Nano Research

Open Access

Research Article

Low-probability events detection using unsupervised multi-prototype clustering for single-molecule electronics

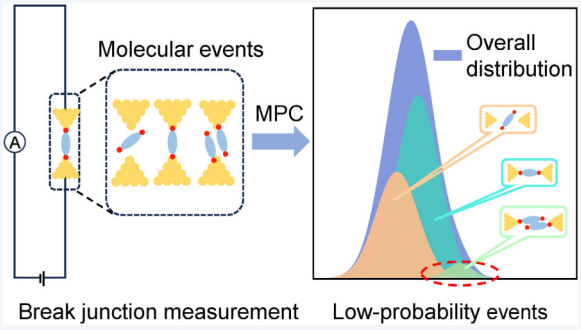
Chi Shang[1], Rigong Te[1], Shenglun Xiong[1], Xipeng Liu[1], Taige Lu[1], Yixuan Zhu[1], Chun Tang[1], Jing Li[1], Yu Zhou[1], Haojie Liu[1,2] ✉, Junyang Liu[1] ✉, and Wenjing Hong[1] ✉

[1] *Institution State Key Laboratory of Physical Chemistry of Solid Surfaces, College of Chemistry and Chemical Engineering and Institute of Artificial Intelligence and Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), Xiamen University, Xiamen 361005, China*

[2] *Research Center of Grid Energy Storage and Battery Application, School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China*

Cite this article: *Nano Research*, **2025**, *18*, 94907276. https://doi.org/10.26599/NR.2025.94907276

ABSTRACT: Artificial intelligence for science (AI4S) has emerged as a new horizon in state-of-the-art scientific research, and single-molecule electronics could be considered an ideal prototype in AI4S due to the opportunities in correlating high-throughput and high-quality data with clear physical mechanisms. Towards using artificial intelligence for single-molecule electronics (AI4SME), the unsupervised extraction of low-probability events from the massive experimental data becomes the key step, which has emerged for accurate detection of different configurations and even structural changes in single-molecule junctions. However, the present algorithms suffer from the "uniform effect", in which the majority events are erroneously allocated to minority ones, resulting in a relatively equal spread of cluster sizes and hindering the investigations for charge transport mechanisms with subtle and complex behaviors in single-molecule electronics. In this work, we propose a new multi-prototype clustering technique for precisely discriminating molecular events during the break junction process, especially those occurring with a probability below 10%, and further precisely extract the product species at the onset of the electric field-driven single-molecule keto-enol reaction with a probability as low as 1.5%. Our work tackles the long-term bottleneck of uniform effect for the precise detection of low-probability single-molecule events.



KEYWORDS: single-molecule electronics, low-probability events detection, machine learning, artificial intelligence for science (AI4S)

1 Introduction

The emergence of artificial intelligence for science (AI4S) has opened new horizons in scientific exploration of complex challenges across catalyst discovery [1], drug screening [2], and materials design [3], being increasingly integrated into scientific discovery to accelerate research [4, 5]. In the realm of single-molecule electronics, scientists selected only a few conductance traces for each device due to the feasibility of experiments at the early stage [6], and later, the break junction (BJ) method facilitated the data acquisition efficiency and promoted the statistical analysis on thousands of conductance traces, by emphasizing the molecular signal without data selection [7–10]. However, many single-molecule events with low-probability less than 10% during the process, such as transient molecular configurations with short lifetime [10–13], molecule-electrode connections with non-classic bonding [14–16], and weak intramolecular interactions with short plateau length and low junction formation probability (JFP) [10, 17–19], are often overlooked and obscured by the traditional statistical analyses with large time and labor consuming. The approach of artificial intelligence for single-molecule electronics (AI4SME) offers the potential for effective extraction and precise detection of low-probability events due to the capability of more fine-grained data mining and pattern recognition.

Existing machine learning (ML) algorithms can be mainly

Received: December 5, 2024; **Revised:** January 26, 2025
Accepted: January 27, 2025

✉Address correspondence to Wenjing Hong, whong@xmu.edu.cn; Junyang Liu, jyliu@xmu.edu.cn; Haojie Liu, hjliu@xmu.edu.cn

categorized into supervised and unsupervised methods depending on whether real data labeling is needed or not. For instance, Fu et al. developed a supervised convolutional neural network (CNN)-based model to identify molecular junction characteristics with higher accuracy and robustness compared to non-deep learning models [20]. Although great progress is achieved, the performance of supervised learning relies heavily on the quality and quantity of labeled data, which is a time-consuming and labor-intensive in practice. In contrast, unsupervised machine learning algorithms have shown advantages in discovering hidden patterns without prior knowledge [21–25]. For instance, Liu et al. developed a hybrid grid-based density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm to dissect the time-evolved conductance behavior of single-molecule junctions, revealing hidden patterns and dynamics [26]. New experimental phenomena, such as the σ–σ interactions between non-conjugated molecules and the catalytic cycle of formate dehydrogenase, have been discovered, benefiting the ability to identify these events [27, 28]. Although significant advancements are achieved, those prevailing algorithms often overlook the intrinsic distributions of experimental events, which often suffer from the "uniform effect", obscuring the low-probability data points [29–31]. Therefore, it is of great significance to pursue more accurate algorithms to ensure the true distributions of diverse molecular events, which enables us to break through the long-term bottleneck for the AI4SME and even AI4S.

In this study, we propose a multi-prototype clustering (MPC) algorithm to detect low-probability single-molecule events in three typical scenarios, including the quantification of junction formation probability, recognition of low-probability stretching configuration changes by *in situ* mechanical regulation and investigation of structural changes by electric field-driven reaction. The results demonstrate that our algorithm not only accurately distinguishes the events between molecular junction and pure tunneling for the evaluation of junction formation probability but also accurately quantifies the events originating from molecular conformational changes that occur with a probability of less than 10%. Moreover, the MPC algorithm could extract the product species at the onset of the electric field-driven single-molecule keto-enol reaction with only 1.5% probability. The advance of the MPC algorithms significantly promotes the quantitative evaluation of low-probability single-molecule events, which demonstrates a new strategy in AI4SME.

2 Results and discussion

2.1 The multi-prototype clustering algorithm

Low-probability events, such as configuration transformations, are often obscured by traditional conductance-distance histogram analysis and misattributed by existing clustering algorithms, as illustrated by the green curves shown in Figs. 1(a) and 1(b). The

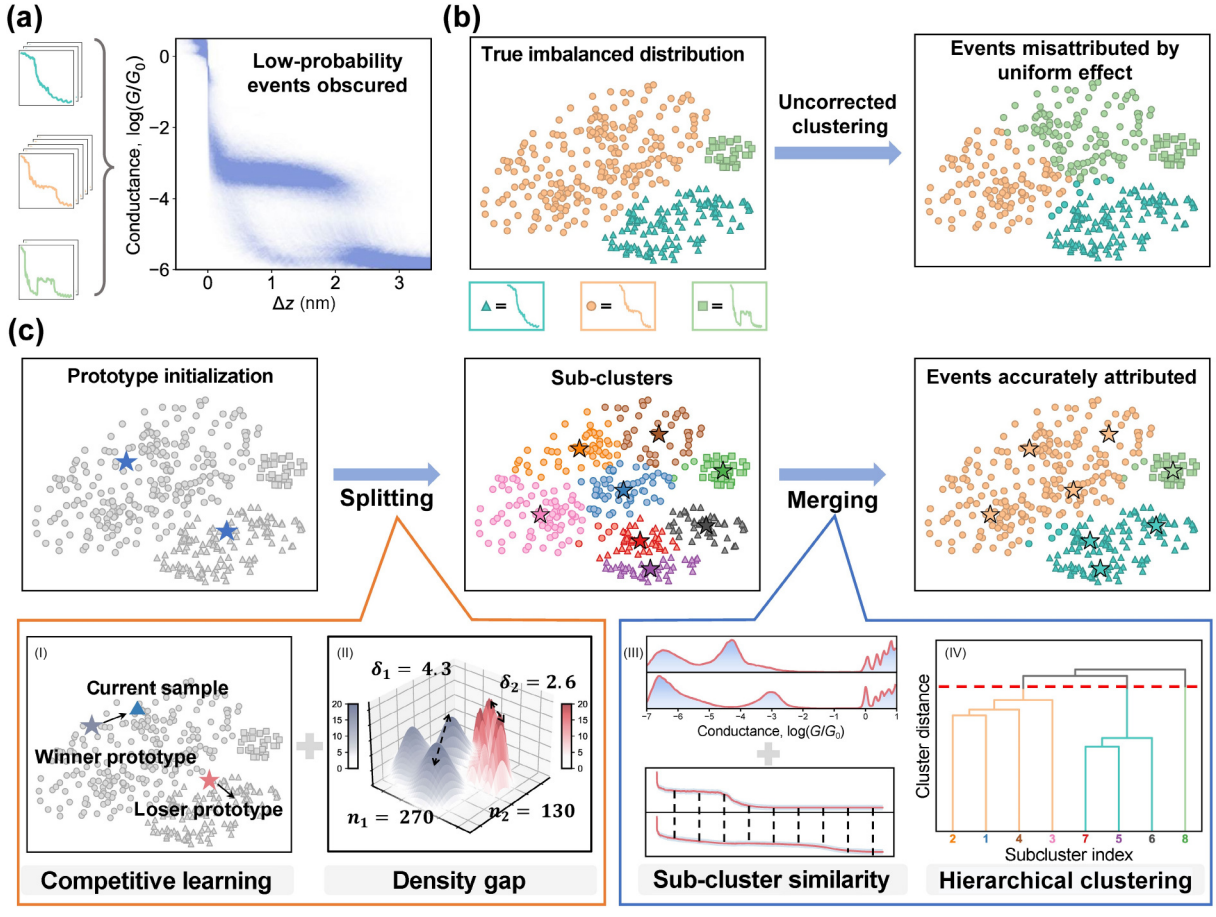


Figure 1 The current challenge in low-probability events detection and the overview of the MPC algorithm. (a) Low-probability events are obscured by traditional statistical analysis. (b) Low-probability events are misattributed by existing clustering algorithms. (c) Schematic of the MPC algorithm, which mainly consists of the splitting and the merging stages.

challenge of misattribution mainly arises from the following reasons. Firstly, the existing algorithms assume that clusters are spherical and uniformly distributed, and they typically start with one prototype per cluster for subsequent iterations. This assumption can be problematic when the experimental datasets have non-spherical shapes and varying distributions [32]. If the initial prototypes are not representative of the actual cluster centers, the algorithm may converge to a suboptimal solution that creates clusters of similar size [33]. Secondly, the optimization often seeks to minimize local within-cluster variance at each step, which tends to produce clusters of similar sizes [30, 31]. These reasons make the existing methods often fall into the pitfall of "uniform effect" by allocating majority events to minority ones, as shown in Fig. 1(b), which has been a long-term bottleneck for accurately identifying low-probability event.

To resolve the uniform effect, we designed the MPC algorithm that used multiple prototypes to represent one cluster, thereby accommodating more complex data with clusters of varying shapes and sizes [34]. We sought a global optimum by combining the advantages of partitioning and hierarchical methods to accurately uncover the minority molecular junction events within imbalanced data distributions [35]. As shown in Fig. 1(c), there are two main steps in the MPC algorithm. To better capture the shape and distribution of each cluster, we first adopted a splitting strategy to separate data into small sub-clusters and simplify the datasets by generating sufficient prototypes. Subsequently, a hierarchical clustering strategy was designed to merge the sub-clusters with high similarity and generate the final clusters with different shapes and sizes.

To enhance the differentiation of sample curves, specific segments were delineated based on the conductance range of interest and then subjected to min–max normalization. This will also reduce the interference of noise points and outliers on distance calculations. After such preprocessing, the MPC finely splits the entire dataset by generating multiple prototypes. The majority class generates more prototypes, while the minority class generates fewer prototypes. This approach reveals latent minority patterns and accurately models the data distribution. To autonomously determine the optimal number of prototypes and to ensure their representativeness, we adopted a strategy that combines competitive learning with the gradual addition of prototypes. Competitive learning is responsible for updating existing prototypes, where for each sample, the nearest prototype serves as the winner and moves closer to the sample to learn its patterns, while other prototypes are penalized by moving away from the sample (Fig. 1(c)(I)) [36]. Prototype generation is determined by the size of the sub-clusters and the distribution of their density peaks [37] (Fig. 1(c)(II)). Larger sub-clusters and those with more distant density peaks are more likely to contain more prototypes. Prototype updating and generation were alternately performed. When a prototype with zero wins is encountered, the splitting phase ends [38]. Ultimately, each prototype belongs to only one sub-cluster.

The splitting stage was designed to capture subtle data patterns but with a potential risk of over-segmentation for larger sub-clusters. To restore the true distribution of the data, sub-clusters were merged appropriately according to their similarity. The assessment of sub-cluster similarity is based on two pivotal characteristics of break junction data traces: the conductance value and the plateau shape (Fig. 1(c)(III)). Specifically, we employed the envelopes of conductance histograms to perform a quantitative evaluation of the similarity in conductance features. This approach transcends the most probable peak values by integrating the distribution information of all samples within the sub-clusters, thus facilitating a more comprehensive and robust assessment. For the plateau shape similarity, we utilized the Euclidean distance between prototypes, which proves to be remarkably effective in assessing the similarity of time series signals [39]. The prototypes learned through the process are much smoother than the original traces, while also capturing the conductance variation patterns of the subclusters. Using these prototypes to assess the similarity between subclusters further reduces the interference from noise and outliers within the subclusters. Then, we adopted a bottom-up hierarchical agglomerative approach to merge sub-clusters based on the joint similarity (Fig. 1(c)(IV)), preventing the minority classes from being mixed with the majority classes again [40].

To evaluate the algorithm, we conducted systematic experiments with many representative unsupervised algorithms on three typical single-molecule event detection datasets with varying complexities, including K-means (the most typical K-centers clustering technique) [41], spectral clustering (SC, classic graph-based algorithm) [22], multi-parameter vector-based classification (MPVC) [42]. The first dataset comprises pure solvent tunneling traces and molecular conductance traces, with various relative proportions to simulate the JFP of single-molecule electrical measurement. The second dataset contains conductance traces from two molecular conformations with different occurrence probabilities in addition to the conductance of pure tunneling. The final dataset originated from the reaction equilibrium established by voltage-driven keto-enol tautomerism. Together, these datasets encompass a wide variety of distributions and conductance characteristics, ranging from highly imbalanced to balanced scenarios, with diverse conductance change behaviors. This diversity ensures a comprehensive evaluation of our method across different experimental conditions.

We assessed the clustering results by the adjusted Rand index (ARI) metric, which is not dependent on the type of categorical data or cluster size and is well-suited for the scenarios of unsupervised learning [43, 44]. Since ARI computation requires true labels, we implemented a rigorous manually annotation process on the first two datasets involving multiple annotators and cross-validation to minimize human bias.

2.2 Quantification of junction formation probability

The quantification of JFP in single-molecule electrical measurements was calculated through the ratio between molecular tunneling and pure tunneling without a molecular junction, which is crucial for advancing the scientific understanding of single-molecule electronics [45–47]. Therefore, the MPC algorithm was first used in the scenario of JFP quantification on a two-pattern dataset, which was acquired from a scanning tunneling microscope BJ (STM-BJ) experiment of the target molecule MPP (4-[4-(methylthio)phenyl]pyridine, $C_{12}H_{11}NS$) concentration of 0.01 mM. The single-molecule conductance of MPP was measured under a bias voltage of 0.1 V between the gold tip and substrate. Due to the dynamic stochastic nature of break junction process, two typical events occur, including pure tunneling and molecular tunneling, depending on whether the target molecule is connected between the electrodes or not, as shown in Fig. 2(a), the corresponding conductance traces and the statistical histogram of the whole dataset are presented in Fig. 2(b) [48]. The MPP data traces clearly

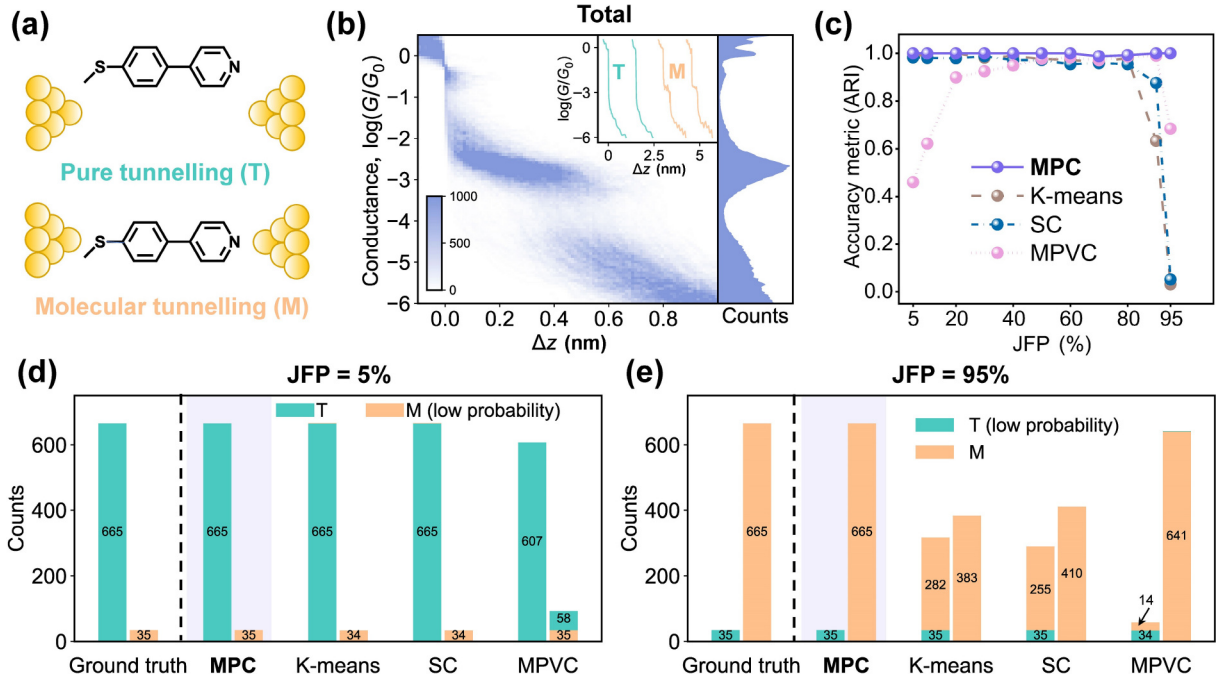


Figure 2 The quantification results of junction formation probability. (a) Typical junction configurations of pure tunneling (labeled as T) and MPP molecular tunneling (labeled as M). (b) The statistical results of the MPP molecule. Left: two-dimensional (2D) conductance-distance histogram; right: 1D conductance histogram; and the inset shows the individual conductance traces of pure tunneling (blue) and molecular tunneling (orange). (c) The ARI values of each clustering algorithm under different JFP scenarios, from the molecular tunneling being the minority class to the majority class. (d) The results comparison of four clustering algorithms against the ground truth when JFP = 5%. (e) The results comparison of four clustering algorithms against the ground truth when JFP = 95%.

display a plateau with the conductance value of $10^{-2.7}$ G_0, where G_0 represents the conductance quantum and equals $2e^2/h$ (77.5 μS, h is Planck's constant). For this whole dataset, the JFP calculated from the results of the MPC clustering is 93%, which is essentially consistent with the 96% obtained through step length filtering, indicating that the MPC method is effective. To further evaluate the generality of the MPC method, we simulated different scenarios of various JFPs by manually constructing 11 datasets, each consisting of 700 conductance-distance traces, including pure tunneling signals (labeled as T) and the MPP tunneling signals (labeled as M), with the JFP ranging from 5% to 95%. The clustering results of four algorithms in different scenarios are shown in Fig. 2(c) and Table S3 in the Electronic Supplementary Material (ESM). Notably, the MPC algorithm consistently achieves the highest ARI scores across all scenarios, with an average score of 0.998. In contrast, the scores of the other three algorithms are consistently lower than that of the MPC, especially under the imbalanced conditions when the JFP is 10% (molecular tunneling is the low-probability event) and 90% (pure tunneling is the low-probability event). For better comparison, the detailed clustering results from each algorithm under the circumstances of JFP = 5% and JFP = 95% are shown in Figs. 2(d) and 2(e) (other clustering results are shown in Fig. S1 in the ESM). Under both conditions, the results of the MPC algorithm are consistent with the ground truth, while the other three algorithms show varying levels of misclassification. Particularly in the case of JFP = 95%, the K-means, SC, and MPVC algorithms cluster many samples from major clusters into minor clusters, demonstrating the characteristic of "uniform effect" [30]. In the case of JFP = 5%, the K-means and SC perform better than in the case of JFP = 95%, because pure tunneling data is naturally denser and has lower within-class variance than that of the molecular tunneling data.

The results show that the MPC algorithm emerges as the superior clustering solution across all test datasets, with an average ARI improvement of 15.4%, 13.4%, and 16.2% compared to those of the K-means, SC, and MPVC algorithms, respectively. Specifically, on imbalanced data (where the minority class constitutes less than 20%), the average ARI improvement increase to 51.5%, 38.4%, and 45.5%. The results demonstrate that our MPC algorithm can precisely differentiate molecular tunneling events from pure tunneling events in conditions of extreme data imbalance, providing an important and reliable method for accurately quantifying the JFP of single-molecule electronics.

2.3 Recognition of low-probability stretching configurations by *in situ* mechanical regulation

Based on the JFP determination, the junction configuration evolution in the break junction process, especially the configurations with low probability but easily being neglected, is the next crucial issue for single-molecule electronics. *In situ* regulation of molecular junction configurations offers a unique way to control and monitor the charge transport through molecular electronic devices. Therefore, distinguishing different molecular junction configurations is vital for understanding the fundamental properties and functions of molecules [8, 49]. We further validated the effectiveness of the MPC algorithm in uncovering the latent configurations in the scenarios exhibiting more intricate single-molecule events with different probabilities. The core of the target molecule composition is indacenodithiophene (IDT), flanked by two 1,3-diethyl-2-thiobarbituric anchoring groups (IDT-T). The molecule has multiple binding sites on the anchoring groups, leading to different events of junction configurations during the break junction process.

As reported in our previous work, there might be at least three

events during the break junction process of the IDT-T molecule [50]. For the convenience of elaboration, we labeled these three types of events as A, B, and C, respectively, as shown in Fig. 3(a). Event A suggests the pure tunnelling without conductance plateau (blue curve in the inset of Fig. 3(b)). Event B shows that the two electrodes are bonded with the sulfur (S) atoms at both ends with a plateau of high conductance (orange curve in the inset of Fig. 3(b)). Event C represents a process during stretching where a structure initially forms with one end bonded to nitrogen atom and the other to sulfur, then transitions to a state where both ends are bonded to sulfur. Accordingly, its conductance trace (green curve in the inset of Fig. 3(b)) exhibits a conductance jumping feature. The statistical conductance histogram of the whole dataset is shown in Fig. 3(b). The histogram explicitly illustrates a plateau correlating to a conductance value of $10^{-3.42}$ G_0, while the different events are difficult to distinguish. For quantitative evaluation, we manually labeled the data and removed noise traces with a final dataset including 324 traces of event A, 1667 traces of event B, and 149 traces of event C (only 6.96% in all traces with low probability). Then, we applied the MPC algorithm to distinguish between different molecular events, and the statistical results of different clusters are shown in Fig. 3(c). It is obvious that the three junction configuration events are clearly separated by the MPC algorithm.

For comparison, the ARI values of the four results obtained from different clustering algorithms were calculated. As shown in Fig. 3(d), MPC achieved the highest ARI score of 0.95, which is 236%, 348%, and 135%, which are higher than those of the K-means, SC, and MPVC, respectively. For further investigation, we analyzed the actual number of samples of each type in the clustering results. As shown in Fig. 3(e), the MPC algorithm demonstrates the most accurate alignment with the ground truth in identifying the low-probability Event C from the unbalanced dataset. Such effectiveness is attributed to the fact that the MPC algorithm divided the dataset into 13 sub-clusters in the splitting phase (Fig. S2 in the ESM), capturing three distinctly different conductance change patterns. The MPC then adaptively determine the optimal number of clusters to be 3 (Fig. S3 in the ESM). In contrast, varying degrees of uniform effects are exhibited in the clustering results of the other three algorithms, with an obvious misclassification among samples across these three categories. The conductance statistics for each cluster obtained by these algorithms are shown in Fig. S4 in the ESM. The results demonstrate that the MPC algorithm can well quantify the probability of different molecular junction configurations during stretching, which provides guidance for research on junction configuration evolution and precise regulation.

2.4 Investigation of structural changes by electric field-driven reaction

In a confined nanospace, the reactions were characterized by an extremely low concentration of reactant molecules, particularly at

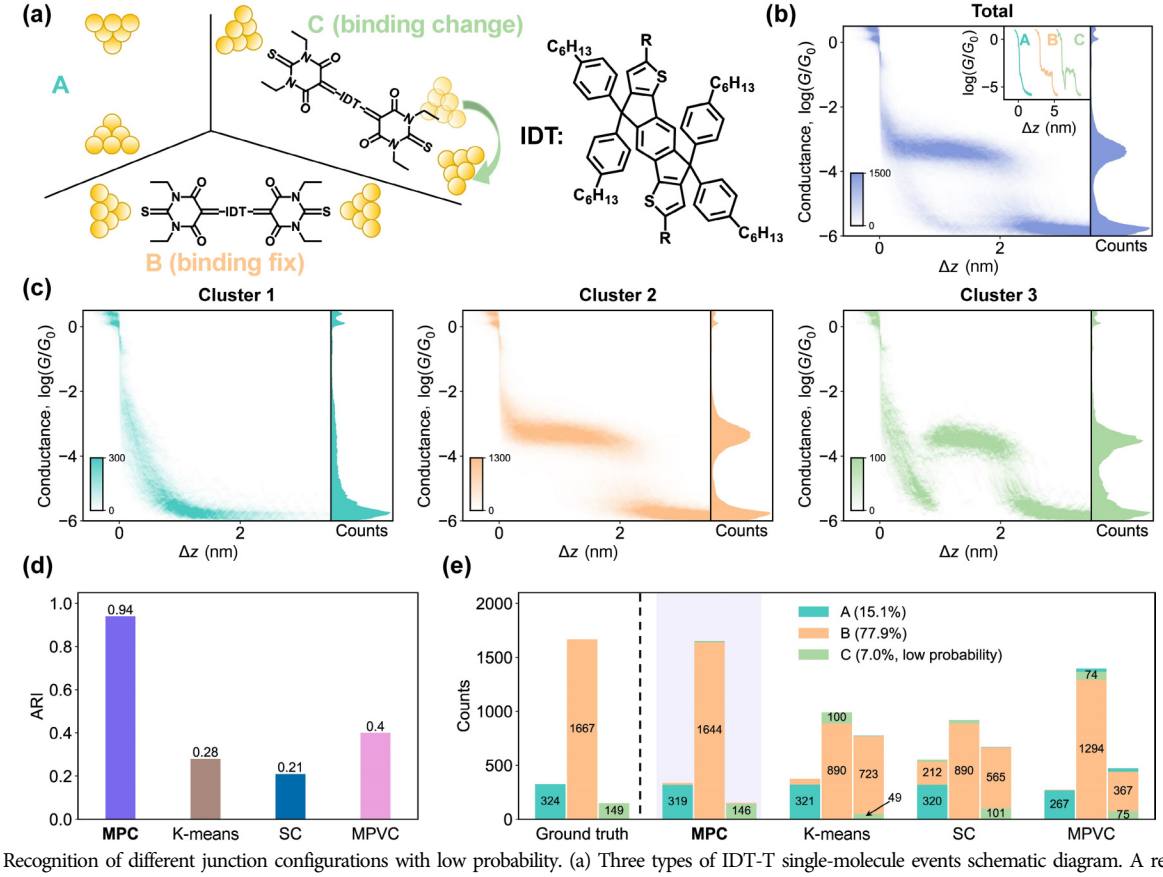


Figure 3 Recognition of different junction configurations with low probability. (a) Three types of IDT-T single-molecule events schematic diagram. A represents no molecules captured between the electrodes, B represents electrodes connected to sulfur at each end (corresponding to high conductance), and C represents a change in the anchor position during stretching (corresponding to a change from low conductance to high conductance). (b) The statistical results of the entire dataset. Left side: 2D conductance-distance histogram; right side: 1D conductance histogram; and inset: three typical conductance traces. (c) 2D conductance-distance histograms and 1D conductance histograms of clusters 1, 2, and 3 from the clustering result of MPC. (d) ARI scores for the four algorithms. (e) The results comparison of four clustering algorithms against the ground truth, with each color representing a molecular event.

the initial stage, making it challenging to perform accurate quantification for reaction dynamics [51, 52], which offers a potential solution for quantitative assessments of reactants and products within such limited volumes [53]. However, conventional classification methods are hindered by the uniform effect, which makes it difficult to accurately distinguish the formed products during the initial stages, particularly at the onset of the reaction. To address this issue, we considered employing our MPC method to extract information about low-probability species during the chemical reaction, which potentially allows us to better understand and quantify the thermodynamics in a confined nanospace. The evaluation was conducted by an electric field-driven single-molecule keto-enol reaction (Fig. 4(a)), and single-molecule measurement of a keto-enol equilibrium was achieved by adjusting the bias voltage between the electrodes at room temperature [54]. The typical conductance traces of the two molecular forms are shown in Fig. 4(b). It has been reported that the enol form appears with low probability at low bias but becomes more and more pronounced with increased bias. However, traditional one-dimensional (1D) conductance histogram can only allow for qualitative observation and cannot quantitatively describe the transition process, as shown in Fig. S5 in the ESM.

To evaluate the reaction process quantitatively, we applied the above four clustering algorithms to characterize the state distribution under different biases. The results of the MPC algorithm are shown in Fig. 4(c), and the results of the other three algorithms are shown in Fig. S6 in the ESM. In this evaluation, the MPC algorithm achieves an impressive result in classifying low-probability high conductance events, which account for only 1.5% when the bias is 0.1 V. In contrast, the proportion of high-conductance events classified by the SC algorithm is 14.9% at the same bias condition. The K-means algorithm exhibits a nearly equal distribution of different conductance stages, where the proportion of low conductance for keto state ("L") and high conductance for enol state ("H") events are classified to be 50.1% and 49.9%, respectively. On the other hand, the results of the MPVC algorithm show that the proportion of high-conductance events (66.9%) is higher than that of low-conductance events (33.1%), which obviously contradicts the results of qualitative analysis. It can be clearly seen from the statistical results that there are different degrees of misclassification in the high-conductivity events classified by the other three comparison algorithms.

Further investigation results of the clustering effects of the four algorithms under different biases is shown in Fig. 4(d). The results show that the MPC algorithm indicates an increasing trend in the proportion of high-conductance data with increasing bias voltage, consistent with the results of qualitative analysis by histogram statistical methods. Specifically, the proportion of the "H" state remains below 10% at the bias voltages ranging from 0.1 to 0.3 V. When the bias increases beyond 0.4 V, the proportion of the "H" state starts to increase significantly, reaching nearly 60% at 0.6 V. In comparison, spectral clustering shows a higher proportion of the "H" state, consistently around 15%, when the bias is below 0.4 V. For K-means, the proportions of the "H" state at the bias ranging from 0.1 to 0.3 V are around 50%, which indicates that the algorithm tends to distribute samples into two clusters evenly. Moreover, the proportions of MPVC show a fluctuation, not reflecting the bias-dependent trend. The results demonstrate the valuable capability of MPC in the identification of rare events for quantitatively evaluating the single-molecule reaction process.

Given the clustering results, a quantitative comparison based on the change in free energy difference ΔG (the free energy difference between "L" and "H" states) at different biases was further made. A detailed theoretical investigation into the enolization isomerization of acetone, utilized as a simplified model, has been thoroughly performed in the presence of an external electric field (EEF) [48]. Although the results are specific to the acetone structure, they are

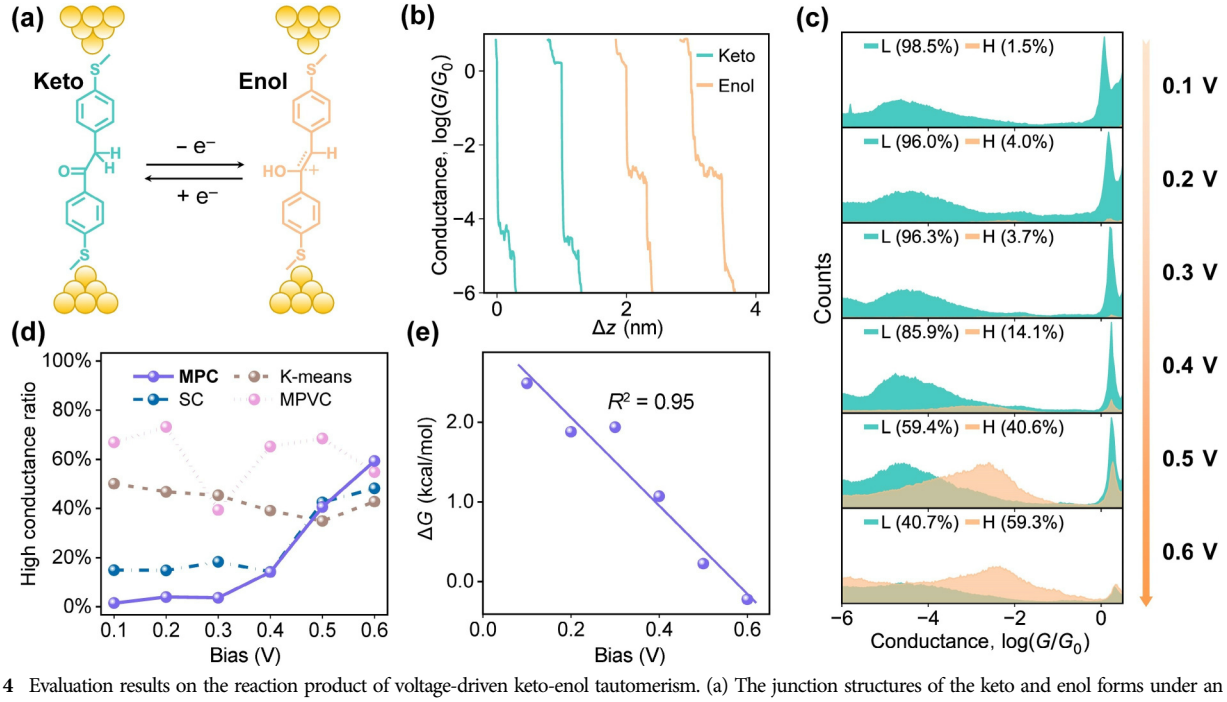


Figure 4 Evaluation results on the reaction product of voltage-driven keto-enol tautomerism. (a) The junction structures of the keto and enol forms under an electric field. (b) The typical conductance traces for the keto form (blue, with low conductance plateau) and enol form (orange, with high conductance plateau). (c) Distribution probabilities of states "L" (blue) and "H" (orange) at different bias voltages based on clustering results using the MPC algorithm. (d) Percentages of high-conductance data in each clustering result under different bias voltages. (e) The linear fitting between ΔG and bias voltage based on clustering results from MPC.

expected to provide insights that may enhance the understanding of keto-enol tautomerism behavior. The calculations reveal that the reaction energy of keto-enol isomerization in EEF exhibits approximately linearly related to the electric field strength (Fig. S7 in the ESM). In particular, when a stronger uniform electric field is applied along the positive direction of the z-axis (F_z (V/nm) < 0), it leads to a reduction in the value of the activation energy and ΔG, indicating that higher EEF strength is more favorable for the formation of enol. Therefore, the relationship between the equilibrium constant (K_{L-H}) and the bias voltage can be defined through ΔG as $-RT\ln K_{L-H} = \Delta G \approx L \times V + \Delta G_0$, where L is the slope, V is the electric field strength, and ΔG_0 is the Gibbs free energy change under standard condition in the absence of EEF [55, 56]. We estimated the K_{L-H} based on the ratio of "H" to "L" obtained by different clustering algorithms. The estimated ΔG values at different biases using MPC are shown in Fig. 4(e) (the results summarized by the other three algorithms are shown in Fig. S8 in the ESM). The coefficient of determination (R^2) obtained from linear regression on the scatter plot can assess the reliability of the computational results. The MPC algorithm exhibits a good linear relationship with an R^2 value of 0.95. In contrast, the SC algorithm exhibits a lower R^2 value of 0.70. Moreover, both the K-means and MPVC algorithms even yield positive slopes, which fail to reflect the experimental observation that higher field strength favors the generation of the enol form compound. Overall, the MPC algorithm yields the results that are closest to the theoretical calculations in both qualitative and quantitative analysis. This suggests that the MPC algorithm has the potential to evaluate reaction thermodynamics by digging deep into single-molecule conductance data. This suggests that the MPC algorithm provides the potential to evaluate the reaction thermodynamics by characterizing single-molecule conductance data.

To ensure the comprehensiveness of the performance validation, we also conducted additional experiments on host-guest interactions and π-stacked dimers datasets. The results are shown in Figs. S9–S12 in the ESM, which demonstrate that the MPC algorithm can accurately distinguishing between different molecular events and configurations across diverse scenarios. Moreover, more comparisons with two representative algorithms, including DBSCAN [57] and Gaussian mixture model (GMM) [58], were conducted (see Figs. S13–S16 in the ESM), which further validates the superior performance of the MPC algorithm in identifying low-probability events in single-molecule electronics. Additionally, the MPC algorithm exhibits a certain level of robustness to noise, as illustrated in Fig. S17 in the ESM.

3 Conclusions

In this study, we propose an MPC algorithm to address the inherent challenges of identifying low-probability events in single-molecule electronics and validate it across several typical scenarios. For the quantification of different junction formation probabilities of an MPP molecular junction, MPC consistently achieved the best performance across 11 different sub-datasets, with an average ARI score of 0.998 higher than other compared algorithms. For the junction configuration recognition during the dynamic stretching of the IDT-T molecular junction, the MPC algorithm's ARI score are 236%, 348%, and 135% higher than those of the three benchmark methods, K-means, SC, and MPVC, respectively. For the structural changes detection in the electric field-driven single-molecule keto-enol reaction process, the MPC algorithm can precisely extract the product species at the onset of the electric field-driven single-molecule keto-enol reaction with only 1.5% probability, enabling the calculation of the equilibrium constant and the quantitative evaluation of reaction thermodynamics. The significant advancements in the MPC algorithm demonstrate its applicability to diverse data distributions, particularly overcoming the uniformity effect on imbalanced data, paving the way for more nuanced and precise scientific insights in the realm of single-molecule electronics. It is envisioned that this algorithm will have broad applications across various scientific domains, helping to uncover underlying physical and chemical mechanisms that were previously elusive.

Electronic Supplementary Material: Supplementary material (details of the algorithm design, theoretical calculations, Figs. S1–S17, and Tables S1–S3) is available in the online version of this article at https://doi.org/10.26599/NR.2025.94907276.

Data availability

All data needed to support the conclusions in the paper are presented in the manuscript and the Electronic Supplementary Material. Additional data related to this paper may be requested from the corresponding author upon request.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2024YFA1208103), the National Natural Science Foundation of China (Nos. 22403079, 22173075, 22325303, 21933012, and 22250003), the Fujian Provincial Department of Science and Technology (Nos. 2022H6014 and 2023H6002), the Fundamental Research Funds for the Central Universities (Nos. 20720220020 and 20720200068).

Declaration of competing interest

All the contributing authors report no conflict of interests in this work.

Author contribution statement

C. S.: Algorithm design, code implementation, experimental design and analysis, writing – original draft, writing – review & editing. R. G. T.: Dataset construction, experiment execution, writing – editing. S. L. X.: Theoretical calculations, writing – review & editing. X. P. L.: Dataset construction, experiment execution. T. G. L.: Data curation. Y. X. Z.: Data curation. C. T.: Data curation, writing – review & editing. J. L.: Data analysis. Y. Z.: Data analysis. H. J. L.: Funding acquisition, resources, writing – review & editing. J. Y. L.: Funding acquisition, resources, writing – review & editing. W. J. H.: Funding acquisition, resources, supervision. All the authors have approved the final manuscript.

Use of AI statement

None.

References

[1] Suvarna, M.; Pérez-Ramírez, J. Embracing data science in catalysis research. *Nat. Catal.* **2024**, *7*, 624–635.

[2] Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

[3] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

[4] Wang, H. C.; Fu, T. F.; Du, Y. Q.; Gao, W. H.; Huang, K. X.; Liu, Z. M.; Chandak, P.; Liu, S. C.; Van Katwyk, P.; Deac, A. et al. Scientific discovery in the age of artificial intelligence. *Nature* **2023**, *620*, 47–60.

[5] Xu, Y. J.; Wang, F.; An, Z. L.; Wang, Q.; Zhang, Z. Artificial intelligence for science-bridging data to wisdom. *Innovation* **2023**, *4*, 100525.

[6] Reed, M. A.; Zhou, C.; Muller, C. J.; Burgin, T. P.; Tour, J. M. Conductance of a molecular junction. *Science* **1997**, *278*, 252–254.

[7] Xu, B. Q.; Tao, N. J. Measurement of single-molecule resistance by repeated formation of molecular junctions. *Science* **2003**, *301*, 1221–1223.

[8] 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*, 904–907.

[9] Liu, H. J.; Chen, L. J.; Zhang, H.; Yang, Z. Q.; Ye, J. Y.; Zhou, P.; Fang, C.; Xu, W.; Shi, J.; Liu, J. Y. et al. Single-molecule photoelectron tunnelling spectroscopy. *Nat. Mater.* **2023**, *22*, 1007–1012.

[10] Wang, L.; Gong, Z. L.; Li, S. Y.; Hong, W. J.; Zhong, Y. W.; Wang, D.; Wan, L. J. Molecular conductance through a quadruple-hydrogen-bond-bridged supramolecular junction. *Angew. Chem., Int. Ed.* **2016**, *55*, 12393–12397.

[11] Li, R. H.; Lu, Z. Y.; Cai, Y. T.; Jiang, F.; Tang, C.; Chen, Z. X.; Zheng, J. T.; Pi, J. C.; Zhang, R.; Liu, J. Y. et al. Switching of charge transport pathways via delocalization changes in single-molecule metallacycles junctions. *J. Am. Chem. Soc.* **2017**, *139*, 14344–14347.

[12] Giacalone, F.; Herranz, M. Á.; Grüter, L.; González, M. T.; Calame, M.; Schönenberger, C.; Arroyo, C. R.; Rubio-Bollinger, G.; Vélez, M.; Agraït, N. et al. Tetrathiafulvalene-based molecular nanowires. *Chem. Commun.* **2007**, 4854–4856.

[13] Géranton, G.; Seiler, C.; Bagrets, A.; Venkataraman, L.; Evers, F. Transport properties of individual C_{60}-molecules. *J. Chem. Phys.* **2013**, *139*, 234701.

[14] Sadeghi, H.; Sangtarash, S.; Lambert, C. Robust molecular anchoring to graphene electrodes. *Nano Lett.* **2017**, *17*, 4611–4618.

[15] Hong, W. J.; Li, H.; Liu, S. X.; Fu, Y. C.; Li, J. F.; Kaliginedi, V.; Decurtins, S.; Wandlowski, T. Trimethylsilyl-terminated oligo(phenylene ethynylene)s: An approach to single-molecule junctions with covalent Au–C σ-bonds. *J. Am. Chem. Soc.* **2012**, *134*, 19425–19431.

[16] 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.

[17] Yoshioka, K.; Katayama, I.; Minami, Y.; Kitajima, M.; Yoshida, S.; Shigekawa, H.; Takeda, J. Real-space coherent manipulation of electrons in a single tunnel junction by single-cycle terahertz electric fields. *Nat. Photonics* **2016**, *10*, 762–765.

[18] Zhou, C.; Li, X. X.; Gong, Z. L.; Jia, C. C.; Lin, Y. W.; Gu, C. H.; He, G.; Zhong, Y. W.; Yang, J. L.; Guo, X. F. Direct observation of single-molecule hydrogen-bond dynamics with single-bond resolution. *Nat. Commun.* **2018**, *9*, 807.

[19] Zhong, X. L.; Pandey, R.; Karna, S. P. Stacking dependent electronic structure and transport in bilayer graphene nanoribbons. *Carbon* **2012**, *50*, 784–790.

[20] Fu, T. R.; Zang, Y. P.; Zou, Q.; Nuckolls, C.; Venkataraman, L. Using deep learning to identify molecular junction characteristics. *Nano Lett.* **2020**, *20*, 3320–3325.

[21] Huang, F. F.; Li, R. H.; Wang, G.; Zheng, J. T.; Tang, Y. X.; Liu, J. Y.; Yang, Y.; Yao, Y.; Shi, J.; Hong, W. J. Automatic classification of single-molecule charge transport data with an unsupervised machine-learning algorithm. *Phys. Chem. Chem. Phys.* **2020**, *22*, 1674–1681.

[22] Lin, L. C.; Tang, C.; Dong, G.; Chen, Z. X.; Pan, Z. C.; Liu, J. Y.; Yang, Y.; Shi, J.; Ji, R. R.; Hong, W. J. Spectral clustering to analyze the hidden events in single-molecule break junctions. *J. Phys. Chem. C* **2021**, *125*, 3623–3630.

[23] Vladyka, A.; Albrecht, T. Unsupervised classification of single-molecule data with autoencoders and transfer learning. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 035013.

[24] Magyarkuti, A.; Balogh, N.; Balogh, Z.; Venkataraman, L.; Halbritter, A. Unsupervised feature recognition in single-molecule break junction data. *Nanoscale* **2020**, *12*, 8355–8363.

[25] Zheng, J. T.; Gao, W. L.; Lu, T. G.; Chen, L. J.; Lin, L. C.; Huang, R. Y.; Tang, Y. X.; Dong, G.; Liu, J. Y.; Pan, Y. F. et al. Structure identification for force-induced reaction using single-molecule conductance measurement. *CCS Chem.* **2023**, *5*, 1888–1895.

[26] Liu, B.; Murayama, S.; Komoto, Y.; Tsutsui, M.; Taniguchi, M. Dissecting time-evolved conductance behavior of single molecule junctions by nonparametric machine learning. *J. Phys. Chem. Lett.* **2020**, *11*, 6567–6572.

[27] Zhang, A. H.; Zhuang, X. Y.; Liu, J.; Huang, J. C.; Lin, L. C.; Tang, Y. X.; Zhao, S. Q.; Li, R. H.; Wang, B. J.; Fang, B. S. et al. Catalytic cycle of formate dehydrogenase captured by single-molecule conductance. *Nat. Catal.* **2023**, *6*, 266–275.

[28] Feng, A. N.; Zhou, Y.; Al-Shebami, M. A. Y.; Chen, L. C.; Pan, Z. C.; Xu, W.; Zhao, S. Q.; Zeng, B. F.; Xiao, Z. Y.; Yang, Y. et al. σ–σ stacked supramolecular junctions. *Nat. Chem.* **2022**, *14*, 1158–1164.

[29] 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.

[30] Xiong, H.; Wu, J. J.; Chen, J. K-means clustering versus validation measures: A data-distribution perspective. *IEEE Trans. Syst. Man Cybern. B* **2009**, *39*, 318–331.

[31] Zhou, K. L.; Yang, S. L. Effect of cluster size distribution on clustering: A comparative study of K-means and fuzzy c-means clustering. *Pattern Anal. Appl.* **2020**, *23*, 455–466.

[32] Krishnapuram, R.; Kim, J. A note on the Gustafson–Kessel and adaptive fuzzy clustering algorithms. *IEEE Trans. Fuzzy Syst.* **1999**, *7*, 453–461.

[33] Fränti, P.; Sieranoja, S. K-means properties on six clustering benchmark datasets. *Appl. Intell.* **2018**, *48*, 4743–4759.

[34] Guha, S.; Rastogi, R.; Shim, K. CURE: An efficient clustering algorithm for large databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, Seattle, USA, 1998, pp 73–84.

[35] Salvador, S.; Chan, P. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* **2007**, *11*, 561–580.

[36] Lu, Y.; Cheung, Y. M.; Tang, Y. Y. Self-adaptive multiprototype-based competitive learning approach: A K-means-type algorithm for imbalanced data clustering. *IEEE Trans. Cybern.* **2021**, *51*, 1598–1612.

[37] Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492–1496.

[38] Xu, L.; Krzyzak, A.; Oja, E. Rival penalized competitive learning for clustering analysis, RBF net, and curve detection. *IEEE Trans. Neural Netw.* **1993**, *4*, 636–649.

[39] Serrà, J.; Arcos, J. L. An empirical evaluation of similarity measures for time series classification. *Knowl.-Based Syst.* **2014**, *67*, 305–314.

[40] Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview. *WIREs Data Min. Knowl. Discov.* **2012**, *2*, 86–97.

[41] Ikotun, A. M.; Ezugwu, A. E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **2023**, *622*, 178–210.

[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] Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218.
[44] Chen, Z. L.; Goldwasser, J.; Tuckman, P.; Liu, J.; Zhang, J.; Gerstein, M. Forest fire clustering for single-cell sequencing combines iterative label propagation with parallelized Monte Carlo simulations. *Nat. Commun.* **2022**, *13*, 3538.
[45] Liu, J. Y.; Zhao, X. T.; Al-Galiby, Q.; Huang, X. Y.; Zheng, J. T.; Li, R. H.; Huang, C. C.; Yang, Y.; Shi, J.; Manrique, D. Z. et al. Radical-enhanced charge transport in single-molecule phenothiazine electrical junctions. *Angew. Chem., Int. Ed.* **2017**, *56*, 13061–13065.
[46] Limburg, B.; Thomas, J. O.; Holloway, G.; Sadeghi, H.; Sangtarash, S.; Hou, I. C. Y.; Cremers, J.; Narita, A.; Müllen, K.; Lambert, C. J. et al. Anchor groups for graphene-porphyrin single-molecule transistors. *Adv. Funct. Mater.* **2018**, *28*, 1803629.
[47] Zhao, S. Q.; Wu, Q. Q.; Pi, J. C.; Liu, J. Y.; Zheng, J. T.; Hou, S. J.; Wei, J. Y.; Li, R. H.; Sadeghi, H.; Yang, Y. et al. Cross-plane transport in a single-molecule two-dimensional van der Waals heterojunction. *Sci. Adv.* **2020**, *6*, eaba6714.
[48] Tang, C.; Su, M. L.; Lu, T. G.; Zheng, J. T.; Wang, J. J.; Zhou, Y.; Zou, Y. L.; Liu, W. Q.; Huang, R. Y.; Xu, W. et al. Massive acceleration of S_N2 reaction using the oriented external electric field. *Chem. Sci.* **2024**, *15*, 13486–13494.
[49] Su, T. A.; Li, H. X.; Steigerwald, M. L.; Venkataraman, L.; Nuckolls, C. Stereoelectronic switching in single-molecule junctions. *Nat. Chem.* **2015**, *7*, 215–220.
[50] Zhu, Y. X.; Zhou, Y.; Ren, L.; Ye, J. Y.; Wang, H. C.; Liu, X. Y.; Huang, R. Y.; Liu, H. J.; Liu, J. Y.; Shi, J. et al. Switching quantum interference in single-molecule junctions by mechanical tuning. *Angew. Chem., Int. Ed.* **2023**, *62*, e202302693.
[51] Qiu, K. P.; Fato, T. P.; Yuan, B.; Long, Y. T. Toward precision measurement and manipulation of single-molecule reactions by a confined space. *Small* **2019**, *15*, 1805426.
[52] Liu, W.; Yang, Z. L.; Yang, C. N.; Ying, Y. L.; Long, Y. T. Profiling single-molecule reaction kinetics under nanopore confinement. *Chem. Sci.* **2022**, *13*, 4109–4114.
[53] Huang, X. Y.; Tang, C.; Li, J. Q.; Chen, L. C.; Zheng, J. T.; Zhang, P.; Le, J. B.; Li, R. H.; Li, X. H.; Liu, J. Y. et al. Electric field-induced selective catalysis of single-molecule reaction. *Sci. Adv.* **2019**, *5*, eaaw3072.
[54] Tang, C.; Stuyver, T.; Lu, T. G.; Liu, J. Y.; Ye, Y. L.; Gao, T. Y.; Lin, L. C.; Zheng, J. T.; Liu, W. Q.; Shi, J. et al. Voltage-driven control of single-molecule keto-enol equilibrium in a two-terminal junction system. *Nat. Commun.* **2023**, *14*, 3657.
[55] Shaik, S.; Ramanan, R.; Danovich, D.; Mandal, D. Structure and reactivity/selectivity control by oriented-external electric fields. *Chem. Soc. Rev.* **2018**, *47*, 5125–5145.
[56] Yang, C.; Liu, Z. T.; Li, Y. W.; Zhou, S. Y.; Lu, C. X.; Guo, Y. L.; Ramirez, M.; Zhang, Q. Z.; Li, Y.; Liu, Z. R. et al. Electric field-catalyzed single-molecule Diels-Alder reaction dynamics. *Sci. Adv.* **2021**, *7*, eabf0689.
[57] Ester, M.; Kriegel, H. P.; Sander, J.; Xu, X. W. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, USA, 1996, pp 226–231.
[58] Williams, C. K. I.; Rasmussen, C. E. Gaussian processes for regression. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, Denver, USA, 1995, pp 514–520.