Supplementary material

Damien Cabosart,[1] Maria el Abbassi,[1] Davide Stefani,[1] Riccardo Frisenda,[2] Michel Calame,[3]
Herre S. J. van der Zant,[1] and Mickael L. Perrin[3]

[1] *Kavli Institute of Nanoscience, Delft University of Technology, 2600 GA Delft, The Netherlands*

[2] *Instituto Madrileo de Estudios Avanzados en Nanociencia (IMDEA-nanociencia), Campus de Cantoblanco, E-28049, Madrid, Spain*

[3] *Empa, Swiss Federal Laboratories for Materials Science and Technology, Transport at Nanoscale Interfaces Laboratory, 8600 Dbendorf, Switzerland*

I. Influence of reference vector in method proposed by Lemmers et al. Nat. Comm. 7, 2016.

In the following, the influence of the reference vector on the clustering outcome is investigated. Here, we use a procedure which is similar to the work by Lemmers et al. (Nat. Comm. 7, 2016), but with one difference, as will be explained below. As reference vector, a tunneling trace with a slope of 7 decades per nanometer is used. Using this trace, the length of the difference vector |Y| = |X-R|, with X being the measured trace, and R the reference vector is used. In addition, the angle between R and Y is used as second coordinate in the feature vector. Finally, as third coordinate, in contrast to the work by Lemmers et al. the trace length is used, which is defined as the length in nanometer between the breaking of the last gold-gold contact to the first conductance value below 5e-6 G0. Altogether, the feature vector of each breaking trace consists of three dimensions. Using these three coordinates, clustering is performed using the Gustafson–Kessel fuzzy clustering algorithm. The clusters obtained after application of this method on sample 2b are shown in Figure S1. Cluster 1 consist of traces which appear to be vacuum tunneling. Cluster 2 is populated with very short traces in which the gold-gold contact snaps very quickly below the noise floor of the setup, while cluster 3 consist of a variety of traces which appear to be molecular in origin.

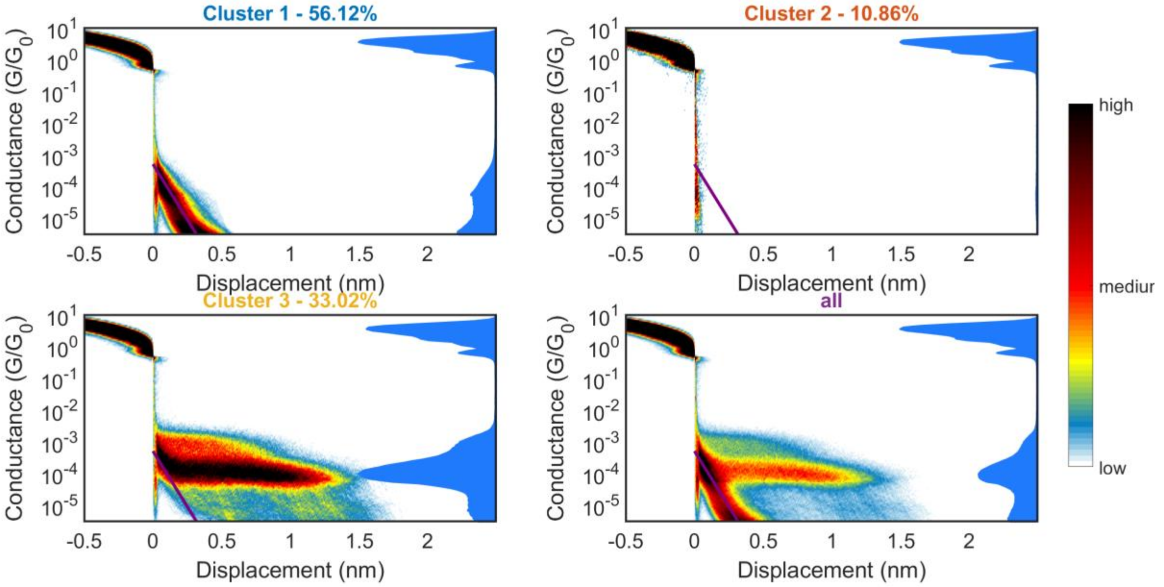

Figure S1: (a) Two-dimensional and conductance histograms built from all the breaking curves recorded at different bias voltages (*i.e.* sample 6a-f), corresponding to 11 124 traces. (b-c) Two-dimensional and conductance histograms built from the breaking curves of classes 1, 2 and 3, respectively, obtained thanks to the clustering method.

To investigate the influence of the reference vector, we varied its slope between 0.2 and 15 orders of magnitude per nanometer. Figure S2 shows the resulting population of the three clusters as a function of the slope of the reference. As shown in the figure, the population of the clusters varies with choice of the reference vector, with variation of the population up to 5% for cluster 1 and 3. Its influence on the clustering outcome is unwanted, and renders the method subjective.

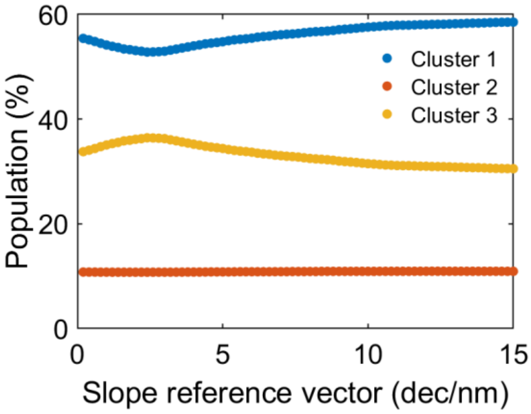

Figure S2: Population of the three clusters as a function of the slope of the reference vector.

II. OPE3 datasets

Table S1 is a summary of the OPE3 datasets. Six samples were measured in different experimental conditions (bias voltage, breaking speed and molecular yield). The total number of curves is 51 040. The detail of the calculation of the molecular yield for the datasets recorded at V = 100mV is presented in section III.

Sample number	**Sample name**	**Bias voltage (mV)**	**Breaking speed (nm/s)**	**Molecular yield (%)**	**Total number of traces**
1a	scan160720_17	100	3.0	30.96	10000
1b	scan160721_11	100	6.0	63.13	6346
2a	scan160512_50	100	1.0	10.38	3180
2b	scan160513_30	100	1.0	3.69	10000
3	scan160719_15	100	2.0	30.9	1440
4	scan161119_23	100	3.0	62.35	2000
5a	scan161123_25	100	3.0	25.45	2000
5b	scan161124_6	100	2.0	30.05	2000
5c	scan161124_10	100	6.0	28.55	2000
5d	scan161125_0	100	1.5	39.36	950
6a	scan161125_47	100	2.0	36.25	2000
6b	scan161126_0	150	2.0	28.50	2000
6c	scan161126_1	200	2.5	36.85	2000
6d	scan161126_2	250	2.5	45.30	2000
6e	scan161127_0	300	3.0	51.95	2000
6f	scan161127_1	50	5.0	29.81	1124

Table S1 : Details about the OPE3 datasets used for this work.

III. Extraction of the most probable conductance using the unfiltered data

For every dataset presented in Table S1 (section I), the most probable conductance (G_M) was extracted by fitting a log-normal distribution to the prominent peak in the associated conductance histogram. Figure S3 shows the fit results as well as the extracted G_M. The graph of the most probable conductance of the unfiltered data as a function of the molecular yield is displayed in Figures 1 and 4a in the main manuscript.

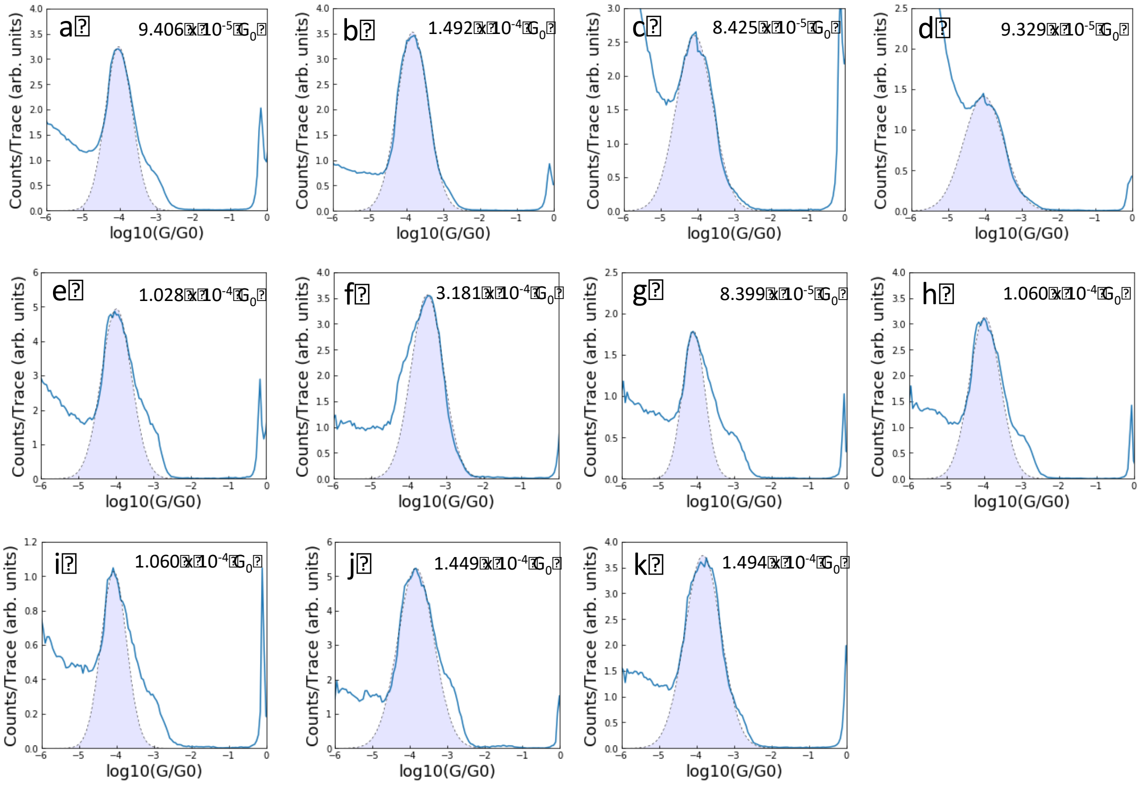

Figure S3 : (a-k) Conductance histograms built from the unfiltered data of samples 1a, 1b, 2a, 2b, 3, 4, 5a, 5b, 5c, 5d and 6a (see Table S1 in section I), respectively. For every histogram, the shaded region is a log-normal distribution fit to the data. The mean value of the log-normal distribution fit corresponds to the extracted most probable conductance (the obtained value is displayed inside every graph).

IV. Calculation of the molecular yield for the datasets at *V* = 100 mV

For a given dataset recorded at V = 100 mV in Table S1 (section I), the molecular yield is defined as the fraction of breaking curves belonging to the created classes 2 and 3 in Figures 3(c) and 3(d) in the main manuscript. In other words, it corresponds to the percentage of curves with plateau-like features. For example, sample 1a has 1154 and 1942 traces in classes 2 and 3, respectively, giving a molecular yield of 30.96% (= ((1154+1942)/10000) x 100).

V. Determination of the most probable conductance of classes 2 and 3 (molecular yield dependence)

The molecular yield dependence of the most probable conductance was investigated using the datasets in Table S1 (section I) for a constant bias voltage, *i.e.*, at V = 100 mV. It corresponds to eleven datasets of breaking curves associated with different molecular yields (*i.e.*, samples 1a, 1b, 2a, 2b, 3, 4, 5a, 5b, 5c, 5d and 6a). All the eleven datasets were merged to create a unique set of more than 40 000 traces and then split into three classes using the clustering method (see Figure 3 in the main manuscript). Since one class is a mixture of curves belonging to the different initial datasets, one can select all the curves of one of the initial datasets inside a specific class. By fitting a log-normal distribution to the prominent peak in the associated 1D histogram, one extract the most probable conductance related to one class for a given molecular yield. The results of the G_M extraction are presented in Figure S4 in the case of classes 2 and 3 in Figures 3(c) and 3(d), respectively, in the main manuscript.

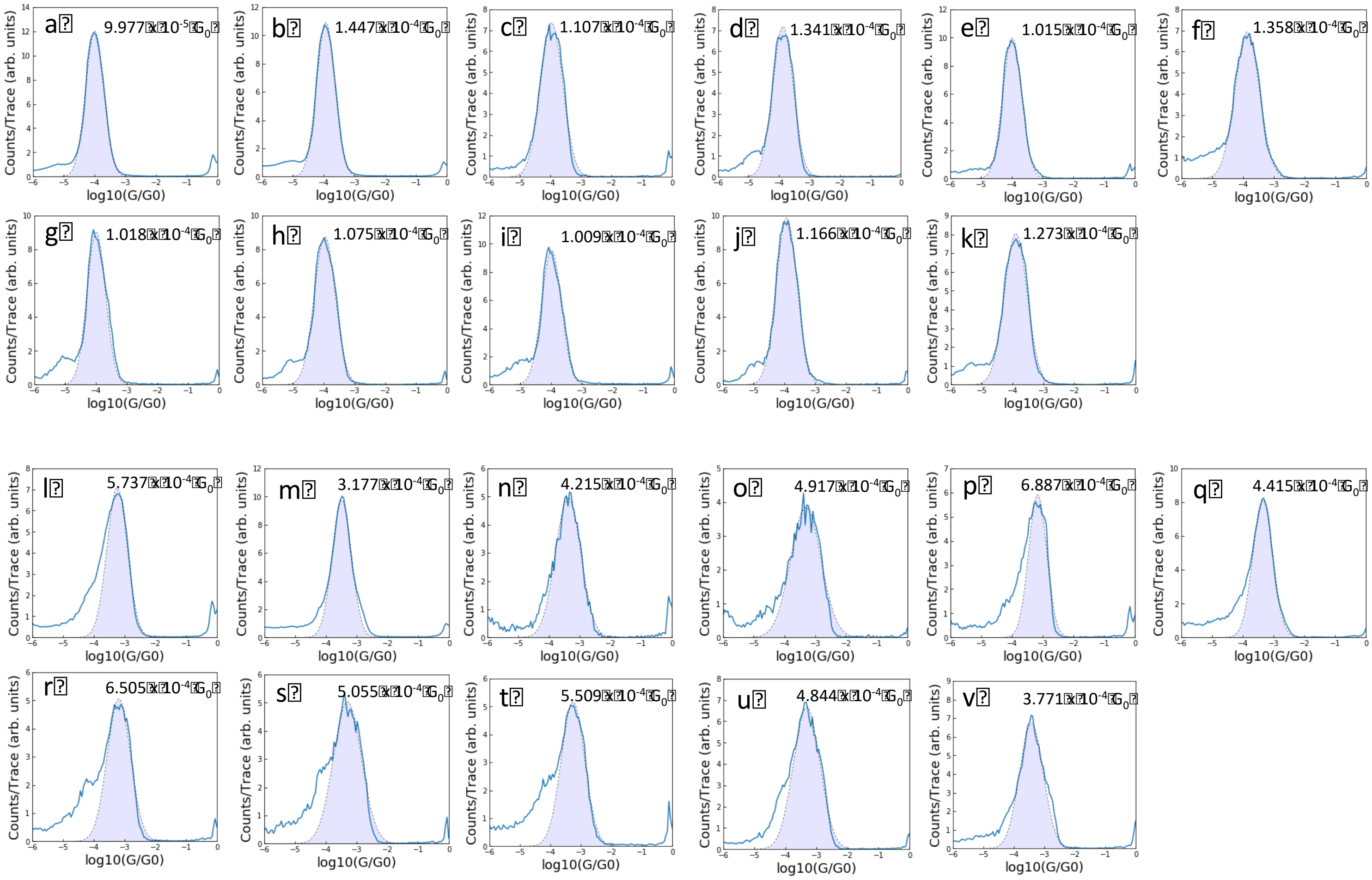

Figure S4: (a-k) Conductance histograms built from breaking curves belonging to class 3 and related to samples 1a, 1b, 2a, 2b, 3, 4, 5a, 5b, 5c, 5d and 6a (see Table S1 in section I), respectively. (l-v) Conductance histograms built from breaking curves belonging to class 2 and related to samples 1a, 1b, 2a, 2b, 3, 4, 5a, 5b, 5c, 5d and 6a, respectively. For every histogram, the shaded region is a log-normal distribution fit to the data. The mean value of the log-normal distribution fit corresponds to the extracted most probable conductance (the obtained value is displayed inside every graph.).

VI. Conductance and 2D histograms of the classes 1, 2 and 3 as well as the unfiltered data (bias voltage dependence)

The bias voltage dependence of the most probable conductance was investigated using the datasets in Table S1 (section I) for different bias voltages (V = 50, 100, 150, 200, 250 and 300 mV). It corresponds to six datasets of breaking curves recorded with the same break junction (*i.e.*, sample 6a-f). All the six datasets were merged to create a unique set of more than 10 000 traces, and then split into three classes using the clustering method (see Figure S5).

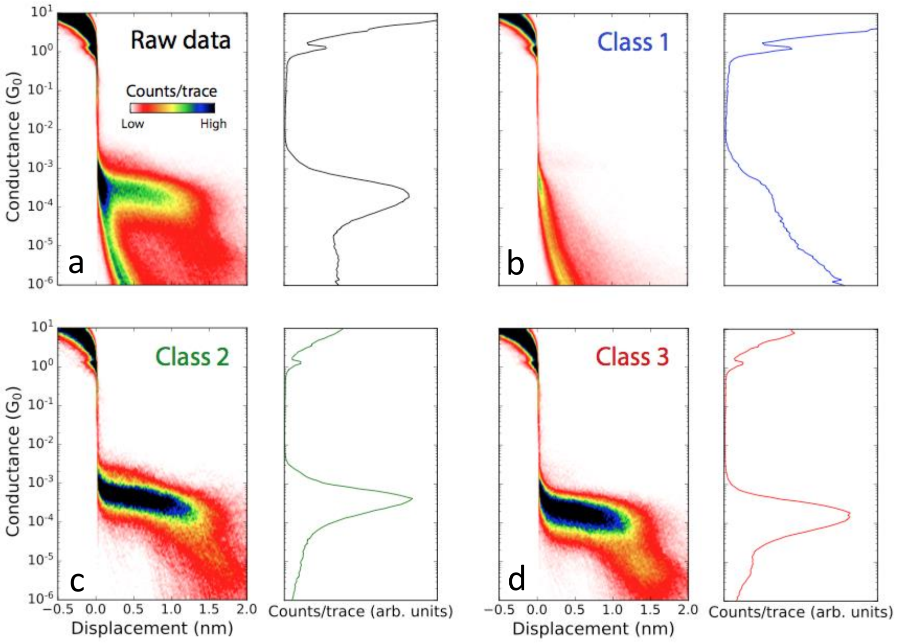

Figure S5: (a) Two-dimensional and conductance histograms built from all the breaking curves recorded at different bias voltages (*i.e.* sample 6a-f), corresponding to 11 124 traces. (b-c) Two-dimensional and conductance histograms built from the breaking curves of classes 1, 2 and 3, respectively, obtained thanks to the clustering method.

VII. Reduced feature spaces associated with the created datasets for the molecular yield and bias voltage dependence analysis

Figures S6(a) and S6(b) show the classification results of the datasets used for the molecular yield and bias voltage dependence analysis, respectively, in the reduced feature space. The three-dimensional reduced representations were created thanks to the principal component analysis (PCA) technique [1]. The latter method consists in projecting each feature vector onto the first three eigenvectors of the covariance matrix related to the dataset. The blue, green and red clusters are associated with the classes 1, 2 and 3, respectively.

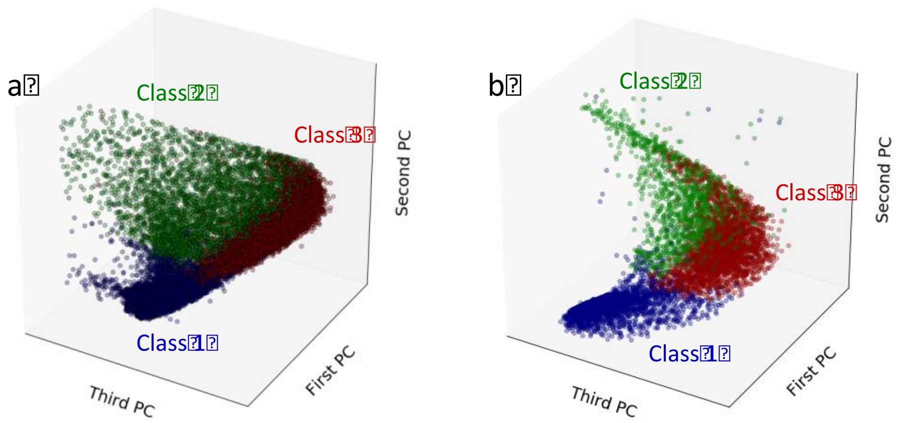

Figure S6: Reduced feature vector distributions associated with the created datasets for (a) the molecular yield and (b) bias voltage dependence analysis. The blue, green and red clusters correspond to the classes 1, 2 and 3, respectively.

VIII. Principle of the K-means algorithm

For a given feature space, one can apply a clustering algorithm in order to group the breaking curves into classes according to relevant trace features. For our investigation, we used the K-means++ algorithm (from the Scikit-Learn Python library) that is a variant of the standard K-means clustering method. The only free parameter to define is the number of final classes (K).

Because of its popularity, a detail description of the K-means technique is easily accessible and widely discussed in the literature (for instance, see Ref. [2] or [3] for an introduction chapter and video, respectively, about the method). The workflow of the standard K-means algorithm can be summarized as follows:

1. Initialization : every feature vector is randomly assigned to one class
2. For each class, we compute the mean value of the associated set of feature vectors, also called centroid or cluster center.
3. Each feature vector is reassigned to the class related to the closer centroid.
4. The centroid value is updated for every class.
5. The steps 3 and 4 are repeated until the assignments no longer change.

Note that, in the case of the K-means++ algorithm, the only difference with respect to the standard K-means method is the improvement of the centroid initialization (*i.e.*, steps 1 and 2).

It can be shown that the K-means algorithm aims to minimize an objective function that computes the sum of squared distances from each feature vector to its centroid. This means that, depending on the investigated dataset, the final results can vary from one initialization to another, *i.e.*, it can end up with a different local optimum of the objective function. Therefore, to optimize the classification task, the K-means algorithm is run for 100 different initializations. The final classification result is the one corresponding to the smallest objective function value.

IX. Summary of the results obtained with different clustering algorithms

Table S2 gives a summary of the results obtained with different clustering algorithms applied in our created high-dimensional feature space.

Name of the clustering method	Results
K-means++ [4]	✓ Fast to run and can deal with high-dimensional space.
Density-based spatial clustering of applications with noise (DBSCAN) [5]	✗ For any values of minimum cluster size, the curves with plateau-like features always belong to class -1 corresponding to outliers.
Hierarchical Density-based spatial clustering of applications with noise (HDBSCAN) [6]	✗ Same results as DBSCAN.
Gaussian mixture model [7]	✗ Computationally expensive and creates just one single cluster.
Gustafson-Kessel clustering model [8,9]	✗ Creation of only one cluster (all the centroids have the same position)
Maximum a-posteriori Dirichlet process mixtures (MAP-DP) [10]	✗ For a feature space with more than 40 dimensions, the algorithm does not work because of matrix singularity problems.

Table S2 : Summary of the results obtained with different clustering methods in the case of our created high-dimensional feature space.

X. Influence of the number of classes and bins on the determination of the molecular conductance

Here, we test the influence of the number of classes and bins on the determination of the molecular conductance for the dataset of sample 5a (see Table S1 in section I), *i.e.*, the one related to the clustering results presented in Figure 2(d) in the main manuscript.

Figure S7(a) shows the results of the most probable conductance values extracted from the conductance histograms of every created class (see, *e.g*, section II for the method) as a function of the number of classes K. For K = 1 and 2, only one conductance value is extracted (corresponding to just one class with plateau-like features). However, from K = 3 to 7, one can identify two values for the most probable conductance, *i.e.*, around 7x10^{-4} and 1x10^{-4} G_0. The latter values are associated with the set of breaking curves of classes 2 and 3, respectively, in Figure 2(d) in the main manuscript. Note that increasing the number of classes above K = 7 leads to the split of the classes 2 and 3 into subclasses. This is due to the fact that, for higher values of K, it is more favorable to create subclasses in order to minimize the objective function of the K-means++ algorithm. As observed on Figure S7(a), the most probable conductance values of the subclasses related to class 3 are very close to each other, while those of class 2 are clearly different. One possible explanation of the last observation could be that the breaking curves related to class 3 mainly correspond to a unique conformation of the OPE3 molecule (*i.e.*, fully stretched molecule) and therefore, creating subclasses leads to very similar conductance values. On the other hand, the breaking curves of class 2 might be a mixture of different molecular behaviors. Increasing the number of classes K enables to get more insights into the composition of one class.

Figure S7(b) is a graph of the most probable conductance associated with classes 2 and 3 as a function of the number of bins (M and N). To extract the most probable conductance values of every class, we first used the clustering algorithm in order to split the data into 3 classes (for a given number of bins M and N). Then, the most probable conductance is obtained using the conductance histogram related to every class (see, *e.g*, section II for the method). As observed in Figure S7(b), the extracted values are independent of the number of bins between 15 and 100. However, decreasing M and N below 15 leads to a modification of the extracted conductance values, especially in the case of class 2 showing a decrease of the conductance. This is mainly due to the fact that the resolution of the created individual 2D histograms is too poor to make a good detection of the relevant features related the breaking curves. Therefore, some curves from class 3 are 'picked up' by class 2, leading to a decrease of the most probable conductance of class 2. However, increasing the resolution does not modify the final results.

The investigation of the influence of the number of classes and bins on the determination of the molecular conductance show the reproducibility and robustness of the identification of 2 sets of breaking curves with plateau-like features, *i.e.*, classes 2 and 3 in Figure 2(d) in the main manuscript.

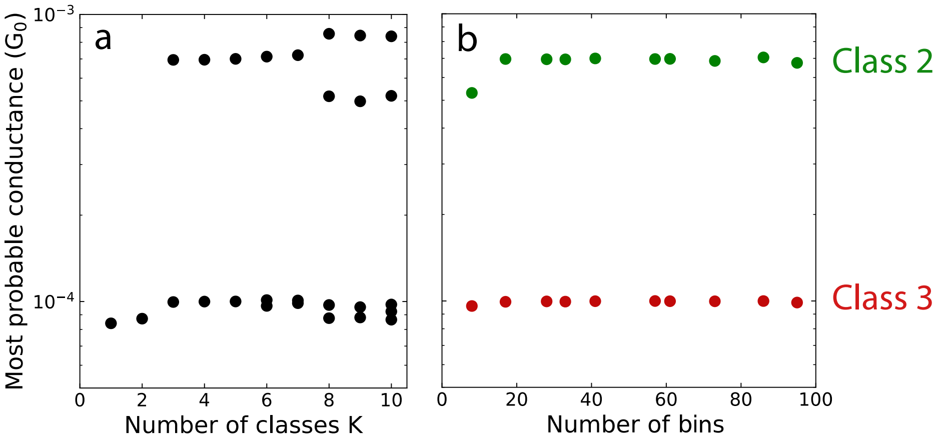

Figure S7 : Most probable conductance as a function of (a) the number of classes K and (b) the number of bins (M and N, where M=N) in the case of the OPE3 dataset of sample 5a. The number of bins was chosen randomly between 1 and 100.

XI. Subclasses of class 2 related to the molecular yield dependence analysis

Figure S8 shows the classification results after applying the clustering method on class 2 in Figure 3(c) in the main manuscript. As one can see, the two created classes show well-defined behaviors. Subclass A contains breaking curves that are longer than subclass B. It is worth noting that the occurrence of breaking traces belonging to subclass A is only significant at high molecular yield.

The identification of a new type of breaking curves (*i.e.*, belonging to subclass A) at high molecular yield (*i.e.*, > 40%) could be one of the reasons related to the change of the ratio between the occurrence of classes 2 and 3 observed in Figure 4(b) in the main manuscript. In other words, an increase of the molecular yields may lead to new types of molecular behavior.

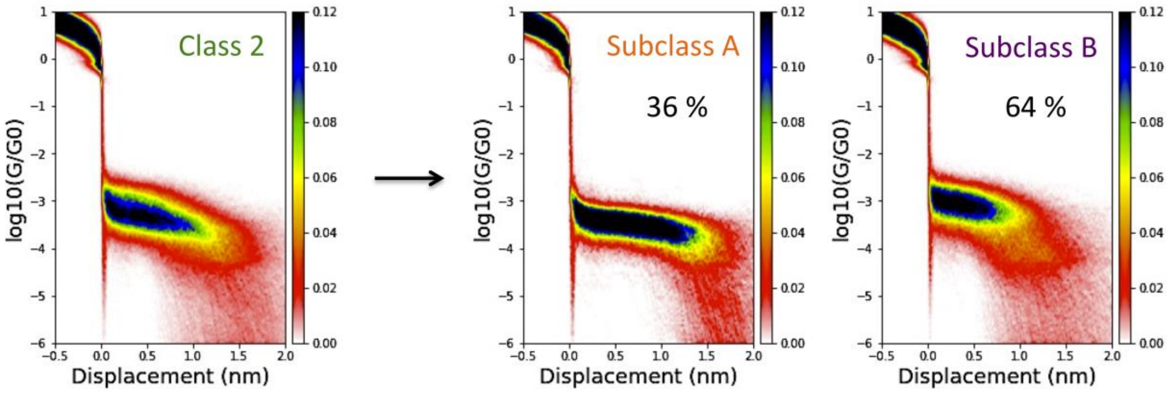

Figure S8 : Classification results after applying the clustering method for K=2 on class 2 presented in Figure 3(c) in the main manuscript. Two classes with different types of breaking traces are created.

References

[1] A. Géron, *O'Reilly* (2017)

[2] G. James et al., *Springer* (2013)

[3] https://fr.coursera.org/learn/machine-learning

[4] D. Arthur et al., *SODA*, 1027 (2007)

[5] M. Ester et al., *Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining* (1996)

[6] R. Campello et al., *Proc. of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2013)

[7] C. Bishop, *Springer*, 2007

[8] D. Gustafson et al., *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes* 761 (1978)

[9] M. Lemmer et al., *Nature Communications*, 7 (2016)

[10] Y. Raykov et al., *PLOS One*, 11(9) (2016)