

Article

Not peer-reviewed version

---

# Identifying Clean and Contaminated Atomic-Sized Gold Contacts Under Ambient Conditions Using a Clustering Algorithm

---

Guillem Pellicer and [Carlos Sabater](#) \*

Posted Date: 12 June 2025

doi: [10.20944/preprints202506.0993.v1](https://doi.org/10.20944/preprints202506.0993.v1)

**Keywords:** Molecular Electronics; Atomic-sized contacts; Conductance trace classification; DBSCAN clustering; Environmental contamination; Break-junction technique; Room temperature measurements; trace of conductance; Histogram; Density plots



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

## Article

# Identifying Clean and Contaminated Atomic-Sized Gold Contacts under Ambient Conditions Using a Clustering Algorithm

Guillem Pellicer  and Carlos Sabater  \*

Departamento de Física Aplicada and Instituto Universitario de Materiales de Alicante (IUMA), Universidad de Alicante, Campus de San Vicente del Raspeig, E-03690 Alicante, Spain

\* Correspondence: carlos.sabater@ua.es

**Abstract:** Molecular electronics studies have advanced from early, simple single-molecule experiments at cryogenic temperatures to complex and multifunctional molecules under ambient conditions. However, room-temperature environments increase the risk of contamination, making it essential to identify and quantify clean and contaminated rupture traces (i.e., conductance versus relative electrode displacement) within large datasets. Given the high throughput of measurements, manual analysis becomes unfeasible. Clustering algorithms offer an effective solution by enabling automatic classification and quantification of contamination levels. Despite the rapid development of machine learning, its application in molecular electronics remains limited. In this work, we present a methodology based on the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to extract representative traces from both clean and contaminated regimes, providing a scalable and objective tool to evaluate environmental contamination in molecular junction experiments.

**Keywords:** molecular electronics; atomic-sized contacts; conductance trace classification; DBSCAN clustering; environmental contamination; break-junction technique; room temperature measurements; trace of conductance; histogram; density plots

## 1. Introduction

The core idea behind molecular electronics [1–3] is to use the smallest possible components—individual atoms or molecules—as active elements in electronic devices. A common strategy involves connecting a single atom or molecule between two electrodes. Break-Junction (BJ) techniques offer an excellent platform to achieve this level of control. The most widely used methods to measure electronic transport in such junctions are *Scanning Tunneling Microscopy Break Junctions* (STM-BJ) [4] and *Mechanically Controllable Break Junctions* (MCBJ) [5–7]. These techniques allow measurements electrical conductance  $G$  the conductance (defined as  $G = \frac{1}{R} = \frac{I}{V_{bias}}$ ). Where  $I$  is the current that follows the junctions and the  $V_{bias}$  is the voltage applied to the junction. According to Landauer's formalism [8], the conductance of atomic and molecular junctions is quantized and can be expressed as  $G = G_0 \sum_i T_i$ , where  $T_i$  is the transmission probability of the  $i$ -th conduction channel, and  $G_0 = \frac{2e^2}{h}$  is the quantum of conductance [9]. Here,  $e$  is the elementary charge,  $h$  is Planck's constant, and the factor of 2 accounts for spin degeneracy.

In the case of single-atom metallic contacts—such as gold—the transmission probability is typically close to one, resulting in a conductance near  $G \approx G_0 \approx \frac{1}{12906} \Omega^{-1} \approx 7.75 \times 10^{-5} \text{ S}$  [10]. However, when the gold contact is further stretched, the junction breaks and the conductance abruptly drops to the tunneling regime, where the current decreases exponentially with the distance that separates the electrodes. When a molecule bridges the electrodes, the transmission is usually significantly reduced compared to metallic contacts (if the molecule is a poor conductor). The specific value depends on the molecular structure and its electronic coupling to the electrodes. In general, molecular junctions exhibit conductance values lower than  $G_0$ , often several orders of magnitude smaller [11].

The field of molecular electronics has undergone remarkable progress in recent decades [1–3], evolving from early demonstrations of single-molecule junctions at cryogenic temperatures [12–14] to room-temperature molecular bridges with various functional properties [15–20]. When measuring electronic transport in atomic-scale contacts under ambient conditions using BJ techniques, one of the main risks is sample degradation due to environmental exposure. Even brief contact with the laboratory atmosphere can lead to contamination of the junctions. Contamination that in form unknown molecule is captured between the leads when they are stretched or compressed.

In parallel with advancements in molecular electronics, the past decade has seen a transformative rise in machine learning and artificial intelligence. While these tools have revolutionized numerous scientific fields, their application in molecular electronics started to grow in the last six years [21–25]. Our manuscript also is contributin to bring a new utilities for the molecular electronics using DB-SCAN[26–28] clustering algorithm. This approach enables the automatic extraction of representative conductance-versus-displacement that correspond to both clean and contaminated regimes in atomic-sized gold contacts measured under ambient conditions. By providing a robust statistical framework, our method allows for the distinction between pure metallic junctions and those compromised by environmental contamination, enhancing both the reproducibility and interpretability of room-temperature molecular electronics experiments.

Given the known challenges of obtaining ultra-clean atomic-scale junctions, one strategy to mitigate contamination is the *in situ* cleaning of samples, for example through plasma treatment protocols. However, such cleaning techniques typically require disassembling the setup and removing the sample from the experimental chamber, which is not always feasible, especially during high-throughput or time-sensitive measurements.

In this article, we propose an alternative approach. Rather than physically cleaning the samples, we leverage a clustering algorithm to automatically classify conductance versus displacement traces—commonly referred to as conductance traces—according to whether they are clean (i.e., characteristic of pure metallic junctions) or contaminated (i.e., altered by molecular adsorption or ambient impurities). Our manuscript presents both the clustering protocol employed and its successful application to experimental data, clearly revealing the presence of two distinct classes of traces: those corresponding to clean gold junctions and those affected by contamination. This offers a powerful tool for assessing the cleanliness of the junctions during measurements, enabling the estimation of the proportion of clean traces in a dataset. In turn, this allows researchers to define thresholds for initiating sample cleaning or discarding data segments.

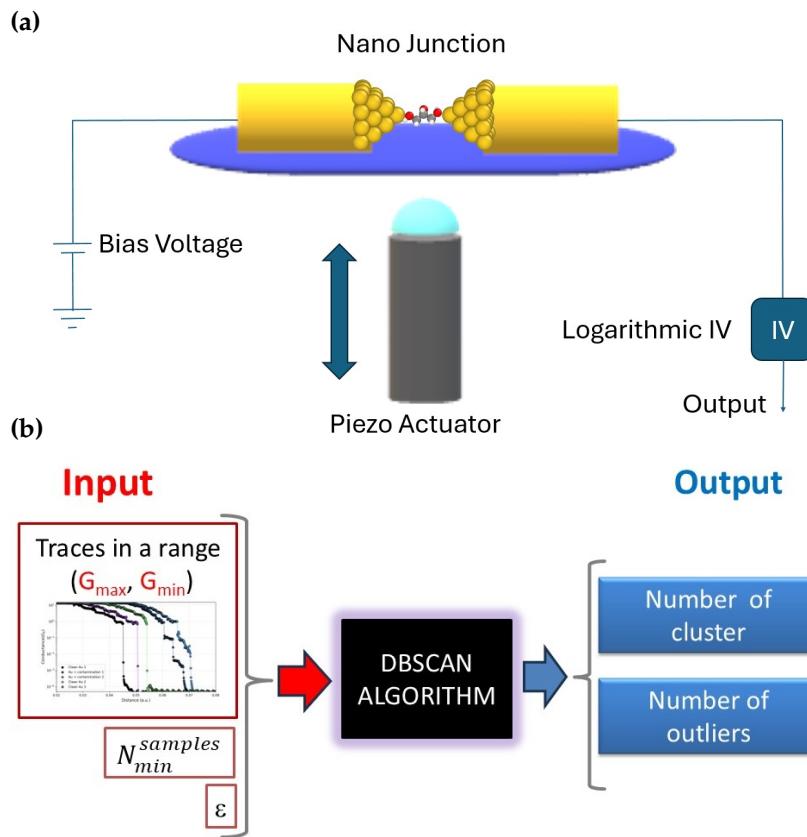
## 2. Materials and Methods

### 2.1. Molecular Electronics Based on BJ Experiments

To investigate atomic-scale electronic transport under ambient conditions, we employed a mechanically controllable break junction (MCBJ) setup (see Figure 1a). This illustration shows a single molecule captured between two gold electrodes. A bias voltage is applied on the left side, and the current flows through a molecule trapped between the electrodes. These electrodes are mounted on a flexible polylactic acid (PLA) substrate. This substrate is bent using a piezoelectric actuator which, when pushed, causes the gold wire to break, and upon retraction, allows the flexible substrate to relax, bringing the electrodes back into contact and forming the junction again.

The current flowing through the molecule is extremely small, typically in the nanoampere or sub-nanoampere range. To accurately detect these low currents, a custom-built logarithmic I/V converter was used [29], allowing the amplification and recording of conductance signals as low as  $10^{-4} G_0$ .

The system consists of a notched gold wire (Goodfellow, 0.1 mm diameter) carefully deposited and glued onto the flexible PLA substrate [30]. This configuration is integrated into a three-point bending mechanism, where a piezoelectric element provides precise mechanical control of the electrode separation, enabling the reproducible formation and rupture of atomic-scale junctions.



**Figure 1.** (a) Illustration and basic circuit of a MCBJ. (b) Input and output sketch of the DBSCAN algorithm.

During the experiments, a constant bias voltage of 100 mV was applied across the junction. The resulting current was measured using the custom logarithmic I/V converter and recorded with a data acquisition (DAQ) system. This setup enables the acquisition of conductance traces, defined as the conductance versus relative electrode displacement—measured in volts or, when calibrated, in angstroms.

In this study, we focus exclusively on rupture traces. A total of approximately 5024 traces were acquired and analyzed under ambient conditions.

## 2.2. Classification Method: Use of the DBSCAN Algorithm

Although there are various approaches to applying *clustering* techniques to the classification of conductance traces, in this manuscript, we have chosen to use the DBSCAN algorithm. While the *k-means* algorithm may be faster and more computationally efficient, we have preferred to sacrifice speed in favor of greater classification fidelity. DBSCAN is particularly effective when clusters exhibit relatively uniform density and there is a clear distinction between dense and sparse regions. Its operation is based on two key parameters: the neighborhood radius  $\epsilon$ , which defines the maximum distance between two points to be considered neighbors, and the minimum number of samples  $N_{\min}^{\text{samples}}$  required for a point to be considered a core point.

In our case, the goal of clustering is not to classify individual points, but rather to identify and classify segments or ranges within each conductance trace. To this end, each segment is transformed into a feature vector, so the algorithm operates on these vectors rather than on the entire trace. This is illustrated in Figure 1b, where the *inputs* to the algorithm are the conductance ranges to be analyzed in each individual trace. It is important to note that, under this formulation, the parameters  $\epsilon$  and  $N_{\min}^{\text{samples}}$  are applied directly to the feature vector space, not to the individual points contained in the original traces.

As shown in Figure 1b, the left side indicates the *inputs*: the segmentation range, the number of variables considered, and the value of  $N_{\min}^{\text{samples}}$ . Once a set of trace segments is processed (in our case, 5024 segments), and specific values of  $\epsilon$  and  $N_{\min}^{\text{samples}}$  are set, the algorithm returns the number of identified *clusters*, as well as the number of traces that are not assigned to any cluster (classified as *outliers*). One of the main advantages of DBSCAN is precisely its ability to automatically identify these *outliers*.

Although DBSCAN does not require prior knowledge of the number of *clusters*, having this information can be advantageous. In our case, we already know that the data ideally separates into two *clusters*: one corresponding to clean traces and the other to contaminated ones. For a given segmentation range and specific parameters, we assess whether DBSCAN identifies these two groups. Any data that does not fit well into either *cluster* is classified as an *outlier*. However, the optimal values of  $\epsilon$  and  $N_{\min}^{\text{samples}}$  are not known in advance.

Our goal is to distinguish between two types of conductance traces, or two *clusters*: one corresponding to clean traces and the other to contaminated ones. As shown in Figure 1b, the left panel illustrates the *inputs*: the segmentation range, the number of variables considered, and the value of  $N_{\min}^{\text{samples}}$ . Once a set of trace segments (in our case, 5024) is processed and specific values of  $\epsilon$  and  $N_{\min}^{\text{samples}}$  are set, DBSCAN returns the number of identified *clusters* and the *outliers*. In this work, we propose a systematic methodology to determine the optimal values of  $\epsilon$  and  $N_{\min}^{\text{samples}}$ . To do so, we construct a meshgrid over the parameter space and run DBSCAN for all combinations of values. For each parameter pair, we record the number of *clusters* and *outliers*. To automate the process, we developed a Python script that explores the parameter space and generates a 3D map, where the z-axis indicates the number of *clusters* and the color scale encodes the number of *outliers*. This visual tool provides an intuitive way to select an optimal parameter set for reliable classification. Once this plot is obtained, we refer to it as the result of a completed iteration.

However, in our protocol we must perform two iterations of DBSCAN in two distinct ranges to successfully classify the traces. This strategy has allowed us to develop a robust protocol that accurately distinguishes between ultraclean traces and those contaminated by the environment, even enabling estimation of the relative percentage of each type. It is worth noting that the complete application of the protocol takes no more than ten minutes per data set.

### 3. Results

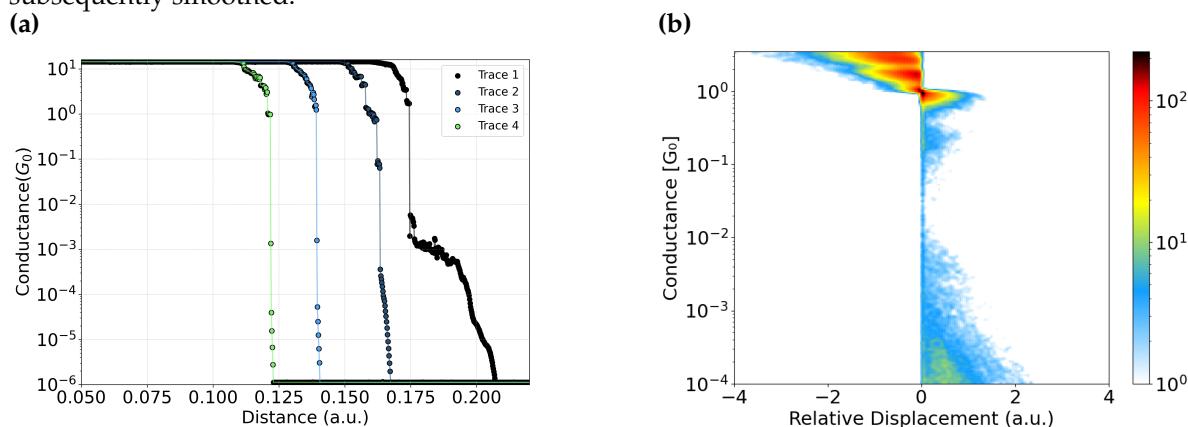
#### 3.1. Data Raw Atomic-Sized Contacts of Gold at Room Conditions

Figure 2a shows a selection of traces from the same experiment, plotted in logarithmic scale relative to displacement. It is well known that when an atomic contact is clean, the conductance drops sharply to the tunneling regime after the final atomic plateau at  $1 G_0$  upon further elongation [31–33]. In contrast, when a molecule is present between the electrodes, a plateau often appears in the sub-quantum conductance range due to its low transmission [34]. The traces 1 and 2 of the Figure 2a drop abruptly from  $1 G_0$  to below  $10^{-4} G_0$ , without exhibiting any intermediate plateaus; these ones can be classified as clean. In contrast, traces 3 and 4 display conductance plateaus between  $0.1 G_0$  and  $10^{-4} G_0$ , despite the absence of any intentionally deposited molecules. This behavior suggests the presence of contamination, potentially due to unintentional molecular junctions or residual species influencing the conductance.

An alternative statistical representation is the density color map, which can be constructed by aligning all traces with respect to the onset of the last plateau corresponding to the single-atom contact (in the case of gold  $1.0 G_0$ ). Specifically, each trace is shifted such that the first point at  $1.0 G_0$  is set to zero displacement. This procedure is applied to all traces, allowing them to share a common reference point.

Using this alignment method, a density map is generated, as shown in Figure 2b, representing the full dataset composed of 5024 conductance traces. The map displays the number of events using a color scale: warm colors (dark red) indicate high occurrence density, while cool colors (blue) denote

low density. The plot is constructed with 245 bins along the x-axis and 250 bins along the y-axis and is subsequently smoothed.



**Figure 2.** (a) Representative conductance rupture traces plotted on a logarithmic scale. Traces 3 and 4 show direct transitions from  $1 G_0$  to below  $10^{-4} G_0$ , while traces 1 and 2 exhibit intermediate plateaus, indicating possible contamination. (b) Density map of 5024 gold rupture traces recorded at room temperature. The color scale reflects the number of events, with warm colors indicating high density and cool colors low density.

### 3.2. Automated Selection of Optimal DBSCAN Parameters for Trace Classification

As explained earlier, DBSCAN depends on two parameters:  $\epsilon$  and  $N_{\min}^{\text{samples}}$ . To choose them automatically, we created a Python script that scans the parameter space. For each pair of values, it runs DBSCAN and records the number of clusters and outliers. The results are shown in a 3D plot where the  $x$  and  $y$  axes represent the parameters, the  $z$ -axis shows the number of clusters, and the color scale indicates the number of outliers. This makes it easy to spot the most suitable region, as shown both panels Figure 3.

However, as previously noted in the *Materials and Methods* section, a single DBSCAN iteration is not sufficient to accurately discriminate between fully clean and contaminated traces. For this reason, our protocol involves two successive DBSCAN runs, each applied to a different segment of the conductance traces. In the first iteration, the conductance traces are segmented in the range  $[10^{-1}, 10^{-3}] G_0$ . This segment is used as input for a parameter sweep over  $\epsilon$  and  $N_{\min}^{\text{samples}}$  using a meshgrid approach. For each parameter pair, we compute the number of clusters and outliers identified by DBSCAN. The results are visualized in Figure 3a.

From the 3D map, we identify an optimal region where the number of clusters is two and  $N_{\min}^{\text{samples}}$  is between 4 and 7. Selecting parameters in this region helps the algorithm detect well-defined clusters associated with clean traces. Table 1 shows a brief summary of the four selected possible DBSCAN parameter combinations explored. The first column shows the number of the combination. Second and third columns show the values of  $\epsilon$  and  $N_{\min}^{\text{samples}}$ , which yield exactly two clusters—matching the classification objective. The last three columns indicate the number of outliers, clean traces, and contaminated traces.

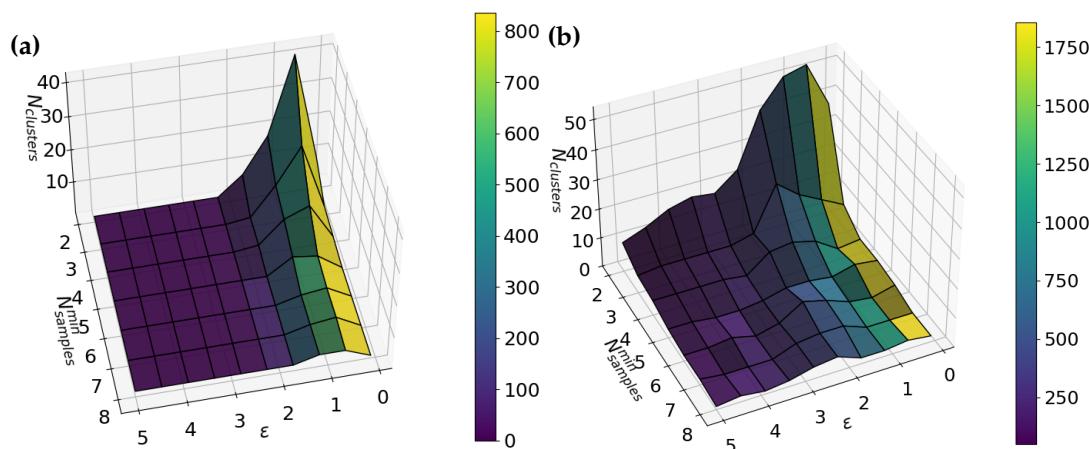
**Table 1.** Summary of the possible DBSCAN parameter combinations explored in the first iteration.

Combination	$\epsilon$	$N_{\min}^{\text{samples}}$	Clusters	Outliers	Clean Traces	Contaminated Traces
1	1.733	7	2	76	4188	760
2	1.733	6	2	71	4188	765
3	4.456	7	2	1	4188	835
4	4.456	6	2	1	4188	835

As shown in Table 1, for  $\epsilon$  values between approximately 1.5 and 4.5 and  $N_{\min}^{\text{samples}}$  between 6 and 7, the number of traces classified as clean remains constant at 4188. Within this range, the number of outliers varies between 1 and 76, while the number of contaminated traces ranges from 760 to 836,

always over a total of 5024 traces. In this first iteration, we selected combinations 3 and 4, as both yield the minimum number of *outliers* (1) and allow the remaining traces to be classified as contaminated without affecting the number of clean traces. Since both combinations produce equivalent results, we adopted option 3 (or 4) as the reference for the next iteration. However, as clearly shown in Figure 4a, a noticeable cloud of counts remains in the region around  $10^{-1}G_0$ , indicating that the separation is not entirely clean. In fact, the data raw density plot in Figure 2b, which includes all unfiltered traces (including contaminated ones), appears strikingly similar to the distribution obtained after the first DBSCAN filtering shown in Figure 4a. Therefore, a second iteration is warranted, as previously outlined in the Materials and Methods section of this manuscript.

In the second iteration, the same procedure is applied to a different segment of the conductance traces,  $[10^0, 10^{-1}]G_0$ , over the 4188 clean traces. The results obtained are shown in Figure 3b. This second analysis serves to refine the classification and helps distinguish really clean traces or traces with contaminants, completing the overall protocol.



**Figure 3.** Three-dimensional plots showing the number of clusters (z-axis) as a function of the DBSCAN parameters  $\epsilon$  and  $N_{\text{samples}}^{\text{min}}$ . The color scale indicates the number of outliers detected. **(a)** corresponds to the first iteration and **(b)** to the second iteration of DBSCAN.

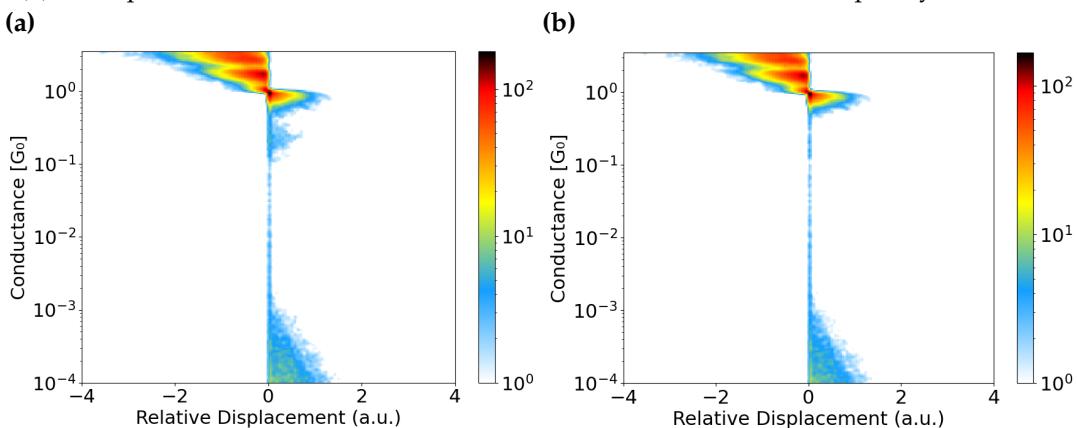
As in the first iteration, we compile a summary table to evaluate different DBSCAN parameter combinations that result in two clusters. Table 2 presents these combinations. Based on this table, we observe that for very small values of  $\epsilon$ , the number of outliers increases significantly. Upon inspection, this behavior appears to misclassify some genuinely clean traces as outliers. On the other hand, when  $\epsilon$  is too large, the number of outliers drops drastically. However, visual inspection of the supposedly clean traces reveals the presence of residual sub-plateaus in certain regions, suggesting that some contaminated traces are being misclassified as clean. Therefore, we select an intermediate value of  $\epsilon = 3.5$ , which results in a reasonable number of outliers and a sufficiently high number of clean traces. Specifically, out of the initial 4188 clean traces, 3912 are retained after this second filtering step.

**Table 2.** Summary of the possible DBSCAN parameter combinations explored in the **second** iteration.

Combination	$\epsilon$	$N_{\text{samples}}^{\text{min}}$	Clusters	Outliers	Clean Traces	Contaminated Traces
1	0.644	6	2	1103	3077	6
2	3.400	7	2	218	3912	56
3	5.000	6	2	91	4087	8

### 3.3. Clustering Based on DBSCAN to Identify Pure Metallic Atomic-Scale Gold Contacts and Traces with Contamination

Once the DBSCAN algorithm has been applied to classify the traces as clean or contaminated, one of the most effective ways to evaluate its performance is to represent both iterations using density plots. Figure 4 displays the results: panel (a) shows the density plot from the first iteration, while panel (b) corresponds to the second iteration, in which traces identified as completely clean are shown.



**Figure 4.** Panel (a) shows a density plot constructed from the 4188 clean traces identified in the first iteration using combination 3–4 from Table 1. Panel (b) displays a density plot based on the 3912 clean traces identified in the second iteration using combination 2 from Table 2. The number of bins used for the density plots are 250 and 300, respectively.

As can be observed, the contaminated traces shown in Figure 4 panel (a) behave quite differently. Around  $10^{-1} G_0$ , a bluish cloud is clearly visible, indicating contamination. In contrast, the density plot in panel (b) exhibits very few counts within a noticeable gap between the atomic contact of gold ( $\sim 1 G_0$ ) and the tunneling regime (which starts around  $10^{-3} G_0$  and is characterized by its slope on the logarithmic scale). Although a small number of points remain in this clear gap, the number of counts in this gap is three orders of magnitude compared to the maximum. These isolated points are typically attributed to artifacts from the data acquisition system (DAQ).

These qualitative differences between the two density plots confirm that they reflect distinct physical behaviors. A more detailed discussion supporting the validity of our classification algorithm is provided in the following section.

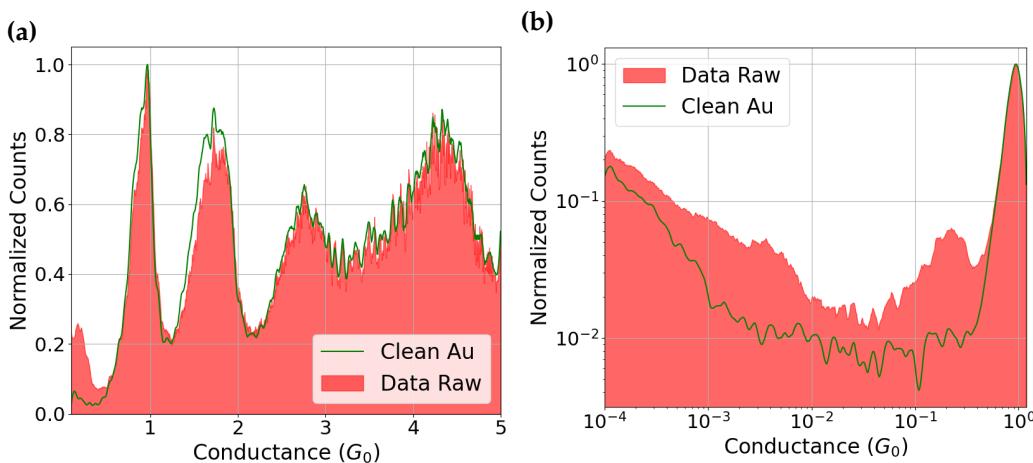
## 4. Discussion

Thanks to the DBSCAN clustering algorithm applied following the protocol described above, we have been able to classify conductance traces into two well-defined groups: those confidently labeled as *clean*—even under ambient conditions—and those interpreted as affected by *contamination*. This classification is clearly reflected in the density plots shown in Figure 4. Panel (a) corresponds to the first iteration, which fails to correctly identify clean traces, while panel (b) shows the result of the second iteration, where clean traces are clearly isolated.

Moreover, we want to bring clarity to our message and to be more conclusively demonstrate the cleanliness of the selected traces. Therefore, we propose an additional statistical analysis based on normalized conductance histograms (to one quantum of conductance), presented in both linear and logarithmic scales (see Figure 5). In these histograms, we compare the raw dataset with the subset of 3912 traces identified as clean. The color code indicates red for the data raw and green for the classified as clean.

The linear-scale histogram allows us to observe high conductance features and directly compare the overall profile of the raw(red) and clean datasets (green). In contrast, the log-log histogram (logarithmic in both counts and conductance) enhances the visibility of features typically associated with environmental contamination, such as peaks around  $\sim 1 \cdot 10^{-1} G_0$  or even down to  $\sim 1 \cdot 10^{-4} G_0$ .

As expected, the log-log representation emphasizes all differences, especially in the low-conductance region. Finally, we provide a comparative table showing the number of traces classified as clean and contaminated within our dataset.



**Figure 5.** Panel (a) shows the histogram on a linear scale, while panel (b) presents it on a logarithmic scale (log-log). The red curves correspond to the raw dataset (5024 traces), and the green curves represent the subset of clean 3912 traces identified by the DBSCAN algorithm.

Further assess the robustness of our clustering procedure, we analyze the conductance histograms in the range from 0.1 to  $5 G_0$  for both the raw data (red color) and the clean subset (green line). As shown in Figure 5a, the histogram of the full dataset exhibits a strong peak centered around  $1 G_0$ , consistent with the typical conductance of single-atom Au contacts. Focusing on the peaks at 1, 2, and  $3 G_0$ , we observe that, for both the raw data and the clean dataset, the peak positions remain essentially the same, with only subtle differences. A slight decrease in the count density is noticeable near  $2 G_0$  for the clean traces. Interestingly, and in contrast with previous studies where molecules are deposited on gold atomic contacts—where a significant broadening of the  $1 G_0$  peak is typically observed [35,36]—such broadening is absent in our data, suggesting that environmental contamination does not produce a comparable effect.

However, a notable difference emerges in the low-conductance regime, around  $0.1 G_0$ . In this region, a clear peak appears in the raw dataset but is entirely absent in the clean traces. This highlights the importance of representing conductance histograms in a logarithmic scale, which allows better resolution in the low-conductance region and helps reveal the true positions and nature of such peaks, potentially masked in linear representations.

The distinction between clean and contaminated traces is clearly observed in the log-log scale histograms in Figure 5b. In the clean dataset (represented by the green histogram), there's a marked peak deep in the count density, specifically between approximately  $8 \cdot 10^{-1} G_0$  and  $1 \cdot 10^{-3} G_0$ . Below this threshold, the histogram exhibits a linear trend on the log-log scale, which is a characteristic behavior of the tunneling regime. Conversely, the raw data (red area) displays a prominent distribution of conductance values within the same range. Here, a significant concentration of events stands out, centered around  $2 \cdot 10^{-1} G_0$ . This peak is commonly associated with contamination from environmental sources or hydrocarbons, as confirmed in our previous study [35]. It's crucial to note that the count density of this peak in the contaminated histogram, close to  $2 \cdot 10^{-1} G_0$ , is almost an order of magnitude higher compared to the clean traces. The latter, in contrast, show no distinct peak in this region, only baseline counts. Furthermore, another prominence exists in the contaminated data, which also disrupts the linear trends observed in the clean traces within the  $10^{-2} G_0$  to  $10^{-3} G_0$  range. This deviation is also attributed to contamination, given that no molecules were intentionally deposited and, under ideal conditions, no counts should appear in this range.

With all this data, we can quantitatively determine the percentage of clean and contaminated traces obtained after applying our classification process. This statistical analysis allows us to evaluate whether

our samples progressively degrade over time due to environmental exposure, or if contamination occurs immediately upon exposure.

The following Table 3 summarizes the total number of traces analyzed, along with the relative proportions of clean and contaminated traces:

**Table 3.** Statistical summary of trace classification.

Total traces	Clean traces (%)	Contaminated traces (%)
5024	3912 (77.8%)	1112 (22.2%)

Altogether, the combined visual and statistical analyses support the effectiveness of our DBSCAN-based classification. The subset of clean traces not only excludes contamination artifacts but also reproduces features known from high-purity, low-temperature measurements. This reinforces the reliability of our approach and demonstrates that meaningful structural and electronic information can be extracted from ambient-condition experiments, provided that appropriate data-cleaning procedures are applied.

## 5. Conclusions

Therefore, we conclude that we have developed a robust trace classification procedure capable of separating clean and contaminated traces through a two-step DBSCAN-based protocol. This methodology enables a quantitative evaluation of the proportion of clean versus contaminated traces within a dataset. In our case, we determined that 77.8% of the traces are completely clean, meaning they do not exhibit any counts or plateaus in the range of  $[8 \cdot 10^{-1}, 1 \cdot 10^{-3}] G_0$ .

Importantly, this constitutes the first compact device that, via electronic transport measurements, can quantify the presence of a single contaminant molecule. Although it does not provide information about the chemical identity of the contaminant, it reliably assesses its abundance.

These findings demonstrate that a large fraction of the traces are of high purity. Furthermore, our methodology could be extended to evaluate how different metals degrade upon exposure to environmental conditions. This, in turn, would support decisions about whether molecular deposition is justified or help define the minimum contamination threshold required for reproducible and reliable molecular-scale measurements.

**Author Contributions:** G.P. contributed to the methodology, software development, investigation, and data analysis. C.S. contributed to the conceptualization, methodology, validation, formal analysis, investigation, resources, supervision, writing—original draft, writing—review and editing, and funding acquisition.

**Funding:** This research received no external funding

**Acknowledgments:** This work received financial support from the Generalitat Valenciana through CIDEVG/2022/45 and the Spanish Government by PID2023-146660OB-I00. This research is an integral part of the Advanced Materials program, supported by MCIN with funding from the European Union NextGenerationEU (PRTR-C17.I1) and the Generalitat Valenciana (MFA/2022/045). We also acknowledge funding from MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF/EU) under project PID2023-146660OB-I00. Finally, the authors extend their gratitude for insightful discussions with Prof. Carlos Untiedt, Dr. Wynand Denam, Dr. Tamara de Ara, Andrés Martínez, Juan Pablo Cuenca, and Enrique Guzmán.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

MCBJ	Mechanical Controllable Break Junctions
DBSCAN	stands for Density-Based Spatial Clustering of Applications with Noise

## References

1. Agrait, N.; Yeyati, A.L.; Van Ruitenbeek, J.M. Quantum properties of atomic-sized conductors. *Phys. Rep.* **2003**, *377*, 81–279.
2. Cuevas, J.C.; Scheer, E. *Molecular Electronics*, 2nd ed.; WORLD SCIENTIFIC, 2017. <https://doi.org/10.1142/10598>.
3. Evers, F.; Korytár, R.; Tewari, S.; van Ruitenbeek, J.M. Advances and challenges in single-molecule electron transport. *Rev. Mod. Phys.* **2020**, *92*, 035001–035065. <https://doi.org/10.1103/RevModPhys.92.035001>.
4. Pascual, J.I.; Méndez, J.; Gómez-Herrero, J.; Baró, A.M.; García, N.; Binh, V.T. Quantum contact in gold nanostructures by scanning tunneling microscopy. *Phys. Rev. Lett.* **1993**, *71*, 1852–1855. <https://doi.org/10.1103/PhysRevLett.71.1852>.
5. Krans, J.M.; Muller, C.J.; Yanson, I.K.; Govaert, T.C.M.; Hesper, R.; van Ruitenbeek, J.M. One-atom point contacts. *Phys. Rev. B* **1993**, *48*, 14721–14724. <https://doi.org/10.1103/PhysRevB.48.14721>.
6. Krans, J.M.; van Ruitenbeek, J.M. Subquantum conductance steps in atom-sized contacts of the semimetal Sb. *Phys. Rev. B* **1994**, *50*, 17659–17661. <https://doi.org/10.1103/PhysRevB.50.17659>.
7. Krans, J.M.; van Ruitenbeek, J.M.; Fisun, V.V.; Yanson, I.K.; de Jongh, L.J. The signature of conductance quantization in metallic point contacts. *Nature* **1995**, *375*, 767–769. <https://doi.org/10.1038/375767a0>.
8. Landauer, R. Spatial Variation of Currents and Fields Due to Localized Scatterers in Metallic Conduction. *IBM Journal of Research and Development* **1957**, *1*, 223–231. <https://doi.org/10.1147/rd.13.0223>.
9. van Wees, B.J.; van Houten, H.; Beenakker, C.W.J.; Williamson, J.G.; Kouwenhoven, L.P.; van der Marel, D.; Foxon, C.T. Quantized conductance of point contacts in a two-dimensional electron gas. *Phys. Rev. Lett.* **1988**, *60*, 848–850. <https://doi.org/10.1103/PhysRevLett.60.848>.
10. Agrait, N.; Rodrigo, J.G.; Vieira, S. Conductance steps and quantization in atomic-size contacts. *Phys. Rev. B* **1993**, *47*, 12345–12348. <https://doi.org/10.1103/PhysRevB.47.12345>.
11. Pan, X.; Qian, C.; Chow, A.; Wang, L.; Kamenetska, M. Atomically precise binding conformations of adenine and its variants on gold using single molecule conductance signatures. *The Journal of Chemical Physics* **2022**, *157*, 234201. <https://doi.org/10.1063/5.0103642>.
12. Smit, R.H.M.; Noat, Y.; Untiedt, C.; Lang, N.D.; van Hemert, M.C.; van Ruitenbeek, J.M. Measurement of the conductance of a hydrogen molecule. *Nature* **2002**, *419*, 906–909. <https://doi.org/10.1038/nature01103>.
13. Kim, Y.; Pietsch, T.; Erbe, A.; Belzig, W.; Scheer, E. Benzenedithiol: A Broad-Range Single-Channel Molecular Conductor. *Nano Letters* **2011**, *11*, 3734–3738, [\[https://doi.org/10.1021/nl201777m\]](https://doi.org/10.1021/nl201777m). PMID: 21805977, <https://doi.org/10.1021/nl201777m>.
14. Tewari, S.; Sabater, C.; van Ruitenbeek, J. Identification of vibration modes in single-molecule junctions by strong inelastic signals in noise. *Nanoscale* **2019**, *11*, 19462–19467. <https://doi.org/10.1039/C9NR05774A>.
15. Reed, M.; Zhou, C.; Muller, C.; Burgin, T.; Tour, J. Conductance of a molecular junction. *Science* **1997**, *278*, 252–254. Cited by: 3424, <https://doi.org/10.1126/science.278.5336.252>.
16. Xu, B.; Tao, N.J. Measurement of Single-Molecule Resistance by Repeated Formation of Molecular Junctions. *Science* **2003**, *301*, 1221–1223, [\[https://www.science.org/doi/pdf/10.1126/science.1087481\]](https://www.science.org/doi/pdf/10.1126/science.1087481). <https://doi.org/10.1126/science.1087481>.
17. Herrer, I.L.; Ismael, A.K.; Milán, D.C.; Vezzoli, A.; Martín, S.; González-Orive, A.; Grace, I.; Lambert, C.; Serrano, J.L.; Nichols, R.J.; et al. Unconventional Single-Molecule Conductance Behavior for a New Heterocyclic Anchoring Group: Pyrazolyl. *The Journal of Physical Chemistry Letters* **2018**, *9*, 5364–5372, [\[https://doi.org/10.1021/acs.jpclett.8b02051\]](https://doi.org/10.1021/acs.jpclett.8b02051). PMID: 30160491, <https://doi.org/10.1021/acs.jpclett.8b02051>.
18. Montenegro-Pohlhammer, N.; Sánchez-de Armas, R.; Calzado, C.J.; Borges-Martínez, M.; Cárdenas-Jirón, G. A photo-induced spin crossover based molecular switch and spin filter operating at room temperature. *Dalton Trans.* **2021**, *50*, 6578–6587. <https://doi.org/10.1039/D1DT00078K>.
19. de Ara, T.; Hsu, C.; Martínez-García, A.; Baciu, B.C.; Bronk, P.J.; Ornago, L.; van der Poel, S.; Lombardi, E.B.; Guijarro, A.; Sabater, C.; et al. Evidence of an Off-Resonant Electronic Transport Mechanism in Helicenes. *The Journal of Physical Chemistry Letters* **2024**, *15*, 8343–8350, [\[https://doi.org/10.1021/acs.jpclett.4c01425\]](https://doi.org/10.1021/acs.jpclett.4c01425). PMID: 39110695, <https://doi.org/10.1021/acs.jpclett.4c01425>.
20. Singh, A.K.; Martin, K.; Mastropasqua Talamo, M.; Houssin, A.; Vanthuyne, N.; Avarvari, N.; Tal, O. Single-molecule junctions map the interplay between electrons and chirality. *Nature Communications* **2025**, *16*, 1759. <https://doi.org/10.1038/s41467-025-56718-9>.
21. Cabosart, D.; El Abbassi, M.; Stefani, D.; Frisenda, R.; Calame, M.; van der Zant, H.S.J.; Perrin, M.L. A reference-free clustering method for the analysis of molecular break-junction mea-

surements. *Applied Physics Letters* **2019**, *114*, 143102, [[https://pubs.aip.org/aip/apl/article-pdf/doi/10.1063/1.5089198/13147123/143102\\_1\\_online.pdf](https://pubs.aip.org/aip/apl/article-pdf/doi/10.1063/1.5089198/13147123/143102_1_online.pdf)]. <https://doi.org/10.1063/1.5089198>.

- 22. Liu, B.; Murayama, S.; Komoto, Y.; Tsutsui, M.; Taniguchi, M. Dissecting Time-Evolved Conductance Behavior of Single Molecule Junctions by Nonparametric Machine Learning. *The Journal of Physical Chemistry Letters* **2020**, *11*, 6567–6572, [<https://doi.org/10.1021/acs.jpcllett.0c01948>]. PMID: 32668163, <https://doi.org/10.1021/acs.jpcllett.0c01948>.
- 23. Lin, L.; Tang, C.; Dong, G.; Chen, Z.; Pan, Z.; Liu, J.; Yang, Y.; Shi, J.; Ji, R.; Hong, W. Spectral Clustering to Analyze the Hidden Events in Single-Molecule Break Junctions. *The Journal of Physical Chemistry C* **2021**, *125*, 3623–3630, [<https://doi.org/10.1021/acs.jpcc.0c11473>]. <https://doi.org/10.1021/acs.jpcc.0c11473>.
- 24. Bro-Jørgensen, W.; Hamill, J.M.; Bro, R.; Solomon, G.C. Trusting our machines: validating machine learning models for single-molecule transport experiments. *Chem. Soc. Rev.* **2022**, *51*, 6875–6892. <https://doi.org/10.1039/D1CS00884F>.
- 25. Komoto, Y.; Ryu, J.; Taniguchi, M. Machine learning and analytical methods for single-molecule conductance measurements. *Chem. Commun.* **2023**, *59*, 6796–6810. <https://doi.org/10.1039/D3CC01570J>.
- 26. 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 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996, KDD'96, p. 226–231.
- 27. Ester, M. Density-Based Clustering. In *Data Clustering*, 1st ed.; Chapman and Hall/CRC, 2014; p. 17.
- 28. Schubert, E.; Sander, J.; Ester, M.; Kriegel, H.P.; Xu, X. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems (TODS)* **2017**, *42*, 19:1–19:21. <https://doi.org/10.1145/3068335>.
- 29. Ornago, L. Complexity of Electron Transport in Nanoscale Molecular Junctions. Dissertation (tu delft), Delft University of Technology, Delft, Netherlands, 2023. Supervisors: H.S.J. van der Zant, F.C. Grozema.
- 30. Cuenca, J.P.; de Ara, T.; Martinez-Garcia, A.; Guzman, F.; Sabater, C. Exploring Three-Atom-Thick Gold Structures as a Benchmark for Atomic-Scale Calibration of Break-Junction Systems. *Arxiv* **2025**, *X*, X. <https://doi.org/X>.
- 31. Gimzewski, J.K.; Möller, R. Transition from the tunneling regime to point contact studied using scanning tunneling microscopy. *Phys. Rev. B* **1987**, *36*, 1284–1287. <https://doi.org/10.1103/PhysRevB.36.1284>.
- 32. Untiedt, C.; Caturla, M.J.; Calvo, M.R.; Palacios, J.J.; Segers, R.C.; van Ruitenbeek, J.M. Formation of a Metallic Contact: Jump to Contact Revisited. *Phys. Rev. Lett.* **2007**, *98*, 206801. <https://doi.org/10.1103/PhysRevLett.98.206801>.
- 33. Sabater, C.; Caturla, M.J.; Palacios, J.J.; Untiedt, C. Understanding the structure of the first atomic contact in gold. *Nanoscale Research Letters* **2013**, *8*, 257. <https://doi.org/10.1186/1556-276X-8-257>.
- 34. Xu, B.; Tao, N.J. Measurement of Single-Molecule Resistance by Repeated Formation of Molecular Junctions. *Science* **2003**, *301*, 1221–1223, [<https://www.science.org/doi/pdf/10.1126/science.1087481>]. <https://doi.org/10.1126/science.1087481>.
- 35. de Ara, T.; Sabater, C.; Borja-Espinosa, C.; Ferrer-Alcaraz, P.; Baciu, B.C.; Guijarro, A.; Untiedt, C. Signature of adsorbed solvents for molecular electronics revealed via scanning tunneling microscopy. *Materials Chemistry and Physics* **2022**, *291*, 126645. <https://doi.org/https://doi.org/10.1016/j.matchemphys.2022.126645>.
- 36. Martinez-Garcia, A.; de Ara, T.; Pastor-Amat, L.; Untiedt, C.; Lombardi, E.B.; Dednam, W.; Sabater, C. Unraveling the Interplay between Quantum Transport and Geometrical Conformations in Monocyclic Hydrocarbons' Molecular Junctions. *The Journal of Physical Chemistry C* **2023**, *127*, 23303–23311, [<https://doi.org/10.1021/acs.jpcc.3c05393>]. <https://doi.org/10.1021/acs.jpcc.3c05393>.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.