THE JOURNAL OF PHYSICAL CHEMISTRY C

pubs.acs.org/JPCC Article

Unsupervised Segmentation-Based Machine Learning as an Advanced Analysis Tool for Single Molecule Break Junction Data

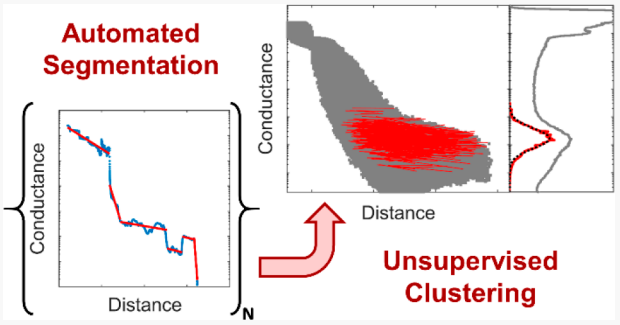
Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

Nathan D. Bamberger, Jeffrey A. Ivie, Keshaba N. Parida, Dominic V. McGrath, and Oliver L. A. Monti*

Cite This: *J. Phys. Chem. C* 2020, 124, 18302−18315 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Improved understanding of charge-transport in single molecules is essential for harnessing the potential of molecules, e.g., as circuit components at the ultimate size limit. However, interpretation and analysis of the large, stochastic data sets produced by most quantum transport experiments remain an ongoing challenge to discovering much-needed structure−property relationships. Here, we introduce *segment clustering*, a novel unsupervised hypothesis generation tool for investigating single molecule break junction distance−conductance traces. In contrast to previous machine learning approaches for single molecule data, segment clustering identifies groupings of similar *pieces of traces* instead of *entire traces*. This offers a new and advantageous perspective into data set structure because it facilitates the identification of meaningful local trace behaviors that may otherwise be obscured by random fluctuations over longer distance scales. We illustrate the power and broad applicability of this approach with two case studies that address common challenges encountered in single molecule studies: First, segment clustering is used to extract primary molecular features from a varying background to increase the precision and robustness of conductance measurements, enabling small changes in conductance in response to molecular design to be identified with confidence. Second, segment clustering is applied to a known data mixture to qualitatively separate distinct molecular features in a rigorous and unbiased manner. These examples demonstrate two powerful ways in which segment clustering can aid in the development of structure−property relationships in molecular quantum transport, an outstanding challenge in the field of molecular electronics.

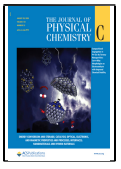


1. INTRODUCTION

Ever since 1974, when Aviram and Ratner proposed using a single molecule to rectify current,[1] the nanoscale transport community has pursued the goal of molecular-based circuitry to take advantage of the small size, enormous design space, and potential low manufacturing costs of circuit components composed of individual molecules.[2] However, in order to create functional devices that can capitalize on these advantages, it is first necessary to understand the fundamental physics and design principles underlying charge transport in single molecule systems. This understanding is most commonly gained using either mechanically controlled break junctions (MCBJs)[3−10] or scanning tunneling microscope break junctions (STM-BJs),[11−15] techniques which pull apart a thin metal bridge, typically made from gold, to form a nanogap, while simultaneously applying a small bias across the bridge or gap and recording the resulting current. The changes in current when individual molecules bridge the gap provide insight into the electrical nonequilibrium properties of single-molecule circuit components.

Most commonly, such experiments yield "breaking traces", in which the junction conductance $G = I/V$ is recorded as a function of stretching distance during the breaking process. Figure 1a contains example breaking traces collected using our MCBJ setup with the molecule OPV3-2BT-H (Chart 1), plotted on a log−linear scale in order to capture the large dynamic range of possible molecular conductances, as is standard in the field. These examples illustrate three characteristic features of breaking traces: (1) Just before rupture, a plateau occurs at the conductance value corresponding to a single atomic point contact. For Au electrodes, this value is 77.48 μS,[16] and denoted 1 G_0; (2) when no molecule is bound in the junction (blue traces), the conductance is solely due to

Received: April 24, 2020
Revised: July 3, 2020
Published: July 22, 2020



https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315

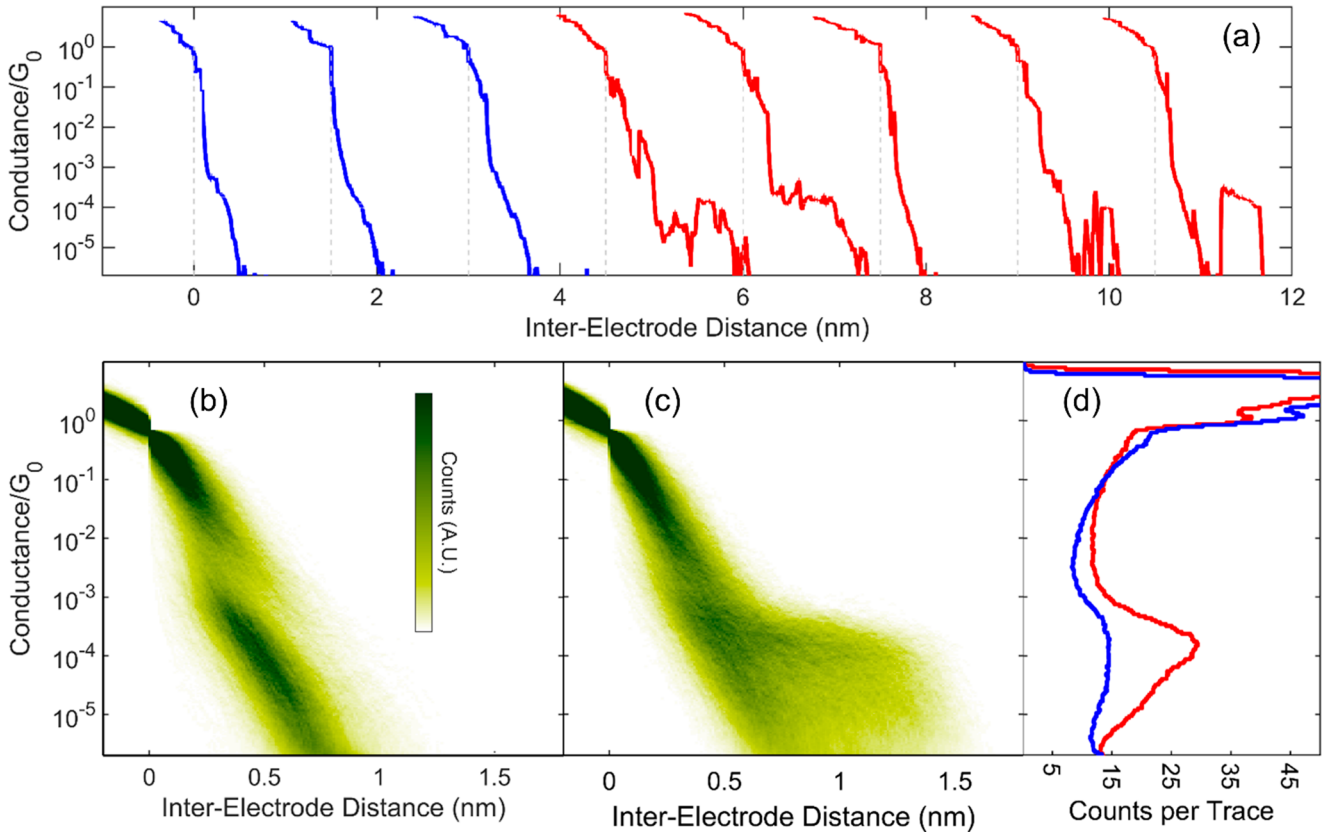
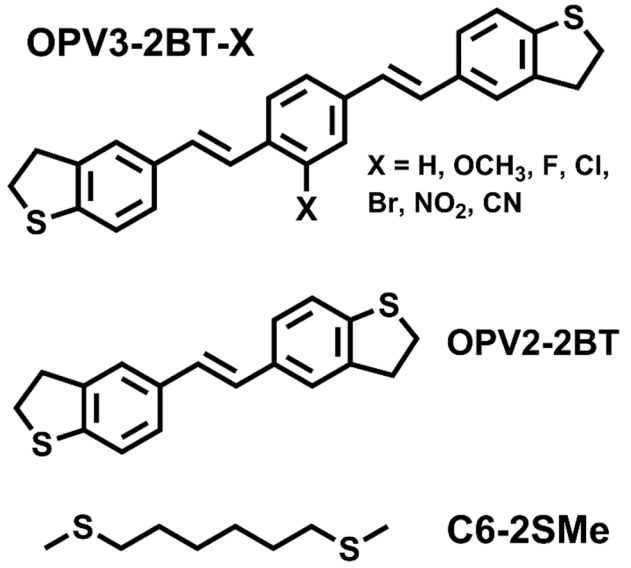


Figure 1. Break junction data collected with the molecule OPV3-2BT-H. (a) Selected breaking traces from before (blue) and after (red) the addition of molecules, offset by 1.5 nm for clarity. The blue traces illustrate exponential tunneling decay in an empty nanogap (linear on a logarithmic scale), while the red traces illustrate molecular plateaus and their variability. (b, c) 2D histograms of 7122 and 6280 consecutive breaking traces collected before and after the addition of molecules, respectively. Part b exhibits a clear tunneling decay feature below 10^{-3} G_0, while part c exhibits a pronounced molecular feature extending out to ∼1.5 nm at $\sim 10^{-4} G_0$. (d) 1D histograms for the data sets in (b) (blue) and (c) (red). While both histograms display a sharp peak at $\sim 1 G_0$ from the single gold point contact plateaus, only the histogram collected after molecular addition displays a broad peak at $\sim 10^{-4} G_0$ due to the presence of molecules.

Chart 1. Structures of Molecules Considered in This Work and Their Associated Abbreviations



tunneling and decays exponentially; and (3) when a molecule *is* bound in the junction (red traces), the conductance is roughly constant (though potentially fluctuating) over the length of the molecule, forming a "molecular plateau".

However, because of the stochastic nature of the breaking process, molecular conformation, and molecular diffusion in and out of the junction, individual molecular traces are highly variable. In particular, plateaus for the same molecule can vary by over an order of magnitude in conductance (e.g., first two red traces in Figure 1a); some traces collected in the presence of molecules do not display any molecular plateau at all (e.g., third red trace in Figure 1a); and molecular plateaus may break off and re-form within the same trace (e.g., last two red traces in Figure 1a). In order to capture this variability, thousands of traces are collected under the same experimental conditions. A set of traces can then be summarized by a 2D histogram (Figure 1b,c), which shows the frequency of observing each pair of interelectrode distance and log(conductance) values; or a 1D histogram (Figure 1d), which is obtained from the 2D histogram by integrating out the interelectrode distance dimension to "collapse" all of the data onto the log-(conductance) axis.

While such histograms usefully summarize the ensemble of single molecule conductance behaviors, they obscure likely meaningful differences within and among different molecular constructs that could be harnessed to advance a host of intriguing molecular electronics research directions. At present, 1D histograms are often used to determine a single "peak" or

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315

"most probable" conductance for a given molecule,[17−27] and 2D histograms have been used to separate molecular features that may correspond to distinct physical phenomena, such as different binding modes.[9,28−32] However, the broad features found in these histograms make it difficult to confidently separate features without introducing bias, and the complex "background" signature, composed of tunneling decay and broken molecular plateaus, makes it hard to robustly fit molecular peaks. These inter-related challenges have motivated several research groups to develop automated clustering and data-sorting methods for analyzing breaking traces[33−41] and related data.[15,42−44] Broadly speaking, the goal of these approaches is to partition a large data set of highly varied traces into separate groupings in order to improve the robustness of peak conductance measurements and/or to identify distinct junction behaviors. Using an automated algorithm to identify clusters of data helps eliminate bias toward seeing only the types of groupings that are expected *a priori*. The clustering approaches developed so far are based on techniques ranging from principal component analysis[34,38] to neural networks,[35,37,38,41,43] and they have found success in separating known features in experimental or simulated data[34,35,37,42,43] and in detecting intriguing subfeatures for further quantitative or qualitative analysis.[33,34,38,41,42,45]

Nearly every published clustering approach applied to breaking traces treats each entire trace as one single object.[15,33−35,37−42,44] This choice implicitly assumes that the overall trajectories that traces follow are nonrandom, and hence such algorithms are best suited for traces that exhibit few unpredictable fluctuations. However, our own experimental data and many published examples suggest that this is often only true over distances much shorter than most molecular lengths. Over longer distances, there are often sudden and unpredictable conductance shifts between mostly linear sections,[32,46−50] and in some instances, such traces constitute the majority of all molecular "plateaus". Whole-trace focused methods can thus easily miss a meaningful subfeature, even one conserved across many traces, if the other parts of those traces differ significantly due to random and uncorrelated behavior. We therefore designed a new approach, "segment clustering", based on the idea of defining *pieces of traces* as the objects to be clustered and, in particular, linearly approximated segments. This definition better matches the empirical structure of trace trajectories in most systems studied so far,[13,51−59] ranging from *in situ* chemical reactions to photoswitching. Segment clustering is thus able to identify the truly conserved features in highly stochastic data sets and has the potential to reveal insights not available to other clustering approaches. Additionally, segment clustering does not require training, like some neural network-based approaches,[35,38,41,43] nor does it rely on criteria that are likely data set-specific, like many filtering-based approaches[15,40,44] and so is expected to be easily generalizable to new data sets.

We emphasize that segment clustering is neither expected nor designed to identify *every* meaningful feature in *every* single molecule data set. Instead, it focuses on one broad category of features—approximately linear trace sections—which are evidently quite common in distance−conductance traces, thus providing a new perspective into data set structure. At the same time, just because segment clustering identifies a given cluster does not, by itself, constitute proof that such a cluster corresponds to a distinct physical behavior. Rather, segment clustering is designed as a *hypothesis generation tool*: by identifying data groupings that may not be obvious to the naked eye and which do not rely on preconceived and potentially flawed notions of meaningful data structure, it can help spawn ideas of what types of behaviors may be present in single-molecule junctions. These ideas can then be tested via additional experiments or targeted data analysis, laying the basis for further insight into the fundamental physics of single-molecule transport.

In the remainder of this paper, we describe our experimental methodology and then explain in detail the motivation and mechanics behind segment clustering. We next present two case studies using our own MCBJ data to illustrate two applications of segment clustering. In the first case study, we show that segment clustering can reliably separate the "primary" molecular feature from a shifting background signal, enabling us to confidently distinguish small changes in conductance across a family of similar molecules. In the second case study, we use a known data mixture to demonstrate that segment clustering can separate molecular features even when they come from overlapping conductance distributions.

2. EXPERIMENTAL SECTION

2.1. Fabrication. Samples for the MCBJ experiment were fabricated by depositing a gold wire on a polyimide-coated phosphor bronze substrate using electron beam evaporation. A 4 nm titanium layer was used to improve adherence of the 80 nm thick gold film. The pattern for gold deposition, including an ∼100 nm wide gold bridge in the center of the wire, was fashioned by electron beam lithography. The gold bridge was then created via O_2/CHF_3 plasma etching of the polyimide to produce an ∼1−2 μm undercut (Figure S1a,b).

2.2. Trace Collection. Samples were clamped and then bent with a push rod placed underneath the gold bridge (Figure S1c). A 100 mV bias was applied across the gold bridge while simultaneously measuring the conductance of the bridge using a custom, high-speed amplifier described previously.[60] A stepper motor (ThorLabs DRV50) was used to move the push rod until the bridge conductance was between 5 and 7 G_0, at which point a linear piezo actuator (ThorLabs PAZ60 or ThorLabs PAS009) was used to break and then re-form the bridge at a rate of 60 μm/s. The motor and the piezo were both controlled with custom LabView software that automatically collected thousands of breaking traces for each sample. The entire bending apparatus is built on a vibrationally isolated laser table to reduce mechanical noise and placed inside a copper Faraday cage to reduce high-frequency electromagnetic noise.

2.3. Molecular Solutions. OPV2-2BT and all OPV3-2BT-X molecules were synthesized on-site, while C6-2SMe was purchased from Sigma-Millipore and used as received. OPV2-2BT and all OPV3-2BT-X molecules were dissolved in dichloromethane (HPLC grade, > 99.8%), and C6-2SMe was dissolved in a mixture of hexanes (reagent grade, > 98.5%). All OPV3-2BT-X solutions were ∼1 μM; both ∼1 μM and ∼10 μM solutions of OPV2-2BT were used (see Supporting Information, Table S3, for details); the C6-2SMe solution was ∼10 μM.

2.4. Running Samples. Each sample was cleaned with O_3/UV immediately before use, and a Kalrez gasket (0.114 in. ID, 0.250 in. OD) was placed around the gold bridge (Figure S1d). Initially, 10 μL of pure dichloromethane or hexanes was deposited inside this gasket using a clean glass syringe for

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315

dichloromethane or a micropipette for hexanes, after which a few thousand breaking traces were collected. Only samples displaying clean breaking and clear tunneling behavior were considered for subsequent experiments. After pausing the LabView program and fully breaking the gold bridge, 10−20 μL of the molecular solution was deposited inside the Kalrez gasket using a clean glass syringe for dichloromethane solutions or micropipette for hexanes solutions, and data acquisition was resumed. For many samples, molecular solution or pure solvent was redeposited multiple times and/or the push rod was fully relaxed prior to restarting the experiment (see Supporting Information, section S.4, for details).

2.5. Initial Data Processing. The voltage applied to the piezo actuator was converted to piezo displacement using a previously performed interferometric calibration. For each sample, the conversion factor between piezo displacement and interelectrode distance was determined by fitting the distribution of tunneling slopes from the traces collected before molecular deposition (see Supporting Information, section S.2, for details), and this conversion factor was applied to all traces collected with that sample. Each breaking trace was aligned at zero interelectrode distance using its last crossing of $0.7G_0$ following the method of Mischenko et al.[61] Breaking traces with no data points between $0.8G_0$ and $1.2G_0$ were excluded from subsequent analysis (typically <1% of the total breaking traces).

3. RESULTS AND DISCUSSION

3.1. Description of Segment Clustering. *3.1.1. Motivation.* A key consideration when deciding how to cluster multidimensional data is what type of object to cluster. In the case of break junction distance−conductance traces, two natural choices are to treat each trace as a single object ("trace clustering", which most approaches[15,33−35,37−42,44] have used so far) or to treat different visited points in distance−conductance space as individual objects ("point clustering", which we used in a previously reported clustering approach[36]). Neither choice is inherently superior to the other. Instead, each has potential advantages that are best understood by considering the question of how much "history" distance−conductance traces have—i.e., how much a trace's behavior at one distance is correlated with its behavior at a previous or future distance. If traces randomly transition between different stable distance/conductance configurations (i.e., traces have "no history"), then point clustering can better identify these stable configurations whereas trace clustering may get confused by the random trajectories. On the other hand, if trace trajectories are highly nonrandom (i.e., traces have "significant history"), then trace clustering can identify groupings of similar trajectories that point clustering will likely miss.

In our experience, however, real experimental traces fall somewhere between these extremes: they display "partial" or "local" history. To illustrate this, we calculate the correlation coefficient between the conductances of all traces at one specific distance with their conductances at a second distance. This is repeated for each pair of distances, and the results are summarized in a "distance correlation histogram", shown in Figure 2 for one of our OPV3-2BT-H data sets. This plot shows that while conductances are strongly correlated at close distances, there is essentially no correlation over longer distances. Similar behavior was found in all of the single molecule data sets we examined, suggesting that trace history is

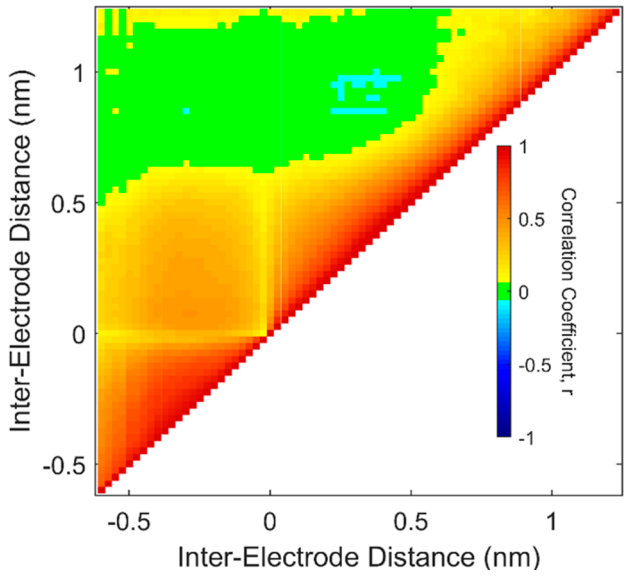


Figure 2. Distance correlation histogram for the OPV3-2BT-H data set from Figure 1c, showing the Pearson's correlation coefficient between the conductances of all traces at each pair of distances. While trace conductances are highly correlated over short distances, this correlation quickly fades with distance, demonstrating that trace "history" is important only locally, not globally.

only relevant over short pieces of an entire trace. This is consistent with investigations of the dynamics of single-molecule junctions held at a fixed distance,[62−67] which have found that junction conductance is relatively stable over short periods of time but jumps unpredictably between different levels over longer time windows. Therefore, both trace clustering and point clustering fail to fully and appropriately capture the empirical balance between predictable and random junction behaviors, limiting the insight they can provide. This motivates the development of a novel clustering approach in which *pieces of traces* are the type of object clustered.

While certain theoretical models predict significantly curved trace features, experimental traces collected from an extremely wide variety of molecular systems[13,51−59] appear (on a logarithmic conductance scale) to be composed mainly of sudden changes between fairly linear sections. Segment clustering is therefore based on using a piecewise-linear approximation to determine where to separate each trace into different sections. This design choice helps ignore noisy high frequency components and instead focuses attention on the principal features of each trace. Additionally, linear segments are a computationally efficient way to represent a trace, since a handful of linear segments can well-approximate thousands of individual data points (e.g., Figure 3a). Implementing segment clustering via this approach consists of four major steps, summarized in Figure 3: segmentation, parametrization, calculating the overall clustering structure, and extracting specific clusters. Where appropriate, we employ established algorithms for these individual steps in order to increase confidence in the robustness of the overall approach, which combines these algorithms in a new way.

3.1.2. Segmentation. The goal of segmentation is to break each trace into consecutive sections such that each section can be well-represented by a linear segment and corresponds to a meaningful piece of the trace structure. Because this goal is common in data-mining applications, several algorithms have

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315

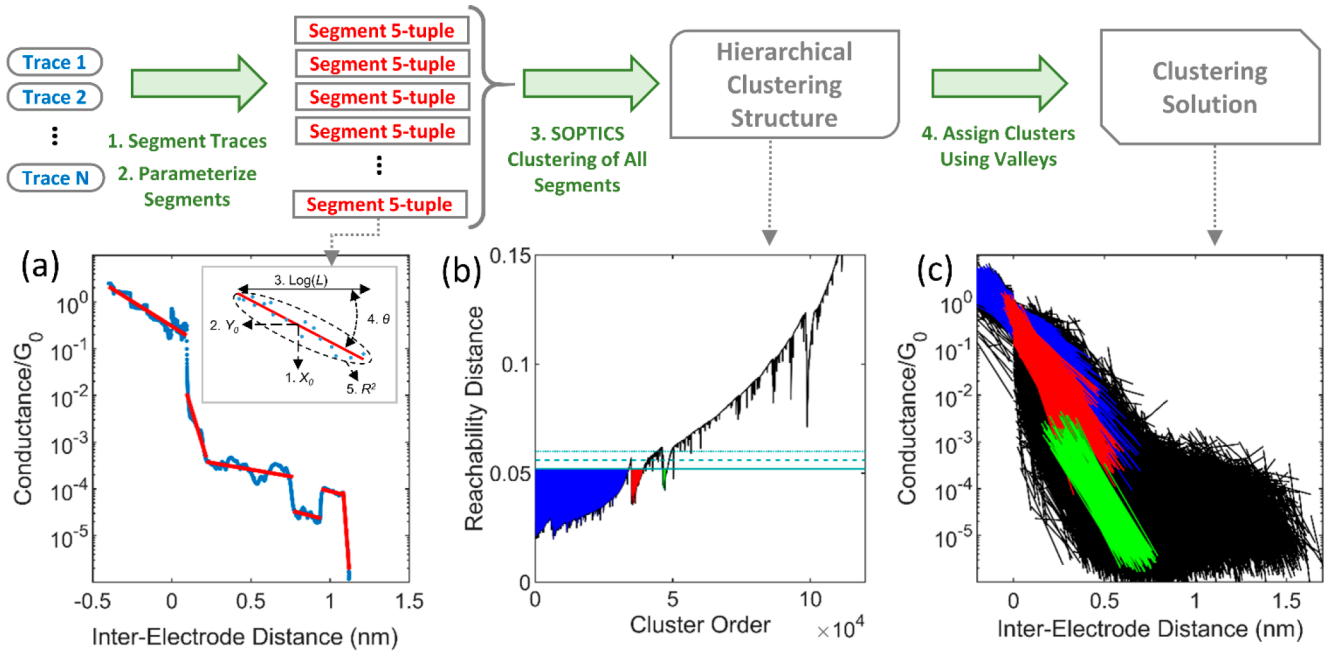


Figure 3. Summary of the segment clustering process. (a) Each breaking trace in a data set is first approximated with a series of linear segments using BUS with the greedy iterative L-method, and then each segment is parametrized to produce a 5-tuple (see inset). (b) Next, the set of 5-tuples for all segments from all traces in the data set are clustered using the SOPTICS algorithm, producing a hierarchical clustering structure that can be visualized using a reachability plot in which valleys correspond to clusters. Finally, a specific clustering solution can be extracted by making a cut through the reachability plot and assigning the points in each valley dipping below that level to a separate cluster, while assigning any points with reachability distances greater than the cut to a catch-all "noise cluster". Extracting at the solid blue line in part b produces the clustering solution in part c, with each valley dipping below the line filled in with color to match its corresponding cluster of segments and the noise cluster segments shown in black.

been developed to try to optimally represent time-series data with a set of piece-wise linear segments.[68] After first applying consistent starting and ending criteria to each trace (see Supporting Information, section S.3.1, for details), we employ the "bottom-up segmentation" (BUS) algorithm because it is conceptually simple and has been found to produce excellent and robust results for data from a variety of contexts.[68,69] Briefly, BUS starts by perfectly representing a time series of n points with $n/2$ two-point segments. Next, BUS iteratively merges the pair of neighboring segments that will least increase the error of the overall segment approximation, repeating until some stopping criteria is met. At each step, every segment is constructed as the linear regression line for the data points it is currently representing, and the error for each segment is taken as the sum of the squared residuals from that regression line.[68]

For our stopping criteria, we use the "greedy iterative L-method", which was found to work well on a wide variety of test data sets.[69] Briefly, this method first performs the merging process to completion, so that a plot of the number of segments remaining vs the error gained at each merge step may be constructed. An iterative fitting process is then used to locate the optimal number of segments by identifying the point at which more segments produce diminishing returns in terms of error reduction. Applying this combination of BUS and the greedy iterative L-method to distance-log(conductance) traces produces convincing segmentation solutions (e.g., Figure 3a). In addition to the examples presented by the developers of the greedy iterative L-method,[69] testing on our own single molecule data demonstrates that this method is quite robust (see Supporting Information, section S.5.5).

3.1.3. Parametrization. Because clustering algorithms need to compute distances between the objects to be clustered, it is necessary to first extract "features" that can be used to represent each object as a point in a metric space. In order to avoid well-known challenges to clustering in high-dimensional spaces (the "curse of dimensionality")—such as increasingly sparse data and a nonintuitive breakdown in the concept of nearest neighbors[70]—it is preferable to choose a minimal set of features while still capturing most of the important information about each trace piece. Our segmentation approach already produces linear segments which capture most trace variation—e.g., 82% for the data set in Figure 1c—and so parametrizing these linear segments produces features that are both efficient and easy to interpret. We therefore convert each segment into a 5-tuple consisting of four parameters that uniquely describe each linear fit and a fifth parameter to describe the fit quality.

The specific parameters chosen to represent each segment are illustrated in Figure 3a. The first two parameters—the center of a segment on the interelectrode distance axis, X_0, and on the log(conductance) axis, Y_0—succinctly represent where each segment is located. Another key segment attribute is its length, L. However, in absolute terms, long segments will tend to differ by more than short ones, making it difficult to form clusters of long segments. We therefore use the logarithm of the length of a segment on the interelectrode distance axis, $\log(L)$, as our third parameter, so that the difference between two segments on this dimension depends on their ratio. To represent how tilted a segment is, the angle that it makes with the horizontal, θ, is used as the fourth parameter. This angle is less sensitive to outliers than a segment's raw slope due to the nature of the *arctan* function. Finally, to represent the linearity of each trace piece, we include the coefficient of determination, R^2, of each segment vis-à-vis the portion of raw data it represents as the fifth parameter. This helps capture additional

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315

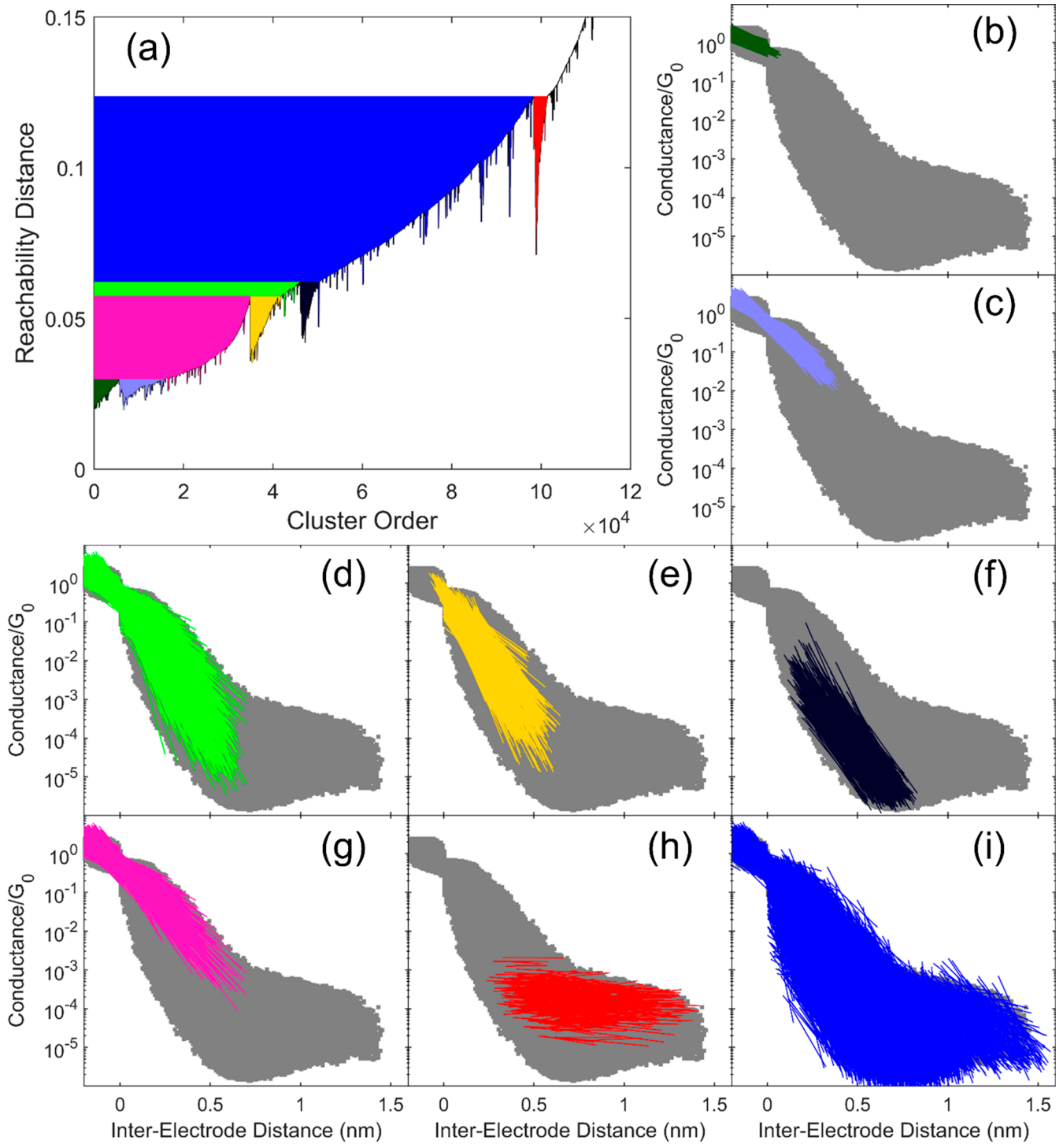


Figure 4. (a) Same reachability plot as in Figure 3b, but color-coded to indicate the maximum size of each valley containing at least 1% of all clustered points. Valleys are filled in hierarchically: the pink valley, e.g., contains the dark green and lavender valleys, the green valley contains the pink and yellow valleys, *etc.* (b–i) "Full-valley clusters" corresponding to each color-coded valley from part a, with segments assigned to the cluster plotted in color on top of the overall data set distribution in gray.

information about mild segment curvature and/or the magnitude of high-frequency noise and is important for differentiating the few segments that are not well-approximated as linear. These five parameters are each measured in different units, so before clustering, each must be standardized so that differences computed along different dimensions are comparable. In order to minimize the influence of outliers, we use the range of the middle 80% for each parameter to carry out this standardization (see Supporting Information, section S.3.2, for details).

3.1.4. Calculating the Overall Clustering Structure. Many clustering algorithms can be applied to a set of 5-tuples, and each has its own advantages and disadvantages.[71] For this work, we employ the ordering points to infer cluster structure (OPTICS) algorithm based on the following advantages relevant to our specific context: (1) it can detect clusters of

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315

arbitrary shape and is not biased toward spherical clusters like other common algorithms;[71,72] we acknowledge that this necessarily brings along a danger that dissimilar groups of data may end up in the same cluster if there is a continuous spread of data between them; (2) it has a limited number of parameters; (3) it does not require the number of clusters to be specified as an input parameter, unlike many popular algorithms such as K-means, BIRCH, *etc.*;[72] and (4) instead of a single partitioning, OPTICS produces a clustering hierarchy in which subclusters are contained within clusters, providing relevant insight into the data structure (see below). To overcome its poor computational scalability on large data sets, we employ a variation called speedy-OPTICS (SOPTICS) in which random projections are used to dramatically reduce the clustering time while producing nearly identical results to the original algorithm.[73]

OPTICS/SOPTICS clustering works by starting at a random data point and then iteratively proceeding to the next unvisited point that is closest to any point visited so far.[36,74] This journey is represented by a "reachability plot" (Figure 3b) in which the distance to the next point (the "reachability distance") is plotted against the order in which the points were visited (the "cluster order"). *Valleys* in the reachability plot intuitively correspond to *clusters* of data points, because the points in a valley are relatively close to each other but relatively far from points outside of the valley.[74] A reachability plot thus visually represents the overall hierarchical structure of a data set, as valleys may contain subvalleys which themselves can contain sub-subvalleys, and so on. We refer to the reachability plot and its associated information as the "clustering output" for a given data set.

In our implementation, SOPTICS relies on four parameters: c_L, c_P, *minSize*, and *minPts*. The first three parameters are related to how SOPTICS approximates the original OPTICS algorithm and, when in a reasonable range, they each have an extremely minimal effect on the clustering results. We thus assign fixed values to each of these parameters (see Supporting Information section S.3.3, for details). The fourth parameter, *minPts*, is the one holdover from OPTICS (SOPTICS does not require the generating distance parameter ε); it is related to how the data density in 5-dimensional standardized parameter space is estimated at each point, and affects how "jagged" the reachability plot is.[74] While *minPts* is the most important parameter for OPTICS/SOPTICS, its abstract definition makes it difficult to assign rationally without a deep understanding of the data under consideration. In acknowledgment of this uncertainty, we recluster each data set using 12 different values of *minPts* (35, 45, 55, 65, 75, 85, 95, 105, 115, 125, 135, and 145). We then use the variation between these 12 clustering outputs as a measure of the uncertainty in the exact boundaries of an extracted cluster. In practice, this variation is quite limited, implying that segment clustering is not overly sensitive to the value of *minPts*. Finally, because OPTICS/SOPTICS is a density-based clustering algorithm and longer segments represent more raw data points than shorter segments, we find that clustering results are improved if, in the density calculations, we weight each segment according to its length (see Supporting Information, section S.3.4, for details).

3.1.5. Extracting Specific Clusters. In order to extract specific clusters from a given clustering output, a cut is made across the reachability plot (e.g., Figure 3b), and the points in each valley dipping below the cut are assigned to a separate cluster, while all points with reachability distances larger than the cut are assigned to a catch-all "noise cluster" (e.g., Figure 3c). We refer to the specific set of clusters generated by a given cut as a "clustering solution". Thus, while the hierarchical nature of OPTICS/SOPTICS is a distinct advantage, it also presents an interpretation challenge, because a single clustering *output* can have many different clustering *solutions* based on different extraction levels.

Meaningful extraction levels can be chosen using the concept of ξ-steepness[74] or by employing an internal cluster validation index,[75,76] but these strategies introduce ambiguity in the form of what value of ξ to use or which index to employ, and many validation indices are expensive to compute. We therefore introduce a new strategy motivated by the observation that the clustering solutions at most extraction levels are extremely similar to one another. For example, Figure 3c shows the clustering solution obtained by extracting at the solid line in Figure 3b. If this extraction level is increased to the dashed line, the only change is that each valley grows slightly, with a few segments moving into those clusters from the noise cluster. The clustering solution will only *qualitatively* change if the extraction level is raised, for example, to the dotted line, where the red and blue valleys/clusters will merge into one. In the context of segment clustering, we are interested in categorizing as many data points as possible, so we extract each individual valley at the highest extraction level before it merges with a neighboring valley to produce what we call "full-valley clusters". If a minimum valley size is then set, an entire clustering output can be efficiently summarized with just a handful of full-valley clusters (Figure 4). This allows us to still examine the hierarchical structure of a clustering output without having to consider an unmanageable number of different solutions. This novel extraction strategy works especially well in the present context because valleys tend to be quite sharp (e.g., Figure 4a), and its robustness is validated by the fact that it successfully identifies equivalent clusters in the multiple clustering outputs for each data set (see Supporting Information, section S.6, for details). However, we note that this extraction approach is not fundamental to segment clustering, and so other methods can be substituted if full-valley clusters were to exhibit shortcomings on new types of data sets. The minimum valley size should be set according to the specific context and what types of clusters a user is interested in; we have found that a minimum size of 1% of the total number of data points (after length-weighting) often works well.

3.2. Using Segment Clustering to Distinguish the Conductances of Similar Molecules. In structure−property investigations of single molecule conductance, it is common to determine a single "most probable" conductance for each molecule by fitting the molecular peak in the 1D histogram.[17−20,22,24,26] The peak value is then identified as the molecular conductance, and it is often compared across different molecules or with first-principles calculations. However, because the molecular signal is necessarily convolved with a "background" signal due to traces in which no molecule was bound or in which the molecule detaches and reattaches multiples times (e.g., Figure 1a), molecular peaks in 1D histograms tend to have complex, asymmetric shapes (e.g., Figure 1d). Fitting these peaks thus requires arbitrary and ill-motivated restrictions and/or background subtraction. Moreover, it has been shown that the molecular peak can vary significantly between repeated measurements under identical

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315

conditions,[33] likely due in part to uncontrolled variation of this "background" signal. Using data collected from a series of OPV3-2BT-X molecules (Chart 1), we show how segment clustering can help address these twin challenges by separating the primary molecular feature from the background signal, enabling subtle conductance differences to be identified with confidence.

3.2.1. Extraction of "Main Plateau Cluster" from Background. In order to perform this background separation, we examined each full-valley cluster for the OPV3-2BT-H data set shown in Figure 4b–i. Of these, the red cluster (Figure 4h) is the unambiguous choice for the primary molecular signature because (1) it most closely corresponds to the dense molecular region in Figure 1c that is not present in Figure 1b and (2) it is composed of relatively long and flat molecular plateaus that approximately match the expected length of the molecule after adding 0.5 nm to account for the "snapback" distance[28,77,78] (see Supporting Information, section S.9, for details). We therefore refer to the cluster in Figure 4h as the "main plateau cluster". In contrast to the raw data, the conductance peak for the main plateau cluster has a simple shape that can be confidently fit with no restrictions by a single Gaussian (Figure 5). This is a direct consequence of segment clustering's novel focus on pieces of traces as the clustering unit, since trace clustering approaches will necessarily produce clusters with complex conductance histogram shapes. However, the main plateau cluster in Figure 4h does not represent *all* of the molecular signature in the data set. In fact, the points in these segments only account for a small fraction of the molecular peak seen in the raw 1D histogram (Figure 5). This may be caused by a majority of molecular traces at room temperature jumping back and forth between tunneling decay and molecular plateaus (e.g., Figure 1a), whereas the segments in the main plateau cluster only originate from the "cleanest" molecular plateaus (i.e., those that are long, unbroken, and relatively constant). We hypothesize that these "cleanest" plateaus will yield the most reliable measure of molecular conductance and the underlying quantum transport, which is otherwise obscured by the large and stochastically visited space of possible junction configurations.

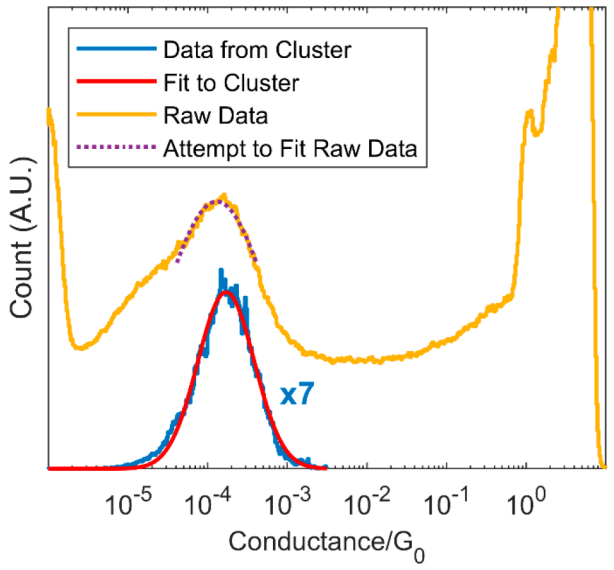


Figure 5. Raw 1D histogram for the OPV3-2BT-H data set from Figure 1c (yellow), along with a restricted Gaussian fit to the molecular peak (dotted purple, see Supporting Information, section S.8, for details). Overlaid in blue is a 1D histogram of the data from just the main plateau cluster (Figure 4h) and an unrestricted Gaussian fit (red), both scaled up by a factor of 7 for clarity. Whereas the complex shape of the raw data peak necessitates arbitrary fitting restrictions to obtain reasonable results, the simple shape of the main plateau cluster peak can be fit without restrictions, leading to a more confident and robust peak value.

To test this hypothesis, we collected nine total OPV3-2BT-H data sets across three different samples run under identical conditions (see Supporting Information, section S.4, for details). Within all but one of these data sets (see Supporting Information, section S.7, for details), a main plateau cluster analogous to the one shown in Figure 4h could be unambiguously identified (Figure 6a–h), providing strong evidence that this type of cluster is a meaningful and reproducible structural element of these data sets. Each of these main plateau clusters can again be effectively fit with an unrestricted single Gaussian (see Supporting Information, section S.8, for details). Comparing the spread of these 8 peaks with the restricted peaks fit to the raw 1D histograms (Figure 6i) reveals a significantly tightened distribution (Table 1), consistent with our hypothesis that segment clustering is aiding the extraction of an inherent molecular feature from a widely varying background.

3.2.2. Quantitative Comparison of Conductances of Similar Molecules. Figures 5 and 6 demonstrate the power of segment clustering: the need for complex and arbitrary fitting criteria is eliminated *and* data set-to-data set reproducibility is improved, allowing us to identify peak molecular conductances with increased precision and confidence. To illustrate the advantages of this increased precision, we used our MCBJ setup to collect multiple sets of breaking traces for a total of seven OPV3-2BT-X molecules (Chart 1; see Supporting Information, section S.4, for details on data sets). For all but two data sets (see Supporting Information, section S.7, for details), we identified a clear and unambiguous choice for the full-valley cluster corresponding to the main plateau feature. Our peak conductance results for all of these OPV3-2BT-X main plateau clusters are summarized in Figure 7, in which the error bars represent the uncertainty introduced by varying the *minPts* parameter (see Supporting Information, section S.6, for details).

Figure 7 shows that, as with OPV3-2BT-H, the peak conductances for each molecule in the series are highly reproducible, further supporting the claim that segment clustering is extracting an inherently molecular feature. Moreover, because of this high reproducibility, we are able to confidently *differentiate* the conductances of these molecules despite their high structural similarity. This makes it possible to search for structure–property relationships to physically explain such conductance differences. Extensive testing confirms that the peak conductances in Figure 7 are not meaningfully affected qualitatively *or quantitatively* by modest changes to the clustering parameters (see Supporting Information, section S.5, for details). Not only does this increase confidence in these specific results but it also provides strong evidence that segment clustering is a highly robust and generalizable tool for unsupervised analysis of potentially subtle variations in molecular conductances.

3.3. Using Segment Clustering to Separate Overlapping Molecular Features. In addition to the extraction of a single "primary" molecular feature in different data sets,

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302–18315

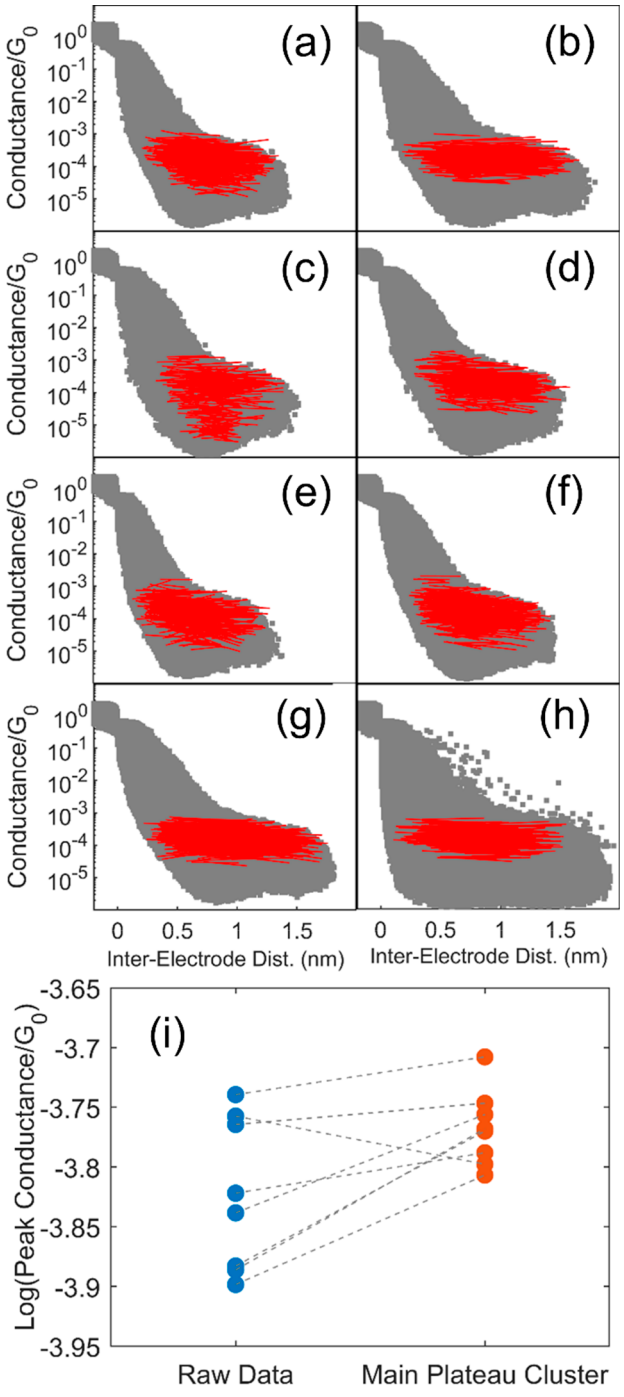


Figure 6. (a–h) Main plateau clusters selected for 8 different OPV3-2BT-H data sets, demonstrating that this feature is a consistent structural element of these data sets. (i) Comparison of peak conductance values from unrestricted Gaussian fits to the main plateau clusters from parts a–h with the peak conductance values from restricted Gaussian fits to the raw 1D histograms (see Supporting Information, section S.8, for details), demonstrating that segment clustering increases the precision of peak conductance measurements.

segment clustering can also be used to distinguish multiple features in a single data set. When 2D histograms of breaking traces display multiple "clouds" of increased density, it is often taken as an impetus to investigate the possibility of different binding modes, molecular configurations, *etc.*[9,28−32] While such clouds can offer tantalizing hints of multiple transport motifs, a major challenge is that it is often quite ambiguous whether density clouds are truly separate or not. This introduces a significant opportunity for bias, and it may also limit the scope of hypotheses considered for further investigation. Because segment clustering is unsupervised and largely model-free, it is a useful tool for *objective* separation of molecular features.

Table 1. Comparison of Different Measures of Spread for the Raw Data Peaks vs the Main Plateau Cluster Peaks for the Eight Different OPV3-2BT-H Datasets (in Figure 6i)[a]

	raw data peaks	main plateau cluster peaks
range	0.159	0.099
standard deviation	0.063	0.032
inter-quartile range	0.121	0.037

[a]All units are decades.

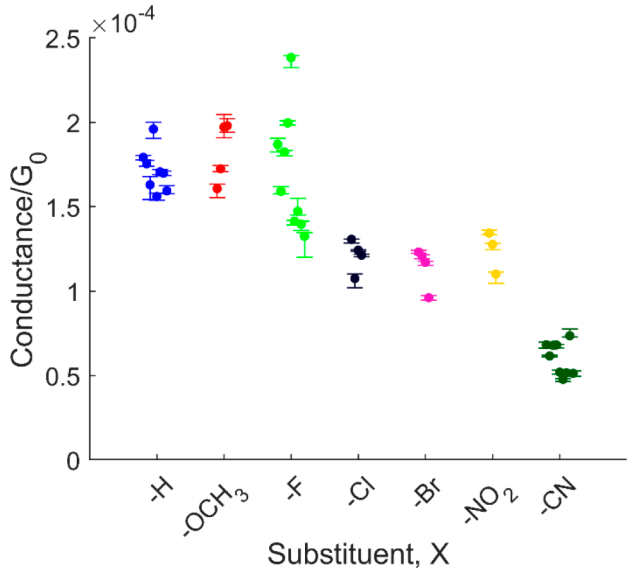


Figure 7. Comparison of peak conductance values from main plateau clusters for each OPV3-2BT-X data set considered in this work. Error bars represent the uncertainty due to clustering with different values of the *minPts* parameter (see Supporting Information, section S.6, for details). Due to the high reproducibility enabled by segment clustering, subtle conductance differences between molecules can be identified with confidence.

To demonstrate this, we constructed a synthetic data set consisting of equal numbers of experimental traces from samples run with two structurally rather different molecules. The first half of traces are taken from a data set collected with the molecule C6-2SMe (Chart 1), which displays a short molecular feature at $\sim 10^{-4} G_0$ (Figure 8a,c). Segment clustering of this data set unambiguously identified a full-valley cluster corresponding to this molecular feature (Figure 8b,c; see Supporting Information, section S.10, for details). The remaining traces for our synthetic mixture are taken from an OPV2-2BT (Chart 1) data set. The histograms of the breaking traces for this molecule reveal a strong high-conductance feature at $\sim 10^{-3} G_0$ as well as a subtler low-conductance feature at $\sim 10^{-4} G_0$ (Figure 8d,f), likely due to molecular stacking or direct π–Au binding.[21,59] While segment clustering identifies a main plateau cluster corresponding to the high-conductance feature (Figure 8e), none of the full-valley clusters matches well with the low-conductance feature (see Supporting Information, section S.10, for details). This shows

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315

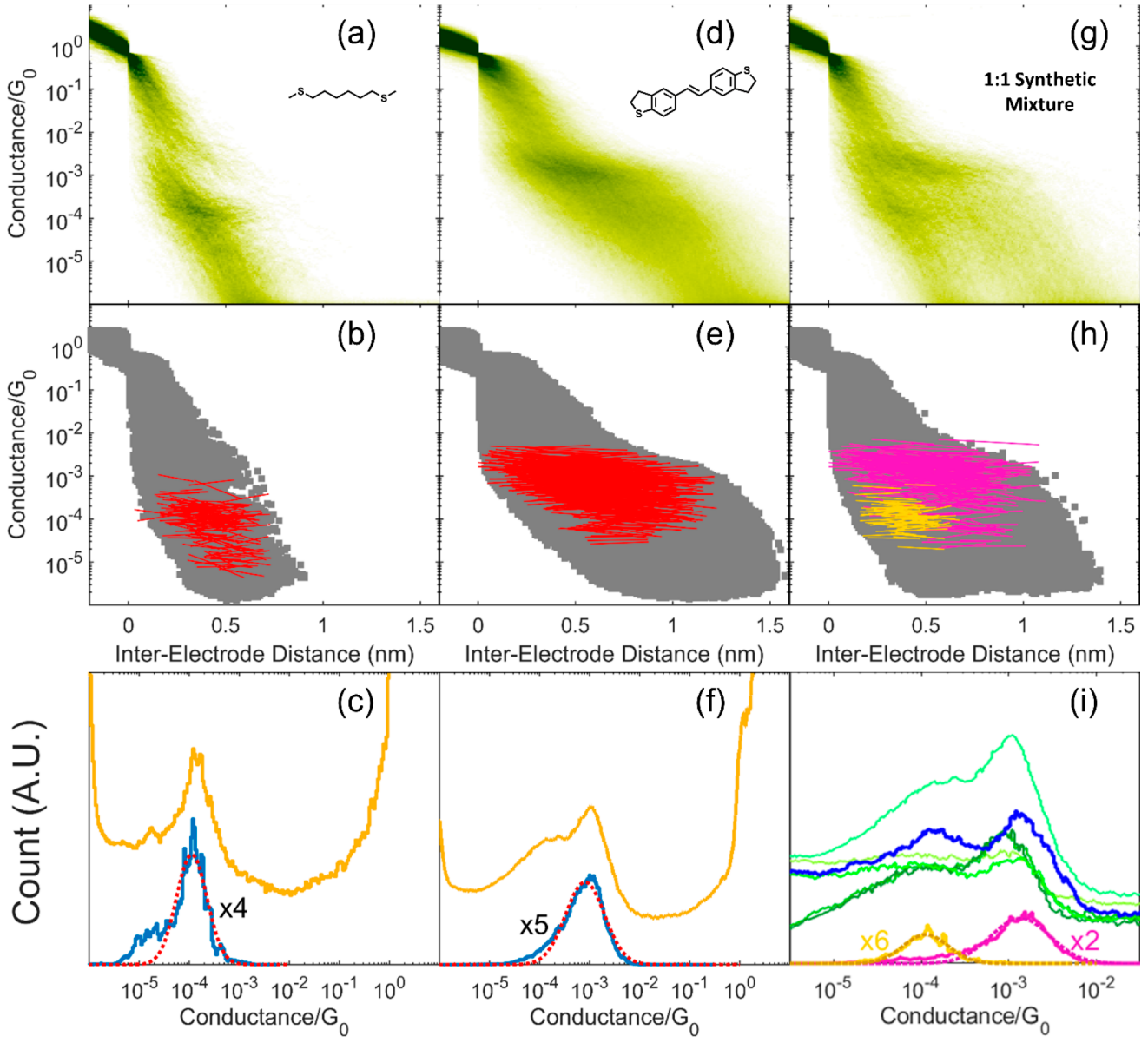


Figure 8. (a) 2D histogram for 1315 consecutive breaking traces collected in the presence of C6-2SMe. (b) Full-valley cluster identified as the main plateau cluster for the data from part a. (c) 1D histogram for the raw data from part a (yellow), overlaid with the 1D histogram for the data from the main plateau cluster in part b (blue) with an unrestricted Gaussian fit (dotted red). (d–f) Analogous plots to parts a–c for a data set containing 5807 consecutive breaking traces collected in the presence of OPV2-2BT. (g) 2D histogram for a synthetic data set constructed by combining equal numbers of traces from the data sets in parts a and d. (h) Two full-valley clusters identified as molecular plateau features for the data from part g. (i) 1D histogram for the data from part g (dark blue), overlaid with the 1D histograms for the two clusters from part h (pink and yellow) and their respective single Gaussian fits (dotted lines). For comparison, 1D histograms for five different raw OPV2-2BT data sets are included (various shades of green), demonstrating that the intensity and location of the peaks in the synthetic mixture lie well within the range of the different pure OPV2-2BT data sets.

that segment clustering will not always extract every meaningful feature from a data set.

However, because the low-conductance feature of OPV2-2BT partially overlaps the primary C6-2SMe feature, our synthetic mixture provides an excellent challenge case for segment clustering. This can be seen in the 2D histogram for our mixture (Figure 8g), which is qualitatively quite similar to the pure OPV2-2BT histogram (Figure 8d) and displays exactly the type of ambiguous dual density cloud often reported in the literature[9,28−31] and is sometimes imbued with speculative microscopic meaning. Moreover, Figure 8i shows that the intensity and location of the lower peak in the 1D histogram of our synthetic mixture falls *within* the variability observed between different pure OPV2-2BT data sets, further illustrating the challenge posed by separating these two molecular distributions.

As shown in Figure 8h, segment clustering of our mixture data set identifies two full-valley clusters that appear to correspond to the main OPV2-2BT and C6-2SMe features (though because both molecular features are "diluted" by mixing, the minimum valley size was lowered below 1% to locate these valleys; see Supporting Information, section S.11, for details). Because this mixture was constructed synthetically, we can quantitatively test this hypothesis. We find that the

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315

separation of molecular features is indeed quite accurate, even though the two clusters partially overlap: 97% of the data in the OPV2-2BT cluster belong to traces taken from the OPV2-2BT data set, and 84% of the data in the C6-2SMe cluster come from C6-2SMe traces. It is not surprising that the C6-2SMe cluster has a higher misidentification rate, because this cluster's shorter segments are much more likely to be found in an arbitrary data set simply by chance. This is evidenced by the fact that a *cluster* of C6-2SMe-like segments *did not exist* in the pure OPV2-2BT data set, indicating that the misassigned segments added to the C6-2SMe cluster from the mixture data set did not form a region of high density by themselves. To further test the robustness of this feature separation, we constructed seven additional 1:1 OPV2-2BT:C6-2SMe synthetic mixture data sets using different combinations of traces from different pure-molecule data sets (see Supporting Information section S.4 for details). As shown in Figure S17 and Table S6 in the Supporting Information, segment clustering successfully extracted both molecular features for all but one of these mixtures (see Supporting Information, section S.12, for details), and each of these separations displayed high quantitative accuracy.

By reliably separating features in an experimental data set, segment clustering contributes to an important goal of single molecule transport research, toward which some progress has already been made. For example, several existing clustering algorithms have a demonstrated ability to extract multiple subfeatures from experimental data sets of one molecular species.[33,34,36,38,42] However, while these studies offer intriguing hints about different binding modes and molecular conformations, such subfeatures are unfortunately difficult to corroborate without extremely trustworthy atomistic simulations. More-testable examples of feature separation have been demonstrated by Hamill et al., whose sorting algorithm successfully separated the features for two molecules in a mixture displaying an "obvious bimodal feature",[34] and recently by Huang et al., whose deep-learning clustering algorithm separated two features from an overlapping molecular mixture.[37] However, because neither mixture was synthetic, these separations could not be quantitatively confirmed for the accuracy of cluster assignments. Finally, Vladyka and Albrecht very recently applied a neural network-based classification algorithm to a synthetic mixture of three different molecules, and while some pairwise separation was qualitatively observed, the combination of all three molecular features could not be separated.[41] The OPV2-2BT/C6-2SMe case study described here is thus a significant advance in that it constitutes a quantitatively validated example of experimental feature separation, and it does so in the challenging case of overlapping features. This provides a powerful demonstration of the usefulness of segment clustering as a hypothesis generation tool.

4. CONCLUSIONS

In this work we presented segment clustering, a novel approach to aid hypothesis generation for data sets of single-molecule breaking traces. Segment clustering is *categorically* different than all previous clustering approaches since it treats, for the first time, pieces of breaking traces as the fundamental clustering unit, allowing behaviors occurring in just part of a trace to be more readily identified. This subtrace focus gives segment clustering the potential to yield new and powerful insights into single-molecule data sets because grouping the data by segments is a better match for the empirical "local history" and piece-wise linear structure of break junction data than grouping by entire traces. This suggests that the segmentation approach described here may be a valuable avenue for future investigations even outside the context of clustering, for example by comparing the distribution of segment lengths between different data sets or exploring the likelihood of certain types of segments to appear in the same traces as others. To encourage such new directions, and to enable the use of the segment clustering in other contexts, we have made our code freely available in a user-friendly open-source package (github.com/LabMonti/SMAUG-Toolbox).

To demonstrate the power and versatility of the full segment clustering approach, we have applied it to two common challenges faced in the analysis of breaking traces. First, to address the related issues of complex peak shapes and varying background signals in conductance histograms, we used segment clustering to extract the "primary" molecular feature in a series of similar molecules. We showed that this increases measurement reproducibility *and* the robustness of peak-fitting, allowing subtle conductance changes to be distinguished with confidence. Second, to address the problem of separating ambiguous or overlapping molecular features, we used segment clustering to search for clusters corresponding to particular features in 1D and 2D histograms. By constructing a synthetic mixture of traces from two different molecules with overlapping conductance distributions, we demonstrated that segment clustering performs this feature separation with high quantitative accuracy even in challenging circumstances. We expect that these two advances in particular, as well as the new perspective offered by segment clustering in general, will aid in the establishment of structure−property relationships in single molecule quantum transport and thus help unlock new paths toward harnessing molecular electronics by design.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.0c03612.

> MCBJ setup, interelectrode distance calibration, additional design criteria for segment clustering, data set collection and construction, robustness of OPV3-2BT-X results to clustering parameters, selecting clusters from multiple clustering outputs for the same data set, selection of main plateau clusters for OPV3-2BT-X data sets, peak fitting, investigating main plateau cluster lengths, selection of main plateau clusters for OPV2-2BT and C6-2SMe, cluster selection for OPV2-2BT/C6-2SMe 1:1 synthetic mixture #1, and clustering of additional synthetic mixtures (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Oliver L. A. Monti − *Department of Chemistry and Biochemistry and Department of Physics, University of Arizona, Tucson, Arizona 85721, United States;* orcid.org/0000-0002-0974-7253; Phone: ++ 520 626 1177; Email: monti@u.arizona.edu

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315

Authors

Nathan D. Bamberger − *Department of Chemistry and Biochemistry, University of Arizona, Tucson, Arizona 85721, United States*

Jeffrey A. Ivie − *Department of Chemistry and Biochemistry, University of Arizona, Tucson, Arizona 85721, United States*

Keshaba N. Parida − *Department of Chemistry and Biochemistry, University of Arizona, Tucson, Arizona 85721, United States*

Dominic V. McGrath − *Department of Chemistry and Biochemistry, University of Arizona, Tucson, Arizona 85721, United States;* orcid.org/0000-0001-9605-2224

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcc.0c03612

Author Contributions

O.L.A.M., J.A.I., and N.D.B. conceived the research ideas. K.N.P. synthesized the OPV2-2BT and OPV3-2BT-X molecules, directed by D.V.M. MCBJ samples were fabricated and run by J.A.I. and N.D.B. Segment clustering was developed and implemented by N.D.B. with advice and input from O.L.A.M. and J.A.I. N.D.B. wrote the manuscript with input and advice from all authors.

Notes

The authors declare no competing financial interest.

The MatLab code used for this work is available free of charge at github.com/LabMonti/SMAUG-Toolbox.

■ ACKNOWLEDGMENTS

The authors would like to acknowledge support from the National Science Foundation, Award No. DMR-1708443, as well as from the Graduate and Professional Student Council at The University of Arizona. Plasma etching was performed in part using a Plasmatherm reactive ion etcher acquired through an NSF MRI grant, Award No. ECCS-1725571. Clustering was performed using High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and RDI and maintained by the UA Research Technologies department. All SEM images and data were collected in the W. M. Keck Center for Nano-Scale Imaging in the Department of Chemistry and Biochemistry at the University of Arizona with funding from a W. M. Keck Foundation grant. The authors would also like to thank R. Himmelhuber for help and advice on sample fabrication, A. Garlant for help with laser interferometry, and C. Raithel and D. Dyer for feedback on the manuscript.

■ REFERENCES

(1) Aviram, A.; Ratner, M. A. Molecular Rectifiers. *Chem. Phys. Lett.* **1974**, *29*, 277−283.

(2) Xiang, D.; Wang, X.; Jia, C.; Lee, T.; Guo, X. Molecular-Scale Electronics: From Concept to Function. *Chem. Rev.* **2016**, *116*, 4318−4440.

(3) Martin, C. A; Ding, D.; van der Zant, H. S J; Ruitenbeek, J. M v. Lithographic Mechanical Break Junctions for Single-Molecule Measurements in Vacuum: Possibilities and Limitations. *New J. Phys.* **2008**, *10*, 065008.

(4) Huisman, E. H.; Trouwborst, M. L.; Bakker, F. L.; van Wees, B. J.; van der Molen, S. J. The Mechanical Response of Lithographically Defined Break Junctions. *J. Appl. Phys.* **2011**, *109*, 104305.

(5) Vrouwe, S. A. G.; van der Giessen, E.; van der Molen, S. J.; Dulic, D.; Trouwborst, M. L.; van Wees, B. J. Mechanics of Lithographically Defined Break Junctions. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2005**, *71*, 035313.

(6) Muller, C. J.; van Ruitenbeek, J. M.; de Jongh, L. J. Conductance and Supercurrent Discontinuities in Atomic-Scale Metallic Constrictions of Variable Width. *Phys. Rev. Lett.* **1992**, *69*, 140−143.

(7) 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.* **2018**, *51*, 1359−1367.

(8) Kaneko, S.; Montes, E.; Suzuki, S.; Fujii, S.; Nishino, T.; Tsukagoshi, K.; Ikeda, K.; Kano, H.; Nakamura, H.; Vázquez, H.; Kiguchi, M. Identifying the Molecular Adsorption Site of a Single Molecule Junction through Combined Raman and Conductance Studies. *Chem. Sci.* **2019**, *10*, 6261−6269.

(9) Moreno-García, P.; Gulcur, M.; Manrique, D. Z.; Pope, T.; Hong, W.; Kaliginedi, V.; Huang, C.; Batsanov, A. S.; Bryce, M. R.; Lambert, C.; Wandlowski, T. Single-Molecule Conductance of Functionalized Oligoynes: Length Dependence and Junction Evolution. *J. Am. Chem. Soc.* **2013**, *135*, 12228−12240.

(10) Zhan, C.; Wang, G.; Zhang, X.-G.; Li, Z.-H.; Wei, J.-Y.; Si, Y.; Yang, Y.; Hong, W.; Tian, Z.-Q. Single-Molecule Measurement of Adsorption Free Energy at the Solid-Liquid Interface. *Angew. Chem.* **2019**, *131*, 14676−14680.

(11) Xu, B.; Tao, N. J. Measurement of Single-Molecule Resistance by Repeated Formation of Molecular Junctions. *Science* **2003**, *301*, 1221−1223.

(12) Arroyo, C. R.; Leary, E.; Castellanos-Gómez, A.; Rubio-Bollinger, G.; González, M. T.; Agraït, N. Influence of Binding Groups on Molecular Junction Formation. *J. Am. Chem. Soc.* **2011**, *133*, 14313−14319.

(13) Zang, Y.; Zou, Q.; Fu, T.; Ng, F.; Fowler, B.; Yang, J.; Li, H.; Steigerwald, M. L.; Nuckolls, C.; Venkataraman, L. Directing Isomerization Reactions of Cumulenes with Electric Fields. *Nat. Commun.* **2019**, *10*, 1−7.

(14) Sun, Y.-Y.; Peng, Z.-L.; Hou, R.; Liang, J.-H.; Zheng, J.-F.; Zhou, X.-Y.; Zhou, X.-S.; Jin, S.; Niu, Z.-J.; Mao, B.-W. Enhancing Electron Transport in Molecular Wires by Insertion of a Ferrocene Center. *Phys. Chem. Chem. Phys.* **2014**, *16*, 2260−2267.

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

(16) Ohnishi, H.; Kondo, Y.; Takayanagi, K. Quantized Conductance through Individual Rows of Suspended Gold Atoms. *Nature* **1998**, *395*, 780−783.

(17) Vonlanthen, D.; Mishchenko, A.; Elbing, M.; Neuburger, M.; Wandlowski, T.; Mayor, M. Chemically Controlled Conductivity: Torsion-Angle Dependence in a Single-Molecule Biphenyldithiol Junction. *Angew. Chem., Int. Ed.* **2009**, *48*, 8886−8890.

(18) Venkataraman, L.; Klare, J. E.; Nuckolls, C.; Hybertsen, M. S.; Steigerwald, M. L. Dependence of Single-Molecule Junction Conductance on Molecular Conformation. *Nature* **2006**, *442*, 904.

(19) Venkataraman, L.; Park, Y. S.; Whalley, A. C.; Nuckolls, C.; Hybertsen, M. S.; Steigerwald, M. L. Electronics and Chemistry: Varying Single-Molecule Junction Conductance Using Chemical Substituents. *Nano Lett.* **2007**, *7*, 502−506.

(20) Jiang, F.; Trupp, D.; Algethami, N.; Zheng, H.; He, W.; Alqorashi, A.; Zhu, C.; Tang, C.; Li, R.; Liu, J.; Sadeghi, H.; Shi, J.; Davidson, R.; Korb, M.; Naher, M.; Sobolev, A. N.; Sangtarash, S.; Low, P. J.; Hong, W.; Lambert, C. Turning the Tap: Conformational Control of Quantum Interference to Modulate Single Molecule Conductance. *Angew. Chem.* **2019**, *131*, 19163−19169.

(21) Martín, S.; Grace, I.; Bryce, M. R.; Wang, C.; Jitchati, R.; Batsanov, A. S.; Higgins, S. J.; Lambert, C. J.; Nichols, R. J. Identifying Diversity in Nanoscale Electrical Break Junctions. *J. Am. Chem. Soc.* **2010**, *132*, 9157−9164.

(22) Chen, Z.; Chen, L.; Liu, J.; Li, R.; Tang, C.; Hua, Y.; Chen, L.; Shi, J.; Yang, Y.; Liu, J.; Zheng, J.; Chen, L.; Cao, J.; Chen, H.; Xia, H.; Hong, W. Modularized Tuning of Charge Transport through Highly Twisted and Localized Single-Molecule Junctions. *J. Phys. Chem. Lett.* **2019**, *10*, 3453−3458.

(23) Inkpen, M. S.; Liu, Z.-F.; Li, H.; Campos, L. M.; Neaton, J. B.; Venkataraman, L. Non-Chemisorbed Gold−Sulfur Binding Prevails in Self-Assembled Monolayers. *Nat. Chem.* **2019**, *11*, 351−358.
(24) Park, Y. S.; Whalley, A. C.; Kamenetska, M.; Steigerwald, M. L.; Hybertsen, M. S.; Nuckolls, C.; Venkataraman, L. Contact Chemistry and Single-Molecule Conductance: A Comparison of Phosphines, Methyl Sulfides, and Amines. *J. Am. Chem. Soc.* **2007**, *129*, 15768−15769.
(25) Venkataraman, L.; Klare, J. E.; Tam, I. W.; Nuckolls, C.; Hybertsen, M. S.; Steigerwald, M. L. Single-Molecule Circuits with Well-Defined Molecular Conductance. *Nano Lett.* **2006**, *6*, 458−462.
(26) Shen, P.; Huang, M.; Qian, J.; Li, J.; Ding, S.; Zhou, X.-S.; Xu, B.; Zhao, Z.; Tang, B. Z. Achieving Efficient Multichannel Conductance in Through-Space Conjugated Single-Molecule Parallel Circuits. *Angew. Chem.* **2020**, *132*, 4611−4618.
(27) Zhang, F.; Wu, X.-H.; Zhou, Y.-F.; Wang, Y.-H.; Zhou, X.-S.; Shao, Y.; Li, J.-F.; Jin, S.; Zheng, J.-F. Improving Gating Efficiency of Electron Transport through Redox-Active Molecular Junctions with Conjugated Chain. *ChemElectroChem* **2020**, *7*, 1337−1341.
(28) Hong, W.; Manrique, D. Z.; Moreno-García, P.; Gulcur, M.; Mishchenko, A.; Lambert, C. J.; Bryce, M. R.; Wandlowski, T. Single Molecular Conductance of Tolanes: Experimental and Theoretical Study on the Junction Evolution Dependent on the Anchoring Group. *J. Am. Chem. Soc.* **2012**, *134*, 2292−2304.
(29) Kaliginedi, V.; Moreno-García, P.; Valkenier, H.; Hong, W.; García-Suárez, V. M.; Buiter, P.; Otten, J. L. H.; Hummelen, J. C.; Lambert, C. J.; Wandlowski, T. Correlations between Molecular Structure and Single-Junction Conductance: A Case Study with Oligo(Phenylene-Ethynylene)-Type Wires. *J. Am. Chem. Soc.* **2012**, *134*, 5262−5275.
(30) Quek, S. Y.; Kamenetska, M.; Steigerwald, M. L.; Choi, H. J.; Louie, S. G.; Hybertsen, M. S.; Neaton, J. B.; Venkataraman, L. Mechanically Controlled Binary Conductance Switching of a Single-Molecule Junction. *Nat. Nanotechnol.* **2009**, *4*, 230−234.
(31) Isshiki, Y.; Fujii, S.; Nishino, T.; Kiguchi, M. Fluctuation in Interface and Electronic Structure of Single-Molecule Junctions Investigated by Current versus Bias Voltage Characteristics. *J. Am. Chem. Soc.* **2018**, *140*, 3760−3767.
(32) Leary, E.; Zotti, L. A.; Miguel, D.; Márquez, I. R.; Palomino-Ruiz, L.; Cuerva, J. M.; Rubio-Bollinger, G.; González, M. T.; Agrait, N. The Role of Oligomeric Gold−Thiolate Units in Single-Molecule Junctions of Thiol-Anchored Molecules. *J. Phys. Chem. C* **2018**, *122*, 3211−3218.
(33) Cabosart, D.; El Abbassi, M.; Stefani, D.; Frisenda, R.; Calame, M.; van der Zant, H. S. J.; Perrin, M. L. A Reference-Free Clustering Method for the Analysis of Molecular Break-Junction Measurements. *Appl. Phys. Lett.* **2019**, *114*, 143102.
(34) Hamill, J. M.; Zhao, X. T.; Mészáros, G.; Bryce, M. R.; Arenz, M. Fast Data Sorting with Modified Principal Component Analysis to Distinguish Unique Single Molecular Break Junction Trajectories. *Phys. Rev. Lett.* **2018**, *120*, 016601.
(35) Lauritzen, K. P.; Magyarkuti, A.; Balogh, Z.; Halbritter, A.; Solomon, G. C. Classification of Conductance Traces with Recurrent Neural Networks. *J. Chem. Phys.* **2018**, *148*, 084111.
(36) Wu, B. H.; Ivie, J. A.; Johnson, T. K.; Monti, O. L. A. Uncovering Hierarchical Data Structure in Single Molecule Transport. *J. Chem. Phys.* **2017**, *146*, 092321.
(37) Huang, F.; Li, R.; Wang, G.; Zheng, J.; Tang, Y.; Liu, J.; Yang, Y.; Yao, Y.; Shi, J.; Hong, W. Automatic Classification of Single-Molecule Charge Transport Data with an Unsupervised Machine-Learning Algorithm. *Phys. Chem. Chem. Phys.* **2020**, *22*, 1674−1681.
(38) Magyarkuti, A.; Balogh, N.; Balogh, Z.; Venkataraman, L.; Halbritter, A. Unsupervised Feature Recognition in Single Molecule Break Junction Data. *Nanoscale* **2020**, *12*, 8355−8363.
(39) Korshoj, L. E.; Afsari, S.; Chatterjee, A.; Nagpal, P. Conformational Smear Characterization and Binning of Single-Molecule Conductance Measurements for Enhanced Molecular Recognition. *J. Am. Chem. Soc.* **2017**, *139*, 15420−15428.
(40) Brooke, R. J.; Szumski, D. S.; Vezzoli, A.; Higgins, S. J.; Nichols, R. J.; Schwarzacher, W. Dual Control of Molecular Conductance through PH and Potential in Single-Molecule Devices. *Nano Lett.* **2018**, *18*, 1317−1322.
(41) Vladyka, A.; Albrecht, T. Unsupervised Classification of Single-Molecule Data with Autoencoders and Transfer Learning. *arXiv.org* 2020; http://arxiv.org/abs/2004.01239 (accessed April 9, 2020).
(42) Lemmer, M.; Inkpen, M. S.; Kornysheva, K.; Long, N. J.; Albrecht, T. Unsupervised Vector-Based Classification of Single-Molecule Charge Transport Data. *Nat. Commun.* **2016**, *7*, 12922.
(43) Albrecht, T.; Slabaugh, G.; Alonso, E.; Al-Arif, S. M. R. Deep Learning for Single-Molecule Science. *Nanotechnology* **2017**, *28*, 423001.
(44) Zhang, Q.; Liu, C.; Tao, S.; Yi, R.; Su, W.; Zhao, C. Z.; Zhao, C.; Dappe, Y. J.; Nichols, R. J.; Yang, L. Fast and Straightforward Analysis Approach of Charge Transport Data in Single Molecule Junctions. *Nanotechnology* **2018**, *29*, 325701.
(45) Zotti, L. A.; Bednarz, B.; Hurtado-Gallego, J.; Cabosart, D.; Rubio-Bollinger, G.; Agrait, N.; van der Zant, H. S. J. Can One Define the Conductance of Amino Acids? *Biomolecules* **2019**, *9*, 580.
(46) Frisenda, R.; Janssen, V. A. E. C.; Grozema, F. C.; van der Zant, H. S. J.; Renaud, N. Mechanically Controlled Quantum Interference in Individual π-Stacked Dimers. *Nat. Chem.* **2016**, *8*, 1099−1104.
(47) Parker, C. R.; Leary, E.; Frisenda, R.; Wei, Z.; Jennum, K. S.; Glibstrup, E.; Abrahamsen, P. B.; Santella, M.; Christensen, M. A.; Della Pia, E. A.; Li, T.; Gonzalez, M. T.; Jiang, X.; Morsing, T. J.; Rubio-Bollinger, G.; Laursen, B. W.; Nørgaard, K.; van der Zant, H.; Agrait, N.; Nielsen, M. B. A Comprehensive Study of Extended Tetrathiafulvalene Cruciform Molecules for Molecular Electronics: Synthesis and Electrical Transport Measurements. *J. Am. Chem. Soc.* **2014**, *136*, 16497−16507.
(48) Leary, E.; La Rosa, A.; González, M. T.; Rubio-Bollinger, G.; Agraït, N.; Martín, N. Incorporating Single Molecules into Electrical Circuits. The Role of the Chemical Anchoring Group. *Chem. Soc. Rev.* **2015**, *44*, 920−942.
(49) Su, T. A.; Widawsky, J. R.; Li, H.; Klausen, R. S.; Leighton, J. L.; Steigerwald, M. L.; Venkataraman, L.; Nuckolls, C. Silicon Ring Strain Creates High-Conductance Pathways in Single-Molecule Circuits. *J. Am. Chem. Soc.* **2013**, *135*, 18331−18334.
(50) Chen, H.; Li, Y.; Chang, S. A Hybrid Molecular Junction Mapping Technique for Simultaneous Measurements of Single Molecule Electronic Conductance and Its Corresponding Binding Geometry in a Tunneling Junction. *Anal. Chem.* **2020**, *92*, 6423−6429.
(51) Johnson, T. K.; Ivie, J. A.; Jaruvang, J.; Monti, O. L. A. Fast Sensitive Amplifier for Two-Probe Conductance Measurements in Single Molecule Break Junctions. *Rev. Sci. Instrum.* **2017**, *88*, 033904.
(52) Mishchenko, A.; Zotti, L. A.; Vonlanthen, D.; Bürkle, M.; Pauly, F.; Cuevas, J. C.; Mayor, M.; Wandlowski, T. Single-Molecule Junctions Based on Nitrile-Terminated Biphenyls: A Promising New Anchoring Group. *J. Am. Chem. Soc.* **2011**, *133*, 184−187.
(53) Xiang, D.; Sydoruk, V.; Vitusevich, S.; Petrychuk, M. V.; Offenhäusser, A.; Kochelap, V. A.; Belyaev, A. E.; Mayer, D. Noise Characterization of Metal-Single Molecule Contacts. *Appl. Phys. Lett.* **2015**, *106*, 063702.
(54) Hasegawa, Y.; Harashima, T.; Jono, Y.; Seki, T.; Kiguchi, M.; Nishino, T. Kinetic Investigation of Chemical Process in Single-Molecule Junction. *Chem. Commun.* **2020**, *56*, 309−312.
(55) Frisenda, R.; Tarkuç, S.; Galán, E.; Perrin, M. L.; Eelkema, R.; Grozema, F. C.; van der Zant, H. S. J. Electrical Properties and Mechanical Stability of Anchoring Groups for Single-Molecule Electronics. *Beilstein J. Nanotechnol.* **2015**, *6*, 1558−1567.
(56) Perrin, M. L.; Prins, F.; Martin, C. A.; Shaikh, A. J.; Eelkema, R.; van Esch, J. H.; Briza, T.; Kaplanek, R.; Kral, V.; van Ruitenbeek, J. M.; van der Zant, H. S. J.; Dulić, D. Influence of the Chemical Structure on the Stability and Conductance of Porphyrin Single-Molecule Junctions. *Angew. Chem., Int. Ed.* **2011**, *50*, 11223−11226.

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315

(57) Gil, M.; Malinowski, T.; Iazykov, M.; Klein, H. R. Estimating Single Molecule Conductance from Spontaneous Evolution of a Molecular Contact. *J. Appl. Phys.* **2018**, *123*, 104303.
(58) Brunner, J.; González, M. T.; Schönenberger, C.; Calame, M. Random Telegraph Signals in Molecular Junctions. *J. Phys.: Condens. Matter* **2014**, *26*, 474202.
(59) Manrique, D. Z.; Huang, C.; Baghernejad, M.; Zhao, X.; Al-Owaedi, O. A.; Sadeghi, H.; Kaliginedi, V.; Hong, W.; Gulcur, M.; Wandlowski, T.; Bryce, M. R.; Lambert, C. J. A Quantum Circuit Rule for Interference Effects in Single-Molecule Electrical Junctions. *Nat. Commun.* **2015**, *6*, 6389.
(60) Tan, Z.; Zhang, D.; Tian, H.-R.; Wu, Q.; Hou, S.; Pi, J.; Sadeghi, H.; Tang, Z.; Yang, Y.; Liu, J.; Tan, Y.-Z.; Chen, Z.-B.; Shi, J.; Xiao, Z.; Lambert, C.; Xie, S.-Y.; Hong, W. Atomically Defined Angstrom-Scale All-Carbon Junctions. *Nat. Commun.* **2019**, *10*, 1748.
(61) Huang, C.; Jevric, M.; Borges, A.; Olsen, S. T.; Hamill, J. M.; Zheng, J.-T.; Yang, Y.; Rudnev, A.; Baghernejad, M.; Broekmann, P.; Petersen, A. U.; Wandlowski, T.; Mikkelsen, K. V.; Solomon, G. C.; Brøndsted Nielsen, M.; Hong, W. Single-Molecule Detection of Dihydroazulene Photo-Thermal Reaction Using Break Junction Technique. *Nat. Commun.* **2017**, *8*, 15436.
(62) Li, Y.; Xiang, L.; Palma, J. L.; Asai, Y.; Tao, N. Thermoelectric Effect and Its Dependence on Molecular Length and Sequence in Single DNA Molecules. *Nat. Commun.* **2016**, *7*, 11294.
(63) Roldan, D.; Kaliginedi, V.; Cobo, S.; Kolivoska, V.; Bucher, C.; Hong, W.; Royal, G.; Wandlowski, T. Charge Transport in Photoswitchable Dimethyldihydropyrene-Type Single-Molecule Junctions. *J. Am. Chem. Soc.* **2013**, *135*, 5974−5977.
(64) Vladyka, A.; Perrin, M. L.; Overbeck, J.; Ferradás, R. R.; García-Suárez, V.; Gantenbein, M.; Brunner, J.; Mayor, M.; Ferrer, J.; Calame, M. In-Situ Formation of One-Dimensional Coordination Polymers in Molecular Junctions. *Nat. Commun.* **2019**, *10*, 262.
(65) Pan, X.; Lawson, B.; Rustad, A. M.; Kamenetska, M. PH-Activated Single Molecule Conductance and Binding Mechanism of Imidazole on Gold. *Nano Lett.* **2020**, *20*, 4687−4692.
(66) Baghernejad, M.; Van Dyck, C.; Bergfield, J.; Levine, D. R.; Gubicza, A.; Tovar, J. D.; Calame, M.; Broekmann, P.; Hong, W. Quantum Interference Enhanced Chemical Responsivity in Single-Molecule Dithienoborepin Junctions. *Chem. - Eur. J.* **2019**, *25*, 15141−15146.
(67) Wu, S.; González, M. T.; Huber, R.; Grunder, S.; Mayor, M.; Schönenberger, C.; Calame, M. Molecular Junctions Based on Aromatic Coupling. *Nat. Nanotechnol.* **2008**, *3*, 569−574.
(68) Keogh, E.; Chu, S.; Hart, D.; Pazzani, M. Segmenting Time Series: A Survey and Novel Approach. *Data Mining in Time Series Databases* **2004**, *57*, 1−21.
(69) Salvador, S.; Chan, P. Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. *In 16th IEEE International Conference on Tools with Artificial Intelligence* **2004**, 576−584.
(70) Verleysen, M.; François, D. The Curse of Dimensionality in Data Mining and Time Series Prediction. *Computational Intelligence and Bioinspired Systems* **2005**, *3512*, 758−770.
(71) Berkhin, P. A Survey of Clustering Data Mining Techniques. *Grouping Multidimensional Data: Recent Advances in Clustering* **2006**, 25−71.
(72) Han, J.; Kamber, M.; Pei, J. Cluster Analysis: Basic Concepts and Methods. *Data Mining: Concepts and Techniques* **2012**, 443−495.
(73) de Amorim, R. C. A Survey on Feature Weighting Based K-Means Algorithms. *J. Classif.* **2016**, *33*, 210−242.
(74) Schneider, J.; Vlachos, M. Scalable Density-Based Clustering with Quality Guarantees Using Random Projections. *Data Min. Knowl. Discovery* **2017**, *31*, 972−1005.
(75) Ankerst, M.; Breunig, M. M.; Kriegel, H.-P.; Sander, J. OPTICS: Ordering Points to Identify the Clustering Structure. *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data* **1999**, *28*, 49−60.
(76) Hämäläinen, J.; Jauhiainen, S.; Kärkkäinen, T. Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering. *Algorithms* **2017**, *10*, 105.
(77) Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J. M.; Perona, I. An Extensive Comparative Study of Cluster Validity Indices. *Pattern Recognit.* **2013**, *46*, 243−256.
(78) Yanson, A. I.; Bollinger, G. R.; van den Brom, H. E.; Agraït, N.; van Ruitenbeek, J. M. Formation and Manipulation of a Metallic Wire of Single Gold Atoms. *Nature* **1998**, *395*, 783−785.

https://dx.doi.org/10.1021/acs.jpcc.0c03612
J. Phys. Chem. C 2020, 124, 18302−18315