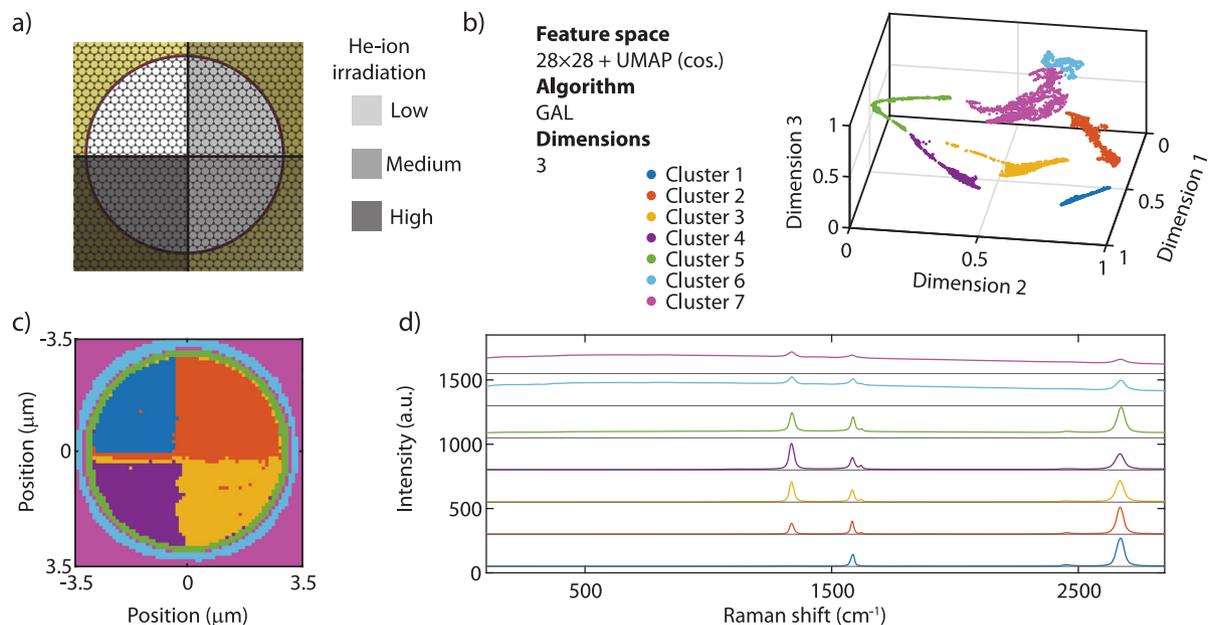
Fig. 5 Application of the method on Raman spectra. a Sample layout: suspended graphene membrane irradiated with four different He-ion doses. **b** Partitioned feature space, constructed with 28×28 + uniform manifold approximation and projection (UMAP) using the cosine (cos.) distance metric and the graph average linkage (GAL) clustering algorithm. **c** Spatial map of the extracted clusters. **d** Average Raman spectrum of each cluster.

construct a 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: 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 from the which the following characteristics can be evoked: 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 Clusters 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-irradiaton[65,66]. 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.

Discussion
In the synthetic data, the t-SNE and UMAP algorithms score equally well in reducing each measurement from a 784

dimensional space (28×28) down to the 3 dimensional feature space. On the experimental datasets, however, UMAP tends to perform better. This difference emphasizes the need for labeled 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[67,68]. While t-SNE reproduces well the local structure of the data, UMAP reproduces both the local and large-scale structure[42]. Moreover, one could also investigate more advanced variants of UMAP[69] 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 28×28 image merit additional research as some of them have proven to be highly accurate on the MNIST and other databases[57], despite their high computational cost.

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 benchmark 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, highlighting that a careful selection of the feature space construction and partition algorithm can significantly improve the classification results. We also provide guidelines for the estimation of the optimal NoC using a wide range of CVIs. 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 behaviors.

Data availability

additional datasets generated for the benchmark are available from the corresponding author upon reasonable request.

Code availability

The code used for this benchmark is freely available online at https://github.com/MickaelPerrin74/ClusteringBenchmark. In addition, we provide a graphical user interface for clustering data in a user-friendly fashion, containing all feature space construction methods and clustering algorithms used in this study. The code of this GUI is freely available online at https://github.com/MickaelPerrin74/DataClustering.

References

1. International Data Corporation (IDC). *Worldwide Spending on Artificial Intelligence Systems Will Be Nearly $98 Billion in 2023* https://www.idc.com/getdoc.jsp?containerId=prUS45481219 (2019).
2. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
3. Sun, Y., Wang, X. & Tang, X. Deep learning face representation from predicting 10,000 classes. In *Proc. 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 (IEEE Computer Society, 2015).
5. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J. & Khudanpur, S. Recurrent neural network based language model. In *Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010* (eds. Kobayashi, T., Hirose, K. & Nakamura, S.) Vol. 2, 1045–1048 (Interspeech, 2010).
6. Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* **29**, 82–97 (2012).
7. Zhang, X., Zhao, J. & LeCun, Y. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, Vol. 28 (eds Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R.), 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. *Comput. Struct. Biotechnol. J.* **13**, 8–17 (2015).
10. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
11. Hutto, C. & Gilbert, E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In *Proc. Eighth International AAAI Conference on Weblogs and Social Media 216* (eds. Eytan, A. & Paul, R.) Vol. 18 (Association for the Advancement of Artificial Intelligence (AAAI Press), 2014).
12. Bojarski, M. et al. End-to-End Learning for Self-Driving Cars (2016).
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 Lett.* **7**, 238–242 (2007).
15. El Abbassi, M. et al. Unravelling the conductance path through single-porphyrin junctions. *Chem. Sci.* **10**, 8299–8305 (2019).
16. Brown, K. A., Brittman, S., Maccaferri, N., Jariwala, D. & Celano, U. Machine learning in nanoscience: big data at small scales. *Nano Lett.* **20**, 2–10 (2020).
17. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
18. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* **111**, 218–230 (2016).
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. *Nat. Commun.* **10**, 1–10 (2019).
20. Lennon, D. T. et al. Efficiently measuring a quantum device using machine learning. *npj Quantum Inf.* **5**, 79 (2019).
21. Masubuchi, S. et al. Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials. *npj 2D Mater. Appl.* **4**, 3 (2020).
22. Lauritzen, K. P. et al. Perspective: theory of quantum transport in molecular junctions. *J. Chem. Phys.* **148**, 84111 (2018).
23. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **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. *Nat. Commun.* **7**, 12922 (2016).
25. Wu, B. H., Ivie, J. A., Johnson, T. K. & Monti, O. L. A. Uncovering hierarchical data structure in single molecule transport. *J. Chem. Phys.* **146**, 92321 (2017).
26. 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.* **120**, 016601 (2018).
27. Cabosart, D. et al. A reference-free clustering method for the analysis of molecular break-junction measurements. *Appl. Phys. Lett.* **114**, 143102 (2019).
28. El Abbassi, M. et al. Robust graphene-based molecular devices. *Nat. Nanotechnol.* **14**, 957–961 (2019).
29. Huang, F. et al. Automatic classification of single-molecule charge transport data with an unsupervised machine-learning algorithm. *Phys. Chem. Chem. Phys.* **22**, 3 (2019).
30. Vladyka, A. & Albrecht, T. Unsupervised classification of single-molecule data with autoencoders and transfer learning. *Mach. Learn.: Sci. Technol.* **1**, 3 (2020).
31. 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* **124**, 18302–18315 (2020).
32. Gustafson, D. E. & Kessel, W. C. Fuzzy clustering with a fuzzy covariance matrix. In *Proc. IEEE Conference on Decision and Control*, 761–766 (IEEE, 1978).
33. Bellman, R. *Dynamic Programming* (Princeton University Press, 2010).
34. Xu, B. Q. & Tao, N. J. Measurement of single-molecule resistance by repeated formation of molecular junctions. *Science* **301**, 1221–1223 (2003).
35. 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).
36. El Abbassi, M. et al. All experimental datasets are available at: https://doi.org/10.6084/m9.figshare.13258640 (2020).
37. LeCun, Y., Cortes, C. & Burges, C. *MNIST Handwritten Digit Database* http://yann.lecun.com/exdb/mnist/ (1998).
38. Van Der Maaten, L., Postma, E. & Van den Herik, J. Dimensionality reduction: a comparative review. *J. Mach. Learn. Res.* **10**, 66–71 (2009).
39. Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **C-18**, 401–409 (1969).
40. Hinton, G. E. & Roweis, S. T. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems* (eds. Becker, S., Thrun, S. & Obermayer, K.) 857–864 (MIT Press, 2003).
41. Van Der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
42. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
43. Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms.* (Kluwer Academic Publishers, 1981).
44. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59–69 (1982).
45. Silla, C. N. & Freitas, A. A. A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov.* **22**, 31–72 (2011).
46. Williams, C. K. I. & Rasmussen, C. E. Gaussian processes for regression *Proc. 8th International Conference on Neural Information Processing Systems.* 514–520 MIT Press: 1995.
47. Zhang, W., Wang, X., Zhao, D. & Tang, X. Graph Degree Linkage: Agglomerative Clustering on a Directed Graph. In *Computer Vision – ECCV 2012. ECCV 2012 (Lecture Notes in Computer Science)* (eds. Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y. & Schmid, C.) Vol. 7572, (Springer, Berlin, Heidelberg, 2012).
48. Zhang, W., Zhao, D. & Wang, X. Agglomerative clustering via maximum incremental path integral. *Pattern Recognit.* **46**, 3056–3065 (2013).
49. Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 888–905 (2000).
50. Ng, A. Y., Jordan, M. I. & Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Proc. 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (eds. Dieterich, T. G., Becker, S. & Ghahramani, Z.) 849–856 (MIT Press, 2001).
51. Ankerst, M., Breunig, M. M., peter Kriegel, H. & Sander, J. Optics: Ordering points to identify the clustering structure. In *Proc. ACM SIGMOD International Conference on Management of Data*, 49–60 (ACM Press, 1999).
52. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of the Second*

International Conference on Knowledge Discovery and Data Mining, 226–231 (AAAI Press, 1996).

53. Campello, R. J. G. B., Moulavi, D., Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science*, (eds Pei J., Tseng V.S., Cao L., Motoda H. & Xu G.) vol 7819, (Springer, Berlin, Heidelberg, 2013).

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

55. Chang, J., Wang, L., Meng, G., Xiang, S. & Pan, C. Deep adaptive image clustering. In *Proc. IEEE International Conference on Computer Vision*, Vol. 2017-October, 5880–5888 (Institute of Electrical and Electronics Engineers Inc., 2017).

56. Haeusser, P., Plapp, J., Golkov, V., Aljalbout, E. & Cremers, D. Associative deep clustering: training a classification network with no labels. In *Pattern recognition. GCPR 2018. Lecture Notes in Computer Science* (eds. Brox, T., Bruhn, A. & Fritz, M.) Vol. 11269 (Springer, Cham, 2019).

57. Ji, X., Vedaldi, A. & Henriques, J. F. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South)* 9864–9873 https://doi.org/10.1109/ICCV.2019.00996 (2019).

58. Fowlkes, E. B. & Mallows, C. L. A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.* **78**, 553–569 (1983).

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

60. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. Nbclust: an R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* **61**, 1–36 (2014).

61. Hämäläinen, J., Jauhiainen, S. & Kärkkäinen, T. Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms* **10**, 105 (2017).

62. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. Nbclust: an r package for determining the relevant number of clusters in a data set. *J. Stat. Softw. Artic.* **61**, 1–36 (2014).

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

64. Perrin, M. L., Eelkema, R., Thijssen, J., Grozema, F. C. & van der Zant, H. S. J. Single-molecule functionality in electronic components based on orbital resonances. *Phys. Chem. Chem. Phys.* **22**, 12849–12866 (2020).

65. 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).

66. Shorubalko, I., Pillatsch, L. & Utke, I. Direct-write milling and deposition with noble gases. In *Helium Ion Microscopy* (eds Hlawacek, G. & Gölzhäuser, A.) 355–393 (Springer Verlag, 2016).

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

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

69. McConville, R., Santos-Rodriguez, R., Piechocki, R. J. & Craddock, I. N2D: (Not Too) deep clustering via clustering the local manifold of an autoencoded embedding. Preprint at: https://arxiv.org/abs/1908.05968 (2019).

Acknowledgements

Author contributions

M.L.P. performed the machine learning analysis, with input from all authors. J.O. and O. B. performed the Raman measurements. M.E., J.O., O.B., M.C., H.S.J.v.d.Z., and M.L.P. discussed the data and wrote the manuscript. M.L.P. supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-021-00549-9.

Correspondence and requests for materials should be addressed to M.L.P.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.