Electronic Supplementary Information

Automatic classification of single-molecule charge transport data with an unsupervised machine-learning algorithm

Feifei Huang,[a] Ruihao Li,[a] Gan Wang,[a] Jueting Zheng,[a] Yongxiang Tang,[a] Junyang Liu,[a] Yang Yang,[a] Yuan Yao,*[b] Jia Shi,*[a] and Wenjing Hong*[a]

[a] State Key Laboratory of Physical Chemistry of Solid Surfaces, College of Chemistry and Chemical Engineering, *i*ChEM, Pen-Tung Sah Institute of Micro-Nano Science and Technology, Xiamen University, Xiamen 361005

[b] Department of Chemical Engineering, National Tsing Hua University, Hsinchu 30013

** Correspondence and requests for materials should be addressed to W. H. (E-mail: whong@xmu.edu.cn) or to J. S. (E-mail: jshi@xmu.edu.cn) or to Y. Y. (E-mail: yyao@mx.nthu.edu.tw)*

Table of contents

1. Mechanically controlled break junction

Mechanically controlled break junction (MCBJ) is a technique to measure the conductance of a molecule which is trapped between two gold electrodes. During the experiment, a sheet of spring steel was used as a substrate and fixed by two supports on both ends. The notched wire and liquid cell were fixed upon the substrate. Then we added the target molecules and solvent into the liquid cell. The sheet was bent and released by the pushing rod which is driven by a combination of a stepping motor and a piezo stack, which results in repeated breaking and connecting of gold wire. By applying a bias voltage of 100 mV, the real-time conductance was recorded by the lab-built I-V converter with a sampling rate of 20 kHz. For each experiment, the breaking and connecting process was performed thousands of times, and through this, we could get thousands of conductance-distance traces for statistical analysis. More details about our MCBJ setup are published elsewhere [1, 2].

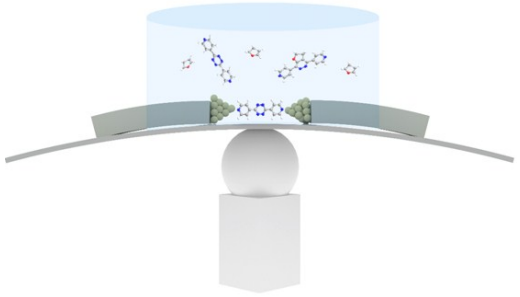

Figure S1. Schematic of our MCBJ setup.

2. Methods

2.1 Pseudo-code of the deep autoencoder K-means (DAK) algorithm

We have described the details of the DAK algorithm in section 2. To get a more intuitive understanding of the algorithm, we give the pseudo-code of the DAK algorithm below (Algorithm 1). The parameters setting of the DAK algorithm for the experiments is shown in Table S1. The neural network architecture of stacked auto-encoder is shown in Table S2.

For the single-molecule charge transport data, there is no universal and feasible method to determine the clustering number automatically. Alternatively, trial and error method is an effective and feasible method because there is only a few types of single-molecule junction events in most cases of single-molecule measurements. In this work, we adopted the trial and error method.

Table S1. DAK algorithm parameters setting

	Learning rate (α)	Batch size (m)	# Clusters (K)
DAK	0.00008	100	4 (Simulation data) 2 (Experiment 1) 3 (Experiment 2) 2 (Experiment 3)

Algorithm 1: Deep Auto-encoder K-Means (DAK)

Require: Stacked auto-encoder (SAE) including the encoder network f_{θ_1} and the decoder network g_{θ_2}

Require: A single-molecule charge transport dataset $G = \{G_{1:M}\}$, learning rate α, batch size m, and the number of clusters K

Training for SAE:

1. Initialize f_{θ_1}, g_{θ_2} with random weights θ_1, θ_2
2. **repeat**:
3. Random sample a batch of the data $\{G_i\}_{i=1}^{m}$ from the dataset $G = \{G_{1:M}\}$.
4. Compute the loss of the SAE according to Eq. (1)

5. Update encoder and decoder parameters (θ_1 , θ_2) via gradient descent algorithm.
6. **until** the parameters (θ_1 , θ_2) of SAE converge to optimal values (θ_1^* , θ_2^*)

Feature extraction:

7. Get the encoder outputs as the features of the dataset according to Eq. (2)

Clustering:

8. Initialize K clustering centroids of K-Means $\{C_j\}_{j=1}^{K}$
9. **repeat**
10. Form K clusters by assigning all points to the closest centroid as Eq. (4)
11. Recalculate the centroid for each cluster by Eq. (5)
12. **until** clustering centroids $\{C_j\}_{j=1}^{K}$ converges.

2.2 Simulation model of the single-molecule break-junction process

In order to obtain the conductance-distance trace data ($G(z)$) of the single-molecule break-junction process, we divide the single-molecule break-junction into four regions, as shown in Figure S2, and build the model for each region as follows:

Region (I) (as shown in Figure S2(b)(I)) represents the tunneling process before the formation of molecular junctions. The value of $\log(G/G_0)$ is calculated by:

$$G(z) = G_0 \times \exp(-\beta_T \times z) \tag{2.1}$$

$$\log(G(z)/G_0) = -0.434\beta_T z = -\beta z \tag{2.2}$$

where β is tunneling decay constant.

Region (II) (as shown in Figure S2(b)(II)) represents charge transport through the molecule, implying that conductance value G remains approximately constant at the conductance value G_p.

Region (III) (as shown in Figure S2(b)(III)) is directed tunneling caused by the rupture of the molecular junction. The tunneling decay constant β of Region (III) and Region (II) are the same.

Region (IV) represents the background noise of our instrument. Let z_b the break-off distance and G_b the conductance value resulted from background noise.

Based on the above model, we define four types of conductance traces by setting the parameters of the models as follows:

Direct tunneling (DT): $\beta = 1.0$ Å^{-1} (STDEV = 0.01)
$G_b = -7.0 \log(G/G_0)$

High conductance (HC): $\beta = 1.0$ Å^{-1} (STDEV = 0.01)
$\log(G/G_0)$ (STDEV = 0.2)
$z_b = 2.0$ nm (STDEV = 0.01)
$G_b = -7.0 \log(G/G_0)$

Low conductance (LC): $\beta = 1.0$ Å^{-1} (STDEV = 0.01)
$G_p = -5.0 \log(G/G_0)$(STDEV = 0.2)
$z_b = 2.3$ nm (STDEV = 0.01)
$G_b = -7.0 \log(G/G_0)$

Sinusoidal plateau (Sine): $\beta = 1.0$ Å^{-1} (STDEV = 0.01)
$G_p = -5.0 \log(G/G_0)$(STDEV = 0.2)
$z_b = 2.3$ nm (STDEV = 0.01)
$G_b = -7.0 \log(G/G_0)$
Sinusoidal amplitude: $0.6\log(G/G_0)$; Sinusoidal frequency: 1.5Hz.

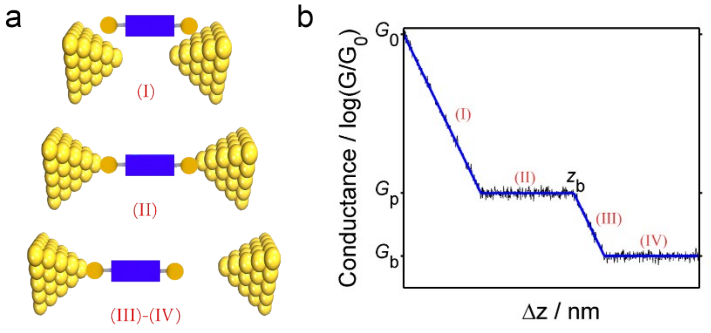

Figure S2. Simulation data. (**a**) Schematic of $G(z)$ during junction formation and rupture. (**b**) Simulated $G(z)$ trace (blue) with noise (black). Conductance at (I)-(IV) corresponds to three configurations of break junction shown in (a).

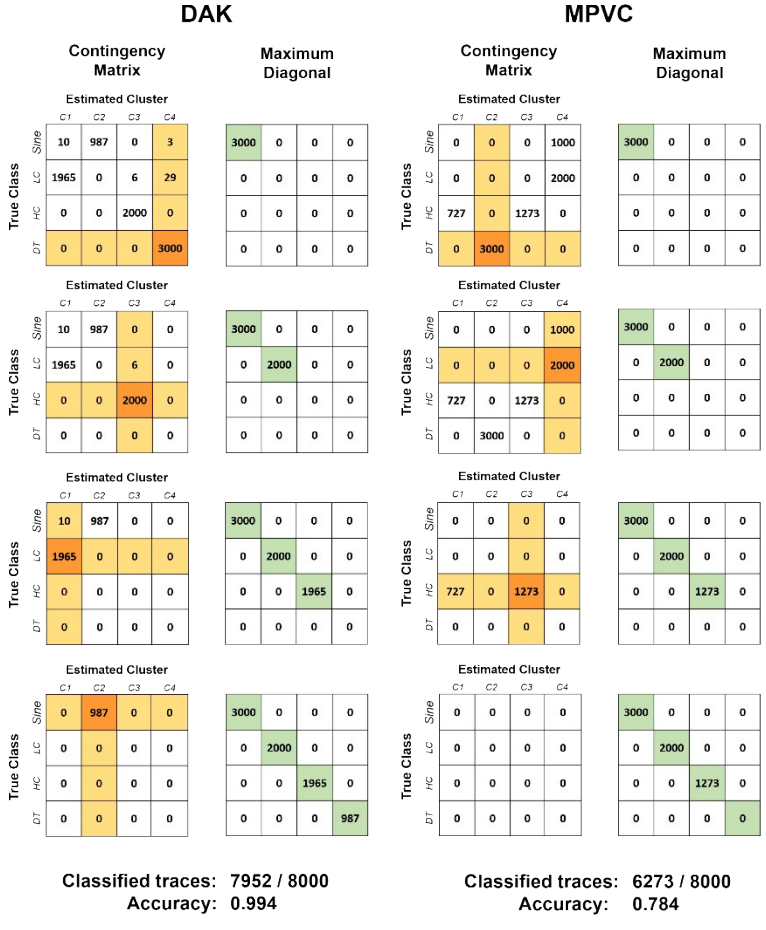

Figure S3. Procedures for the maximum diagonal computation in DAK and MPVC. The contingency matrix computes the intersection cardinality for every true and estimated cluster pair. In order to calculate the maximum diagonal, we select the cell with the maximum value of the contingency matrix and move the selected cell to the upper-left corner of a maximum diagonal matrix, and then other cells in the same row and column are turned to 0. In the first row of the DAK algorithm contingency matrix, when the cells with maximum value 3000 is selected, all cell in the same row and column turned to 0, the 32 left which is unselected is misclassified samples.

2.3 Details of stacked auto-encoder

2.3.1 The network architecture of stacked auto-encoder

Stacked auto-encoder (SAE) is one of the widely used deep unsupervised learning approaches, which is made of encoder and decoder neural networks [3,4]. Both encoder and decoder networks are composed of several dense layers. For each layer the number of the neural and activation

function should be determined before training. The detailed information of the neural network architecture used in this paper is summarized in Table S2.

Table S2. The network architecture of SAE

	Layer name	#Nodes	Activation function
Encoder	Input layer	X	None
	Hidden layer 1	1000	Sigmoid
	Hidden layer 2	500	Sigmoid
Code/Feature	Code layer	3 (Simulation data) 10 (Experiment 1) 20 (Experiment 2) 50 (Experiment 3)	Sigmoid
Decoder	Hidden layer 1	500	Sigmoid
	Hidden layer 2	1000	Sigmoid
	Output layer	X	None

#Node: the value of X is determined according to the dimension of the trace.

2.3.2 Detailed information on neural network training

In order to extract the features from the trace data automatically, the neural network of SAE must be trained to minimize the loss function which is defined by Eq.(1). We initialize the weights and biases of the neural network with a random number with the standard normal distribution. Then the training is conducted by adaptive moment estimation (Adam) algorithm [5]. The learning rate is 0.00008 and the training epoch is set to 50000. The mini-batch size for each epoch is set to 100 (as shown in Table S1). Figures S4-S6 show the detailed information of training results for **Experiment** 1-3, respectively.

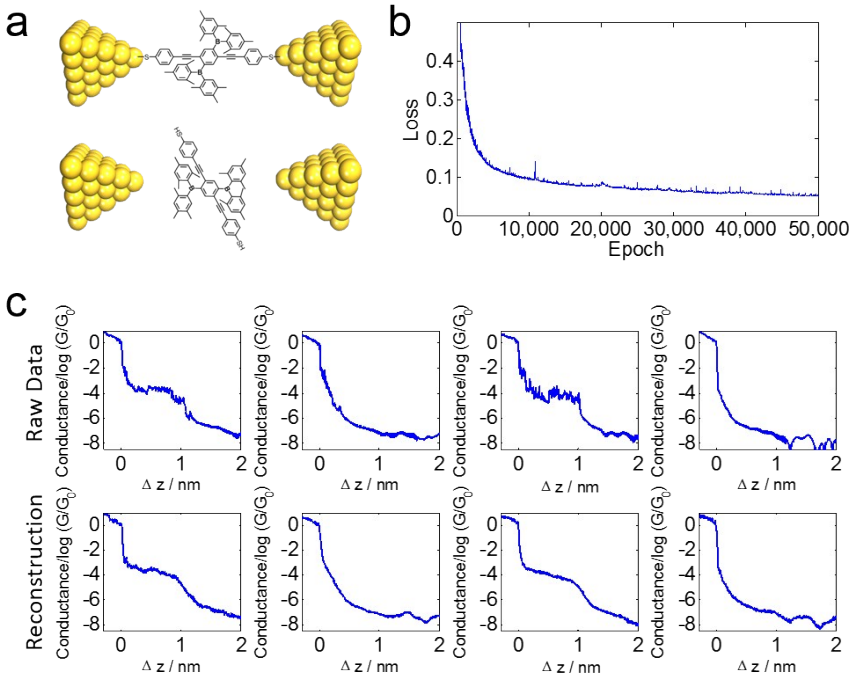

Figure S4. Detailed information of SAE training for Experiment 1. (**a**) Schematic of break junction in Experiment 1. (**b**) The loss curve of the SAE training process, and the training epoch is set to 50000. (**c**) Conductance curves which are randomly selected from raw data (the first row) and the corresponding reconstructed curves of SAE after training (the second row).

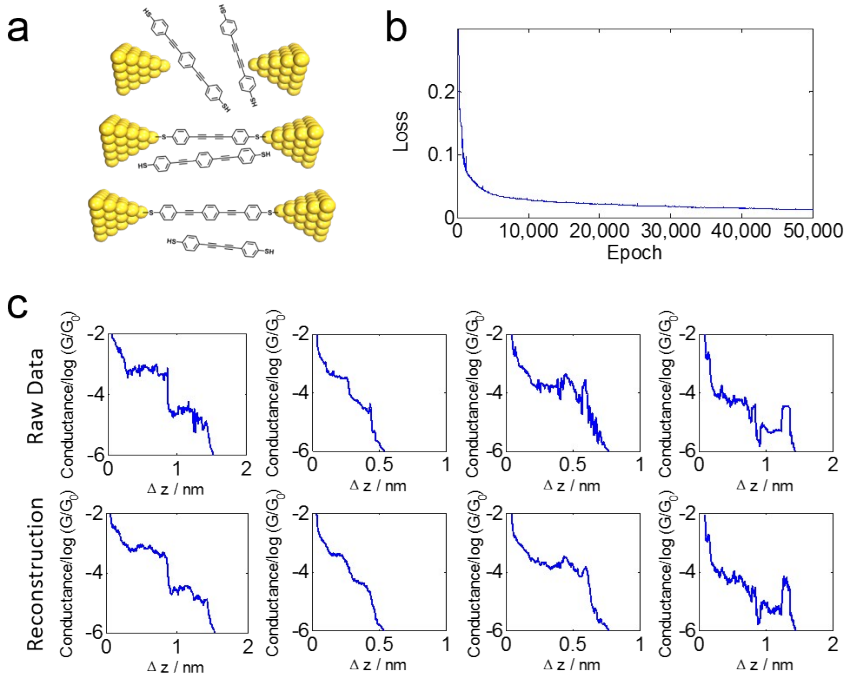

Figure S5. Detailed information on SAE training for Experiment 2. (**a**) Schematic of break junction in Experiment 2. (**b**) The loss curve of SAE training process, and the training epoch is set to 50000. (**c**) Conductance curves which are randomly selected from raw data (the first row) and the corresponding reconstructed curves of SAE after training (the second row).

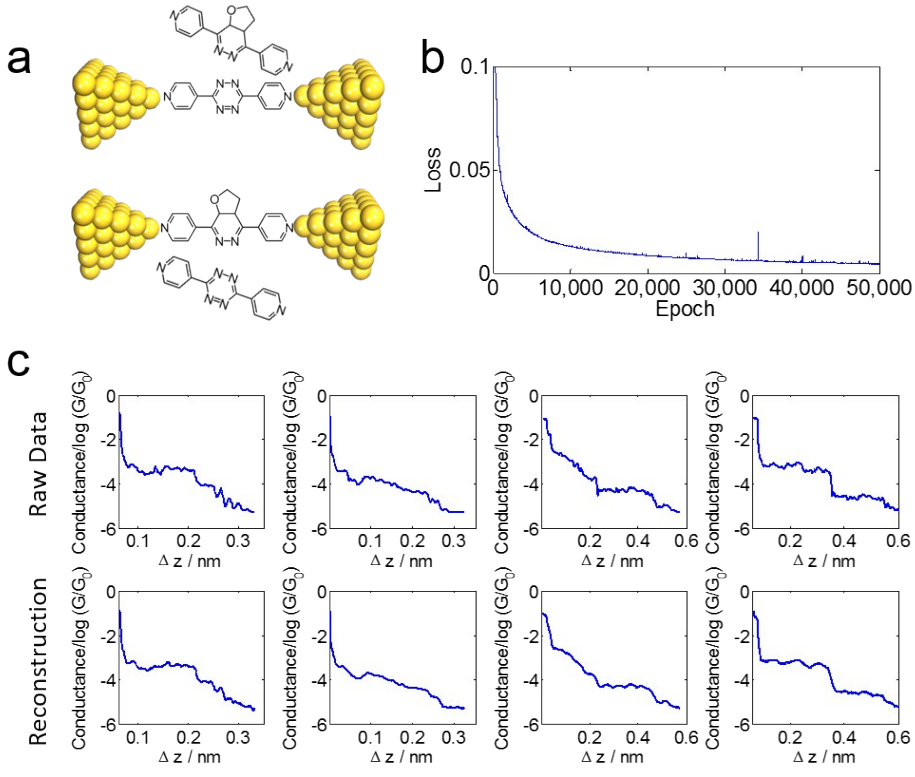

Figure S6. Detailed information on SAE training for Experiment 3. (**a**) Schematic of break junction in Experiment 3. (**b**) The loss curve of the SAE training process, and the training epoch is set to 50000. (**c**) Conductance curves which are randomly selected from raw data (the first row) and the corresponding reconstructed curves of SAE after training (the second row)

2.4 Visualization of the feature space based on simulation data

In order to compare the performances of the proposed DAK algorithm and MPVC algorithm developed by Lemmer et al. [6], the feature space of the two algorithms are visualized, as shown in Figures S7-S8, where Figure S7 (a) and Figure S8 (a) give the visualizations of the real type in the feature space based on the DAK algorithm and the MPVC algorithm, respectively, and Figure S7 (b) and Figure S8 (b) show the classification results of the DAK algorithm and the MPVC algorithm, respectively.

Figure S7 shows that the clustering results (Figure S7c) based on the DAK algorithm are consistent with the true type of data (Figure S7b) except for a very few datapoints classified incorrectly, as indicated by the dash circle in Figure S7c. However, for the feature space determined by MPVC algorithm, the datapoints with the low conductance and the oscillation mode, i.e. the light blue points and dark blue points, overlap seriously and cannot be identified by MPVC algorithm. For the datapoints correspond to the high conductance, i.e. yellow points, MPVC algorithm mistakenly categorized them into two groups. Compared with MPVC algorithm, the DAK algorithm employs a deep neural network for feature extraction from all data traces. Benefiting from that, it leads to a better separability of the datapoints in feature space, and thus provides a more accurate result, as shown in Figure S7, S8.

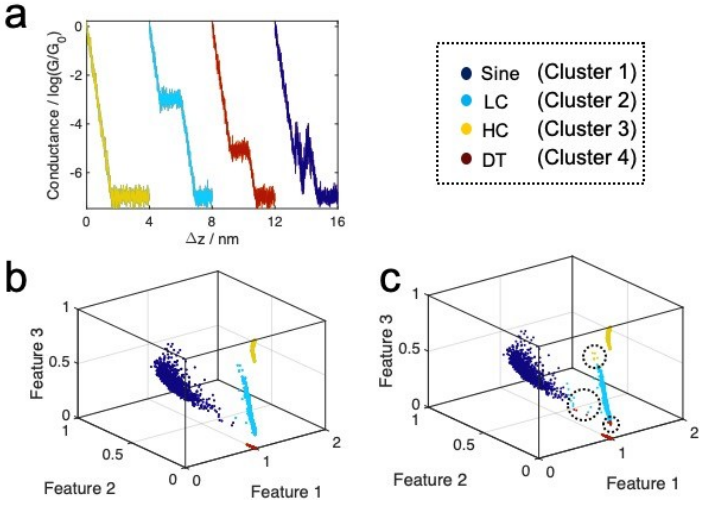

Figure S7. Feature space and classification result of the simulation data based on the DAK algorithm. (a) Four types individual traces of simulation data. (b) True labels of simulation data in feature space. (c) Clustering results based on the DAK algorithm with some error points circled by dash line.

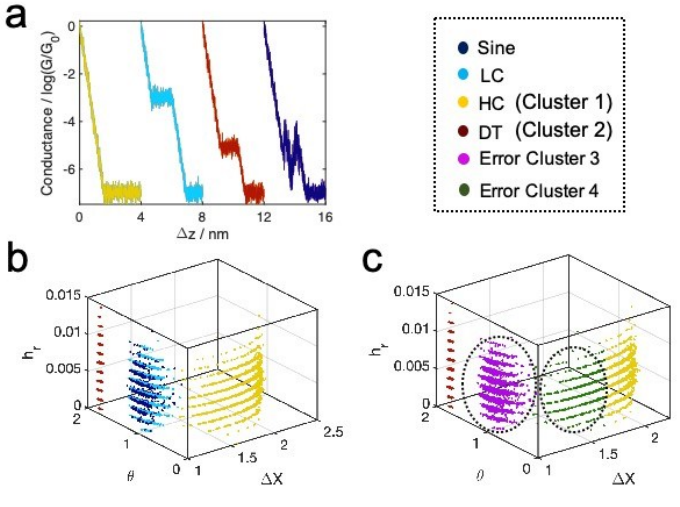

Figure S8. Feature space and classification result of the simulation data based on the MPVC algorithm. (a) Four types of individual traces of simulation data. (b) True labels of simulation data in feature space. (c) Clustering results based on the MPVC algorithm in feature space where ΔX represents the Euclidean distance, θ represents the angle, h_r represents the hamming distance, and the error points circled by dash line.

3. Experimental methods

3.1 Materials and synthetic methods

Materials. 1,3,5-Trimethylbenzene (mesitylene) and 2,3-dihydrofuran were purchased from Sigma-Aldrich. 3,6-Di(4-pyridyl)-1,2,4,5-tetrazine was purchased from Tokyo Chemical Industry. Dichloromethane (DCM) and tetrahydrofuran (THF) were obtained by distillation over sodium. All other chemicals, reagents, and solvents from commercial sources were used as received without further purification unless otherwise noted. 1H and 13C nuclear magnetic resonance (NMR) spectra were recorded on a Bruker AVIII-500 spectrometer (500 MHz and 125 MHz, respectively) and Bruker AVIII-850 spectrometer at 298 K (850 MHz and 213 MHz, respectively). The spectra were referenced to residual proton-solvent references (1H: CD2Cl2: 5.32 ppm; 13C: CD2Cl2: 53.84 ppm). High-resolution mass spectra (HR-MS) were recorded on a Bruker En Apex Ultra 7.0T FT-MS mass spectrometer.

Experiment 1: The OPE type molecule in case 1 was synthesized by Dr. Shi-Xia Liu's group in University of Bern [7]. A mixture of S-4-iodophenyl ethanethioate (64 mg, 0.21 mmol), 1,4-bis(diethynyl)-2,5-bis(dimesitylboryl)benzene (60 mg, 0.1 mmol), CuI (2 mg, 0.01 mmol) and Pd(PPh3)4 (6 mg, 0.005 mmol) in Et3N (7 mL) in a 20 mL vial was purged with N2 for 15 min. The resulting mixture was subjected to microwave irradiation by pre-stirring for 1 min, and reacted at 50 °C for 8 h. After cooling to room temperature, the solvent was removed in vacuum. The crude product was purified on silica gel chromatography using a mixture of hexane and dichloromethane (v/v 3:2) as eluent to afford compound 1 (OPE types molecule). Yield: 40 mg (43%); 1H NMR (300 MHz, CDCl3): δ 7.46 (s, 2H), 7.23-7.21 (d, J = 8.4 Hz, 4H), 7.02-6.99 (d, J = 8.4 Hz, 4H), 6.77 (s, 8H), 2.40 (s, 6H), 2.26 (s, 12H), 2.03 (s, 24H); 13C NMR (75.5 MHz, CDCl3): δ 193.6, 142.4, 141.0, 139.4, 137.7, 133.6, 132.1, 128.5, 127.6, 125.8, 124.4, 93.2, 91.9, 30.3, 23.4, 21.3; HRMS (ESI): m/z calcd for C62H61O2B2S2: 923.4294; found: 923.4338 (M+H+).

Experiment 2: 1,4-bis((4-(methylthio)phenyl)ethynyl)benzene (OPE3-SMe)(c) was synthesize by a reported method [8]. A solution of Pd(PPh3)Cl2 (62 mg, 0.08 mmol) and CuI (28 mg, 0.04 mmol) was added 1-bromo-4-methylthiobenzene (300 mg, 1.48 mmol) in dry THF (30 mL) at room temperature. After the system was purged with argon, 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) (1.35 g, 8.86 mmol) was added by syringe. Argon-sparged 1,4-bis((trimethylsilyl)ethynyl)benzene (200 mg, 0.80 mmol) was injected into the reaction flask and followed immediately by distilled water (0.01 mL, 0.591 mmol). The reaction mixture was stirred at reflux for 18 h. All volatile material was removed under vacuum and the residue was dissolved in CH2Cl2 (50 mL). The organic layer was washed with saturated aqueous NaCl (2 x 50 mL), dried over MgSO4 and the solvent removed under vacuum. The crude product was purified by silica gel column chromatography (dichloromethane: petroleum ether, 1:2 (V:V)) to afford 109 mg (0.29 mmol, 40%) of 1,4-bis((4-(methylthio)phenyl)ethynyl)benzene as a white solid.

Characterization of OPE3-SMe: 1H NMR (500 MHz, CDCl3) δ 7.42(s,4H), 7.37(d,J=8.25Hz,4H), 7.145(d,J=8.25Hz,4H), 2.44(s,6H).

Experiment 3: product B was synthesize according to the published protocol [9]. 3,6-Di(4-pyridyl)-1,2,4,5-tetrazine (Compound A, 23.6 mg, 0.1 mmol) was dissolved in tetrahydrofuran (100 mL). Then excess 2,3-dihydrofuran (7.009 g, 7.56 mL, 100 mmol) was added to a stirred solution of 3,6-Di(4-pyridyl)-1,2,4,5-tetrazine. Monitored by in-situ NMR, the peaks of 3,6-Di(4-pyridyl)-1,2,4,5-tetrazine vanished while the peaks of compound B appeared. As compound B is instable in condition of purification, it could only be characterized by in-situ NMR.

Characterization of B: 1H NMR (500 MHz, CD2Cl2) δ 9.34-9.35 (dd, J = 6.04 Hz, 2H), 9.32-9.34 (dd, J = 6.11 Hz, 2H), 8.56-8.57 (dd, J = 6.18 Hz, 2H), 8.27-8.28 (dd, J = 6.12 Hz, 2H). It was hard to distinguish the signals in the high-field region from reactants and solvents in-situ.

3.2 Characterization data

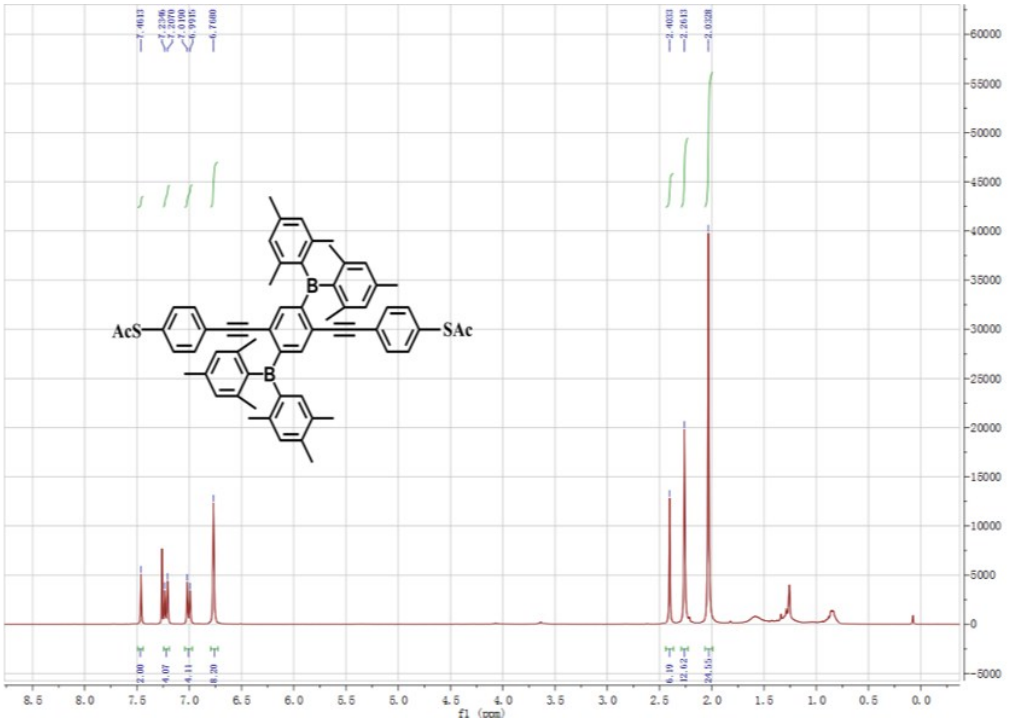

Figure S9. ^{1}H NMR spectrum for compound 1 (OPE types molecule).

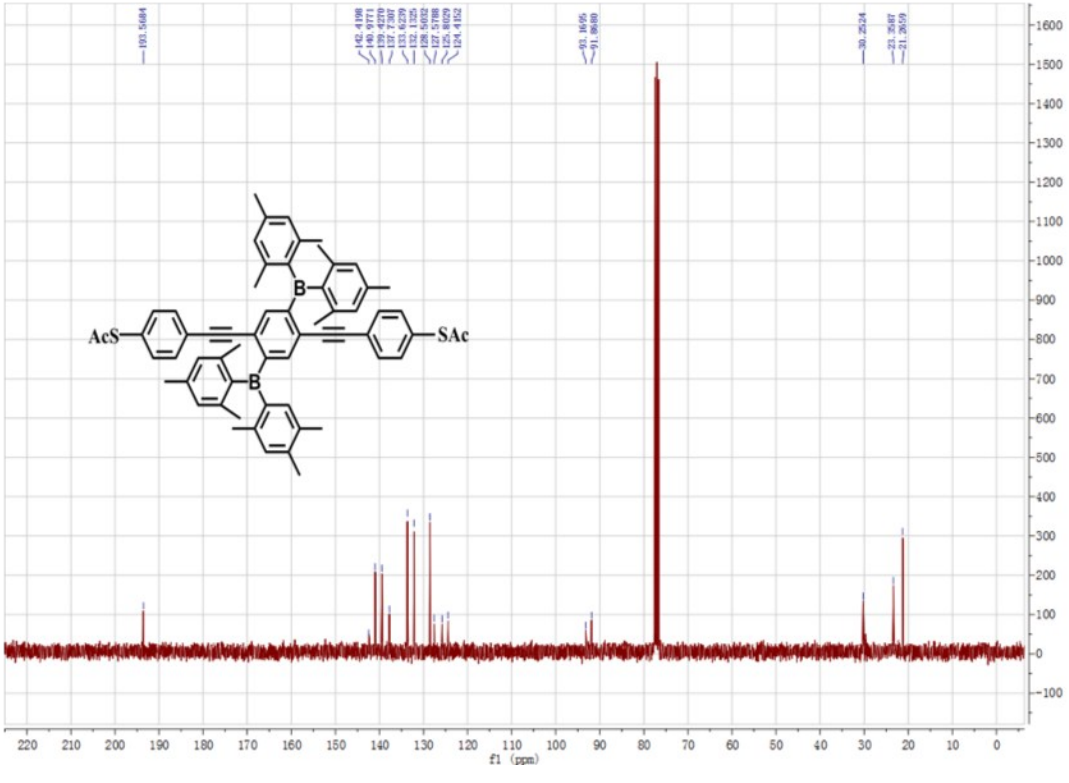

Figure S10. ^{13}C NMR spectrum for compound 1.

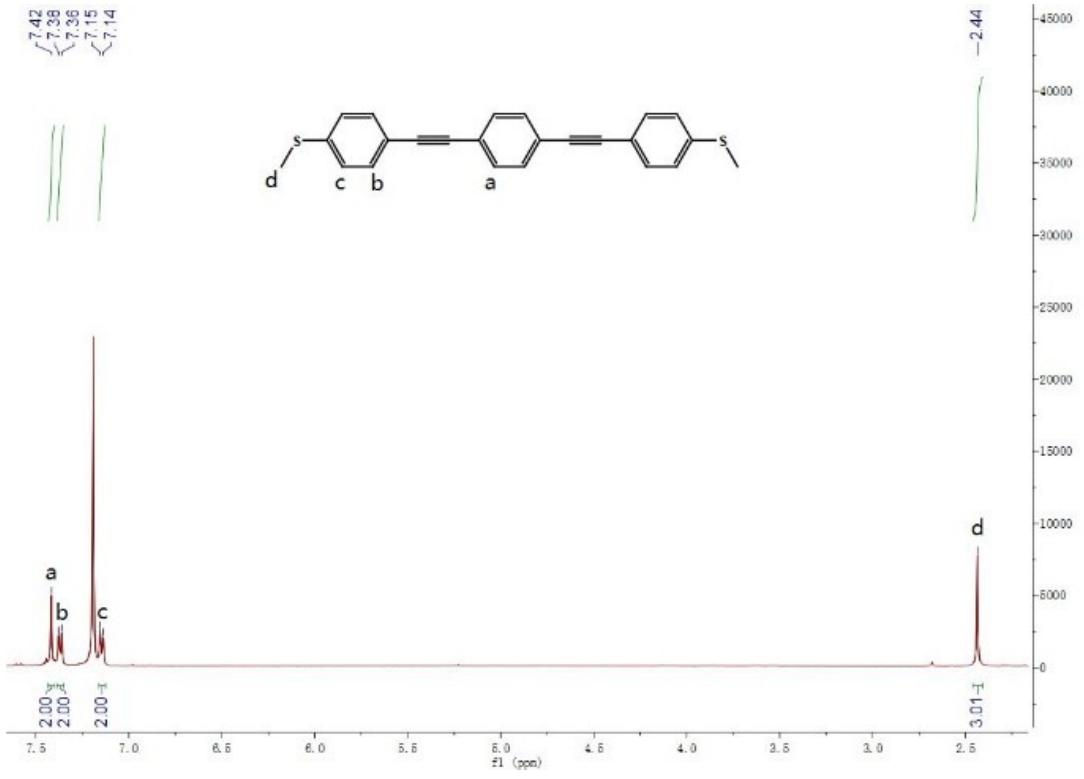

Figure S11. The [1]H NMR (500 MHz, CD2Cl2) spectrum of OPE3-SMe.

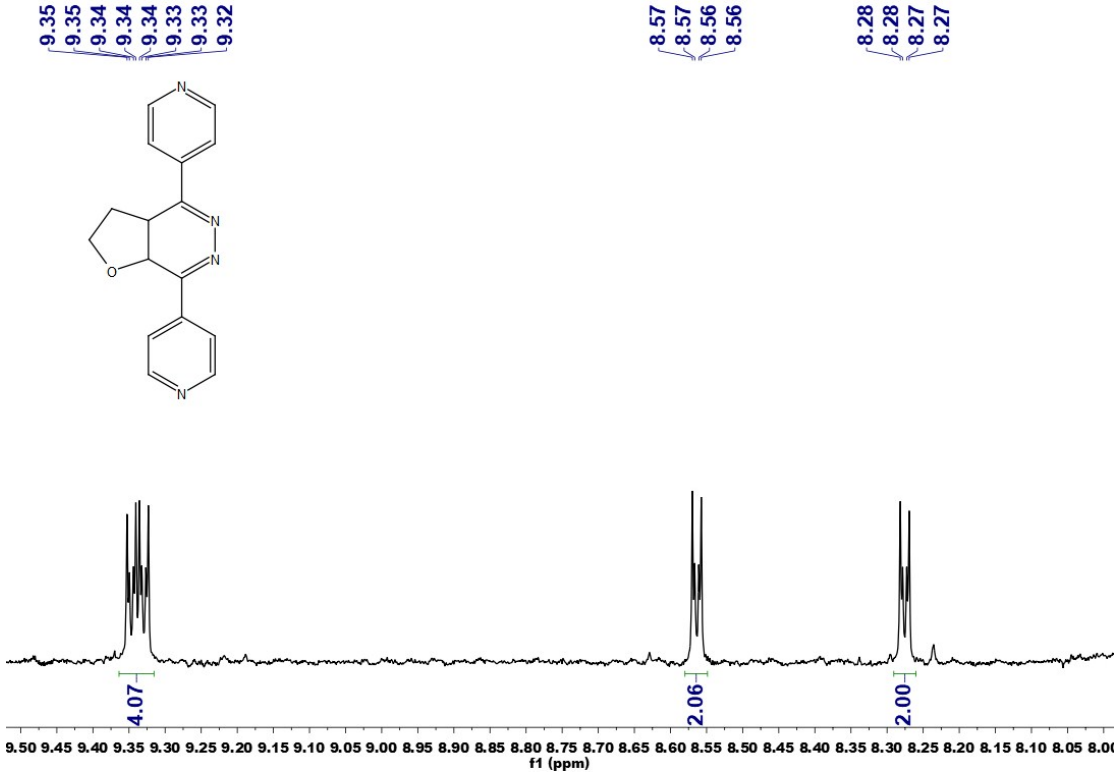

Figure S12. In-situ [1]H NMR spectrum of product B (only showing high field region).

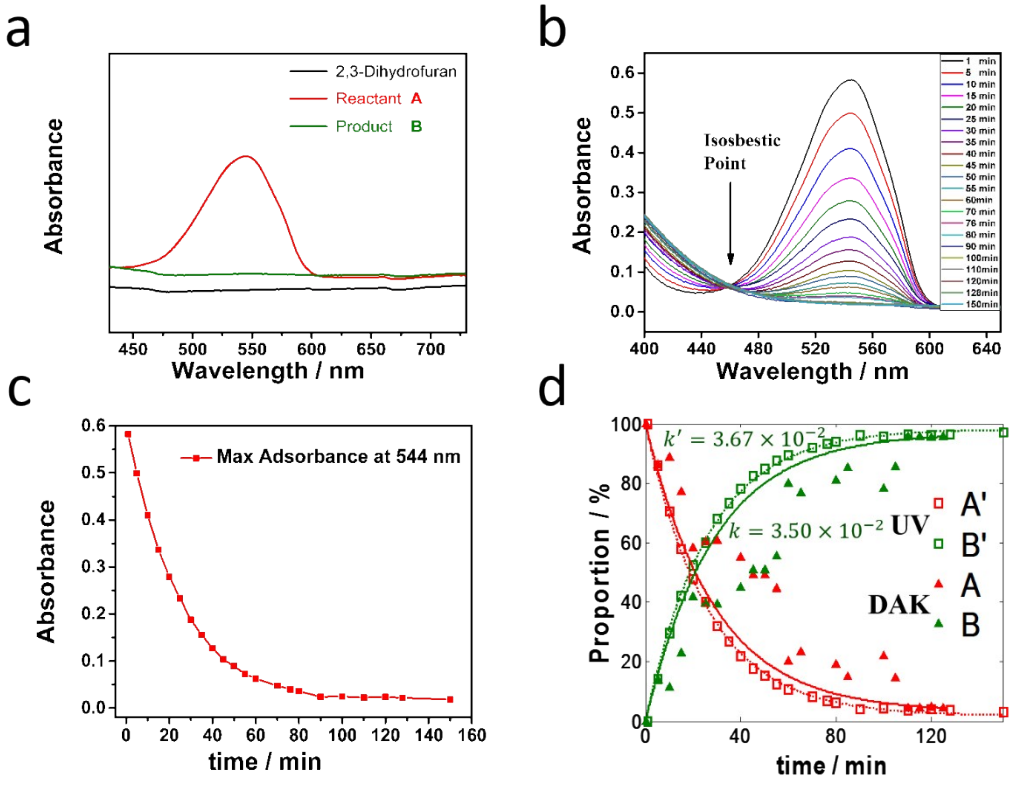

Figure S13. UV-Vis spectra for Diels-Alder reaction and the result of our method DAK. (**a**) UV-Vis characterizations of 2,3-Dihydrofuran, reactant A and product B. (**b**) In-situ UV-Vis monitoring of the Diels-Alder reaction. The spectra were recorded every 5 minutes. The isosbestic point (indicated) is at 458 nm. (**c**) Statistical graph for the maximum absorbance at the absorption band of 544 nm extracted from (b). (**d**) The proportion of A and B versus time via UV-Vis and the result of DAK. k_{UV} is fitted as 3.67×10^{-2} mM/min, which shows a good agreement with DAK algorithm measurements, which is 3.50×10^{-2} mM/min.

3.3 Experimental data preprocessing

Before classification, the datasets should be pre-processed to ensure that all datasets have the same dimension. We scan trace data $G_i(z)$ to find the first data point that is near the value G_0 and re-index it as the first data $G_i(0)$. And then we pick out the same number of data in order from the dataset to form a trace $\{G_i(j); j = 0,1,\mathrm{L}\ ,N\}$, where N indicates the dimension of trace determined to ensure that the charge transport information of the molecule is covered, resulting in trace dataset $\{G_i(j); j = 0,1,\mathrm{L}\ N; i = 1,2,\mathrm{L}\ M\}$, where j represents the distance index between the two electrodes and i is the trace index.

3.4 Visualization of the feature space based on experimental data

Stacked auto-encoder (SAE) is a vital part of the DAK algorithm which can learn the feature or representation of data. After the training, the encoder part of the SAE can be used for feature extraction. It is necessary to observe the distribution of data in the feature space. And by visualizing the feature space, we can more intuitively verify the feasibility of clustering. Here we employ principal component analysis (PCA) to reduce the dimension of features and then visualize the feature in three-dimensional space, as shown in Figures S14-S16

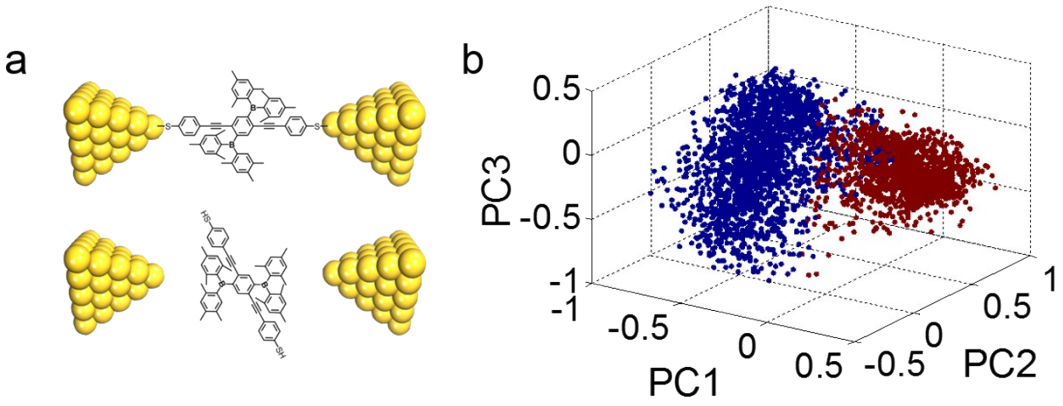

Figure S14. Visualization of the feature space for Experiment 1. (**a**) Schematic of break junction with and without molecules. (**b**) Scatter plot of data in feature space. Red and blue colors indicate the clustering result by the K-means algorithm. PC1, PC2 and PC3 are top 3 principal components computed by the PCA algorithm.

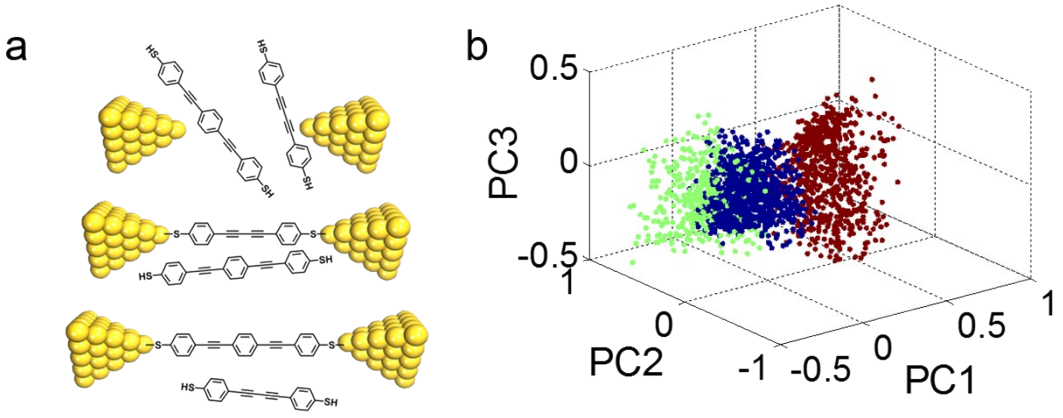

Figure S15. Visualization of the feature space for Experiment 2. (**a**) Schematic of break junction with three events, which include different molecules connected to gold electrodes and tunneling signal. (**b**) Scatter plot of data in feature space. Different colors indicate the clustering result by the K-means algorithm. PC1, PC2 and PC3 are top 3 principal components computed by the PCA algorithm.

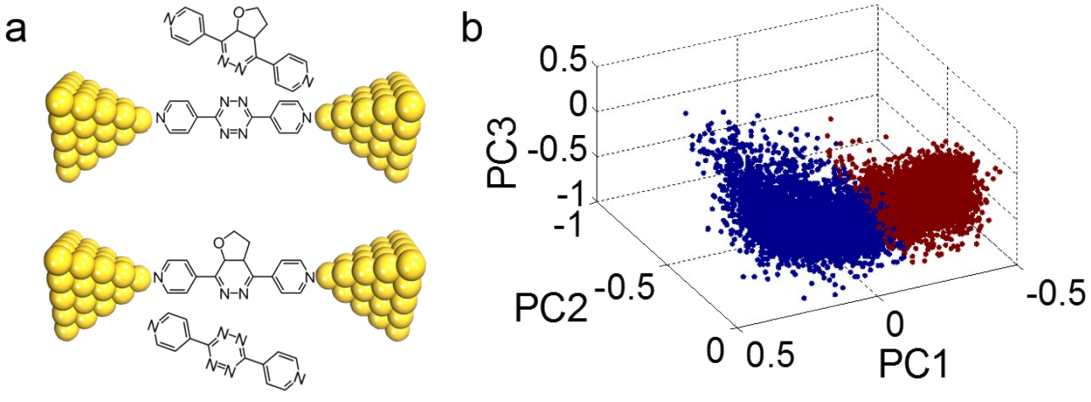

Figure S16. Visualization of the feature space for Experiment 3. (**a**) Schematic of break junction with two events, which include different molecules connected to gold electrodes. (**b**) Scatter plot of data in feature space. Red and blue indicate the clustering result by the K-means algorithm. PC1, PC2 and PC3 are top 3 principal components computed by the PCA algorithm.

References

1. R. Li, Z. Lu, Y. Cai, F. Jiang, C. Tang, Z. Chen, J. Zheng, J. Pi, R. Zhang, J. Liu, Z.-B. Chen, Y. Yang, J. Shi, W. Hong and H. Xia, *J. Am. Chem. Soc.*, 2017, **139**, 14344-14347.
2. J. Liu, X. Zhao, Q. Al-Galiby, X. Huang, J. Zheng, R. Li, C. Huang, Y. Yang, J. Shi, D. Z. Manrique, C. J. Lambert, M. R. Bryce and W. Hong, *Angew. Chem. Int. Ed.*, 2017, **56**, 13061-13065.
3. H. Bourlard and Y. Kamp, *Biol. Cybern.*, 1988, **59**, 291-294.
4. G. E. Hinton and R. S. Zemel, *Advances in neural information processing systems*, 1994, 3-10.
5. D. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, *Arxiv*, 2014.
6. M. Lemmer, M. S. Inkpen, K. Kornysheva, N. J. Long and T. Albrecht, *Nat. Commun.*, 2016, **7**, 12922.
7. X. Liu, X. Li, S. Sangtarash, H. Sadeghi, S. Decurtins, R. Häner, W. Hong, C. J. Lambert and S.-X. Liu, *Nanoscale*, 2018, **10**, 18131-18134.
8. R. Frisenda, S. Tarkuc, E. Galan, M. L. Perrin, R. Eelkema, F. C. Grozema and H. S. J. van der Zant, *Beilstein J. Nanotech.*, 2015, **6**, 1558-1567.
9. J. Sauer, D. K. Heldmann, J. Hetzenegger, J. Krauthan, H. Sichert and J. Schuster, *Eur. J. Org. Chem.*, 1998, **12**, 2885-2896.