Applications of Machine Learning to Single-Molecule Junction Studies

Tianren Fu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

Abstract

Applications of Machine Learning to Single-Molecule Junction Studies

Tianren Fu

The scanning tunneling microscope-break junction (STM-BJ) technique is an ideal platform for single-molecule studies related to the design of molecular electronics. STM-BJ is particularly advantageous for molecular junctions for characterizing key properties of molecular conductance as well as many other related properties, which contribute to a growing understand of the mechanisms of electron transport on the single-molecular level. Prior STM-BJ studies have generally focused on simple systems with only one type of molecule forming one type of junction. However, some systems (such as those involve *in-situ* chemical reactions) are intrinsically complex with multiple molecules and junction structures that can be accessed in the experiment. The analysis of such complex systems requires more powerful analytical methods that can distinguish different junction types. Machine learning has been demonstrated as a powerful tool for the analysis of such large datasets. In this work, we develop tools to analyze, with a high-accuracy, individual junction characteristics using machine learning to classify the data and provide mechanistic understanding of the STM-BJ method.

We start our work by investigating the imidazolyl linker. Imidazole is a five-member aromatic heterocycle with two nitrogen atoms, in which its pyridinic nitrogen can bind to gold electrodes. We study a series of alkanes of different lengths with two terminal 1-imidazolyl linker groups. While the intramolecular transmission across these molecules gives the pyridinic double peak, we find and prove that π-stacking between two imidazole rings is strong enough to form a

third intermolecular conductance peak with higher conductance. This behavior is a good example where multiple types of junction are formed with just one molecule.

Then, we focus on developing a trace-wise classification method using deep learning to resolve the data from such complicated systems of special molecules, mixture solutions, or *in-situ* chemical reactions. Compared to existing methods, ours reduces the loss of information during the data preprocessing and demonstrates better performance by employing a convolutional neural network structure with larger capacity. Benchmarking with several commercially available molecules, we show that our model reaches up to 97% accuracy and outruns all the existing methods significantly. Nevertheless, we also demonstrate that our model can retain high accuracy when two essential parameters, the average conductance and the length of the molecular conductance plateau, are removed. Importantly, this capability has not been seen for the other algorithm designs. We then apply our method to an *in-situ* chemical reaction to realize the monitoring of the reaction process. This excellent performance of our model on the trace classification task demonstrates the capability of machine learning methods on STM-BJ data analysis.

Finally, we also explore the feasibility of utilizing the machine learning toolkit in other types of analysis on molecular junctions. We study the relaxation of gold electrodes after junction rupture (termed "snapback") and its relation to pre-rupture evolution of gold contact. With the assistance of machine learning tools, we reveal that while the snapback can be well explained by this evolution history, the length of molecular conductance plateau is not related to either the snapback or this history. We also discover that the junction formation probability for short molecules is negatively correlated to the extension of single-atomic gold contact. Based on these findings, we conclude that the major mechanism for a molecular junction formation involves a molecule bridging across the junction prior to the rupture of the gold contact, in contrast to the previously-accepted picture where the molecule is captured immediately following the rupture.

As a conclusion, we apply machine learning/deep learning on STM-BJ data analysis by developing a model to efficiently classify STM-BJ traces with high accuracy, which is important for measuring complex systems containing multiple species. We also demonstrate the feasibility of analyzing junction formation mechanisms with the help of machine learning tools.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am incredibly grateful to my advisors Prof. Latha Venkataraman and Prof. Colin Nuckolls, were a constant source of advice and support. I am especially grateful for their patience in teaching me the philosophy and methodology of scientific research. I am also extremely thankful to Dr. Michael Steigerwald, Dr. Fay Ng, Prof. Luis Campos, Prof. Sujun Wei, Dr. María Camarasa-Gómez, Prof. Ferdinand Evers, Prof. Yaping Zang, Prof. Qi Zou, Kathleen Frommer, Rachel Starr, Julia Greenwald, Jang-Hun Choi, Prof. Michael Inkpen, Prof. Xavier Roy, Dr. E-Dean Fung, Dr. Boyuan Zhang, Liang Li, Dr. Nicholas Orchanian, Suman Gunasekaran, Dr. Jingjing Yang, Dr. Eskil Andersen, Dr. Jill Chipman, Dr. Evan O'Brian, Isabel Klein, and many others who guided and helped me over these past years.

My special thanks to my parents, Chun and Jianjin, and my girlfriend, Yidan. Without your support and encouragement, I could have never achieved these accomplishments.

Chapter 1. Introduction

Chemistry describes the characteristics and behaviors of molecules. However, the chemistry of a single molecule can be different from that of an ensemble of molecules, where the properties and reactivities of single molecules are generally more sensitive to the environment.[1-2] Comparing to bulky materials, observing single molecules is much harder. For decades, people have been working towards the probing and characterization on single molecules and have developed multiple methods: single-molecule spectroscopy (SMS) techniques[3-5] can realize molecular level spatial resolution with a rather large vision via fluorescence; scanning probing microscopes, such as atomic force microscope (AFM)[6-10] and scanning tunneling microscope (STM),[11-15] can visualize molecules of very high resolutions.

Scanning tunneling microscope-break junction (STM-BJ) is a specialized STM technique focusing on the characterization of electronic properties on the single-molecular level.[16-17] In a basic STM-BJ experiment, the conductance of a metal-molecule-metal junction bridging across two electrodes is measured. Usually, one molecule forms only one kind of molecular junction and translates into one signal in the STM-BJ results. However, some molecules, like imidazole derivatives with strong intermolecular interactions, can form multiple types of molecular junctions and result in multiple signals.[18] When we study chemical reactions at single-molecular level via STM-BJ techniques, the systems would inevitably contain more than one species and thus are also with mixed signals.

To monitor a chemical reaction process, we need to determine the ratio between reactants and products over time. Specific to a chemical reaction happening in the STM-BJ environment, we need a method to attribute each measured junction to its corresponding molecule/species, and to realize this, we resort to machine learning-based analysis. Machine learning methods, from correlation analysis to neural networks, have been proved effective on data measured with STM-BJ and other single-molecular techniques.[19-25] Inspired by these studies, we design a convolutional neural network-based deep learning model to realize the state-of-the-art accuracy of the classification of STM-BJ traces.[26] Our high-accuracy classification algorithm, along with other evolving methods, allows more precise identification of STM-BJ traces and paves the way towards further detailed research on single-molecular chemistry via STM-BJ.

1.1 Single-Molecule Junctions and Experimental Methods

A molecular junction is a device in which a single molecule binds to two electrodes. By the form of a molecular junction, we can easily measure many properties of a molecule with multiple experimental techniques. For metal electrodes like Au or Ag, they usually bind to the molecules through dative bonds with some lone pair-donating functional groups, such as amino or thiomethyl groups, or covalent bonds with thiol anions.[27-28] Besides metal electrodes, a molecule can also bind to modified carbon electrodes like graphene or carbon nanotubes via amide bonds.[29-30] Other than binding though bonds, molecules can bind to electrodes with intermolecular interactions like π-metal interaction to Au electrodes or π-π interaction to carbon electrodes, *etc*.[31-32] Compared to other electrode materials, Au has better properties like ductility and chemically inertness and thus is more studied.

There are multiple methods developed in recent decades to construct metal-molecule-metal junctions. Among them, mechanically controllable-break junction (MC-BJ)[33-34] and scanning tunneling microscope-break junction (STM-BJ)[16-17] are the most popular methods. In these techniques, the distance between the two metal electrodes is controlled by a piezoelectric actuator, and the molecular junction between these electrodes can form and rupture repeatedly through this motion control. Electro-migration (EM) is another method to construct persisting metal nanogaps and then construct molecular junctions.[35-36] In this work, all the molecular junction studies are conducted with the STM-BJ setup.

In an STM-BJ setup, we usually use one Au tip and one Au-coated steel substrate as the two electrodes. A single STM-BJ measurement starts with the two electrodes smashed together and then we begin to pull them away from each other. As the displacement between electrodes increases, the bulk Au-Au contact gradually yields into a single-atomic Au-Au point contact. Once this point contact breaks, if there is a molecule bridging the two electrodes, the corresponding molecular junction forms. Finally, as the elongation continues to increase, the molecular junction will also break. During this process, the bias and current across the junction are continuously recorded and a conductance trace is generated. This whole process can be done under ambient condition, and we usually conduct our measurements in a solution of the target molecule. Thus, this procedure can be easily set up and repeated thousands of times for statistical analysis.

While measuring the junction conductance, many other properties of the junction can be studied in parallel using customized STM-BJ setups. For example, by putting an AFM cantilever in the position of the upper electrode, the force across the junction can be measured together with the conductance;[37-38] with laser or photon detector align with the junction, its photoconductance effects such as photon-assisted transport or light emission can be examined.[39-40] Nevertheless, other mechanisms such as thermopower, magnetoresistance, *etc*. can also be studied with the STM-BJ technique.[41-50]

Additionally, STM-BJ is also useful for chemistry research. The *in-situ* chemistry taking place in the STM-BJ environment is a recent and fascinating finding. STM-BJ provides an environment of very intense electric field which shows catalytic activity. Chemical reactions can also happen with the electrode material or electrons/holes. In recent studies, people have found redox reactions, electric field-catalyzed intra- and intermolecular reactions taking place *in situ* during break junction experiments.[51-56]

In this work, we use STM-BJ to measure solutions of pure molecules, mixtures, and systems which could undergo *in-situ* chemical reactions. Moreover, we study the mechanisms of STM-BJ experiments by analyzing the collected data.

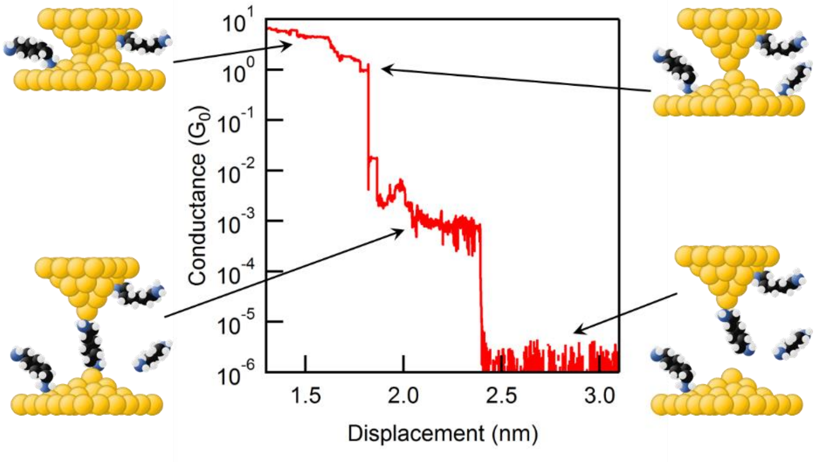


Figure 1.1: An example STM-BJ conductance trace. The drawings illustrate difference phases in an STM-BJ experiment. Top-left: bulk Au-Au contact; top-right: single-atomic Au point contact, whose conductance equals to 1 G0; bottom-left: metal-molecule-metal junction

1.2 Data Analysis in Break Junction

As we discussed in Chapter 1.1, during the measurement of every single STM-BJ trace, the conductance of the junction is calculated and recorded continuously as a function of displacement. In Figure 1.1 we show an example of such conductance versus displacement trace. At the beginning of the trace, when the bulk contact remains, the conductance is high. Then, the conductance drops to an intrinsic number as the interface between electrodes narrows into a single-atomic contact; for Au electrodes, the conductance of this single-atomic Au-Au point contact is 1 G_0, where G_0 is the conductance quantum and 1 $G_0 = 2e^2/h$. After the rupture of this point contact, if there is a molecule bridging the junction, we can observe a conductance plateau typically lower than 1 G_0, which reflects the conductance of the metal-molecule-metal junction. Finally, after junction rupture, the conductance falls to the noise floor.

In STM-BJ research, the most popular way of data analysis is through creation of histogram. Figure 1.2a shows the one-dimensional histogram (1DH) of the trace shown in Figure 1.1 To make this single trace 1DH, we divide the conductance axis into many bins and count the number of data points that fall in each bin. After adding up thousands of single-trace 1DHs from an STM-BJ experiment with a molecule and calculate the average, to obtain the 1DH for the experiment (Figure 1.2b). From the 1DH of the experiment, we can see the characteristic conductance value of the molecular junction. Compared to single-trace histograms, the histogram of the whole experiment shows better-shaped conductance peak (usually Gaussian if logarithmically binned) after averaging over a large number of trace samples. If we construct bins on both the conductance and displacement axes, we get a single-trace two-dimensional histogram (2DH) as shown in Figure 1.2c and then the average 2DH of experiment shown in Figure 1.2d. From a 2DH, we can find the characteristic conductance, the extension of a molecular junction and the junction formation probability.

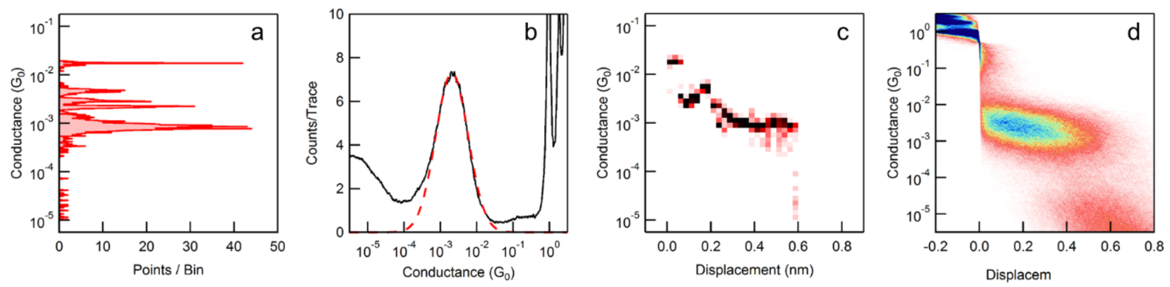


Figure 1.2: Logarithm-binned conductance histograms: (a) a single-trace 1DH and (b) the 1DH of the whole experiment, with the Gaussian fit of the molecular conductance peak in red dotted line; (c) a single-trace 2DH and (d) the 2DH of the whole experiment.

If there is only one kind of molecular junction in the system, there will be only one conductance peak in the histograms. However, for complicated systems with more than one species, there can be more than one peak, and these peaks can overlap with each other and increase the difficulties in distinguishing them. Nevertheless, since the plateau length of species may vary, we cannot simply determine the ratio between species to be the ratio between peak areas in histograms. Thus, individual trace analysis is needed for such studies.

We can break down the workflow of individual trace analysis into two steps: describing a trace and analyzing the description. On the description side, a straightforward method is to manually select several parameters to describe a trace, such as conductance, noise, *etc.* Another popular method is to use single trace histograms to represent a trace; there are several studies making use of single trace 1DHs or 2DHs. Besides these methods, we can also feed the analysis algorithm with the raw trace. This methodology could prevent information loss during construction of histograms, but it requires a specially-designed analysis algorithm to receive it. Speaking of the analysis algorithms, most of the works use clustering, which is a class of unsupervised machine learning algorithms we will discuss later, and some use other methods such as linear decomposition, neural networks, *etc*. With a high-accuracy single trace analysis tool, we can achieve more precise ratio analysis or realize the monitoring on *in-situ* chemical reactions.

1.3 A Brief Introduction to Machine Learning Methods

Machine learning is a methodology to build models and pipelines to automatically make predictions or decisions. A machine learning algorithm generates a model based on reference data, and then we can use this model to generate output on the target data. When the reference data is labeled, which means it comes with corresponding dependence variable values, the learning diagram is called a supervised learning approach. When the reference data is not labeled, it is called an unsupervised learning approach, while in this case, the reference dataset is often the target dataset that we want the model to generate output on.

Most of the well-known machine learning algorithms are for supervised learning. For example, linear regression is one of the simplest supervised algorithms, where we need to provide both the independent variables and the dependent variable together to calculate the model coefficients. Decision tree-based algorithms is an important type of supervised learning algorithms.[57-58] A decision tree is a tree structure where each of the tree nodes contains an if-then-else clause, for example, "if $x < 0$ then go to the left child, else go to the right child", and the outcome of a tree is determined by the whole path. By combining multiple nodes in a series, a decision tree can introduce non-linearity which is important for many problems. The XOR problem illustrated in Figure 1.3a is one of these non-linear problems that cannot be approached with linear models such as linear regression. A decision tree, however, by combining more than one layer, can easily fit this XOR function (shown in Figure 1.3b). This property makes decision trees very useful. Many popular models are some forms of assembly of a series of decision trees, such as random forest and XGBoost.[59-61]

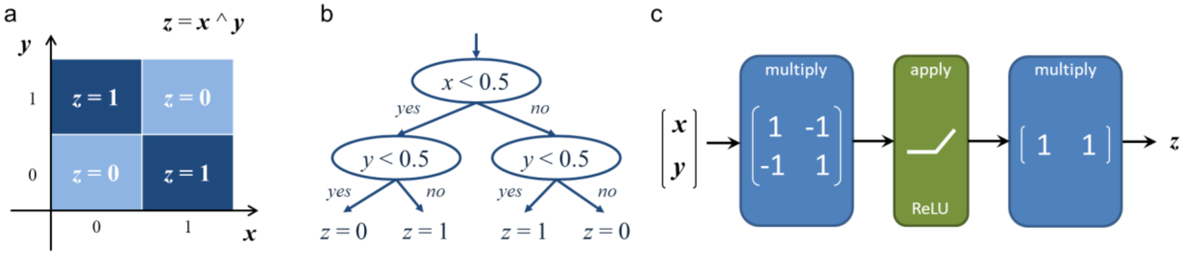


Figure 1.3: (a) The XOR function, z = x ^ y. (b) The decision tree approach towards the XOR function. (c) The neural network approach towards the XOR function. The ReLU function is defined as ReLU(x) = x, x > 0; 0, x ≤ 0.

Unlike supervised learning, in an unsupervised learning task, there is no labeled reference dataset. Many machine learning-based studies on STM-BJ data are by unsupervised learning.

Clustering is one important class of unsupervised learning. In a clustering task, the algorithm aims to partition the given data points into several different clusters, without extra information. For example, given a set of STM-BJ traces, we may deploy a clustering algorithm to separate them into different types, for example, with and without a molecular plateau. The most important clustering algorithm is known as K-Means. The K-Means algorithm minimizes the distance between a point and the geometric center of the class to which this point belongs, and thus all the points within a class are close to each other. Thus, before fitting the model we need to manually choose the number clusters, and then we can fit the model by iteration. Other than K-Means, there are other popular clustering methods such as spectral clustering,[62] DBSCAN.[63] Besides clustering, principal components analysis (PCA), deep autoencoders,[64] *etc*. are also considered as unsupervised learning methods.

Neural network is an important structure of model, which can be applied in both supervised and unsupervised tasks. A big neural network with many layers and a large number of coefficients can approach a very complicated underlying function. One layer in a neural network is usually implemented as a matrix multiplication where the input vector right times the coefficient matrix and results in the output vector. Between layers, this data vector undergoes an element-wise activation function. The rectified linear units (ReLU) is a popular activation function, where ReLU(x) gives x when $x > 0$, and 0 when $x \leq 0$. These non-linear activation functions enable neural networks to approach non-linear functions, for example, the XOR function (shown in Figure 1.3c). When a neural network contains more than three layers, it is usually called a deep learning structure. Among all types of deep learning structures, the convolution neural network (CNN) is very widely-used in image processing[65] and is particularly important for STM-BJ studies. CNN is a neural network structure where some of the layers are replaced by convolutional layers, where a convolutional layer conducts a convolution (cross-correlation) operation between the input and a set of coefficients called “kernels”. The nature of convolution operation in combination with a small kernel size makes a convolutional layer sensitive to local patterns. This structure assumes and exploits the spatial local correlation in the inputs, which is true for images and is also true for STM-BJ traces: the conductance difference between a series of neighboring pixels determines whether it is a conductance drop or a plateau. We find the CNN structure works very well on cases of STM-BJ conductance traces as we can regard them as one-dimension images.

In this work, we introduce a CNN-based deep learning structure that realizes high-accuracy classification on STM-BJ traces. We also demonstrate machine learning methods-assisted analysis on STM-BJ experiment mechanisms.

1.4 Outline

The remainder of this thesis will focus on the machine learning-assisted analysis on STM-BJ data. We will accomplish this by conducting experiments on multiple STM-BJ systems and building models with machine learning algorithms. An outline of the remaining chapters of this thesis will be:

Chapter 2 presents work characterizing a series of molecules with imidazolyl linkers. We show each of these molecules forms three types of junctions. We further reveal that the high-conductance type among these three is a bimolecular junction as the *π-π* interaction between two imidazole rings is strong enough to stabilize this junction.

Chapter 3 introduces a new automatic model for classification of STM-BJ traces. We construct this model employing a convolutional neural network structure. This model performs excellently on experimentally measured mixture of molecules as well as system of *in-situ* chemical reaction, and outruns the existing methods.

Chapter 4 demonstrates work investigating the junction formation mechanism. We study the relation between the evolution of Au contact and some key parameters such as the relaxation of Au electrodes after junction rupture. According to our findings, we conclude that the molecule of a junction is already bridging across the electrodes before the Au contact breaks.

Chapter 2. Enhanced Coupling Through π-Stacking in Imidazole-Based Molecular Junctions

This chapter is based on the manuscript entitled *Enhanced Coupling Through π-Stacking in Imidazole-Based Molecular Junctions* by Tianren Fu, Shanelle Smith, María Camarasa-Gómez, Xiaofang Yu, Jiayi Xue, Colin Nuckolls, Ferdinand Evers, Latha Venkataraman and Sujun Wei published in *Chemical Science*. Shanelle Smith, Xiaofang Yu and Jiayi Xue of Prof. Sujun Wei group synthesized and characterized all the compounds.[18] I performed the conductance measurements and data analysis. Dr. María Camarasa-Gómez of Prof. Ferdinand Evers group and I conducted the theoretical calculation.

In this work, we demonstrate that imidazole based π-π stacked dimers form strong and efficient conductance pathways in single-molecule junctions using the scanning-tunneling microscope-break junction (STM-BJ) technique and with density functional theory-based calculations. We first characterize an imidazole-gold contact by measuring the conductance of imidazolyl-terminated alkanes (im-*N*-im, *N* = 3–6). We show that the conductance of these alkanes decays exponentially with increasing length, indicating that the mechanism for electron transport is through tunneling or super-exchange. We also reveal that π-π stacked dimers can be formed between imidazoles and have better coupling than through-bond tunneling. These experimental results are rationalized by calculations of the molecular junction transmission using non-equilibrium Green's function formalism. This study verifies the capability of imidazole as a Au-binding ligand to form stable single- and π-stacked molecule junctions at room temperature.

2.1 Introduction

Imidazole is an aromatic five-member-ring structure with two nitrogen atoms, one pyridine-like and one pyrrole-like nitrogen (N-3 and N-1 as shown in Figure 2.1). The lone pair electrons on the pyridine-nitrogen coordinates with metals or with protons. Additionally, the electron-rich characteristic of imidazole also enables versatile intermolecular non-covalent interactions, such as accepting hydrogen bond or enhancing π-π interactions. Imidazole thus has varied functionality. For example, as a functional group of the amino acid histidine, it is the active binding site in superoxide dismutases;[66-67] it also acts as a Brønsted base in serine endopeptidases.[68] In metal organic frameworks (MOFs), it is used as a bidentate but non-chelating ligand.[69] Despite these broad functionalities, the electronic characteristics of imidazole as a Au-binding ligand has not yet been tested.

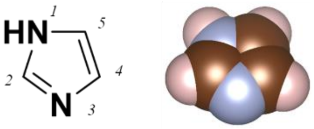


Figure 2.1: The molecular structure of imidazole, and IUPAC numbering of atoms.

2.2 Results and Discussion

Here, we applied the scanning tunneling microscope-based break-junction (STM-BJ) method to create and characterize imidazole-based molecular junctions.[16-17] We synthesized four imidazole-terminated alkane molecules with an 1,ω-di(imidazol-1-yl)alkanes (**im-3-im**, **im-4-im**, **im-5-im** and **im-6-im**) chemical structure, as shown in Figure 2.2a. The synthesis is detailed in Section 2.4.1.

STM-BJ measurements are conducted under ambient condition at room temperature as has been described before.[70] Junctions are formed between a Au substrate and tip from a ~1 mM solution of the target molecule in 1,2,4-trichlorobenzene (TCB). Each individual measurement starts by smashing the tip into the substrate to create a Au-Au contact. The tip is then withdrawn while the conductance, (current/voltage), is measured as a function of the relative tip/substrate displacement and this is repeated at least 5,000 times for each molecule. The individual traces are compiled into logarithmically binned conductance histograms.[71]

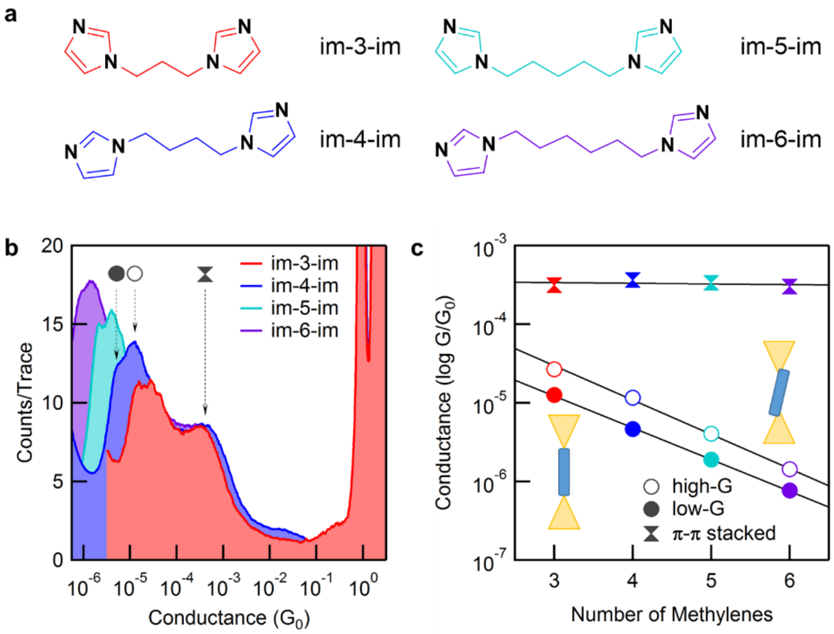


Figure 2.2: (a) The chemical structures of im-N-im molecules. (b) Logarithmically-binned conductance histograms (100 bins per decade) for all four molecules generated from 15000 traces each. The three peaks, two that change with the molecular backbone length and the one that is independent of the backbone length are indicated by the arrows for im-4-im. Histograms are terminated at the noise floor. (c) Molecular junction conductance, determined from a Gaussian fit, is plotted against the number of methylene units in the backbone. The *β* values determined from the fit are 0.93 (low-*G*) and 1.01 (high-*G*), per methylene.

Figure 2.2b shows 1D conductance histograms for all four molecules. 2D conductance-displacement histograms are shown in supplementary Figure 2.5. All measurements are performed at a 900 mV bias. Note that the bias does not affect conductance for these molecules (see supplementary Figure 2.6). A clear peak at ~1 G_0 ($G_0 = 2e^2/h$, the conductance quantum) is seen due to the reproducible formation of a single atom Au contact. The molecular junction conductance peaks occur over a range of 10^{-4} to 10^{-6} G_0 and show a decreasing conductance with increasing backbone length. This can be attributed to a ballistic tunneling transmission through molecular junction. In addition, for every histogram, there is a broad feature at around 10^{-3} G_0 which we attribute to an intermolecular π-π stacked complex that we will discuss in detail further below.[72-77] Further inspection of the molecular conductance peak reveals that it is actually two peaks, similar to what is observed for pyridine-based linkers.[78-80] Since the imidazole linker is binding to Au with its pyridine-nitrogen, it can form a vertical, primarily σ-coupled junction (left) or a tilted σ- and π-coupled junction (right) as illustrated in the inset of Figure 2.2c. As the differences between these two binding configurations has been investigated in detail for pyridine linkers before,[80-81] in the following discussion, we will focus primarily on the lower-conducting σ-coupled configuration.

Figure 2.2c plots the peak conductance value of each molecule against molecular length. The solid circles represent the conductance of the σ-coupled configuration with a lower-conductance (low-G), and the hollow circles represent the conductance of the tilted configuration (high-G). Both series show an exponential decay in conductance with increasing molecular length. We fit these experimental data with tunneling transmission model: $G_N = Ae^{-\beta N}$, and obtain a decay constant, β = 0.93 and 1.01 per methylene for the low-G and high-G configurations respectively. This β value agrees with measurements of alkyl molecules with other linkers,[27, 82] and confirms that these follow a tunneling mechanism. By extending the fit to N = 0, we estimate the conductance of a molecule with no carbon bridging the two imidazole groups; the inverse of this conductance serves as a metric for the linker contact resistance. For the imidazole linker, we obtain a contact resistance of 65 MΩ for the low-G series. As a comparison, the contact resistance for some other common linkers are: -SMe: 0.27 MΩ, $-NH_2$: 0.37 MΩ and $-PMe_2$: 0.13 MΩ.[27] We can also compare imidazole with pyridine which has a contact resistance of 23 MΩ as determined from a direct measurement of 4,4’-bipyridine.[79] Although clearly larger than small linkers, imidazole is comparable to pyridine.

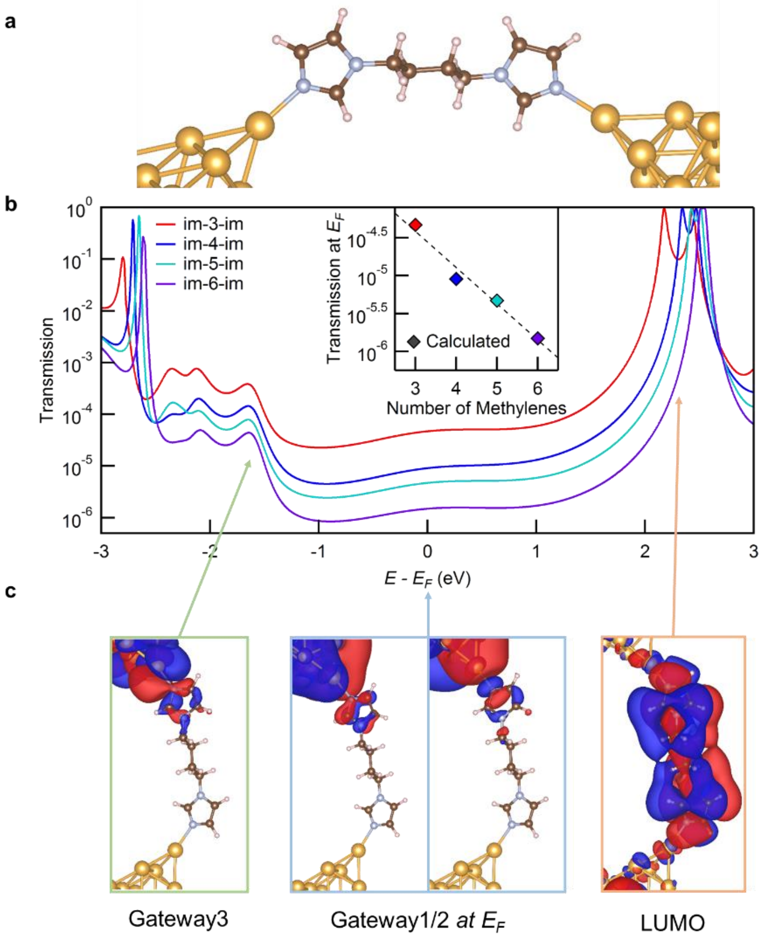


Figure 2.3: (a) The junction geometry of im-4-im. (b) The calculated transmission functions of all the four molecules. Inset: Linear fit of transmission at Fermi energy of each molecular junction. (c) The scattering states for the im-4-im junction determined at the energies corresponding to the two peaks closest to the Fermi energy, as indicated in the figure.

We now turn to transport calculations based on density functional theory (DFT), and compute the electronic transmission through Au-**im-*N*-im**-Au junction models. We employ the FHI-aims package[83-84] with a PBE exchange-correlation functional[85] and apply a non-equilibrium Green’s function formalism implemented within AITRANSS package.[86-87] The structure of a low-G **im-4-im** junction is shown in Figure 2.3a, using the VESTA program.[88] Each electrode consists of a pyramidal cluster of 55 Au atoms, arranged in 6 layers in the (111) direction with closest interatomic distance of 2.88 Å. The **im-*N*-im** molecules are in fully relaxed in an all-*anti* conformation in gas phase and imidazole is bound to the apex atom of Au electrode in a vertical geometry, a structure that represents the lower conducting junction.

The transmission function for all four molecules studied are similar as shown in Figure 2.3b. The transmission at Fermi (E_F) has a strong contribution from the molecular HOMO-4 and HOMO-2 based σ-channels that decay across the molecular backbone, as can be seen from the isosurface plots for **im-4-im** shown in Figure 2.3c (the structures in blue frame). The transmission peak at around –1.6 eV relative to E_F (on the occupied side) represents transmission through the molecular HOMO-8 which is also primarily a σ-based orbital (the structure in green frame). The transmission peak at around +2.3 eV relative to E_F results from the weakly coupled π-based molecular LUMO (the structure in orange frame). The transmission at the E_F decreases exponentially with increasing molecular length, in agreement with the experimental results. The calculated conductance values, obtained by applying the Landauer formula to the transmission at E_F, are plotted against the molecular length in the inset of Figure 2.3b. The calculations overestimate conductance due to known errors with DFT[89] which in turn can also alter the calculated β value. We find that the calculated β value is 1.10/methylene, slightly higher than the experimental one.

We now turn to the molecular conductance peak seen around 10^{-3} G_0 for all the molecules in this series as shown in Figure 2.2b. We attribute this peak to junctions formed by an intermolecular π-π stacked dimer, (see Figure 2.4a and 2.4b). Such a π-π stacked dimer has been observed in aniline derivatives where the contribution of the N-p_z orbital is significant to enhance the intermolecular interaction.[90] The pyrrole nitrogen on the imidazole ring can play a similar role enhancing the electron density in the imidazole π system and augmenting π-π interactions. To confirm this hypothesis, we measure the conductance of 1-methylimidazole (**im-1**) which has only one Au-binding site. The 1D histogram from STM-BJ measurements of **im-1** is shown in Figure 4c, together with the histogram of **im-4-im**. **Im-1** gives a single peak at ~10^{-3} G_0 that overlaps with the peak also observed for all **im-*N*-im** molecules.

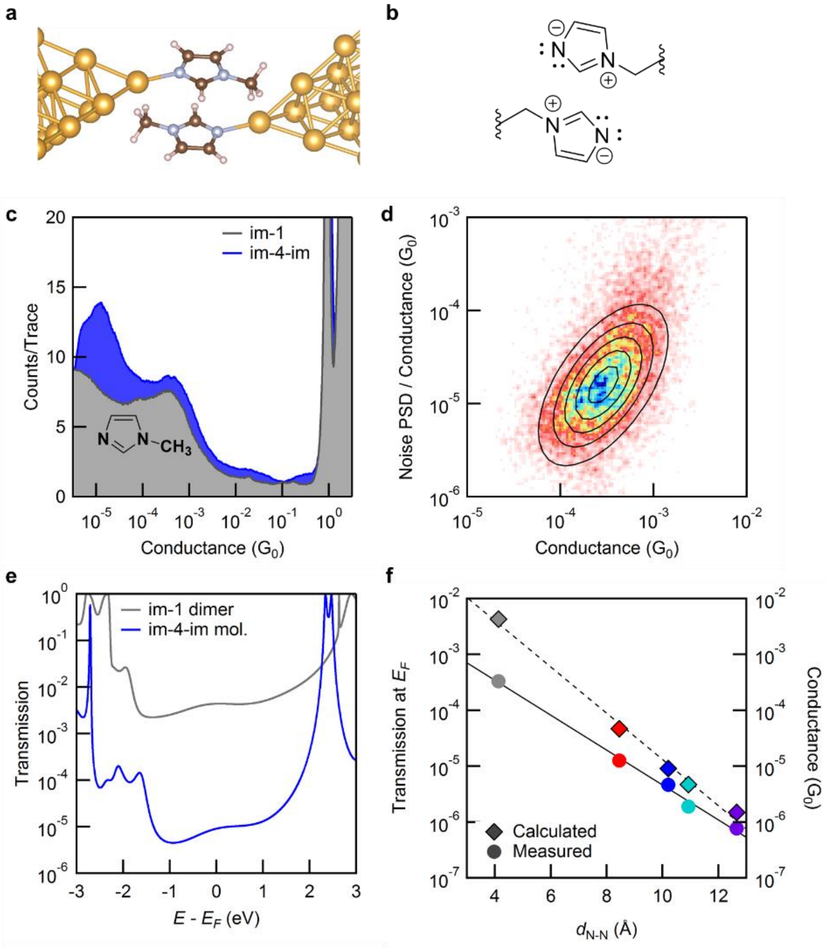


Figure 2.4: (a) The structure of a π-π stacked 1-methylimidazole (im-1) dimer junction used in the DFT calculations. (b) The charge-separated resonance state that stabilizes the π-π stacked dimer. (c) Logarithmically-binned conductance histogram for im-1 and im-4-im measurements. (d) Two-dimensional histogram of PSD/*G* against the average junction conductance *G*. (e) Calculated transmission functions of an im-1 dimer junction along with that of the molecular im-4-im junction. (f) The transmission at Fermi of im-1, together with im-*N*-im plotted against the junction N-N distance (left axis). The corresponding experimental data is also shown (right axis).

Since **im-1** has only one Au-binding site, it can only form a π-π stacked junction. We therefore use flicker noise measurements, which have been used to distinguishes through-space transmission from that of through-bond[91] to confirm this hypothesis. Flicker noise measurements are conducted by first forming an **im-1** dimer junction, holding this for 150 ms and analyzing the conductance, measured with a 100 kHz bandwidth. Two quantities are calculated from the conductance data while the junction is held: the average conductance (G), and the normalized noise power in the form of power spectrum density (PSD). The PSD is obtained from the square of the integral of the discrete Fourier transform of the measured conductance between 100 Hz to 1,000 Hz. The lower frequency limit is constrained by the mechanical stability of the setup. The upper limit is determined by the input noise of the current amplifier. Using these quantities, we create 2D histograms of the normalized noise power against the average conductance from 8,556 traces. The relation between noise power and conductance is extracted by determining the scaling exponent (N) for which PSD/G^N and G are not correlated. We have previously shown that the relationship between flicker noise PSD and conductance G follows a power law dependence (PSD ~ G^N) with the exponent N being indicative of the electronic coupling type. N close to 2 indicates a through-space coupled molecular junction, while an exponent N of 1 indicates a through-bond coupling. Figure 4d shows the 2D histogram of PSD/G against G where a clear positive correlation is visible. For **im-1**, the correlation between PSD/G^N and G goes to zero when $N = 1.9$. This is a clear indication to a through-space transmission. We can therefore attribute the conductance peak at around 10^{-3} G_0 to one that involves a through-space coupled intermolecular imidazole dimer.

To compare this to the calculated transmission, we model the junction as illustrated in Figure 4a; the geometry is optimized within DFT including van der Waals interactions following the methods developed by Tkatchenko and Scheffler.[92] We find that the DFT-relaxed structure has the two **im-1** molecules separated by a ~3.3 Å gap and stabilized by 0.41 eV when compared to two isolated molecules. For the imidazole ring, an electron-rich pyrrole nitrogen and an electron-withdrawing pyridine nitrogen increase the dominance of its charge-separated resonance structure shown in Figure 2.4b. Thus, imidazole π-π stacked dimer is more strongly bound than benzene dimer (which is stabilized by only 0.15 eV). The transmission across this junction is shown in Figure 2.4e. At E_F, the transmission of the dimer **im-1** is ~100 times that of the molecular **im-4-im** junction, in agreement with experiment. For the π-π stacked dimer junction of **im-4-im**, the calculated transmission is close to that of dimer **im-1** junction (see Supplementary Information). The results here confirm that the conductance peaks around 10^{-3} G_0 in Figure 2.2b arise from intermolecular π-π stacked dimer junctions as shown in Figure 2.4b, and thus the length of alkyl chain in **im-*N*-im** series is not important. In Figure 2.4f, we plot the transmission at E_F and the measured conductance against the calculated through-space distance between the two imidazole nitrogens that are directly bound to Au atoms. Interestingly, although the two electrodes are not bridged by one molecule, the π-π stacked dimer structure can still give roughly a same conductance as a single-molecule junction of similar length, in contrast to what is typically expected for such weakly coupled systems.[93]

2.3 Conclusions

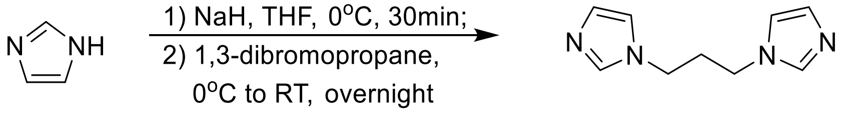
In summary, we have investigated the ability of imidazole to function as an aurophilic linker for molecular junctions using the STM-BJ technique. We find that the conductance of four imidazole-terminated alkanes have a β value that is consistent with that found for other linkers. This provides an outlook to direct measurement on the electronic properties of some imidazole-containing biologically relevant systems. Importantly, we also demonstrated that imidazole can form stable π-π stacked dimers that have a relatively high through-space conductance, which therefore function as the smallest functional group forming stable π-π stacked dimers.

2.4 Supplementary Information

2.4.1 Synthetic Details

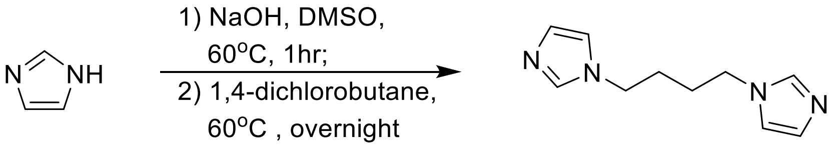
Materials and Instrumentation. All commercially available chemicals, including the 1-methylimidazole (**im-1**), were used as received without further purification unless otherwise noted. All reactions were performed in oven-dried round bottom flasks, unless otherwise noted. The flasks were fitted with rubber septa and reactions were conducted under a positive pressure of nitrogen, unless otherwise noted. Flash column chromatography was performed employing Biotage Isolera One (10 or 25 gram SNAP silica gel column). Thin- layer chromatography (TLC) was performed on silica gel 60 F254 plates (EMD).

^{1}H and ^{13}C nuclear magnetic resonance spectra were recorded at 300 K (unless otherwise noted) on *Bruker* DRX400 (400 MHz) or *Bruker* DRX500 (500 MHz) FT NMR spectrometers at Department of Chemistry and Biochemistry, CUNY Queens College. High-resolution mass spectra were recorded on high resolution mass spectrometers using either electrospray ionization (ESI) or atmospheric pressure chemical ionization (APCI) method at CUNY Hunter College Mass Spectrometry.



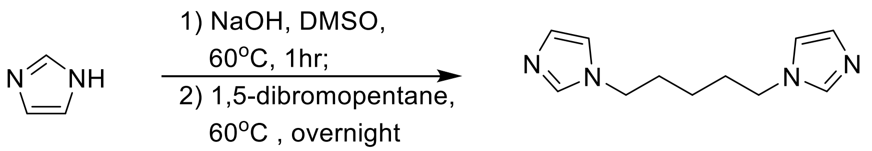
1,3-di(1*H*-imidazol-1-yl)propane (**im-3-im**):[94]

To a stirring solution of dry imidazole (544.8 mg, 8 mmol) in 25 mL anhydrous THF, NaH (320 mg, 60% wt. in mineral, 8mmol) was slowly added at 0 °C under N_2. The resulting mixture was allowed to warm up to room temperature for 30 minutes, and then cooled down to 0 °C again. 1,3-dibromopropane (0.4 mL, 4 mmol) was added as neat to the above solution. The mixture was allowed to slowly warm up to room temperature and stir overnight. The mixture was quenched with water (15 mL); then it was extracted with dichloromethane (50 mL x 3). The combined organic solvents were washed with brine and dried over Na_2SO_4. After removing the solvent, the residue was purified by flash chromatography in 5% of methanol in dichloromethane to give a colorless oil **im-3-im** (162 mg, 23% yield). ^{1}H NMR (400MHz, $CDCl_3$, ppm): δ 7.45 (s, 2H), 7.11 (s, 2H), 6.89 (s, 2H), 3.92 (t, *J* = 8.0 Hz, 4H), 2.29 (m, 2H). ^{13}C NMR (100MHz, $CDCl_3$, ppm): δ 136.9, 129.6, 118.5, 43.2, 31.6.[1] HR-MS *m/z* calcd for $C_9H_{12}N_4$: 176.1063, found: 177.1135 ($M+H^+$).



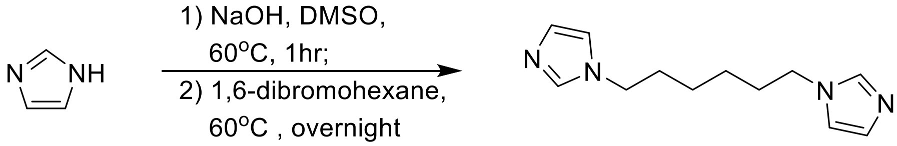
1,4-di(1*H*-imidazol-1-yl)butane (**im-4-im**):[95]

Imidazole (1.7 g, 25 mmol) and NaOH (1 g, 25 mmol) in 10 mL DMSO were heated to 60 °C for 1 hr, followed by addition of 1,4-dichlorobutane (1.40 mL, 12.5 mmol). The resulting mixture was stirred at 60 °C overnight. After cooling down to room temperature, 100 mL DI water was added and the mixture was vigorously stirred for 30 min. The white precipitate was collected by vacuum filtration and washed with plenty of DI water. The solid was dried in air first and then under high vacuum to give 1.2 g of **im-4-im** as white solid in 50% yield. ^{1}H NMR (400MHz, $CDCl_3$, ppm): δ 7.44 (s, 2H), 7.07 (s, 2H), 6.86 (s, 2H), 3.93 (m, 4H), 1.76 (m, 4H). ^{13}C NMR (100MHz, $CDCl_3$, ppm): δ 137.1, 129.9, 118.7, 46.5, 28.2.[2] HR-MS *m/z* calcd for $C_{10}H_{14}N_4$: 190.1221, found: 191.1293 ($M+H^+$).



1,5-di(1*H*-imidazol-1-yl)pentane (**im-5-im**):[96]

Imidazole (1.7 g, 25 mmol) and NaOH (1 g, 25 mmol) in 10mL DMSO were heated to 60 °C for 1 hr, followed by addition of 1,5-dibromopentane (1.7 mL, 12.5 mmol). The resulting mixture was stirred at 60 °C overnight. After cooling, the mixture was diluted with water (100 mL); then it was extracted with dichloromethane (150 mL x 3). The combined organic solvents were washed with water, brine, dried over Na_2SO_4. After removing the solvent, the residue was purified by flash chromatography in 3.3% of methanol in dichloromethane to give a colorless oil **im-5-im** (0.72 g, 28% yield). ^{1}H NMR (500MHz, $CDCl_3$, ppm): δ 7.44 (s, 2H), 7.06 (s, 2H), 6.87 (s, 2H), 3.93-3.90 (m, 4H), 1.81-1.75 (m, 4H), 1.29-1.26 (m, 2H). ^{13}C NMR (125MHz, $CDCl_3$, ppm): δ 137.0, 129.5, 118.7, 46.7, 30.6, 23.7.[3] HR-MS *m/z* calcd for $C_{11}H_{16}N_4$: 204.1376, found: 205.1449 ($M+H^+$).



1,6-di(1*H*-imidazol-1-yl)hexane (**im-6-im**):[97]

Imidazole (1.7 g, 25 mmol) and NaOH (1 g, 25 mmol) in 10mL DMSO were heated to 60 °C for 1 hr, followed by addition of 1,6-dibromohexane (1.89 mL, 12.5 mmol). The resulting mixture was stirred at 60 °C overnight. After cooling, the mixture was diluted with water (100 mL); then it was extracted with dichloromethane (150 mL x 3). The combined organic solvents were washed with water, brine, dried over Na_2SO_4. After removing the solvent, the residue was purified by flash chromatography in 3% of methanol in dichloromethane to give a colorless oil **im-6-im**

(0.96 g, 35% yield). ^{1}H NMR (500 MHz, $CDCl_3$, ppm): δ 7.44 (s, 2H), 7.05 (s, 2H), 6.88 (s, 2H), 3.92-3.89 (m, 4H), 1.77-1.74 (m, 4H), 1.30-1.27 (m, 4H). ^{13}C NMR (125 MHz, $CDCl_3$, ppm): δ 137.0, 129.4, 118.7, 46.8, 30.9, 26.1.[4] HR-MS *m/z* calcd for $C_{12}H_{18}N_4$: 218.1531, found: 219.1603 (M+H$^+$).

2.4.2 Additional Experimental Data

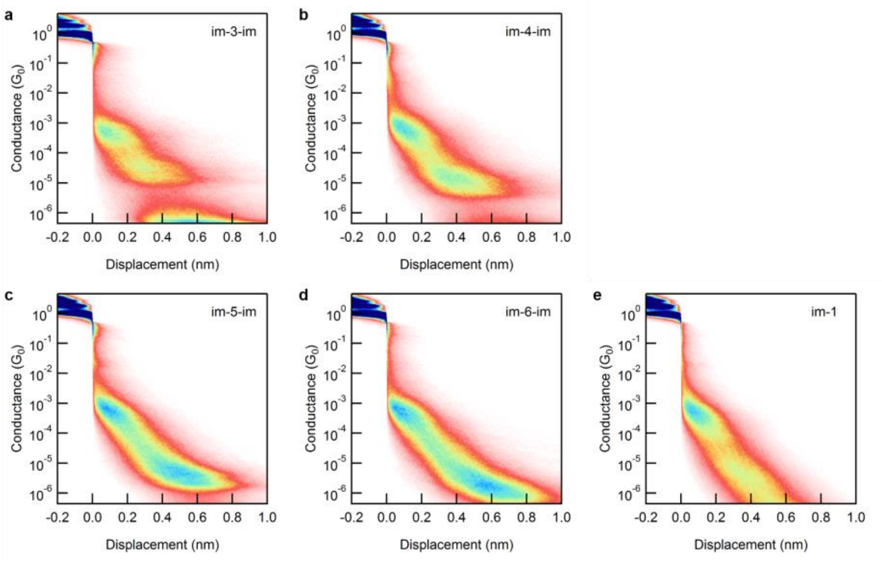


Figure 2.5: 2-Dimensional conductance histograms, log-binned (100 bins per decade) on conductance axis and linear-binned (100 bins per 0.08 nm) on the displacement axis, for (a) im-3-im, (b) im-4-im, (c) im-5-im, (d) im-6-im and (e) im-1.

Figure 2.5 shows the 2-dimensional conductance histograms of the molecules investigated in the main text. From a 2D histogram, an approximate length of junction can be read. Comparing these histograms, we can see the π-π stacked dimer peaks have roughly the same length across this series, while the molecular conductance peaks are longer for the longer molecules.

In the theory section, we applied the simplified Landauer formalism and compared the measured conductance with the calculated transmission at Fermi energy, which is typically valid for low-bias measurement. In our experiment, however, a relatively high bias (900 mV) is applied to increase the signal to noise in our measurements. For **im-*N*-im** junctions, the calculated transmission functions are relatively flat in the ±0.5 eV region, and thus our approximation is still valid. Figure 2.6 shows the conductance histograms of **im-4-im** measured under 90, 180, 360 and 540 mV biases. Compared to the 900 mV measurement, these conductance peaks measurement under different biases are at nearly identical positions (although the low-G peak under 90 mV is below the noise floor). This rationalizes the approximation we applied here.

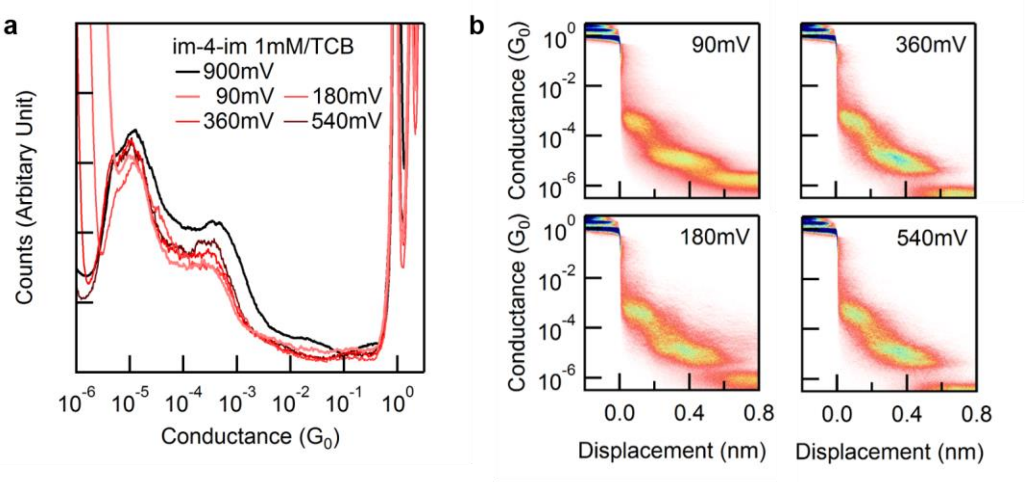


Figure 2.6: (a) 1D histograms of im-4-im under 90, 180, 360 and 540 mV biases, comparing compared to the same molecule measured under at a 900 mV bias (black line, histogram in the main text. (b) 2D histograms of im-4-im under 90, 180, 360 and 540 mV biases); the histogram of 180, 360 and 540 mV are from only 5,000 traces.

2.4.3 Binding Energy

We estimate the binding energy between imidazole's pyridine nitrogen and gold by DFT-based calculations. Here, 20-atom Au pyramids form the electrode. The total energies of an individual Au_{20} pyramid (-10713066.248 eV) and an individual **im-4-im** molecule (-16556.448 eV) are calculated, to compare with the total energy of the **im-4-im** + Au_{20} complex (-10729623.681 eV). The energy difference of 0.98 eV between the complex and the sum of a Au_{20} pyramid and an individual molecule is the binding energy between an imidazole and an Au electrode. In these calculations, the **im-4-im** molecule is allowed to relax fully to determine an optimum geometry, while Au atoms are held fixed. We also estimate the two-site binding energy by calculating the total energy of the Au_{20}- **im-4-im** - Au_{20} complex (-21442690.850 eV). Following the same method, the binding energy per site from this calculation is just slightly lower (0.95 eV). We must keep in mind that these bind energies are estimates given the constraints used.

2.4.4 Additional Theoretical Calculations

For calculations with Au clusters, we applied double-ζ basis set (FHI-aims "light" setup) to make the calculations less expensive. To determine if this will significantly lower the quality of calculation, here we repeat the same calculations on **im-4-im** and **im-1** dimer junctions, with double-ζ plus polarization basis set (FHI-aims "tight" setup) employed on light atoms (H, C, N and O). In Figure 2.7, these calculated transmission functions are compared with transmission functions of **im-4-im** and **im-1** dimer junctions shown in the main text. In the region close to Au Fermi energy level, these variations of theoretical methods show tolerably small difference.

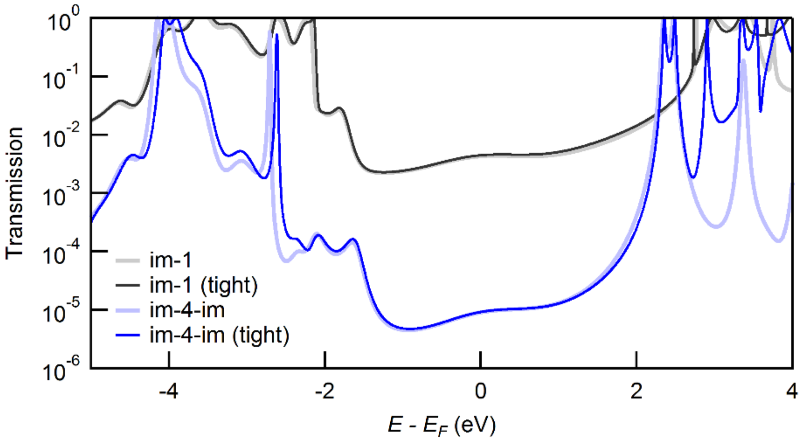


Figure 2.7: Calculated transmission of im-4-im monomer and im-1 dimer junctions using FHI-aims double-ζ basis (light) for all atoms and with FHI-aims double-ζ plus polarization basis (tight) for the molecule and light for Au atoms.

To verify that the dimer junction of **im-*N*-im** has comparable transmission with that of a **im-1** dimer junction, we determine the transmission for a *π*-*π* stacked **im-4-im** dimer. The two **im-4-im** molecules have in *gauche* conformation due to the steric hindrance of the gold electrodes, as shown in Figure 2.8a. The calculated transmission function of this junction is compared with that of an **im-1** dimer junction and a molecular **im-4-im** junction in Figure 2.8b. The dimer junction with either **im-4-im** or **im-1** gives very similar transmission especially around the Fermi energy. The sharp peaks between -1 and -2 eV for **im-4-im** dimer junction are Fano-resonances induced by the alkane side chains. This rationalizes the observation that **im-*N*-im** molecules have the *π-π* stacked dimer peak within the same range of conductance as **im-1**.

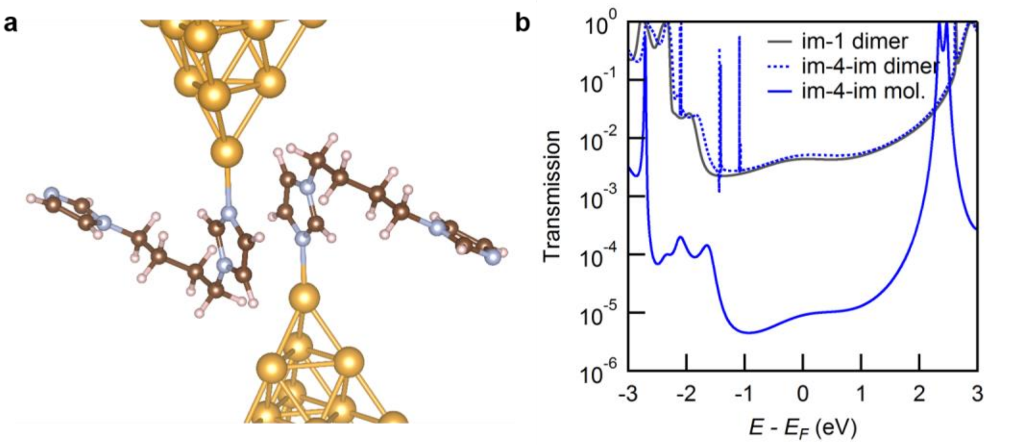


Figure 2.8: (a) The structure of the π-π stacked im-4-im dimer junction with alkane backbones adapting a gauche conformation. (b) The calculated transmissions for this junction (dotted blue line), along with that of the im-1 π-π stacked dimer junction (grey) and molecular im-4-im junction (blue).

2.4.5 Additional Flicker Noise Analysis

To further verify the origin the *π*-stacked conductance peaks and molecular conductance peaks of **im-*N*-im** molecules, we measure and analyze the flicker noise of **im-4-im**. As the same as the measurement on **im-1** described in the main text, the measurements are also conducted by holding an **im-4-im** junction for 150 ms, with a 100 kHz bandwidth. The power spectrum density (PSD) is also defined as the square of the integral of the discrete Fourier transform of the measured conductance between 100 Hz to 1,000 Hz. During the measurement, we cannot selectively form a certain type of junction. We did selection based on conductance and analyse selected traces instead.

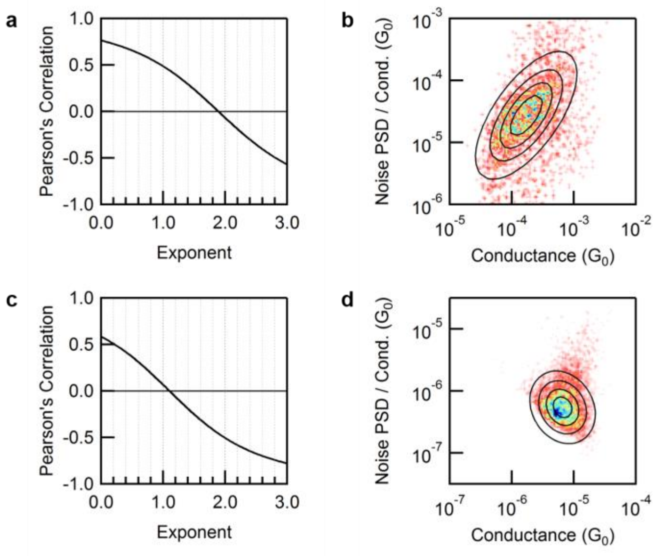
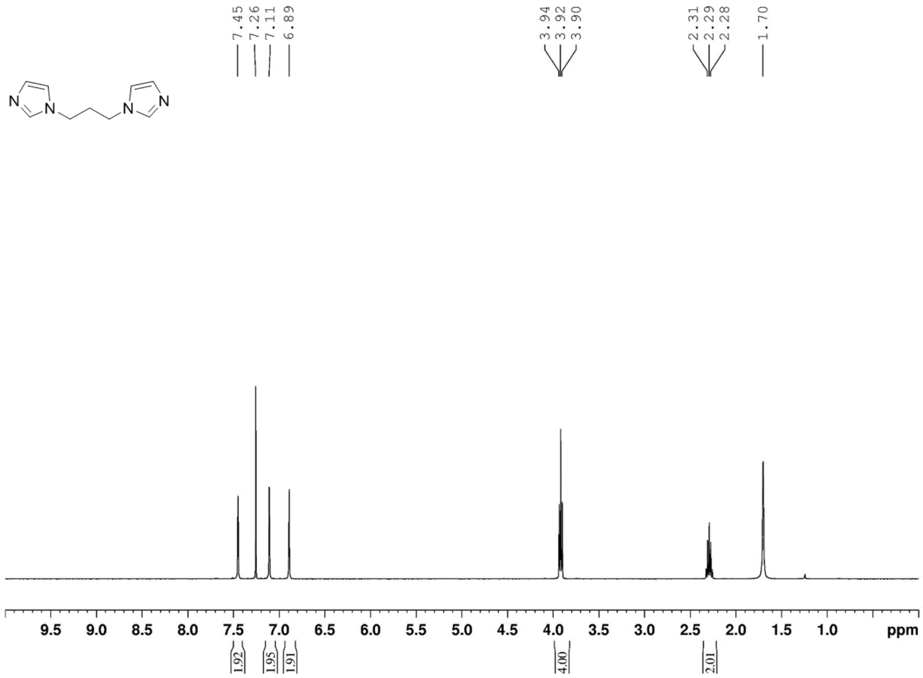


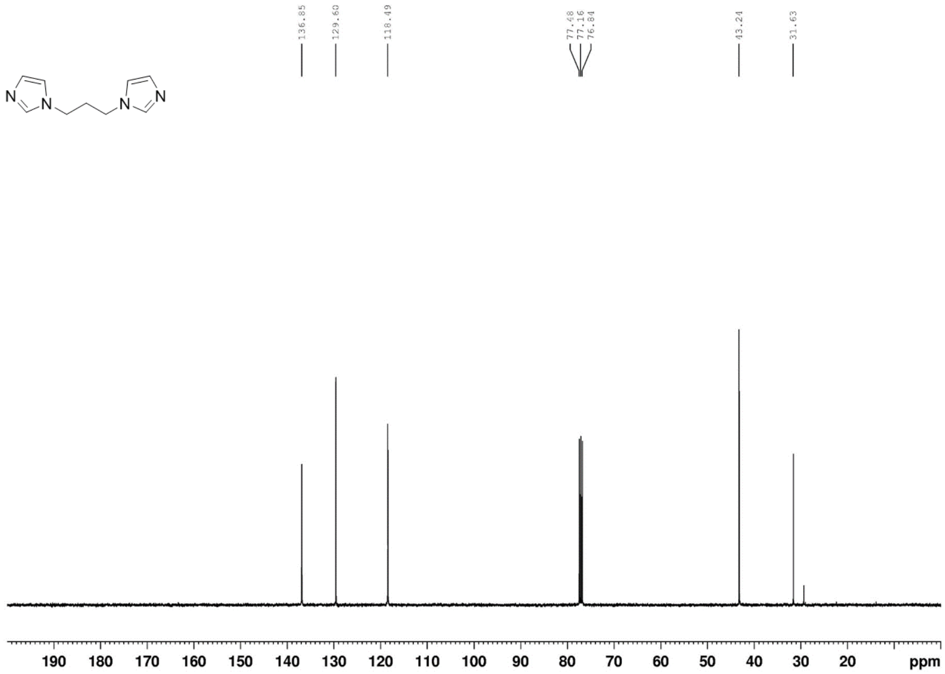
Figure 2.9: (a) The plotted correlation between PSD/G^N and the average junction conductance G versus the scaling exponent N and (b) the 2D histogram of PSD/G against G of the π-stacked peak of im-4-im. (c) The plotted correlation between PSD/G^N and G versus N and (d) the 2D histogram of PSD/G against G of the molecular conductance peaks of im-4-im.

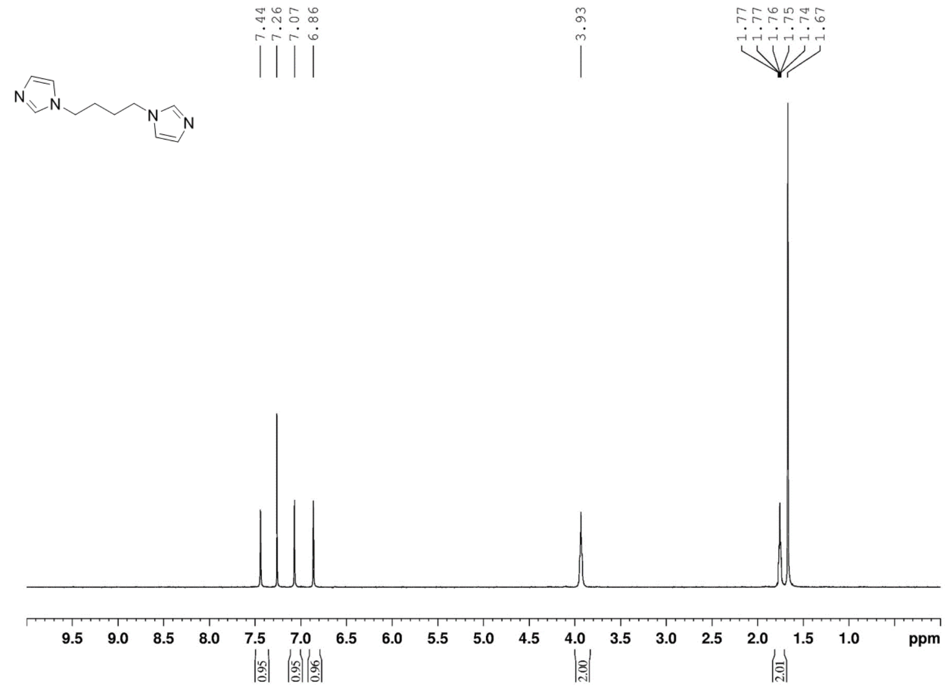
Here, we first select the 2,612 traces with conductance at around 10^{-3} to 10^{-4} G_0 and analyse. These traces consist the π-stacked conductance peaks. We also did the correlation analysis between PSD/G^N and the conductance G, and find they are independent when the scaling exponent N = 1.88, as shown in Figure 2.9a. The N close to 2 indicates this peak to be a through-space transmission peak and verifies our theory. The 2D histogram of PSD/G against G (Figure 2.9b) shows a strong linear correlation at N = 1.

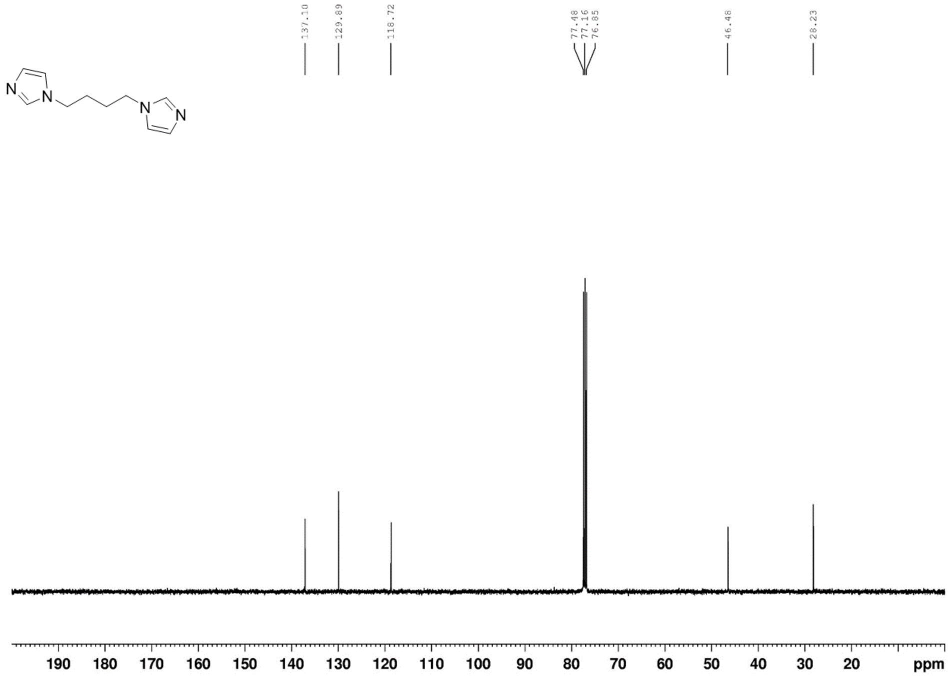
Then, we select the 3,666 traces with conductance at around 10^{-5} and analyse. These traces consist the molecular conductance peaks. Although in 1D histogram, there are a pair of molecular conductance peaks, they are hard to be selected trace by trace according to the conductance, as they are too close in conductance. Thus, we analyze these high-G and low-G molecular conductance peaks together and get N = 1.09 when PSD/G^N and G become independent (Figure 2.9c and 2.9d). The N close to 1 verifies that these peaks are from through-bond transmission.

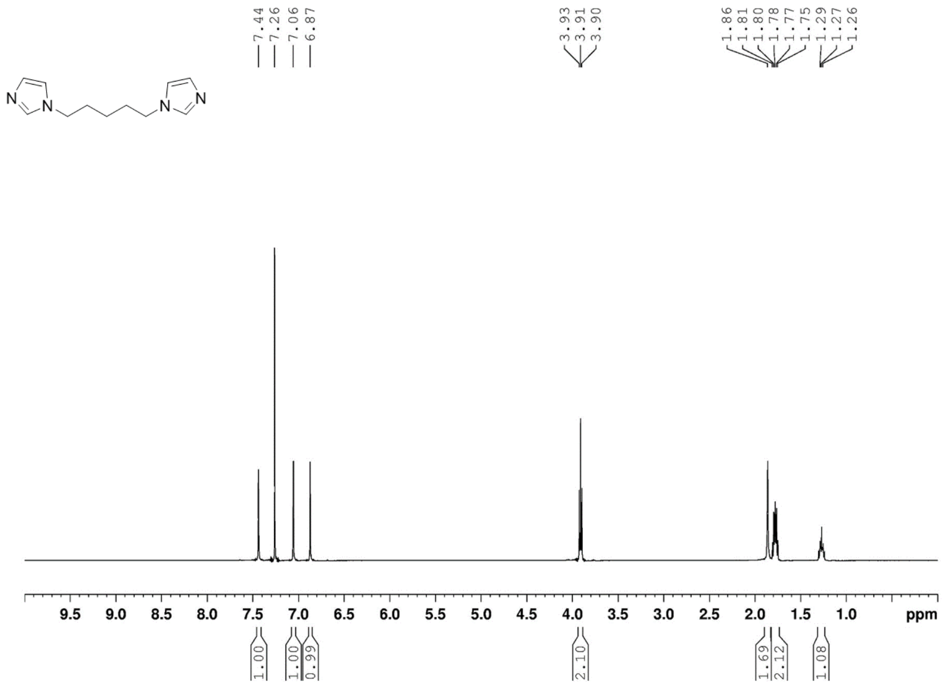
2.4.6 NMR Spectra

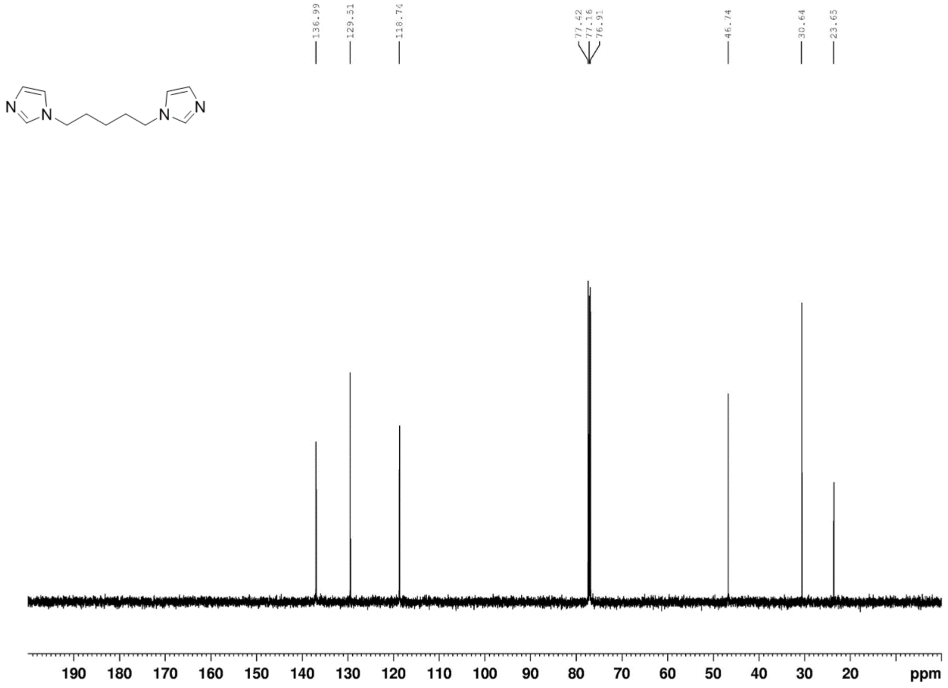


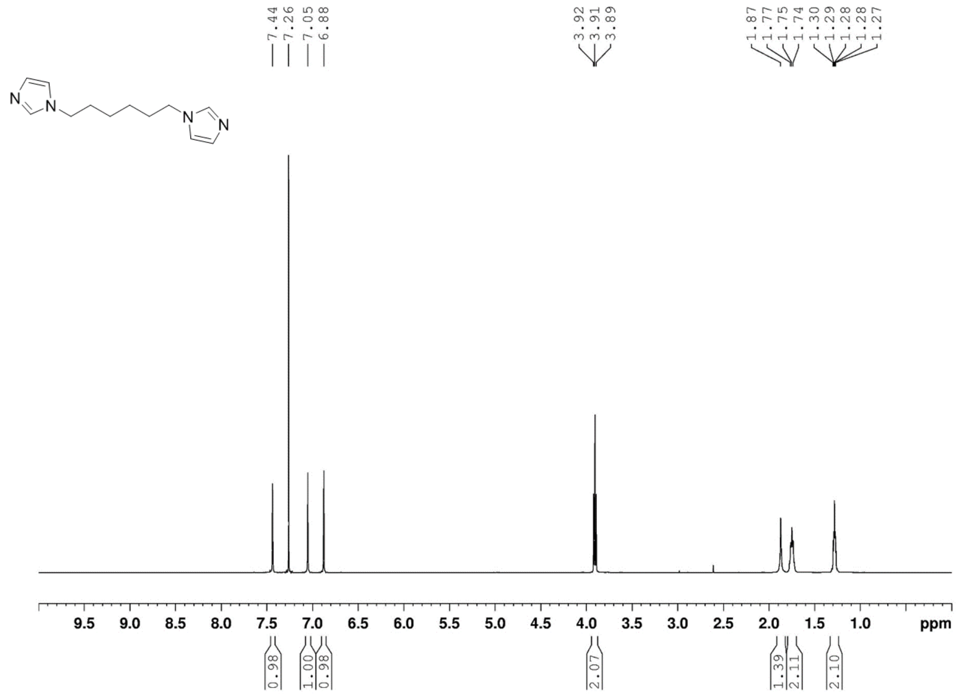


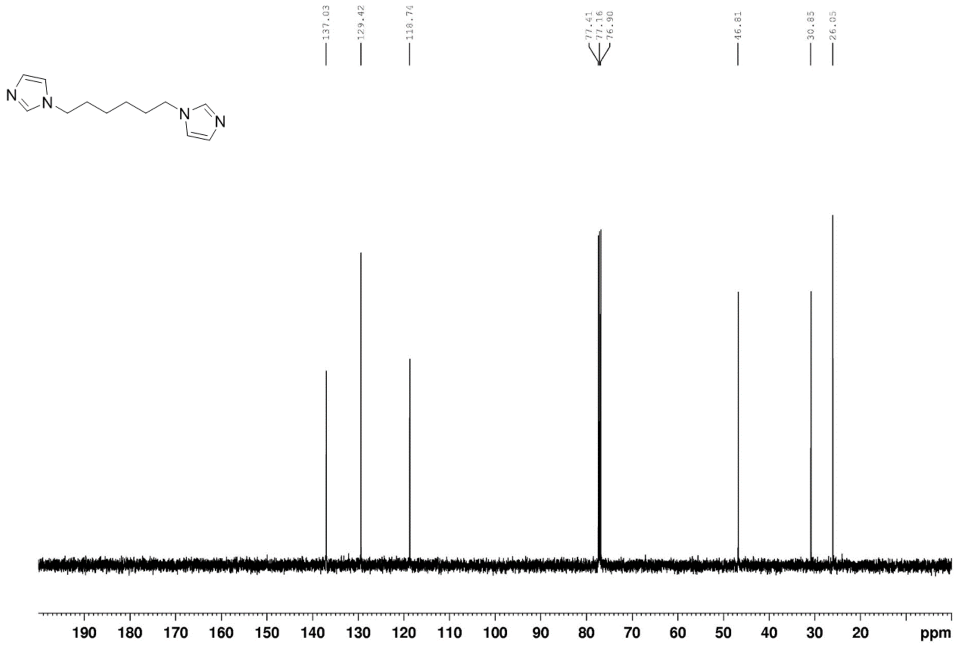

N
N
N
N
7.44
7.26
7.07
6.86
3.93
1.77
1.77
1.76
1.75
1.74
1.67
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
ppm
0.95
0.95
0.96
2.00
2.01


N
N
N
N
137.10
129.89
118.72
77.48
77.16
76.85
46.48
28.23
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
ppm


N
N
N
N
7.44
7.26
7.06
6.87
3.93
3.91
3.90
1.86
1.81
1.80
1.78
1.77
1.75
1.29
1.27
1.26
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
ppm
1.00
1.00
0.99
2.10
1.69
2.12
1.08


N
N
N
N
136.99
129.51
118.74
77.42
77.16
76.91
46.74
30.64
23.65
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
ppm


7.44
7.26
7.05
6.88
3.92
3.91
3.89
1.87
1.77
1.75
1.74
1.30
1.29
1.28
1.28
1.27
N
N
N
N
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
ppm
0.98
1.00
0.98
2.07
1.39
2.11
2.10


137.03
129.42
118.74
77.41
77.16
76.90
46.81
30.85
26.05
N
N
N
N
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
ppm

Chapter 3. Using Deep Learning to Identify Molecular Junction Characteristics

This chapter is based on the manuscript entitled *Using Deep Learning to Identify Molecular Junction Characteristics* by Tianren Fu, Yaping Zang, Qi Zou, Colin Nuckolls and Latha Venkataraman published in *Nano Letters*.[26] Prof. Qi Zou synthesized and characterized all the compounds. Prof. Yaping Zang and I performed the conductance measurements. I conducted the data analysis.

The scanning tunneling microscope-based break junction (STM-BJ) is used widely to create and characterize single metal-molecule-metal junctions. In this technique, conductance is continuously recorded as a metal point-contact is broken in a solution of molecules. Conductance plateaus are seen when stable molecular junctions are formed. Typically, thousands of junctions are created and measured, yielding thousands of distinct conductance versus extension traces. However, such traces are rarely analyzed individually to recognize the types of junctions formed. Here, we present a deep learning-based method to identify molecular junctions and show that it performs better than several commonly used and recently reported techniques. In this work, we demonstrate molecular junction identification from mixed solution measurements with accuracies as high as 97%. We also apply this model to an *in situ* electric-field driven isomerization reaction of a [3]cumulene to follow the reaction over time. Furthermore, we demonstrate that our model can remain accurate even when a key parameter, the average junction conductance, is eliminated from the analysis, showing that our model goes beyond conventional analysis in existing methods.

3.1 Introduction

Break junction techniques, such as the scanning tunneling microscope-based break junction (STM-BJ)[16-17] and mechanically controlled-break junction (MC-BJ),[33-34] are robust and powerful methods to create and characterize well-defined single Au-molecule-Au junctions. In break junction experiments, the electronic properties of these junctions are typically recorded although in addition mechanical, thermoelectric and noise characteristics can also be measured and analyzed.[38, 98-99] Most frequently, conductance data from these measurements are analyzed by looking at averages through histograms. However, a single break-junction measurement with multiple possible junction types requires a junction-by-junction analysis. This is especially true in STM-BJ measurements where *in situ* chemical reactions involve different molecules participating or created during the course of the measurement in one experiment.[51-54] Recently, machine learning methods have been applied to STM-BJ data.[20-21, 100-101] However, these methods still rely on averaging some aspects of the measurements, which results in a loss of information during the data preprocessing and analysis.

Deep learning is a powerful but more complicated machine learning technique which is capable of representing and analyzing multiple aspects of measured data. Recently, deep learning-based analysis have been applied to STM measurements,[24] nano-gap conductance data,[25] and to STM-BJ data using recurrent neural network[102] and deep auto-encoder[103] techniques. Among deep learning techniques, convolutional neural network (CNN) is a particularly powerful and popular method for image recognition.[65] Since STM-BJ data, which records conductance as a function of distance (or equivalently time), can be regarded as a 1D image, CNN can, in principle, be applied to such data. In this study, we develop a CNN-based model that can be applied to single-molecule conductance data collected using an STM-BJ setup and demonstrate its higher accuracy and robustness compared to non-deep learning models. Importantly, we show how this method can be used to characterize junctions where we remove a key parameter, its average conductance, highlighting the rich information available in conductance-time traces beyond what is analyzed using histograms.

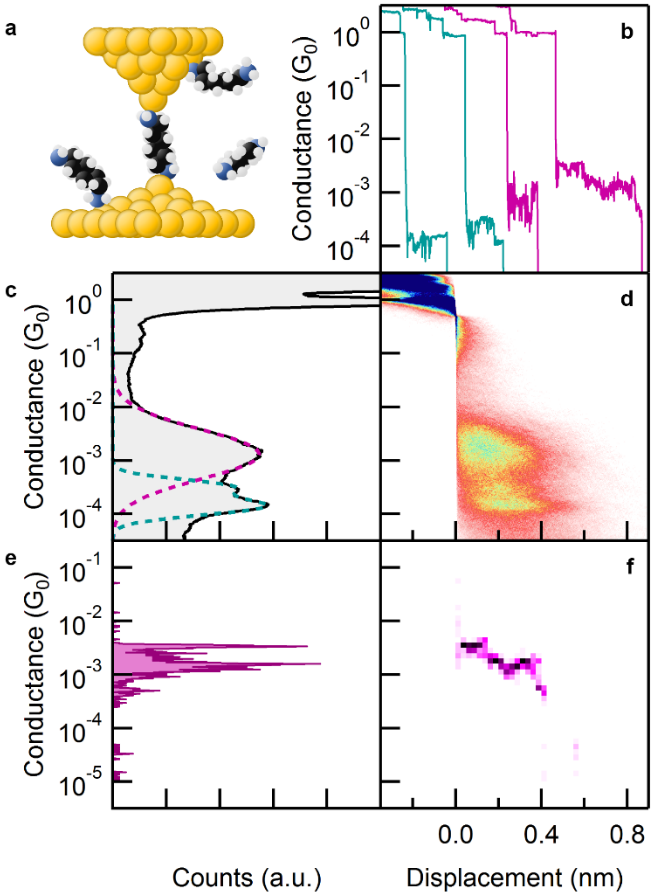


Figure 3.1: (a) Illustration of a molecular junction formed with STM-BJ. (b) Typical STM-BJ traces. (c) The 1D and (d) 2D histograms of a measurement of a mixed solution with 1,6-diaminohexane and 4,4′-bis(methylthiol)biphenyl. (e) The 1D and (f) 2D histograms of the rightmost trace (single trace) shown in (b) showing only the molecular conductance region.

In a single break junction measurement, two gold electrodes start in contact and are gradually pulled apart in a molecular solution, forming molecular junctions as shown in Figure 3.1a. Conductance is recorded as a function of the electrode separation. Plateaus at or above 1 G_0 ($G_0 = 2e^2/h$, the quantum of conductance) correspond to atomic size gold contacts and plateaus below 1 G_0 are attributed to a molecule bridging the gap between the two electrodes. Figure 3.1b shows several example conductance-versus-displacement traces measured in the presences of a mixture of two molecules. Typically, conductance traces are analyzed by creating 1-dimensional (1D) conductance and 2-dimensional (2D) conductance-displacement histograms from all measured traces, as shown in Figure 3.1c and 3.1d. From these histograms we can obtain the average junction conductance and the average junction elongation length.

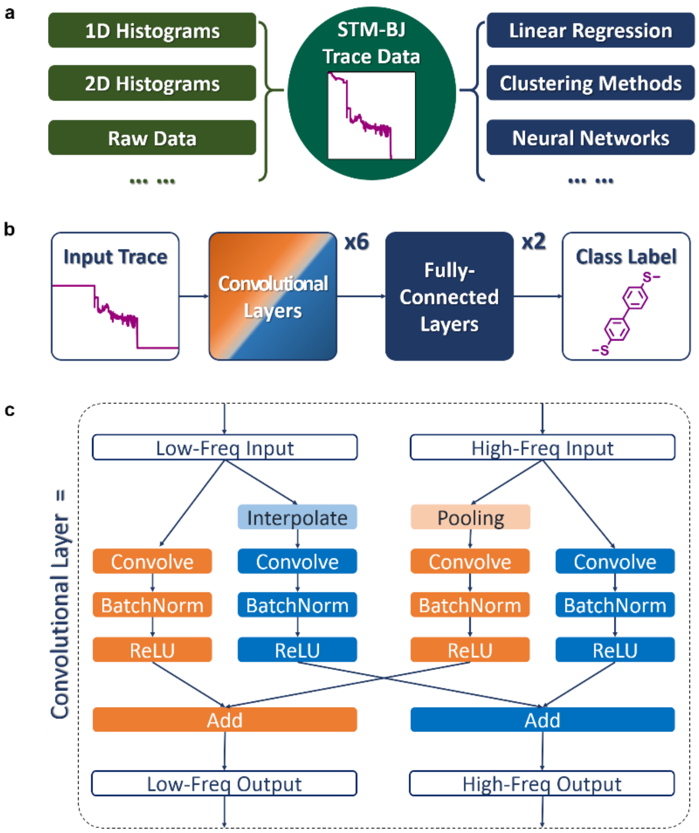


Figure 3.2: (a) Illustration of STM-BJ data analysis methods. On the left are the methods used for data preprocessing to generate an input from original trace. On the right are the models that can be applied to analyze STM-BJ data. (b) A simplified chart showing the flow of data in the CNN model used here. (c) The illustration of one convolutional layer shown in (b).

Single traces can also be converted to individual 1D and 2D histograms (see Figure 3.1e and 3.1f) and then analyzed using machine learning methods. For example, Hamillet *et al*[21] have used the principal component analysis (PCA) method on single-trace 1D histograms (denoted as

PC_1/1DH), while Cabosart *et al*[20] have applied a KMeans++ clustering algorithm[104-105] on single-trace 2D histograms (denoted as KMeans/2DH) to categorize STM-BJ data. However, both these methods lose information that is present in the raw conductance-versus-displacement traces. For example, focusing on the molecular conductance plateau (Figure 3.1b), we see that small fluctuations and oscillations are lost when these are converted into single-trace histograms (Figure 3.1e and 3.1f).

3.2 Results and Discussion

Here, we analyze the original STM-BJ conductance trace, i.e. a 1D array of conductance values. In Figure 3.2a, we summarize some common data analysis methods and show how traces are processed on the left and the classification algorithms used on the right. Among these, keeping all the raw data are likely the best, and this is easiest using a CNN-based analysis method. We therefore then design a CNN-based model as illustrated in Figure 3.2b. In this model, a clipped STM-BJ trace that excludes the gold point contact (data points with a conductance greater than 0.1 G_0) and noise floor (lower than 10^{-5} G_0) is taken as input. This focuses the analysis on the molecular conductance region. After processing the data with 6 convolutional layers and 2 fully-connected layers, the model generates a class label as output, identifying the molecular junction type. The fully-connected layer here has the same structure as a layer in a regular multilayer perceptron, where in each fully-connected layer, the input data matrix is multiplied by a weight matrix and offset by a bias matrix. The result from each of these multiplications undergoes a non-linear activation to break the linearity; here we use a rectified linear unit (ReLU), where the negative values are simply flattened to zero.[106] Dropout is then applied to provide extra robustness by randomly discarding outputs of some neurons during training; this prevents the network from relying on very few neurons.[107] The convolutional layers used in this model are of the octave convolution (OctConv) style,[108] as illustrated in Figure 3.2c. Compared to vanilla convolution, OctConv recognizes data shapes better and remains invariant under scaling (by introducing the low-frequency section in Figure 3.2c). Each of the four columns in Figure 3.2c represent a vanilla convolutional layer, with a 1D convolution operation, batch normalization (BatchNorm)[109] and ReLU. An OctConv layer is broken into four columns of convolutions providing the cross-processing within and between the high-frequency branch and low-frequency branch to keep information shared between the two spatial scales. Nearest neighbor interpolation and average pooling are used to double or half the size of data to match the different data sizes. The structure of OctConv layers is described in detail in Section 3.4.2.

To demonstrate the capabilities of this CNN model in classifying break-junction measurements trace-by-trace as well as those that have been used in the literature, we collect STM-BJ data using three commercial compounds: 1,6-diaminohexane (**1**), 4,4'-bis(methylthiol)biphenyl (**2**) and 1,6-bis(methylthiol)hexane (**3**) (structures shown in Figure 3.3a). We measure each molecule individually and as mixed solutions (**1** with **2**, and **2** with **3)** in 1,2,4-trichlorobenzene (TCB). The 1D and 2D histograms of the **1**/**2** mixture are shown in Figure 1c and 1d (and those of the **2**/**3** mixture are shown in supplementary Figure 3.7c and 3.7d). As an example, we train this CNN model on data obtained from measurements of pure **1** and pure **2**, and an accuracy of 97.6% is achieve on this test dataset (based on analyzing 10% traces that were not used in training). We use this trained model to label the traces from mixed **1**/**2** solution measurement and plot the 1D histogram of all the traces classified to be **1**- and **2**-like by the model in Figure 3.3b. Figure 3.3c shows the corresponding 2D histograms. These histograms are very much like those measured on pure **1** and pure **2** (shown in Figure 3.6a-3.6d). We do not see a peak at the conductance value

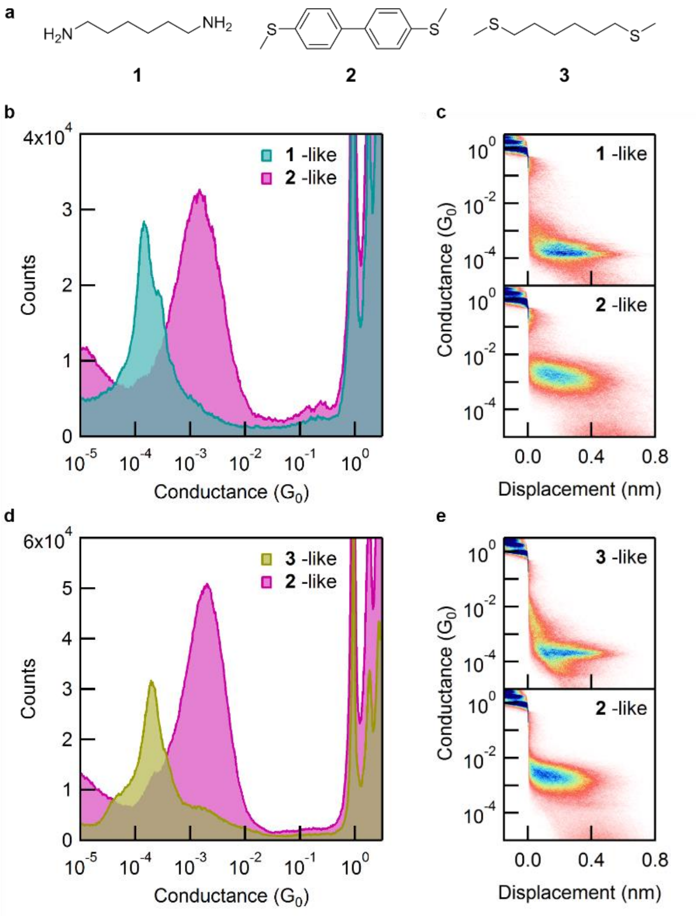


Figure 3.3: (a) Chemical structures of 1, 2, 3. (b) The 1D and (c) 2D histograms of the traces judged to be 1-like (3,406 traces) or 2-like (4,876 traces) by the CNN model from mixed solution measurements. The histograms of all traces are shown in Figure 3.1b and 3.1c, and histograms of measurements on pure solutions are shown in Figure 3.6. (d) The 1D and (e) 2D histograms of the traces judged to be 2-like (7,678 traces) or 3-like (4,098 traces) by the CNN model from mixed solution measurements.

corresponding to **2** in the **1**-like traces and vice versa indicating that the model is highly accurate. The corresponding classification result using model designs reported by others are shown in Figure 3.8; the accuracies of these models on pure molecule-test datasets are significantly lower (Table 3.1). We also train this model in the same way on the **2**/**3** data, and obtain a 95.9% accuracy on the pure molecule test dataset. The 1D and 2D histograms of the algorithm-labeled traces from mixed **2**/**3** solution measurements are shown in Figure 3.3d and 3.3e. We can see this CNN model performs extremely well in sorting data corresponding to molecules that have different backbone structures (alkane versus phenylenes). For molecule pairs with the same backbones (for example two alkanes such as the **1**/**3** pair, shown in Table 3.1), the classification accuracy is lower (89.6% on the test dataset). This indicates that the deep learning algorithm picks out features in the conductance traces that are likely related to the molecular backbone rather than the linker. It is possibly that the backbone contributes more to the trace properties such as the conductance value and plateau length.

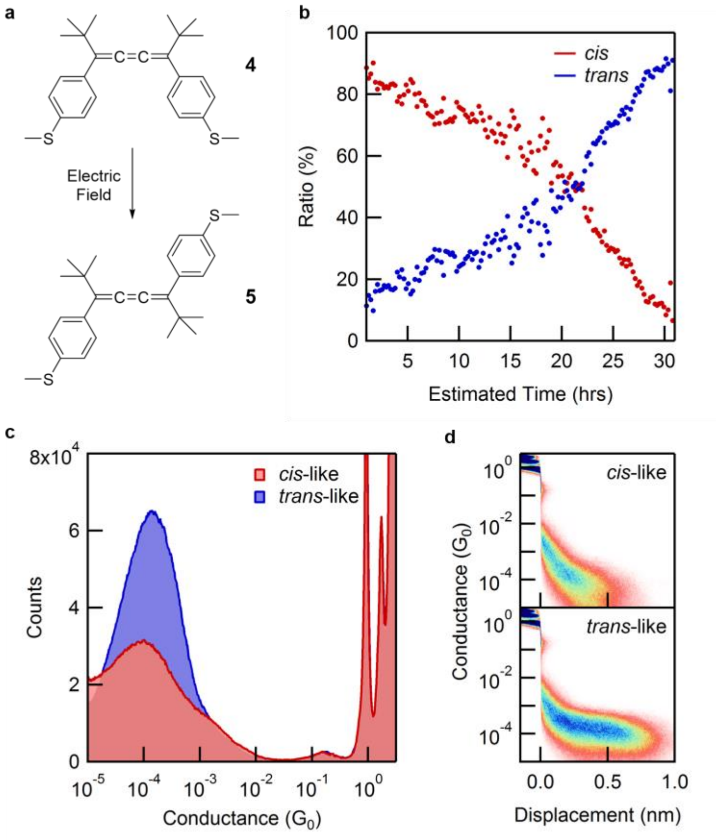


Figure 3.4: (a) Chemical structures of the [3]cumulene derivatives. Under electric field, the *cis*-isomer (4) transforms into the *trans*-isomer (5). (b) The percentage 4 (red dots) and 5 (blue dots) as a function of time as determined by the CNN model. (c) The 1D and (d) 2D histograms of the traces judged to be 4-like (4,997 traces) or 5-like (4,994 traces) by the CNN model from the 10,000 traces measured 22 hrs after starting with a pure 4 solution.

We next apply our CNN model to characterize conductance data measured with [3]cumulene derivatives **4** and **5** (structures shown in Figure 3.4a). We recently discovered and reported that the electric field in STM-BJ setup can isomerize the *cis*-isomer **4** to the *trans*-isomer **5** *in situ*.[54] In this experiment, we recorded more than 100,000 conductance traces over a period of 30-hour. By training the CNN model on measurements of pure **4** and **5** (achieving an 88.4% accuracy on the test dataset) and then applying it to the large data set, we determine the ratio of the *cis*-isomer **4** to the *trans*-isomer **5** as a function of time. Figure 3.4b shows this ratio determined from sets of 1,000 traces. From Figure 3.4b, we can observe the transformation of **4** to **5** during the timescale of the measurement. To demonstrate the performance this classification, we show the 1D and 2D histograms of the algorithm-labeled traces from a set of 10,000 traces measured at about 22 hrs after the start of the measurement in Figure 3.4c and 3.4d. We can see that these histograms have a very similar appearance comparing to the histograms of pure *cis*-isomer **4** and *trans*-isomer **5** (supplementary Figure 3.6g-3.6j), highlighting the accuracy of our model.

In Table 3.1, we show results from applying the alternative models to sort different conductance data. We test the PC_1/1DH and KMeans/2DH models (taken from the literatures[20-21]) and also introduce two additional ones. The first is a "*brute force"* method, which uses individual trace conditional histogram, and then sorts data based on the number of counts within different conductance regions.[110] The second is a naïve logistic regression (LogitR), which does a logistic regression on the raw clipped conductance trace as a series of independent variables; this method is a simple linear model using the same input as the CNN model introduced in this work. We can see from the first column of Table 1 that the CNN model performs significantly better than all these simpler models for the mixed **1**/**2** molecule pair. Thus, although CNN needs more computational power for the training step, its extra complexity yields higher classification accuracy.

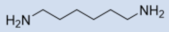
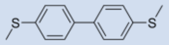
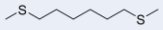
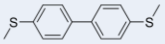
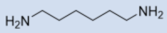
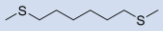
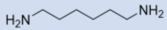
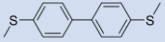
Molecule Pair	CNN on raw	BruteForce	PC$_1$ on 1DH	KMeans on 2DH	LogitR on raw
1  2 	97.6 % (100% training)	87.8 % (88% training)	89.4 % (89% training)	85.7 % (84% training)	77.3 % (81% training)
3  2 	95.9 % (99% training)	87.0 % (87% training)	86.7 % (87% training)	83.9 % (82% training)	77.3 % (81% training)
1  3 	89.6 % (97% training)	60.7 % (61% training)	63.2 % (63% training)	61.5 % (61% training)	54.5 % (63% training)
4 *cis*-Cumu[3] 5 *trans*-Cumu[3]	88.4 % (95% training)	79.0 % (79% training)	75.7 % (76% training)	68.6 % (68% training)	77.2 % (80% training)
1  2  0.4 nm fragment	94.4 % (98% training)	53.8 % (54% training)	60.7 % (63% training)	51.4 % (53% training)	66.0 % (71% training)

Table 3.1: The comparison among the reported and proposed models described in the main text. The accuracies of different models on different molecule pairs are shown in the table. For each experiment shown in each cell, 90% of the labeled dataset are used to train the model, while the remaining 10% are used for testing. In each cell, the accuracy on the test dataset is shown in the center, and the accuracy on the training dataset given in parenthesis.

The accuracy of these four models on all other systems considered here are also shown in Table 1. For sorting the **1**/**3** mixture (the 3rd row) where the backbones are the same and the individual molecular conductances are also similar (both at ~2×10^{-4} G_0), the accuracy is lower for all models when compared to the **1**/**2** and **2**/**3** mixtures. However, the drop in accuracy for the CNN model sorting the **1**/**3** mixture is much smaller than for other models. This implies that the CNN model can identify trace characteristics beyond simply the conductance value. To test if this is indeed the case, we design a reference analysis where the average plateau conductance information is removed (5th row of Table 3.1). Instead of using the clipped conductance trace as input, we use a randomly selected 0.4 nm-long fragment of molecular conductance plateau from the clipped trace and then subtract the average conductance value of this segment from this data, in order to remove the influence from conductance value as well as plateau length. We then test all models using this new input. The accuracies of all the models decrease, but for the CNN model, the accuracy remains reasonably high (94.4% on the test dataset).

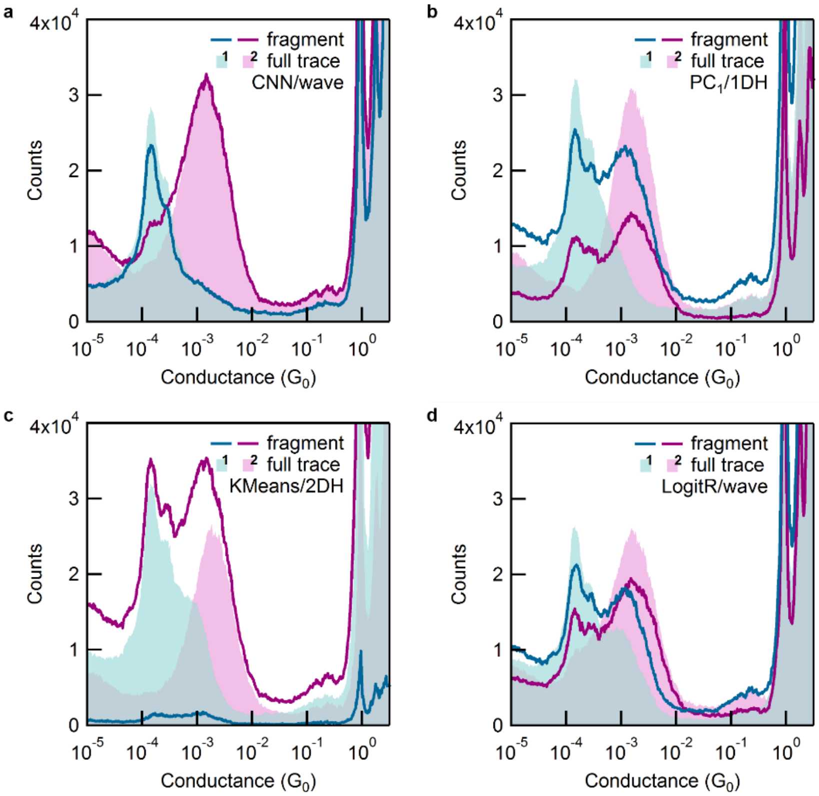


Figure 3.5: The 1D histograms of 1-like (the light blue) or 2-like (the magenta) traces sorted by different models from mixed solution measurements. The classification results based on using 0.4 nm fragments as inputs are shown as solid lines. As a reference, classification results using the clipped trace as input, are reproduced here as shaded regions. (a) The CNN model applied to 0.4 nm fragments yield 3,066 1-like and 5,216 2-like traces (compared with 3,406 1-like and 4,876 2-like traces when using the full trace). (b) The PC$_1$/1DH model applied to 0.4 nm fragments yield 6,053 1-like and 2,229 2-like traces (compared with 4,397 1-like and 3,901 2-like traces when using full trace). (c) The KMeans/2DH model applied to 0.4 nm fragments yield 392 1-like and 7,890 2-like traces (compared with 5,260 1-like and 3,022 2-like traces when using full trace). (d) Logistic regression model applied to 0.4 nm fragments yield 4,730 1-like and 3,553 2-like traces (compared with 4,569 1-like and 3,713 2-like traces when using full trace).

We next demonstrate the classifications of traces excluding the average conductance information on the mixture solution of **1** and **2** in Figure 3.5. The significant result here, shown in Figure 3.5a, is that discarding the average conductance information does not yield very different results when using the CNN model, showing its robustness against the elimination of average conductance information. For the other models, discarding conductance information produces a sorting that is more random. This indicates that these models rely strongly on the average conductance information.

3.3 Conclusions

In conclusion, we have demonstrated a new deep learning-based model to recognize molecular junction measurements performed with the STM-BJ technique that enables an accurate classification and characterization of molecular types. Comparing our model to some widely used and recently reported ones, we show that the CNN-based method achieves a much higher accuracy and importantly is able to sort traces without relying on the average conductance information, a critical innovation of this work. We demonstrate the application of this model to measurements of mixtures of molecules and also apply it to monitor an *in-situ* chemical reaction that is driven by the electric field during STM-BJ experiment. The excellent performance and robustness of this model makes it a favorable algorithm for analyzing such data. Its high-accuracy will enable more detailed investigations on systems with mixture of different kinds of molecular junctions, including, for example *in situ* reaction and surface chemistry.

3.4 Supplementary Information

3.4.1 STM-BJ Experiments in Detail

All the scanning tunneling microscope-break junction (STM-BJ) experiments in this work are conducted in ambient conditions.[17, 111] A gold tip and a gold-coated substrate are used as the two electrodes. To create molecular junctions, a piezo actuator is used to drive the tip and it is moved in and out of contact of the substrate at a rate of 20 nm·s^{-1}. During a measurement, the voltage (V) and current (I) across the junction are continuously recorded, and conductance G is calculated as $G = I/V$. These values are measured and recorded in a sampling frequency of 40 kHz. Hence, a 1 nm displacement corresponds to 2,000 data points.

To form junctions with molecules, a dilute solution of analytes is added on the substrate and measurements are made within this solution environment. For measurements with compounds **1**, **2** and **3**, we use 1,2,4-trichlorobenzene (TCB) as the solvent and apply a bias of 250 mV across the junction. For measurements with compounds **4** and **5**, we use *n*-tetradecane (TD) as the solvent and apply a bias of 100 mV across the junction. All the pure molecule measurements (shown in Figure 3.6) are done with solution concentration of 0.1 mM. The measurement of mixtures of **1** and **2** shown in Figure 3.3b, 3.3c, 3.7a and 3.7b have ~0.05 mM of **1** and **2**. The measurement of mixtures of **2** and **3** shown in Figure 3.3d, 3.3e, 3.7c and 3.7d have 0.01 mM of **2** and 0.0025 mM of **3**. The *in-situ* isomerization experiment of **4** starts with a 0.1 mM TD solution of **4**. The histograms shown in Figure 3.4c, 3.4d, 3.7e and 3.7f are from 10,000 traces measured ~22 hours after the experiment was started.

The compound **1** and **2** are obtained from Aldrich, and **3** from Alfa Aesar. These are used without further purification. The synthesis of compound **4** and **5** is reported in our previous work.[54]

3.4.2 Structure of OctConv Layers

As briefly discussed in Section 3.2, we apply OctConv-style convolutional layers.[108] As shown in Figure 3.2c in Section 3.2, one OctConv layer has a pair of inputs and outputs (a high-frequency and a low-frequency branch). The high-frequency branch uses the original input data, and the low-frequency branch has half the number of data points. Without the low-frequency input and output, the OctConv layer is the same as the vanilla convolutional layer.

Each column in Figure 3.2c is a vanilla convolutional layer structure, where the input matrix goes through a 1D convolution operation, and then ReLU. Batch normalization

(BatchNorm) technique is applied to achieve a better and faster training, by normalizing the output values of convolution into zero mean-unit standard deviation during the training.[109] As illustrated in Figure 3.2c, in one OctConv layer there are four of such vanilla convolutional columns; low-to-low, low-to-high, high-to-low and high-to-high. Among them, the low-to-low and high-to-high do not involve any transformation because the lengths of input and output are the same. For low-to-high, a nearest neighbor interpolation is used to double the data length by duplicating every data point. For high-to-low, an average pooling operation, which replaces two neighboring values by their mean is used to half the length.

The OctConv layers have two pairs of input and output, so when connecting two of them, we connect low-frequency to low-frequency, high-frequency to high-frequency. The other components of the network, however, have only one pair of input and output. As the high-frequency branch is the main stream, it is always retained. The low-frequency input/output is not used when connected to other components of the network. This means that two of the four columns of convolution (low-to-low and either low-to-high or high-to-low) are discarded because the low-frequency is not provided or is not generated.

3.4.3 CNN Model

The convolutional neural network (CNN) model described in the main text is an CNN-based model taking conductance traces as input and determining a class label as output. The model takes a 2,000-point-long 1D vector (segment of the conductance trace) as input. This corresponds to 1 nm of displacement in the measurement. As described in Section 3.2 (and illustrated in Figure 3.2b and 3.2c), the input is first processed by 6 OctConv layers. The underlying convolution operations use 1D kernel sizes of 7, 7, 7, 9, 9, 9, respectively for each layer. The numbers of channels are 32, 32, 32, 64, 128, 256, respectively. If an OctConv layer produces both high-frequency and low-frequency outputs, half of the channel number is distributed to either of them. For example, the third OctConv layer has 16 high-frequency channels and 16 low-frequency channels, giving a total of 32 channels. After all convolutional layers, the data is flattened from 2D (spatial × channels) into 1D vectors, so that they can be further fed into fully-connected layers. The following two fully-connected layers have size of 1,024 and 16, respectively. The dropout[109] rate applied on the fully-connected layers is 20%. Finally, after going through a sigmoid function, the output is generated as 1-bit probability of 0 or 1 providing a probability of the trace having the first label (0) or second label (1).

This model is trained on the TensorFlow platform.[112] For model training, a batch size of 32 and learning rate of 3×10^{-6} are used, with an Adam optimizer[113]. A weight decay[114-115] of 10% learning rate is applied to all the trainable variable in the whole model (except for BatchNorm) to provide extra regularization. These hyperparameters were not deliberately tuned in this work; fine tuning or improvement on the model is left for future work.

Preprocessing of conductance traces.

In order to focus on the molecular plateaus, we remove the gold conductance region of a trace (conductance > 10^{-1} G_0) before feeding it to the model. We also remove the noise floor (conductance < 10^{-5} G_0). In addition, all traces are aligned at close to the point when atomic Au-Au contact breaks (chosen to be 0.5 G_0), as is used in creating 2D conductance-displacement histograms. Finally, we feel only the first 2,000 points of data to the model, which represent the first nanometer of conductance data after Au-Au contact has ruptured.

Modifications analyzing data without including conductance.

In Section 3.2 and Figure 3.5, we discuss a reference analysis when average plateau conductance is removed from the input during the training process. As we described in the main text, here, we use an 800-point-long (0.4 nm) segment of conductance plateau. These segments are randomly cut from molecular conductance plateaus in traces. We first take the logarithm of the values and then subtract the segment average. The input size of the CNN model is set to 800 points. We also change the flatten operation into a global mean operation, where for each channel, it returns the average value. Compared to flatten, global mean does not keep spatial information, and hence more appropriate for this type of input as the absolute values of displacement do not have physical meaning here. While for the training process the segments are randomly cut from plateaus, for the recognition process, the only first 800 points after rupture of Au-Au contact is used to ensure that only one segment is generated from each trace.

3.4.4 Additional STM-BJ Histograms

STM-BJ histograms of pure molecular solutions.

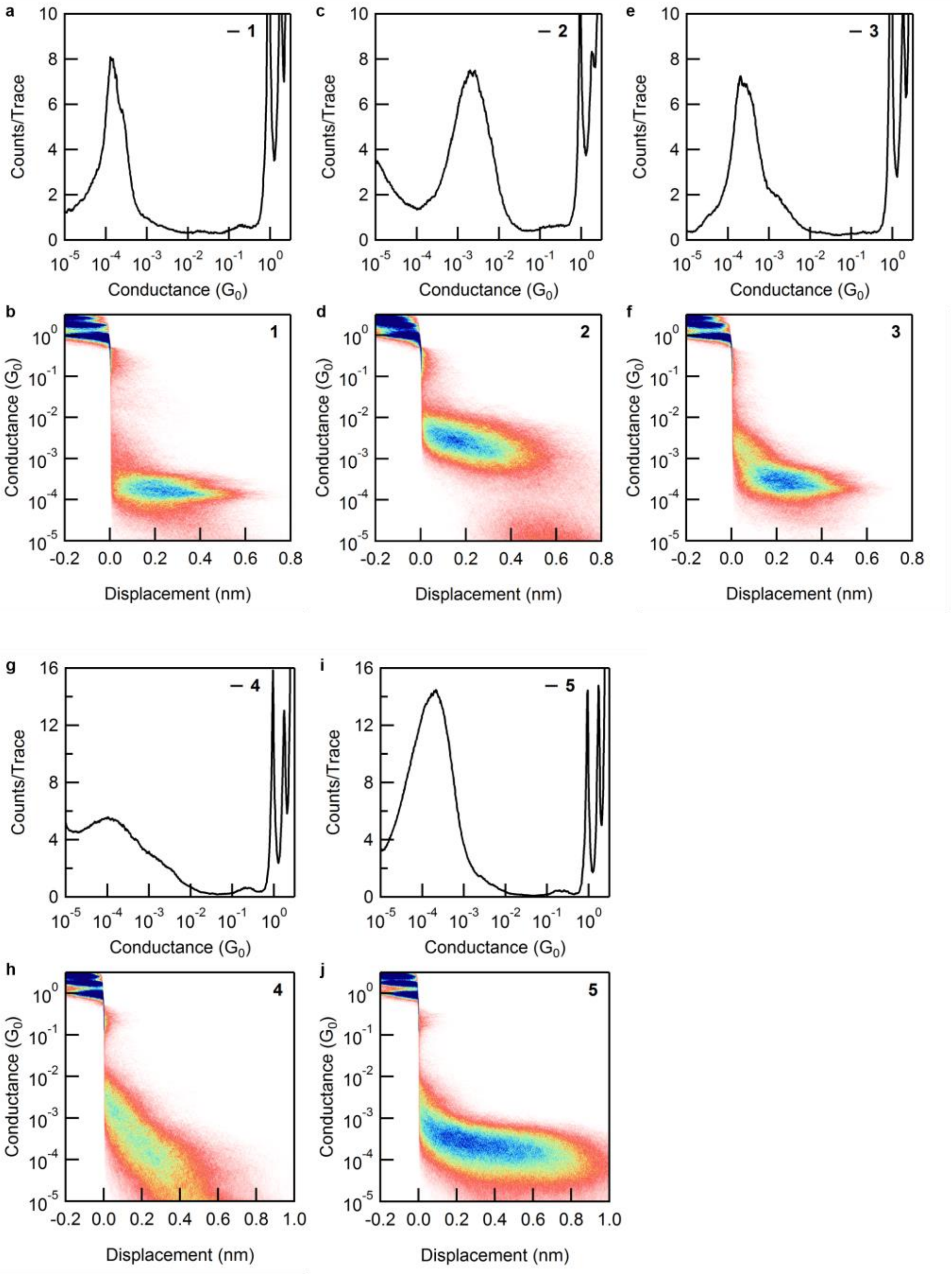


Figure 3.6: (a) 1D and (b) 2D conductance histograms of 1,6-diaminohexane (1). (c) 1D and (d) 2D conductance histograms of 4, 4′-bis(methylthiol)biphenyl (2). (e) 1D and (f) 2D conductance histograms of 1,6-bis(methylthiol)hexane (3). (g) 1D and (h) 2D conductance histograms of 4. (i) 1D and (j) 2D conductance histograms of 5.

STM-BJ Histograms of Mixtures.

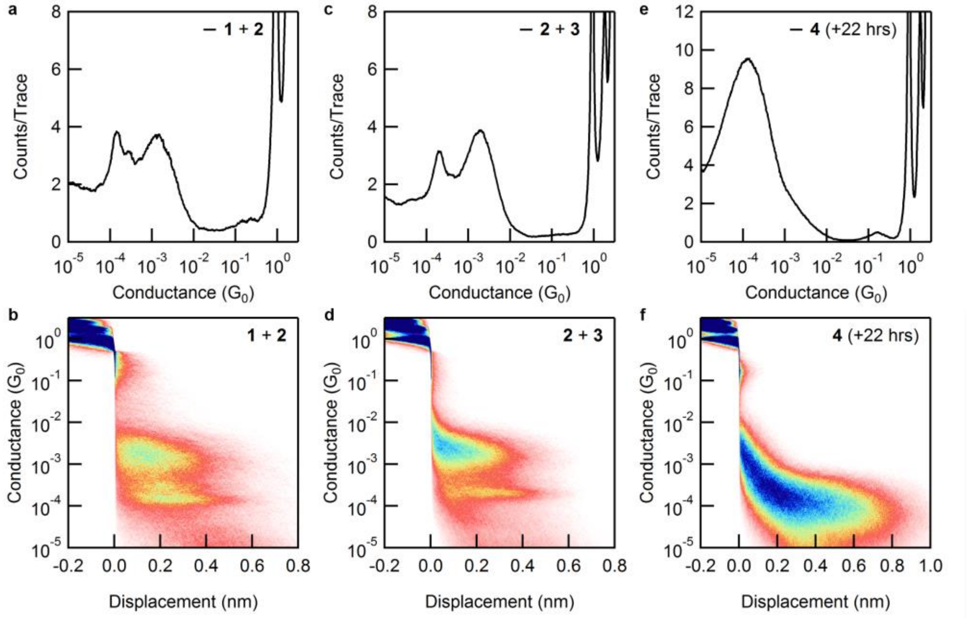


Figure 3.7: (a) 1D and (b) 2D conductance histograms of the mixture of 1 and 2. These are the same histograms as Figure 3.1c and 3.1d except that they are rotated. (c) 1D and (d) 2D conductance histograms of the mixture of 2 and 3. (e) 1D and (f) 2D conductance histograms of the traces measured 22 hours after the experiment starting with pure 4.

3.4.5 Classification Results of All Models

Results from different models with different molecule pairs shown in Table 3.1.

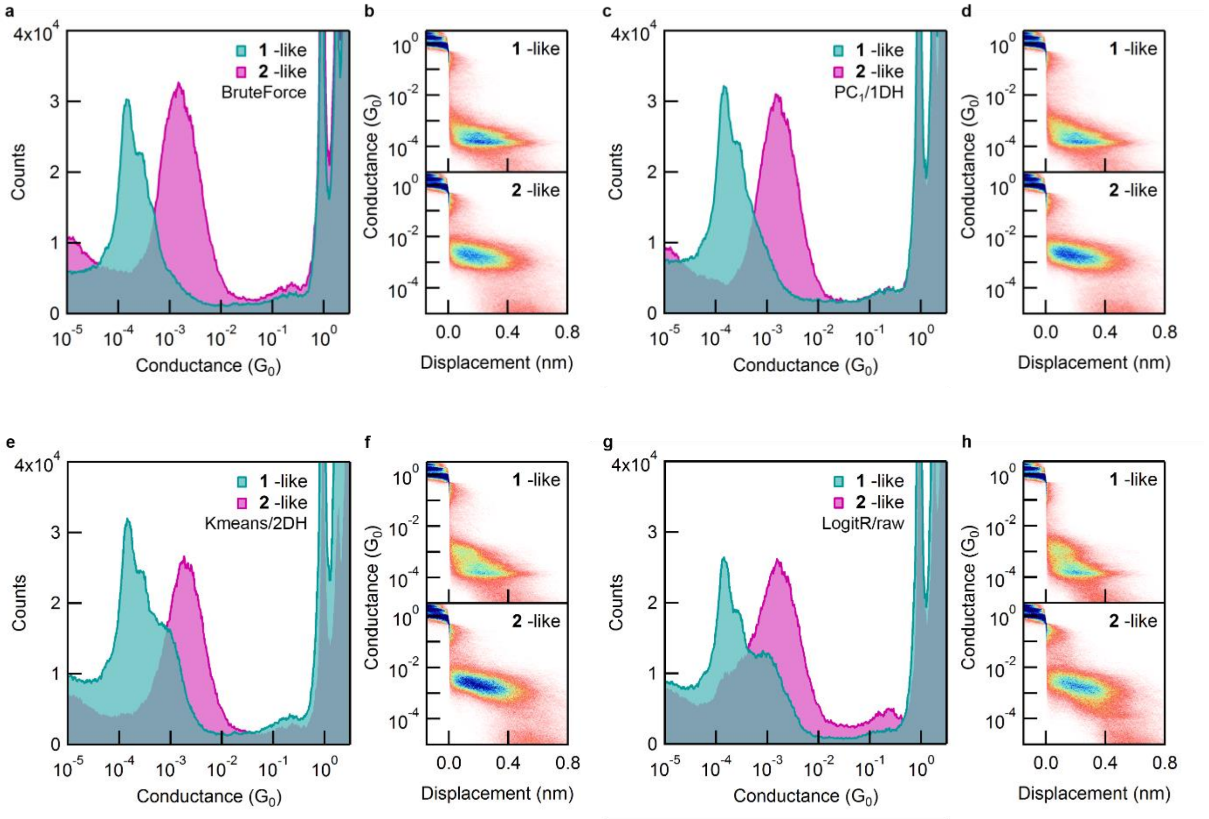


Figure 3.8: Histograms of the traces judged to be 1-like or 2-like from measurements of a mixed solution. (a) The 1D and (b) 2D histograms classified by the *brute force* model: there are 3,782 1-like traces and 4,516 2-like traces. (c) The 1D and (d) 2D histograms classified by the PC$_1$/1DH model: there are 4,397 1-like traces and 3,901 2-like traces. (e) The 1D and (f) 2D histograms classified by the KMeans/2DH model: there are 5,260 1-like traces and 3,022 2-like traces. (g) The 1D and (h) 2D histograms classified by the logistic regression model on raw traces: there are 4,569 1-like traces and 3,901 2-like traces.

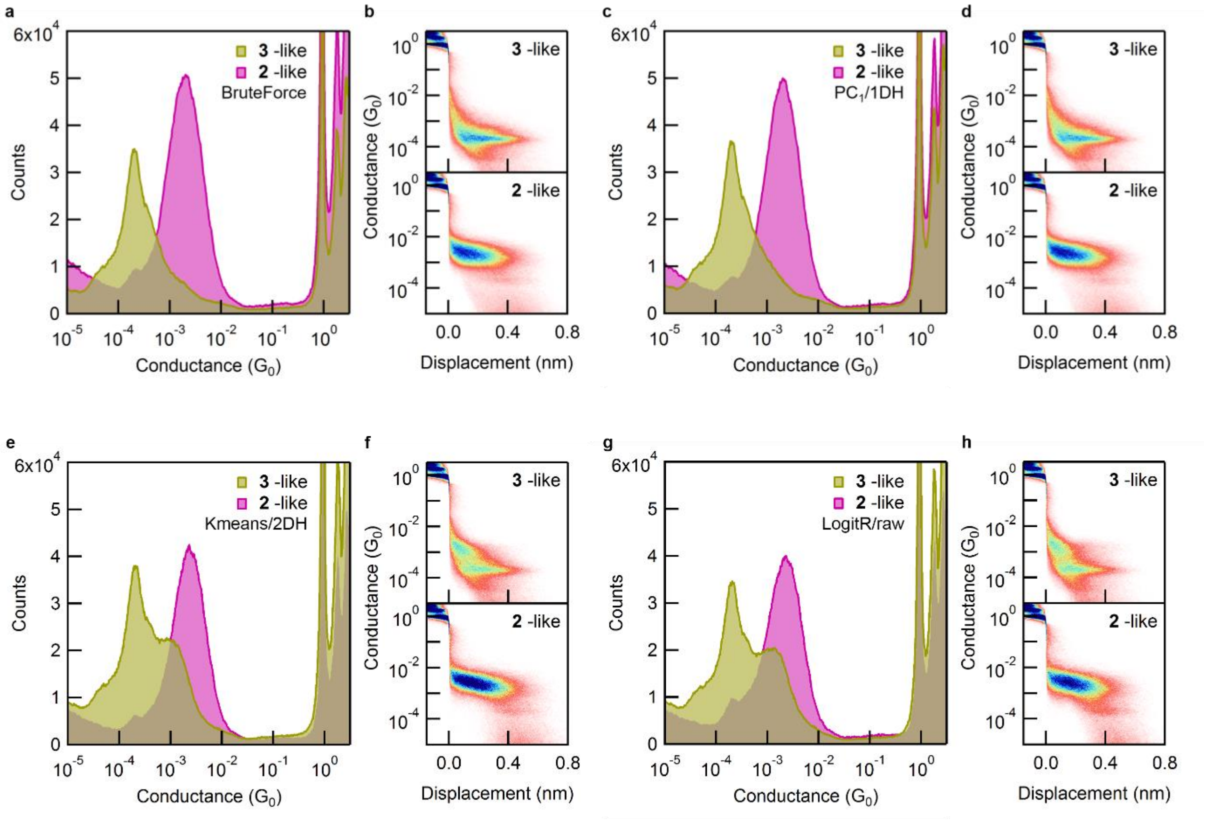


Figure 3.9: Histograms of the traces judged to be 2-like or 3-like from measurements of a mixed solution. (a) The 1D and (b) 2D histograms classified by the *brute force* model: there are 7,062 2-like traces and 4,737 3-like traces. (c) The 1D and (d) 2D histograms classified by the PC$_1$/1DH model: there are 6,549 2-like traces and 5,250 3-like traces. (e) The 1D and (f) 2D histograms classified by the KMeans/2DH model: there are 4,625 2-like traces and 7,151 3-like traces. (g) The 1D and (h) 2D histograms classified by the logistic regression model on raw traces: there are 4,959 2-like traces and 6,817 3-like traces.

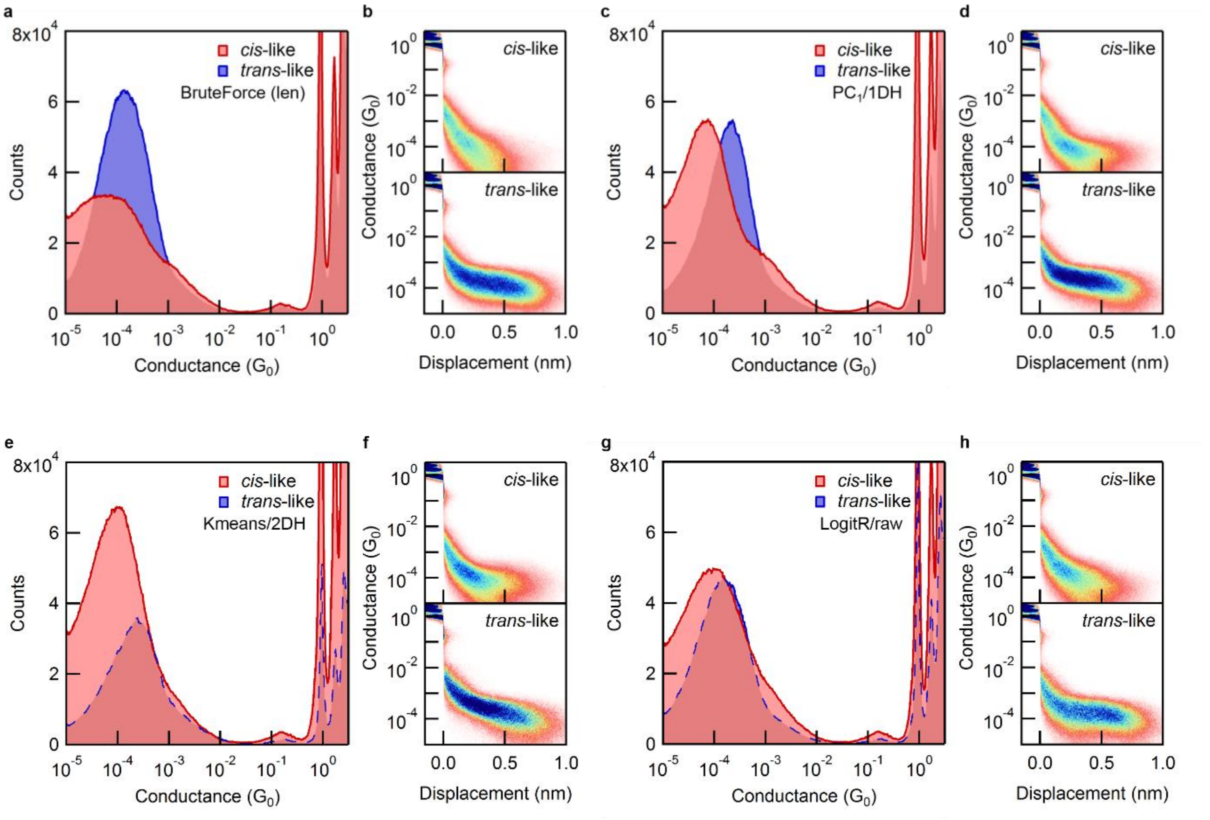


Figure 3.10: Histograms of the traces judged to be 4-like or 5-like from the traces measured after ~22 hrs starting with a pure 4 solution. (a) The 1D and (b) 2D histograms classified by the *brute force* model: there are 5,653 *cis*-like (4-like) traces and 4,338 *trans*-like (5-like) traces; here the *brute force* model judges based on counts of points in the molecular conductance region. (c) The 1D and (d) 2D histograms classified by the PC$_1$/1DH model: there are 6,577 *cis*-like traces and 3,414 *trans*-like traces. (e) The 1D and (f) 2D histograms classified by the KMeans/2DH model: there are 7,552 *cis*-like traces and 2,439 *trans*-like traces. (g) The 1D and (h) 2D histograms classified by the logistic regression model on raw traces: there are 6,691 *cis*-like traces and 3,300 *trans*-like traces.

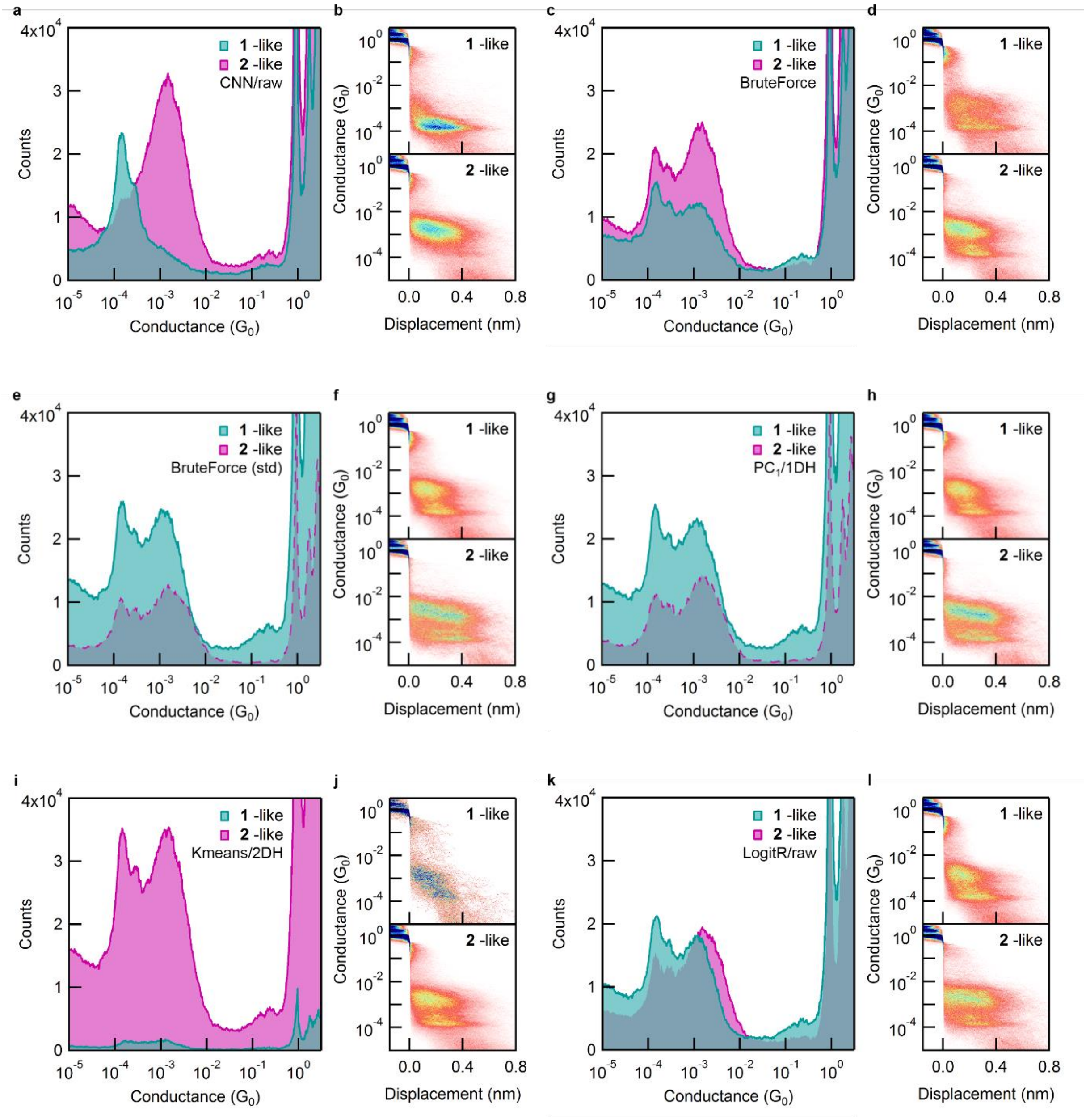


Figure 3.11: Histograms of the traces judged to be 1-like or 2-like with average plateau conductance removed from measurements a mixed solution. (a) The 1D and (b) 2D histograms classified by the CNN model with 3,066 1-like and 5,216 2-like traces. (c) The 1D and (d) 2D histograms classified by the *brute force* model with 3,245 1-like and 5,037 2-like traces. (e) The 1D and (f) 2D histograms classified by modified *brute force* model which uses the standard deviation of the sections with 6,331 1-like and 1,951 2-like traces; (g) The 1D and (h) 2D histograms classified by the PC_1/1DH model with 6,053 1-like and 2,229 2-like traces. (i) The 1D and (j) 2D histograms classified by the KMeans/2DH model with are 392 1-like and 7,890 2-like traces. (k) The 1D and (l) 2D histograms classified by the logistic regression model on raw traces with 4,730 1-like and 3,552 2-like traces.

Comparing performance of models in Figure 3.5 with random selection.

As shown in Figure 3.5, after changing the input from whole trace into segments of molecular conductance plateau with average plateau conductance removed, the performance of models other than the CNN model drops significantly. This reflects that these other models rely on the average conductance information. Figure 3.12 compares histograms made with randomly selected traces with those shown in Figure 3.5. We can see the PC_1/1DH model and KMeans/2DH model are close to a random selection.

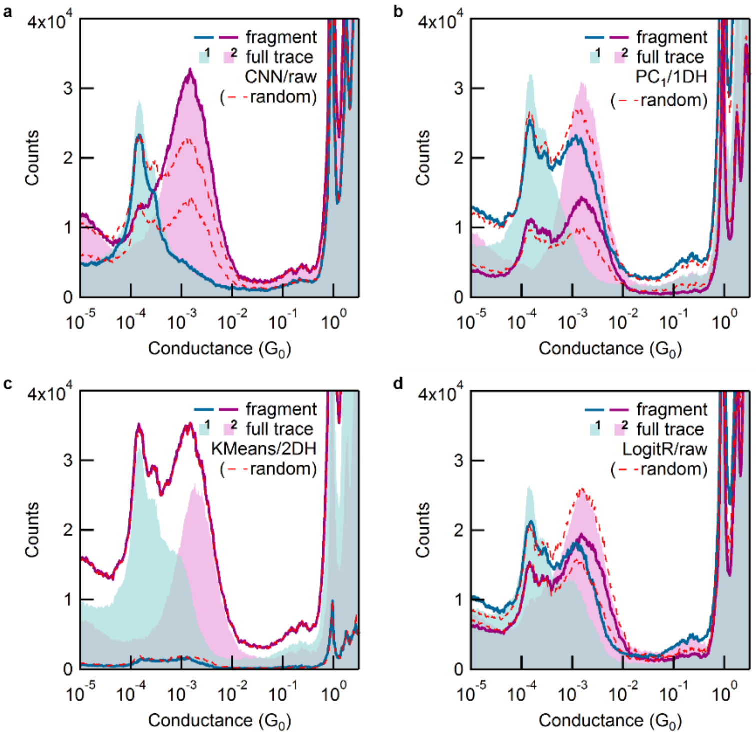


Figure 3.12: We compare the histograms shown in Figure 3.5 with ones generated from selecting the trace class randomly while keeping a similar number of traces as the sorted histograms to maintain a similar histogram height (red dashed lines).

Chapter 4. A Study of Single-Molecule Junction Formation and Evolution

This chapter is based on the working manuscript entitled *A Study of Single-Molecule Junction Formation and Evolution* by Tianren Fu, Kathleen Frommer, Colin Nuckolls and Latha Venkataraman. Kathleen Frommer and I performed the conductance measurements. I conducted the data analysis.

The scanning tunneling microscope-based break-junction (STM-BJ) technique is the most common method used to study electronic properties of single molecule junctions. It relies on repeatedly forming and rupturing an Au contact in an environment of the target molecules. The probability of junction formation is typically very large (~70-95%) prompting questions relating to how the nanoscale structure of the Au electrode before the metal point-contact ruptures alters junction formation. Here analyze conductance traces measured with the STM-BJ setup by combining correlation analysis and multiple machine-learning tools, including gradient boosted trees and neural networks. We show that two key features describing the Au-Au contact prior to rupture determine the extent of the contact relaxation (the snapback) and the probability of junction formation. Importantly, our data indicates strongly that molecular junctions are formed prior to the rupture of the Au-Au contact, explaining the high probability of junction formation observed in room-temperature solution measurements.

4.1 Introduction

The scanning tunneling microscope-based break-junction (STM-BJ) technique has proven to be a unique and versatile tool for investigating the physio-chemical properties of single metal-molecule-metal junctions.[16-17] STM-BJ technique can robustly construct and characterize single molecular junctions of with molecules ranging from organic, inorganic and bio-molecules.[116-121] It is also versatile in that it can be used to measure electronic, mechanical, thermoelectric and photoconducting properties of the junctions.[41-46] In STM-BJ experiments, the impact that the nanoscale electrode structure and its evolution and relaxation upon elongation and rupture play on the molecular junction formation is not well studied or well understood.[122-124] Recently, machine learning-assisted analyses have demonstrated the ability to analyze break-junction data to gain insights into molecular junction properties.[19-23] Here we employ machine learning techniques, from simple correlation analysis to deep neural networks, to comprehensively analyze this problem, and show that we can learn more about the underlying factors that make STM-BJ measurement method robust, reliable and reproducible.

In the STM-BJ method, metal-molecule-metal junctions are repeatedly formed and elongated until they break, while the current across the junction is continuously measured under an applied bias voltage, producing a conductance versus distance trace. At the start of such a trace, the metal electrodes are in contact, resulting in a high conductance, and as the STM tip is pulled away, the conductance drops in steps until a value close to 1 G_0, where $G_0 = 2e^2/h$ is the conductance quantum. This indicates the formation of a single atomic Au-Au contact, which breaks upon further elongation. Following its rupture, a single-molecule junction conductance plateau is often observed indicating that a molecule bridges the gap between the electrodes.[16-17] The average molecular plateau length is related to the molecular backbone length, however the plateau length varies significantly from trace to trace and can depend on the molecular configuration within the junction[122]. It could also be related to the junction formation probability which can depend on the linker groups[125-126]. This average plateau length, however, is not equal to the length of the molecular junction. The difference is often attributed to the fact that Au electrodes relax when point-contact ruptures opening up a gap, known as the “snapback” distance, and this is used to account for the difference between the molecular junction length and the plateau length.[127] Usually, this snapback distance is reported as a single value[28, 125, 127-128]. Here, we show that the snapback distance is affected by the structure of the Au contact formed prior to the formation of the molecular junction which is altered by the solvent, and thus depends on the experimental conditions. Importantly, we also show that for individual traces, the measured snapback is not strongly correlated with the plateau length of a molecular junction, indicating that the plateau length is much less sensitive to the contact formation history.

4.2 Results and Discussion

To probe the structure of the Au contact and determine snapback distances, we modified the standard STM-BJ measurement. The Au contact is initially pulled apart, then pushed back to remake contact, and finally pulled apart again. A sample “pull-push”[28, 129] conductance trace with the accompanying voltage ramp applied to the piezoelectric transducer that controls the substrate motion relative to the tip is shown in Figure 4.1 plotted against time. These measurements are made in a solvent, 1,2,4-trichlorobenzene (TCB), on a Au-coated substrate. As indicated in Figure 1, the time at which the Au-Au contact has the highest conductance (G_{max}) is designated as T_1 and occurs before a single atomic contact forms. This single-atom Au-Au contact breaks at time T_2, and this corresponds to the time when a large conductance drop is seen just below 1 G_0. The displacement at T_2 is denoted as L_{break}. Beyond T_2, conductance drops to the instrument noise floor

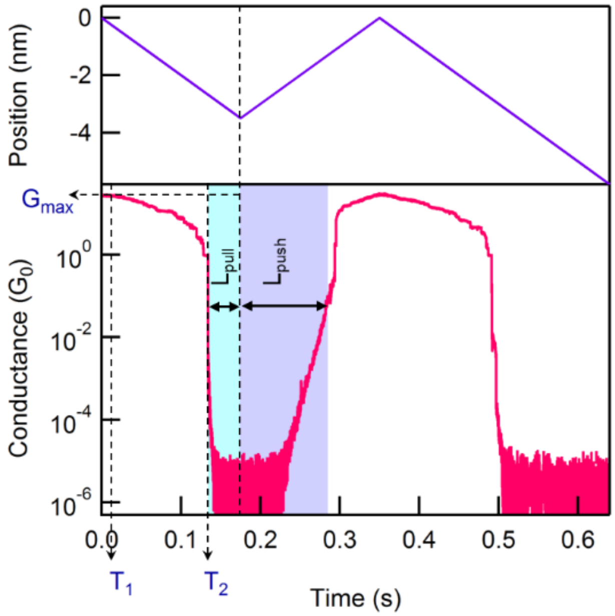


Figure 4.1: Sample piezo ramp (upper panel) and conductance versus time trace (lower panel) for a push-pull trace measured in 1,2,4-trichlorobenzene (TCB) at an applied bias of 100 mV and pulling rate of 15 nm/s. T_1 indicates the time when the highest conductance (G_{max}) is observed; T_2 is the time when the initial Au contact ruptures. L_{pull} indicates the distance pulled after the contact breaks and L_{push} indicates the distance pushed before the contact is reformed.

(~ 10^{-5} G_0) if there is no molecule and remains at this level through the end of the pulling phase (light blue shaded region in Figure 4.1 designated L_{pull}) and into the beginning of the pushing phase of the measurement. On pushing the electrodes further together (dark blue shaded region in Figure 4.1 designated L_{push}), a contact is formed again. We use a conductance threshold of 0.05 G_0 to indicate contact formation (although other thresholds up to 1 G_0 do not alter our findings). We see that the time required to reform a Au-Au contact while pushing is greater than the time the electrodes are pulled apart after breaking the contact (i.e. $L_{push} > L_{pull}$). This is due to the snapback reflecting that the Au electrodes relax after the contact is broken. The snapback is defined as $L_{push} - L_{pull}$. For our analysis, we also consider three additional features related to the Au-Au contact evolution: the slope of the conductance versus distance trace for the Au region (m_{Au}), which is determined by doing a linear regression on the region between T_1 and T_2; the length of the plateau around 1 G_0 (L_1) and length of the plateau around 2 G_0 (L_2). If the measurements are done in a solution of molecules, after T_2, instead of dropping into noise floor, a molecular conductance plateau is observed (see supplementary Figure 4.6 for a sample trace). The length of this plateau in a trace is the distance between the first and last point in the trace that is within the molecular conductance region as determined from a one-dimensional conductance histogram. These five parameters, G_{max}, L_1, L_2, L_{break} and m_{Au} describe evolution of the Au contact which we use to analysis the relation between snapback, molecular plateau length and the Au contact formation history.

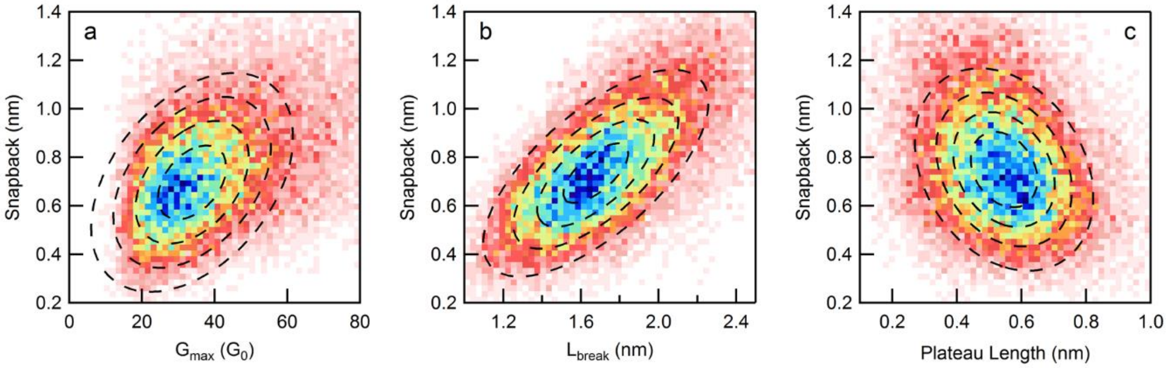


Figure 4.2: Two-dimensional (2D) correlation histograms constructed from 24,880 selected push-pull traces of 4,4''-diamino-*p*-terphenyl 0.1 mM in TCB solution. Black dashed lines are contour lines of 2D Gaussian fits. Snapback versus (a) G_{max} with a Pearson correlation coefficient of 0.412; (b) L_{break} with a Pearson correlation coefficient is 0.679 and (c) Molecular plateau length with a Pearson correlation coefficient is -0.265. See Figure 4.9e-h for histograms of raw data.

To demonstrate the correlation between these five parameters, two-dimensional correlation histograms are constructed from 24,880 measurements with 4,4”-diamino-*p*-terphenyl from a TCB solution. Figures 4.2a and 4.2b shows the correlation between snapback and G_{max} and L_{break}, respectively. We see that snapback is positively correlated with both parameters though the correlation between snapback and L_{break} is much stronger. To rationalize this finding, we note that in addition to structural changes, the force required to elongate a contact also stretches the Au-Au bonds within the tip asperities and the single atomic contact.[122] Like a spring, which will recoil more the more it is stretched, elongating the bonds over a larger distance (a larger L_{break}) will result in a larger relaxation upon rupture, resulting in a larger snapback. Moreover, a larger G_{max} indicates that the contact cross-section has many more atoms, and such a thicker contact will require a longer elongation to break resulting in a larger snapback. Figure 4.2c shows the correlation plot between snapback and molecular plateau length and reveals they are very weakly and negatively correlated. Since a larger snapback results in a wider gap right after the rupture of Au-Au point contact, it is reasonable that larger snapback reduces the further displacement needed to break the molecular junction, and results in a shorter plateau length. The small magnitude of correlation, however, is surprising. It indicates that the Au contacts relax fully only after the rupture of Au-molecule-Au junction. This can be rationalized by considering that the molecule can provide a force necessary to hold the electrodes in a slightly stretched form preventing them from relaxing as illustrated in supplementary Figure 4.9. However, this requires the molecule to be bridging across the electrodes even before the Au-Au point contact breaks, otherwise the relaxation will occur before the molecular bridge forms. This picture is indeed consistent with molecular dynamic simulations.[130-131]

To fully understand the impact of the Au-Au contact evolution history on snapback and plateau length, we plot in Figure 4.3a, the absolute value of their correlations with the five measured parameters: we can see that snapback depends primarily on G_{max} and L_{break}, and not strongly on L_1, L_2 and m_{Au}. However, none of these parameters are strongly correlated with molecular junction plateau length, indicating the plateau length is not determined by the evolution of the Au contact prior to rupture.

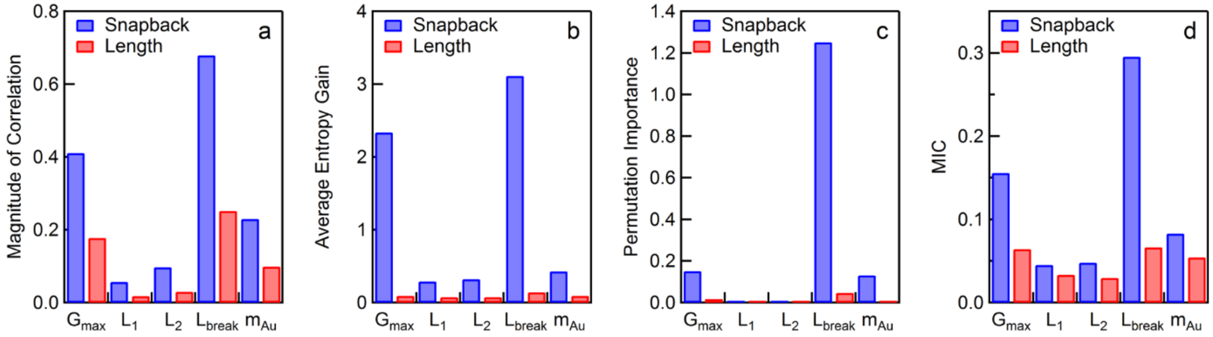


Figure 4.3: Metrics characterizing the importance of different parameters in determining snapback (blue bars) and molecular plateau length (red bars), for measurements in a 0.1 mM solution of 4,4''-diamino-*p*-terphenyl in TCB. The metrics are: (a) the magnitude of the correlation, (b) the total information entropy gain during the training of XGBoost models, (c) the permutation importance according to XGBoost models and (d) the mutual information coefficient (MIC).

The correlation analysis, however, only interprets the linear relations between these parameters. In order to confirm these findings and to see if there might be some nonlinear relations that could change the conclusion, we also apply a few other methods to characterize the importance of these five parameters on snapback and the molecular junction plateau length. The first method, gradient boosted trees (GBT)[132-133] is a machine learning algorithm with high expressivity and generalizability. GBT can find non-linear relations between parameters and has been widely used for feature extraction and selection.[134-136] For the analysis here, we use the XGBoost[60] package, which is one of the most powerful and frequently used implementation of GBT. In a typical GBT model, many decision trees are constructed to determine the dependent variable (say snapback or plateau length) from the independent variables (G_{max}, L_{break}, L_1, L_2 and m_{Au}). Each decision tree is made of many if-then-else decision nodes and the path taken through these nodes determines the output of a tree. During the training process, these nodes are built recursively on the independent variables (a process known as splitting) to satisfy the maximization of information entropy gain after applying the corresponding if-then-else rules.

We show in Figure 4.3b the importance of each parameter in determining snapback or plateau length. This importance is evaluated as the average information entropy gain for a parameter over all the splitting done during the tree construction process. We see that G_{max} and L_{break} are important parameters in predicting snapback and no parameter that describes the gold contact structure predicts the molecular plateau length. With this XGBoost model, we also measure the permutation importance of each feature, shown in Figure 4.3c. The permutation importance[59, 137] is a robust metric against bias on a parameter distribution or model design; permutation importance of one parameter is defined as the performance drop of the model when we randomly shuffle the parameter to make it irrelevant. If the parameter is important, the model will perform worse without it. Here, we find that L_{break} is very helpful in determining the snapback while all other parameters are not critical.

In Figure 4.3d, we show the maximum information coefficients (MIC)[138-140] between the parameters and snapback or plateau length using the *minepy* package.[138] MIC measures the dependence between two parameters that are either linearly or non-linearly related; it reflects the noise level in the data regardless of what the actual underlying relation between the parameters is. Again, we can see that L_{break} has a high importance in determining the snapback, G_{max} has some importance, but the other three are negligibly important. For the plateau length, however, none of the parameters are important. This confirms our earlier hypothesis that molecular junctions are formed prior to the rupture of the Au-Au contact.

To determine if our results are limited by the fact that we use just five manually-selected parameters, we try next to see if we can predict the snapback and plateau lengths by exploiting all the information on the Au contact evolution history, i.e. the trace through T_2. For this, we build a convolutional neural network (CNN)-based deep learning model. Deep learning algorithms on conductance traces have been proven effective and to be able to extract information other than basic features like length and mean conductance.[26, 102, 141] By constructing two models with the identical structure to predict the snapback and plateau length, we find that the correlation between prediction and the actual value is 73.1% for snapback, and only 32.4% for plateau length (see Section 4.4.3 for details). This indicates that CNN algorithm also recognizes the weak correlation between the molecular junction plateau length and the Au contact evolution prior to junction formation.

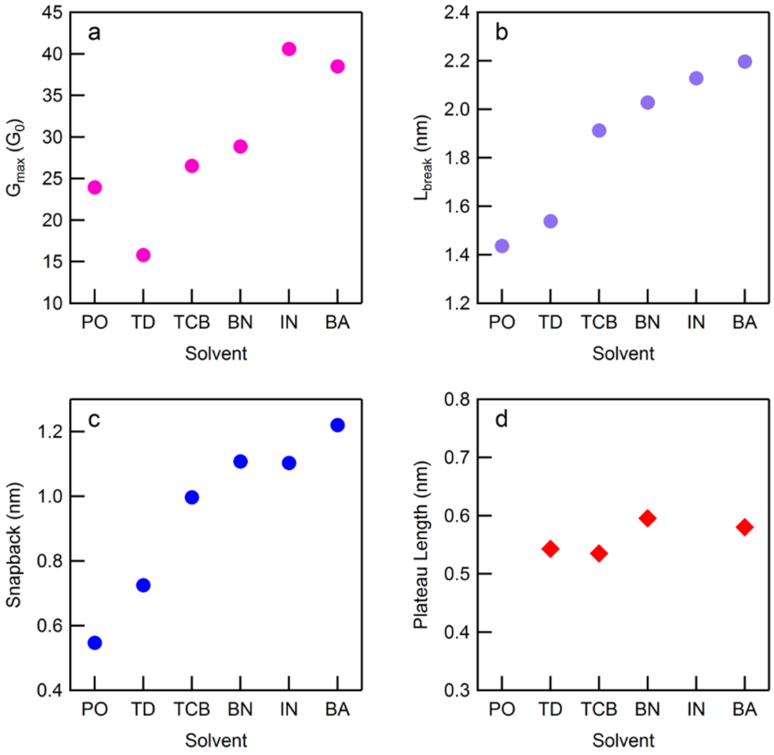


Figure 4.4: The most probable (a) G_{max}, (b) L_{break}, and (c) snapback for 6 different solvents determined from Gaussian fits to histogram data, from pure solvent measurements; (d) The molecular plateau length for 4,4''-diamino-*p*-terphenyl solutions of 4 different solvents. Abbreviations for solvents used are as follows: PO = 1-phenyloctane, TD = tetradecane, TCB = 1,2,4-trichlorobenzene, BN = 1-bromonaphthalene, IN = 1-iodonaphthalene, BA = 4-bromoanisole. See Figure 4.7 and 4.9 for raw data.

We now turn to measurements made in different solvents to understand how the environment affects the elongation and rupture of Au contacts. Figure 4.4a shows G_{max} determined from a series of measurements in different solvents commonly used for STM-BJ measurements, all obtained from commercial sources (see Figure 4.4 caption for list). Figure 4.4b shows L_{break} for these same solvents and Figure 4.4c shows the measured snapback. From Figures 4.4a-c we see that G_{max}, L_{break} and snapback follow the same trends. We find that solvents with a low snapback value, such as phenyloctane (PO) and tetradecane (TD), are those which interact weakly with Au electrodes, and solely through Van der Waals interactions. By contrast, solvents with high snapback values, such as 1-bromonaphthalene (BN), 1-iodonaphthalene (IN) and 4-bromoanisole (BA), interact more strongly with the soft Au atoms through their soft halide group. Solvent-Au binding energy calculations[142] confirm this finding. Since these solvent molecules bind to undercoordinated Au atoms that are pulled out of the surface, they stabilize the newly-formed Au surface, and thus decrease the energy required to elongate the Au contact. This in turn allows for a longer L_{break}. Since we have shown above that L_{break} is positively correlated with the snapback distance, solvents that passivate undercoordinated Au atoms are likely to lead to a longer snapback. In Figure 4.4d we show the measured molecular plateau length of 4,4''-diamino-*p*-terphenyl solution in TD, TCB, BN and BA. We find the plateau lengths are similar across different solvents and do not follow the trend seen with the snapback, likely because the linker-Au interaction is much less affected by the solvent effect. Additionally, this is consistent with our finding that the plateau length is very weakly correlated to the G_{max}, L_{break} and snapback.

For long molecules like 4,4''-diamino-*p*-terphenyl, nearly every Au-contact that is ruptures forms a molecular junction, so we turn to 1,3-diaminopropane to see if the junction formation probability differs in shorter molecules. We repeat the modified break-junction measurement illustrated in Figure 4.1 in a solution of 1,3-diaminopropane in TCB and analyze our data. Figure 4.5a and 4.5b shows the correlation (absolute values) and tree-splitting importance (average information entropy gain of the XGBoost model) of junction formation probability versus G_{max}, L_{break}, L_1, L_2 and B_{Au}. We can see that the junction formation probability is negatively correlated with the L_1 (the length of the $1G_0$ plateau), but is much less correlated with any other parameter. This indicates that junctions with smaller L_1 have a higher chance of forming a molecular junction while those with a longer L_1 are less likely to form a molecular junction. This is consistent with our earlier hypothesis that the molecules bind to the two electrodes in parallel to the gold point-contact. For junctions with short L_1, a pre-existing molecular bridge is likely to survive after the Au point-contact ruptures as illustrated in the upper pathway in Figure 4.5c. For junctions with longer L_1, the molecular bridge is likely to rupture before the Au-Au point contacts ruptures as illustrated in the lower pathway in Figure 4.5b. Together, these findings confirm our hypothesis that molecular junctions form prior to the rupture of the metal-contact in STM-BJ measurements.

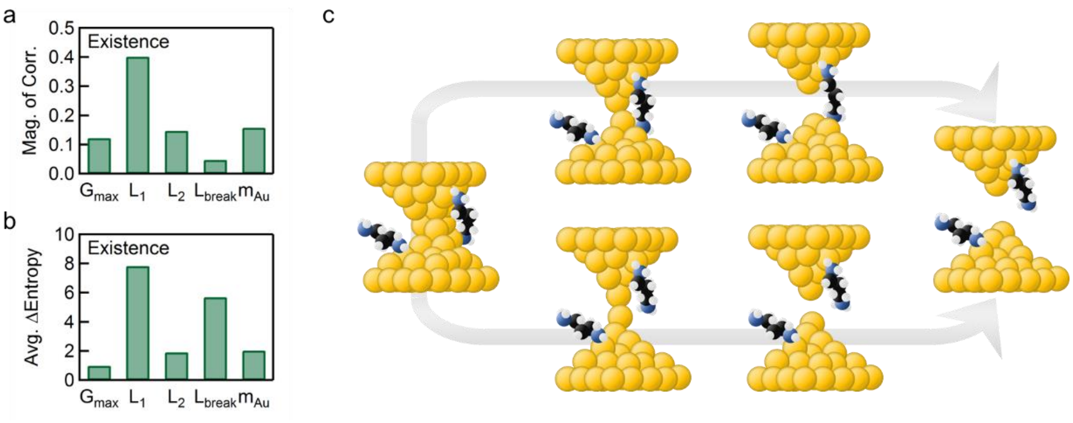


Figure 4.5: (a) The importance of different features in determining the existence of molecular junction (green bars) for the TCB solution of 1,3-propanediamine, as the magnitude of correlation (upper) and the tree-splitting importance according of the XGBoost models (lower). (b) Illustration of the pathway for the rupture of a short (upper) and long (lower) $1G_0$ contact with a short molecule in parallel to the contact. A molecular junction remains only in the upper path with a short Au-Au contact.

4.3 Conclusions

In conclusion, through modified STM-BJ experiments we have shown that the relaxation of Au electrodes after breaking a point contact depends on the environment around the gold electrode. We show that this snapback is mainly dependent on two parameters that describe the Au contact prior to rupture: G_{max} (highest conductance) and L_{break} (displacement until rupture). We however find that the molecular plateau length is only weakly and negatively correlated to the snapback, and is nearly independent on the parameters describing the Au contact. We find that the molecular junction plateau length and the junction formation probability for short molecules is mostly independent on the Au contact structure prior to rupture but we do find that it is negatively correlated with the length of the 1-G_0 plateau. These results indicate that the molecules are bound to the Au electrode before the Au point contact ruptures. A complete relaxation of the electrodes happens only after the molecular junction also ruptures. These findings provide key insights into the versatility of STM-BJ measurements to form and characterize molecular conductance signatures in a ranger of solvents and environments.

4.4 Supplementary Information

4.4.1 Additional Data

Sample trace from a measurement with a molecule.

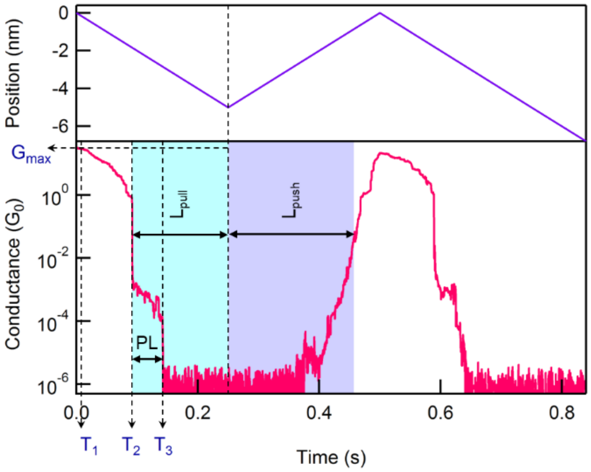


Figure 4.6: Sample piezo ramp (upper panel) and conductance versus time trace (lower panel) for a push-pull trace measured in a 1,2,4-trichlorobenzen (TCB) solution of 4,4″-diamino-*p*-terphenyl at an applied bias of 100 mV and pulling rate of 15 nm/s. T_1 indicates the time when the highest conductance (G_{max}) is observed; T_2 is the time when the initial Au contact ruptures; T_3 is the time when the Au-molecule-Au junction breaks. L_{pull} indicates the distance pulled after the contact breaks and L_{push} indicates the distance pushed before the contact is reformed. PL is the length of molecular junction plateau. For an experiment with a solution of the target molecules, the snapback is determined in the same way as in a measurement with solvent alone, i.e., $L_{push} – L_{pull}$, and the molecular plateau length is defined as the distance between the first and last point of the conductance trace that falls in the molecular conductance region. We only consider traces that have a molecular plateau if the plateau is longer than 0.01 nm.

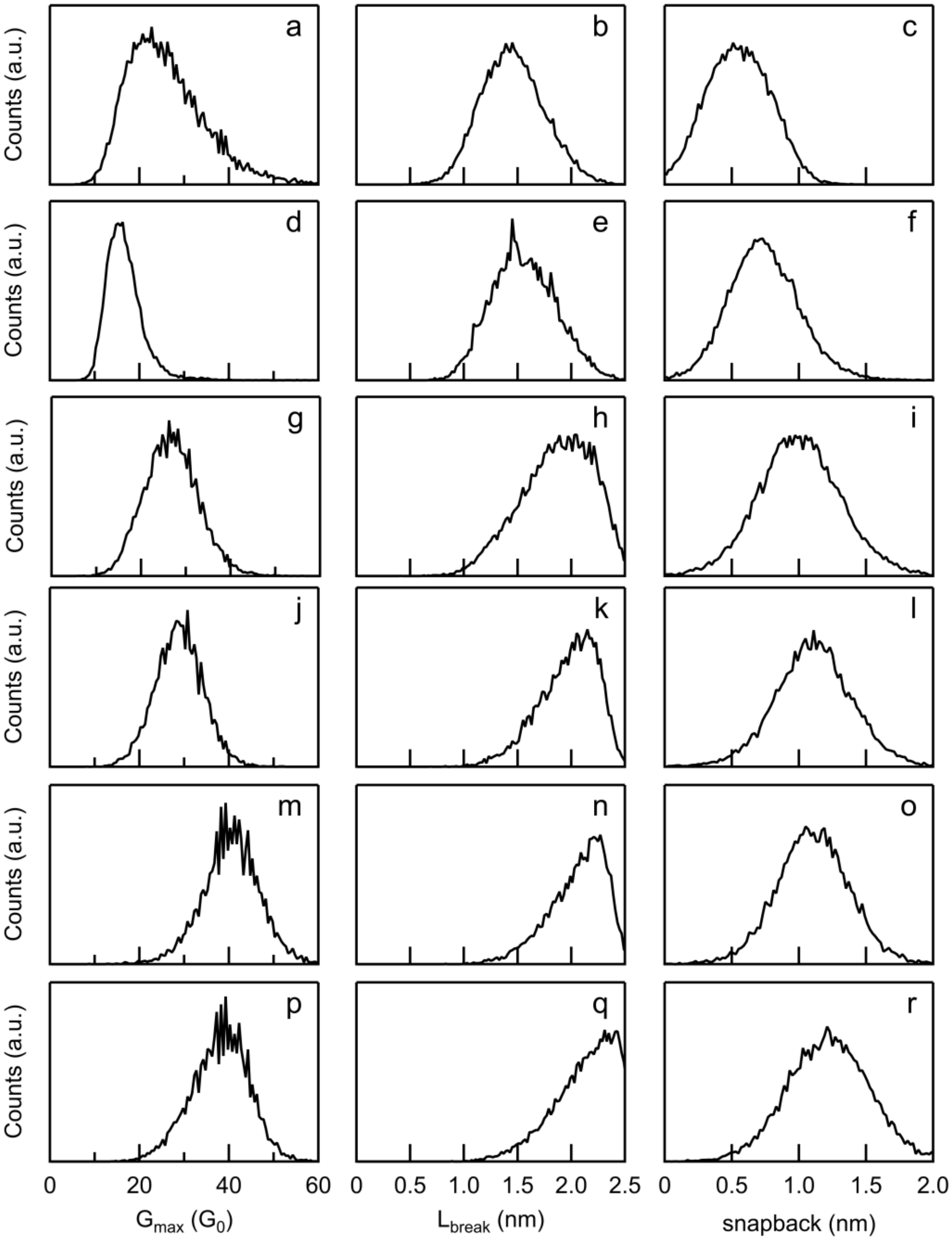


Figure 4.7: Distribution histograms of parameters for pure solvent experiments. From top to bottom: different solvents, (a-c) 1-phenyloctane (PO), (d-f) tetradecane (TD), (g-i) 1,2,4-trichlorobenzene (TCB), (j-l) 1-bromonaphthalene (BN), (m-o) 1-iodonaphthalene (IN) and (p-r) 4-bromoanisole (BA).

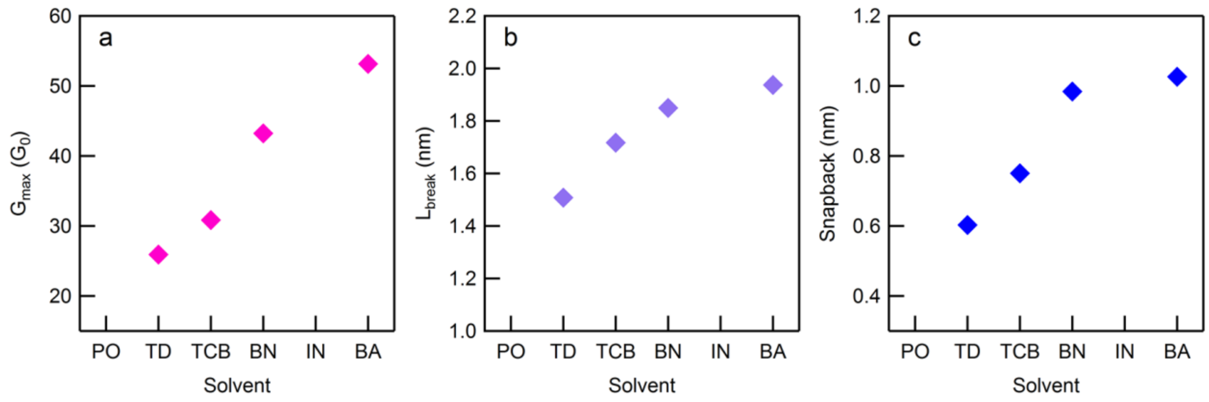


Figure 4.8: (a) The most probable G_{max}, (b) L_{break}, and (c) snapback for 4,4''-diamino-*p*-terphenyl solutions measured in four different solvents. Abbreviations for solvents used are as follows: TD = tetradecane, TCB = 1,2,4-trichlorobenzene, BN = 1-bromonaphthalene, BA = 4-bromoanisole.

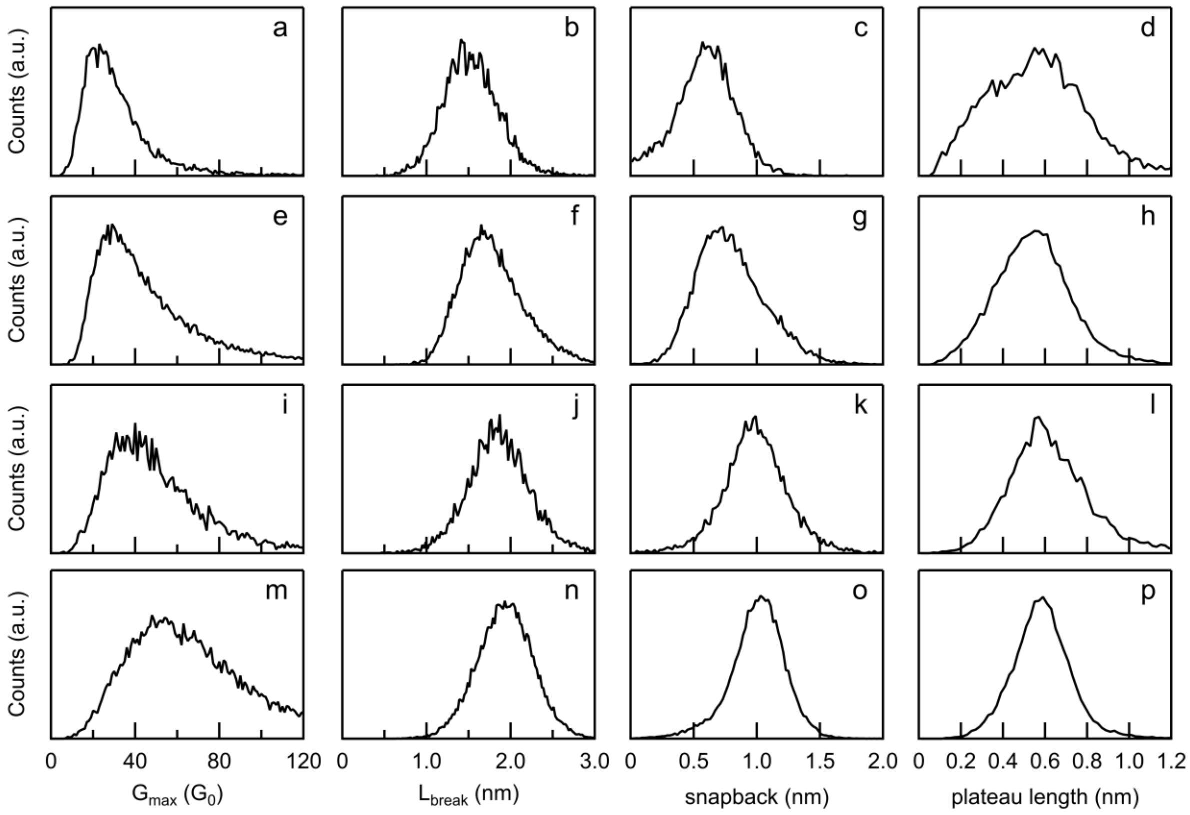


Figure 4.9: Distribution histograms of parameters for experiments of 4,4''-diamino-*p*-terphenyl solutions. From top to bottom: different solvents, (a-d) tetradecane (TD), (e-h) 1,2,4-trichlorobenzene (TCB), (i-l) 1-bromonaphthalene (BN) and (m-p) 4-bromoanisole (BA).

4.4.2 Snapback Experiment Details

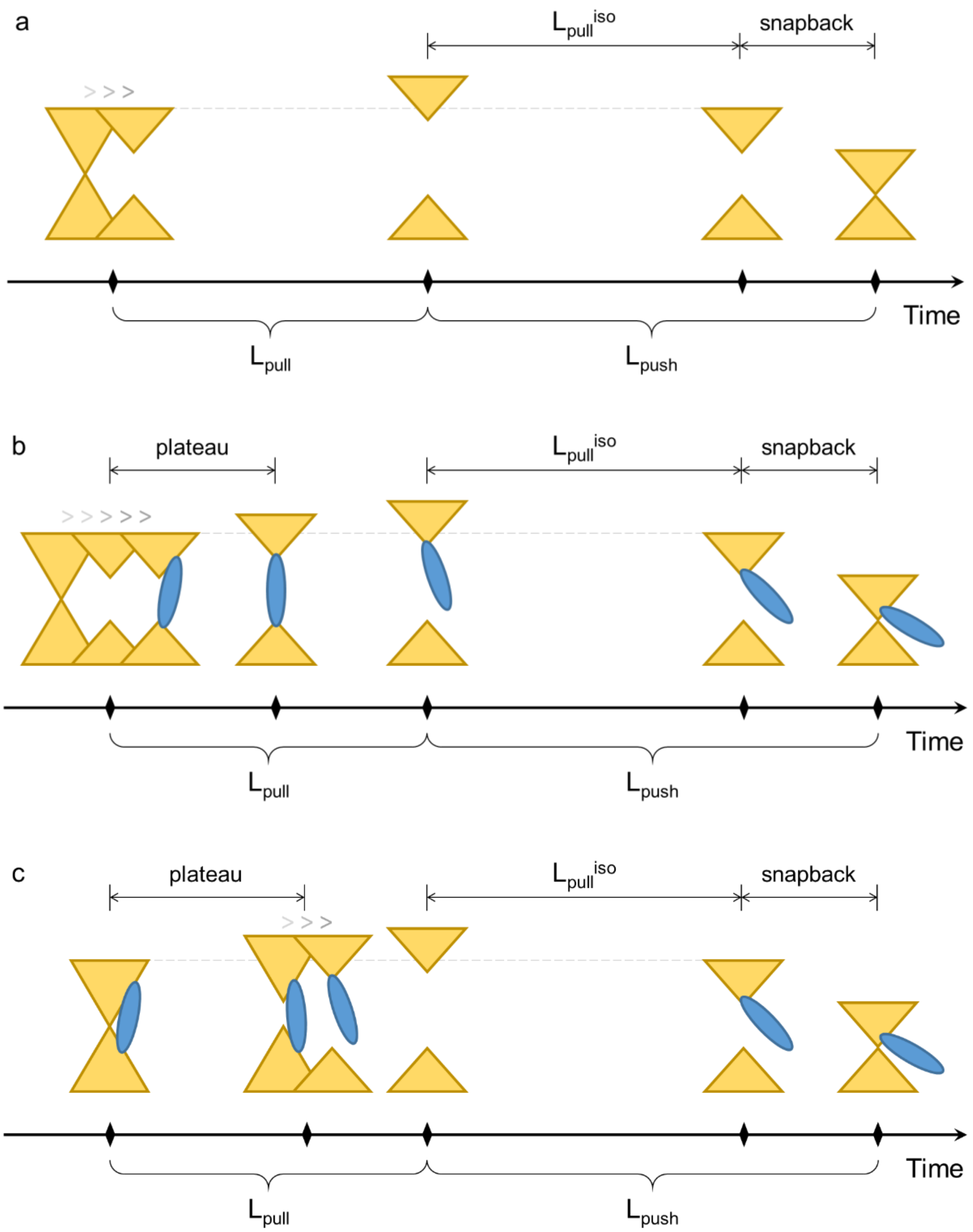


Figure 4.10: Illustration of a *push-pull* STM-BJ experiment in (a) on pure solvent, (b) solvent with a target molecule where the molecule is captured after the rupture of Au-Au contact, and (c) solvent with a target molecule where the molecule bridges the Au electrodes in parallel to the point-contact. The yellow triangles represent Au electrodes, and the blue ovals represent the molecules.

As explained in the main text, we design the modified push-pull STM-BJ experiment to measure the snapback distance of the Au electrodes after the junction ruptures. This distance is $L_{push} – L_{pull}$ from these push-pull measurements and determines the relaxation over a millisecond timescale. Figure 4.10a shows a simplified illustration of this experiment in a solvent, and compliments Figure 4.1.

Going from left to right, Figure 4.10a starts with an Au-Au contact that ruptures then snapbacks back. After the contacts are pulled apart by a distance L_{pull}, we start to push the electrodes towards each other until a contact is formed. The distance the electrodes are pushed together is denoted as L_{push}. L_{push} includes the distance that the electrodes are withdrawn (denoted as L_{pull}^{iso}) and the amount that the electrodes relax, which is the snapback. Note that most of the relaxation occurs as soon as the contacts rupture, while some slower process that reorganize the electrodes also lead to an enlargement of the gap between the electrodes.[28]

For experiments with molecules, the snapback is measured in the same way. In both cases, the threshold for Au-Au contact is chosen to be 0.05 G_0 which is higher than the molecular conductance of all the molecules studied in this work. Hence, the snapback measurement does not interfere with the molecular junction plateau. This process is illustrated in Figure S5b and S5c, where we hypothesize that the snapback happens at the same stage as in the pure solvent experiments.

Figure 4.10b illustrates a path where the molecular junction is formed after the contacts rupture. Here, the snapback occurs immediately after the Au-Au point contact ruptures as in the case of the pure solvent experiments illustrated in Figure 4.10a. A molecule is captured into this gap after the electrodes relax.

Figure 4.10c shows a junction where a molecule is already bound bridging the electrodes in parallel to the Au point-contact, the scenario that we argue here dominates in the measurements. Since the molecule is already present when the point contact breaks, the extent of the snapback is very much decreased since the two electrodes are still held together by the molecule, and thus under some tension. The electrodes relax fully only after the Au-molecule-Au junction finally breaks. We can still measure this snap-back as we do in the case without molecules, as $L_{push} – L_{pull}$.

Irrespective of whether a molecule is present before or after the electrodes relax, the maximum extension of a Au-molecule-Au should be a stochastic value with a molecule-dependent constant mean. For the path detailed in Figure 4.10b, the molecular plateau length should be this maximum extension minus the snapback, because the molecular junction starts with an initial displacement of the snapback. This would result in a strong negative correlation between the plateau length and the snapback, which is not observed in our experiments. For the path shown in Figure 4.10c however, since most of the snapback takes place after the molecular junction breaks, the measured plateau length is still bound by the maximal molecule-dependent value but depends on the details of the junction configuration, not on the snapback.

To rationalize the small negative correlation seen in the experiments between the molecular plateau length and the snapback, this likely indicates the electrodes could relax partially after the Au-Au contact is ruptured. This explanation is reasonable because the strength of the Au-Au bond is greater than that of the linker-Au bond.[38]

4.4.3 Details of the Neural Network Method

We build two convolutional neural network (CNN)-based models that predict snapback and molecular plateau length from the Au conductance region of an STM-BJ conductance trace. Based on the performance of these two models, we can learn the connection between the snapback or plateau length and the Au contact conductance evolution with length.[26, 102, 141] Using this method, we eliminate the errors that could be there if instead of using the entire trace, we simply model the relations using a few key parameters. The drawback of this CNN structure is that it is hard to figure out which characteristics of the traces are important features in determining snapback or plateau lengths.

In these two models, we use as input, the Au conductance region of the raw STM-BJ conductance trace in its initial pulling phase. The other parts of the trace are cut off and set to zero to prevent the model from directly reading out the snapback from the complete trace. Our model consists of 3 convolutional layers followed by a global average pooling and then a fully-connected layer. Each of these convolutional layers is a convolutional-batch normalization[109]-dropout[107] structure, where for the convolutional part, the kernel size = 21 points, the stride = 6 points, and the number of channels is 32. The width of the fully-connected layer is 8. Activation using rectified linear units (ReLU) are applied after each of these layers.

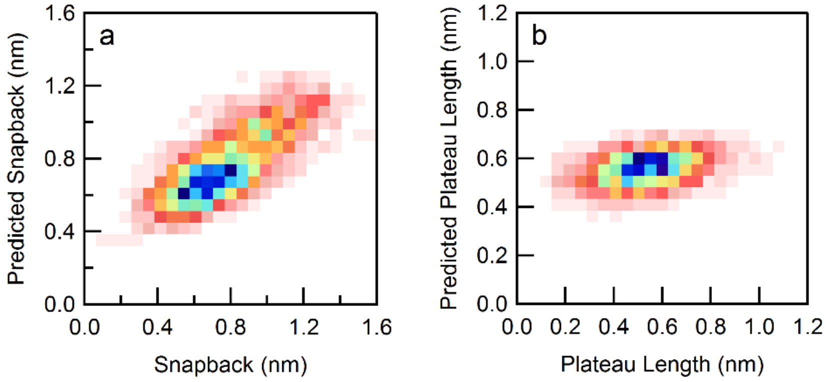


Figure 4.11: The 2D dimensional correlation histograms between the measured and the model-predicted values for (a) snapback and (b) molecular plateau length, according to the CNN-based models.

We study the dataset of 24,880 traces from the experiment in a 1,2,4-trichlorobenzen (TCB) solution of 4,4"-diamino-*p*-terphenyl. We use 90% of these traces to train the models and the other 10% for model validation. Training of these models was fulfilled by TensorFlow.[112]

Using the validation dataset, which is not used during the training process, we obtain a correlation of 73.1% for snapback and 32.4% for the plateau length (Figure S6). The magnitude of this correlation can indicate how much the Au contact evolution history can determine the snapback or plateau length. We see that the snapback is determined by the evolution of the Au contact while the plateau length is not.

References

1. Ho, W., Single-molecule chemistry. *The Journal of Chemical Physics* **2002,** *117* (24), 11033-11061.
2. Moerner, W. E., A Dozen Years of Single-Molecule Spectroscopy in Physics, Chemistry, and Biophysics. *The Journal of Physical Chemistry B* **2002,** *106* (5), 910-927.
3. Tamarat, P.; Maali, A.; Lounis, B.; Orrit, M., Ten Years of Single-Molecule Spectroscopy. *The Journal of Physical Chemistry A* **2000,** *104* (1), 1-16.
4. Barkai, E.; Jung, Y.; Silbey, R., Theory of single-molecule spectroscopy: beyond the ensemble average. *Annu Rev Phys Chem* **2004,** *55*, 457-507.
5. Plakhotnik, T.; Donley, E. A.; Wild, U. P., Single-molecule spectroscopy. *Annu Rev Phys Chem* **1997,** *48*, 181-212.
6. Binnig, G.; Quate, C. F.; Gerber, C., Atomic force microscope. *Phys Rev Lett* **1986,** *56* (9), 930-933.
7. Giessibl, F. J., Atomic resolution of the silicon (111)-(7x7) surface by atomic force microscopy. *Science* **1995,** *267* (5194), 68-71.
8. Giessibl, F. J., Advances in atomic force microscopy. *Reviews of Modern Physics* **2003,** *75* (3), 949-983.
9. Neuman, K. C.; Nagy, A., Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nat Methods* **2008,** *5* (6), 491-505.
10. Rief, M.; Oesterhelt, F.; Heymann, B.; Gaub, H. E., Single Molecule Force Spectroscopy on Polysaccharides by Atomic Force Microscopy. *Science* **1997,** *275* (5304), 1295-7.
11. Binnig, G.; Rohrer, H., Scanning tunneling microscopy. *Surface Science* **1983,** *126* (1-3), 236-244.
12. Moresco, F.; Meyer, G.; Rieder, K. H.; Tang, H.; Gourdon, A.; Joachim, C., Conformational changes of single molecules induced by scanning tunneling microscopy manipulation: a route to molecular switching. *Phys Rev Lett* **2001,** *86* (4), 672-5.
13. Ohtani, H.; Wilson, R. J.; Chiang, S.; Mate, C. M., Scanning tunneling microscopy observations of benzene molecules on the Rh(111)-(3 x 3) (C6H6+2CO) surface. *Phys Rev Lett* **1988,** *60* (23), 2398-2401.
14. Poirier, G. E., Characterization of Organosulfur Molecular Monolayers on Au(111) using Scanning Tunneling Microscopy. *Chem Rev* **1997,** *97* (4), 1117-1128.
15. Repp, J.; Meyer, G.; Stojkovic, S. M.; Gourdon, A.; Joachim, C., Molecules on insulating films: scanning-tunneling microscopy imaging of individual molecular orbitals. *Phys Rev Lett* **2005,** *94* (2), 026803.
16. 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* (3), 458-62.
17. Xu, B.; Tao, N. J., Measurement of single-molecule resistance by repeated formation of molecular junctions. *Science* **2003,** *301* (5637), 1221-3.
18. Fu, T.; Smith, S.; Camarasa-Gomez, M.; Yu, X.; Xue, J.; Nuckolls, C.; Evers, F.; Venkataraman, L.; Wei, S., Enhanced coupling through pi-stacking in imidazole-based molecular junctions. *Chem Sci* **2019,** *10* (43), 9998-10002.
19. Bamberger, N. D.; Ivie, J. A.; Parida, K. N.; McGrath, D. V.; Monti, O. L. A., Unsupervised Segmentation-Based Machine Learning as an Advanced Analysis Tool for Single Molecule Break Junction Data. *The Journal of Physical Chemistry C* **2020,** *124* (33), 18302-18315.

20. 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. *Applied Physics Letters* **2019,** *114* (14).
21. Hamill, J. M.; Zhao, X. T.; Meszaros, 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* (1), 016601.
22. 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.
23. Makk, P.; Tomaszewski, D.; Martinek, J.; Balogh, Z.; Csonka, S.; Wawrzyniak, M.; Frei, M.; Venkataraman, L.; Halbritter, A., Correlation analysis of atomic and single-molecule junction conductance. *ACS Nano* **2012,** *6* (4), 3411-23.
24. Albrecht, T.; Slabaugh, G.; Alonso, E.; Al-Arif, S., Deep learning for single-molecule science. *Nanotechnology* **2017,** *28* (42), 423001.
25. Korol, R.; Segal, D., Machine Learning Prediction of DNA Charge Transport. *J Phys Chem B* **2019,** *123* (13), 2801-2811.
26. Fu, T.; Zang, Y.; Zou, Q.; Nuckolls, C.; Venkataraman, L., Using Deep Learning to Identify Molecular Junction Characteristics. *Nano Lett* **2020,** *20* (5), 3320-3325.
27. 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* (51), 15768-15769.
28. Kim, T.; Vazquez, H.; Hybertsen, M. S.; Venkataraman, L., Conductance of molecular junctions formed with silver electrodes. *Nano Lett* **2013,** *13* (7), 3358-64.
29. Xu, Q.; Scuri, G.; Mathewson, C.; Kim, P.; Nuckolls, C.; Bouilly, D., Single Electron Transistor with Single Aromatic Ring Molecule Covalently Connected to Graphene Nanogaps. *Nano Lett* **2017,** *17* (9), 5335-5341.
30. Guo, X.; Small, J. P.; Klare, J. E.; Wang, Y.; Purewal, M. S.; Tam, I. W.; Hong, B. H.; Caldwell, R.; Huang, L.; O'Brien, S.; Yan, J.; Breslow, R.; Wind, S. J.; Hone, J.; Kim, P.; Nuckolls, C., Covalently bridging gaps in single-walled carbon nanotubes with conducting molecules. *Science* **2006,** *311* (5759), 356-9.
31. 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* (1), 1748.
32. Meisner, J. S.; Ahn, S.; Aradhya, S. V.; Krikorian, M.; Parameswaran, R.; Steigerwald, M.; Venkataraman, L.; Nuckolls, C., Importance of direct metal-pi coupling in electronic transport through conjugated single-molecule junctions. *J Am Chem Soc* **2012,** *134* (50), 20440-5.
33. Reed, M. A., Conductance of a Molecular Junction. *Science* **1997,** *278* (5336), 252-254.
34. Smit, R. H.; Noat, Y.; Untiedt, C.; Lang, N. D.; van Hemert, M. C.; van Ruitenbeek, J. M., Measurement of the conductance of a hydrogen molecule. *Nature* **2002,** *419* (6910), 906-9.
35. Osorio, E. A.; O'Neill, K.; Stuhr-Hansen, N.; Nielsen, O. F.; Bjørnholm, T.; van der Zant, H. S. J., Addition Energies and Vibrational Fine Structure Measured in Electromigrated Single-Molecule Junctions Based on an Oligophenylenevinylene Derivative. *Advanced Materials* **2007,** *19* (2), 281-285.

36. Noguchi, Y.; Nagase, T.; Kubota, T.; Kamikado, T.; Mashiko, S., Fabrication of Au–molecule–Au junctions using electromigration method. *Thin Solid Films* **2006,** *499* (1-2), 90-94.
37. Frei, M.; Aradhya, S. V.; Hybertsen, M. S.; Venkataraman, L., Linker dependent bond rupture force measurements in single-molecule junctions. *J Am Chem Soc* **2012,** *134* (9), 4003-6.
38. Frei, M.; Aradhya, S. V.; Koentopp, M.; Hybertsen, M. S.; Venkataraman, L., Mechanics and chemistry: single molecule bond rupture forces correlate with molecular backbone structure. *Nano Lett* **2011,** *11* (4), 1518-23.
39. Fung, E. D.; Venkataraman, L., Too Cool for Blackbody Radiation: Overbias Photon Emission in Ambient STM Due to Multielectron Processes. *Nano Lett* **2020,** *20* (12), 8912-8918.
40. Fung, E. D.; Adak, O.; Lovat, G.; Scarabelli, D.; Venkataraman, L., Too Hot for Photon-Assisted Transport: Hot-Electrons Dominate Conductance Enhancement in Illuminated Single-Molecule Junctions. *Nano Lett* **2017,** *17* (2), 1255-1261.
41. Aradhya, S. V.; Venkataraman, L., Single-molecule junctions beyond electronic transport. *Nat Nanotechnol* **2013,** *8* (6), 399-410.
42. Huang, C.; Rudnev, A. V.; Hong, W.; Wandlowski, T., Break junction under electrochemical gating: testbed for single-molecule electronics. *Chem Soc Rev* **2015,** *44* (4), 889-901.
43. Rincon-Garcia, L.; Evangeli, C.; Rubio-Bollinger, G.; Agrait, N., Thermopower measurements in molecular junctions. *Chem Soc Rev* **2016,** *45* (15), 4285-306.
44. Gehring, P.; Thijssen, J. M.; van der Zant, H. S. J., Single-molecule quantum-transport phenomena in break junctions. *Nature Reviews Physics* **2019,** *1* (6), 381-396.
45. Moreno-Garcia, 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* (33), 12228-40.
46. Reddy, P.; Jang, S. Y.; Segalman, R. A.; Majumdar, A., Thermoelectricity in molecular junctions. *Science* **2007,** *315* (5818), 1568-71.
47. Aragones, A. C.; Aravena, D.; Valverde-Munoz, F. J.; Real, J. A.; Sanz, F.; Diez-Perez, I.; Ruiz, E., Metal-Controlled Magnetoresistance at Room Temperature in Single-Molecule Devices. *J Am Chem Soc* **2017,** *139* (16), 5768-5778.
48. Chang, W. B.; Mai, C.-K.; Kotiuga, M.; Neaton, J. B.; Bazan, G. C.; Segalman, R. A., Controlling the Thermoelectric Properties of Thiophene-Derived Single-Molecule Junctions. *Chemistry of Materials* **2014,** *26* (24), 7229-7235.
49. Li, J. J.; Bai, M. L.; Chen, Z. B.; Zhou, X. S.; Shi, Z.; Zhang, M.; Ding, S. Y.; Hou, S. M.; Schwarzacher, W.; Nichols, R. J.; Mao, B. W., Giant single-molecule anisotropic magnetoresistance at room temperature. *J Am Chem Soc* **2015,** *137* (18), 5923-9.
50. Zhou, J.; Wang, K.; Xu, B.; Dubi, Y., Photoconductance from Exciton Binding in Molecular Junctions. *J Am Chem Soc* **2018,** *140* (1), 70-73.
51. Aragones, A. C.; Haworth, N. L.; Darwish, N.; Ciampi, S.; Bloomfield, N. J.; Wallace, G. G.; Diez-Perez, I.; Coote, M. L., Electrostatic catalysis of a Diels-Alder reaction. *Nature* **2016,** *531* (7592), 88-91.
52. Huang, X.; Tang, C.; Li, J.; Chen, L. C.; Zheng, J.; Zhang, P.; Le, J.; Li, R.; Li, X.; Liu, J.; Yang, Y.; Shi, J.; Chen, Z.; Bai, M.; Zhang, H. L.; Xia, H.; Cheng, J.; Tian, Z. Q.; Hong, W.,

Electric field-induced selective catalysis of single-molecule reaction. *Sci Adv* **2019,** *5* (6), eaaw3072.

53. Zang, Y.; Pinkard, A.; Liu, Z. F.; Neaton, J. B.; Steigerwald, M. L.; Roy, X.; Venkataraman, L., Electronically Transparent Au-N Bonds for Molecular Junctions. *J Am Chem Soc* **2017,** *139* (42), 14845-14848.
54. 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), 4482.
55. Starr, R. L.; Fu, T.; Doud, E. A.; Stone, I.; Roy, X.; Venkataraman, L., Gold-Carbon Contacts from Oxidative Addition of Aryl Iodides. *J Am Chem Soc* **2020,** *142* (15), 7128-7133.
56. Zang, Y.; Stone, I.; Inkpen, M. S.; Ng, F.; Lambert, T. H.; Nuckolls, C.; Steigerwald, M. L.; Roy, X.; Venkataraman, L., In Situ Coupling of Single Molecules Driven by Gold-Catalyzed Electrooxidation. *Angew Chem Int Ed Engl* **2019,** *58* (45), 16008-16012.
57. Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J., *Classification and Regression Trees*. Wadsworth and Brooks: Monterey, CA, 1984.
58. Quinlan, J. R., Induction of decision trees. *Machine Learning* **1986,** *1* (1), 81-106.
59. Breiman, L., Random Forests. *Machine Learning* **2001,** *45* (1), 5-32.
60. Chen, T.; Guestrin, C., XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery: San Francisco, California, USA, 2016; pp 785-794.
61. Liaw, A.; Wiener, M., Classification and Regression by RandomForest. *Forest* **2001,** *23*.
62. Ng, A. Y.; Jordan, M. I.; Weiss, Y., On spectral clustering: analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic,* MIT Press: Vancouver, British Columbia, Canada, 2001; p 849вЂ“856.
63. Ester, M.; Kriegel, H.-P.; Sander, J. r.; Xu, X., 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,* AAAI Press: Portland, Oregon, 1996; p 226вЂ“231.
64. Kramer, M. A., Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal* **1991,** *37* (2), 233-243.
65. Rawat, W.; Wang, Z., Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Comput* **2017,** *29* (9), 2352-2449.
66. Antonyuk, S. V.; Strange, R. W.; Marklund, S. L.; Hasnain, S. S., The structure of human extracellular copper-zinc superoxide dismutase at 1.7 A resolution: insights into heparin and collagen binding. *J. Mol. Biol.* **2009,** *388* (2), 310-26.
67. Tainer, J. A.; Getzoff, E. D.; Richardson, J. S.; Richardson, D. C., Structure and mechanism of copper, zinc superoxide dismutase. *Nature* **1983,** *306* (5940), 284-7.
68. Polgar, L., The catalytic triad of serine peptidases. *Cell Mol. Life Sci.* **2005,** *62* (19-20), 2161-72.
69. Park, K. S.; Ni, Z.; Cote, A. P.; Choi, J. Y.; Huang, R.; Uribe-Romo, F. J.; Chae, H. K.; O'Keeffe, M.; Yaghi, O. M., Exceptional chemical and thermal stability of zeolitic imidazolate frameworks. *Proc Natl Acad Sci U S A* **2006,** *103* (27), 10186-10191.
70. 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* (3), 458 - 462.

71. Gonzalez, M. T.; Wu, S.; Huber, R.; van der Molen, S. J.; Schonenberger, C.; Calame, M., Electrical conductance of molecular junctions by a robust statistical analysis. *Nano Lett* **2006,** *6* (10), 2238-42.
72. Wu, S.; Gonzalez, M. T.; Huber, R.; Grunder, S.; Mayor, M.; Schonenberger, C.; Calame, M., Molecular junctions based on aromatic coupling. *Nat Nanotechnol* **2008,** *3* (9), 569-74.
73. Martin, 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* (26), 9157-64.
74. Yoshida, K.; Pobelov, I. V.; Manrique, D. Z.; Pope, T.; Meszaros, G.; Gulcur, M.; Bryce, M. R.; Lambert, C. J.; Wandlowski, T., Correlation of breaking forces, conductances and geometries of molecular junctions. *Sci Rep* **2015,** *5*, 9002.
75. González, M. T.; Leary, E.; García, R.; Verma, P.; Herranz, M. Á.; Rubio-Bollinger, G.; Martín, N.; Agraït, N., Break-Junction Experiments on Acetyl-Protected Conjugated Dithiols under Different Environmental Conditions. *The Journal of Physical Chemistry C* **2011,** *115* (36), 17973-17978.
76. Hong, W.; Valkenier, H.; Mészáros, G.; Manrique, D. Z.; Mishchenko, A.; Putz, A.; García, P. M.; Lambert, C. J.; Hummelen, J. C.; Wandlowski, T., An MCBJ case study: The influence of π-conjugation on the single-molecule conductance at a solid/liquid interface. *Beilstein J. Nanotechnol.* **2011,** *2*, 699-713.
77. Zheng, J.-T.; Yan, R.-W.; Tian, J.-H.; Liu, J.-Y.; Pei, L.-Q.; Wu, D.-Y.; Dai, K.; Yang, Y.; Jin, S.; Hong, W.; Tian, Z.-Q., Electrochemically assisted mechanically controllable break junction studies on the stacking configurations of oligo(phenylene ethynylene)s molecular junctions. *Electrochimica Acta* **2016,** *200*, 268-275.
78. Borges, A.; Fung, E. D.; Ng, F.; Venkataraman, L.; Solomon, G. C., Probing the Conductance of the σ-System of Bipyridine Using Destructive Interference. *Journal of Physical Chemistry Letters* **2016,** *7* (23), 4825-4829.
79. Kamenetska, M.; Quek, S. Y.; Whalley, A. C.; Steigerwald, M. L.; Choi, H. J.; Louie, S. G.; Nuckolls, C.; Hybertsen, M. S.; Neaton, J. B.; Venkataraman, L., Conductance and geometry of pyridine-linked single-molecule junctions. *J Am Chem Soc* **2010,** *132* (19), 6817-21.
80. Kim, T.; Darancet, P.; Widawsky, J. R.; Kotiuga, M.; Quek, S. Y.; Neaton, J. B.; Venkataraman, L., Determination of energy level alignment and coupling strength in 4,4'-bipyridine single-molecule junctions. *Nano. Lett.* **2014,** *14* (2), 794-8.
81. 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* (4), 230-234.
82. Li, C.; Pobelov, I.; Wandlowski, T.; Bagrets, A.; Arnold, A.; Evers, F., Charge transport in single Au vertical bar alkanedithiol vertical bar Au junctions: Coordination geometries and conformational degrees of freedom. *J Am Chem Soc* **2008,** *130* (1), 318-326.
83. Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M., Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications* **2009,** *180* (11), 2175-2196.
84. Havu, V.; Blum, V.; Havu, P.; Scheffler, M., Efficient integration for all-electron electronic structure calculation using numeric basis functions. *J. Comp. Phys.* **2009,** *228* (22), 8367-8379.
85. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996,** *77* (18), 3865-3868.

86. Arnold, A.; Weigend, F.; Evers, F., Quantum chemistry calculations for molecules coupled to reservoirs: Formalism, implementation, and application to benzenedithiol. *J. Chem. Phys.* **2007,** *126* (17), 174101.

87. Bagrets, A., Spin-Polarized Electron Transport Across Metal-Organic Molecules: A Density Functional Theory Approach. *J Chem Theory Comput* **2013,** *9* (6), 2801-15.

88. Momma, K.; Izumi, F., VESTA 3for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Cryst.* **2011,** *44* (6), 1272-1276.

89. Koentopp, M.; Burke, K.; Evers, F., Zero-bias molecular electronics: Exchange-correlation corrections to Landauer's formula. *Phys. Rev. B* **2006,** *73* (12), 121403.

90. Magyarkuti, A.; Adak, O.; Halbritter, A.; Venkataraman, L., Electronic and mechanical characteristics of stacked dimer molecular junctions. *Nanoscale* **2018,** *10* (7), 3362-3368.

91. Adak, O.; Rosenthal, E.; Meisner, J.; Andrade, E. F.; Pasupathy, A. N.; Nuckolls, C.; Hybertsen, M. S.; Venkataraman, L., Flicker Noise as a Probe of Electronic Interaction at Metal-Single Molecule Interfaces. *Nano Lett.* **2015,** *15* (6), 4143-4149.

92. Tkatchenko, A.; Scheffler, M., Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009,** *102* (7), 73005-73005.

93. Frisenda, R.; Janssen, V. A.; Grozema, F. C.; van der Zant, H. S.; Renaud, N., Mechanically controlled quantum interference in individual pi-stacked dimers. *Nat. Chem.* **2016,** *8* (12), 1099-1104.

94. McKie, R.; Murphy, J. A.; Park, S. R.; Spicer, M. D.; Zhou, S. Z., Homoleptic crown N-heterocyclic carbene complexes. *Angew Chem Int Ed Engl* **2007,** *46* (34), 6525-8.

95. Ding, L.; Wang, S.; Liu, Y.; Cao, J.; Fang, Y., Bispyrene/surfactant assemblies as fluorescent sensor platform: detection and identification of Cu2+ and Co2+ in aqueous solution. *Journal of Materials Chemistry A* **2013,** *1* (31).

96. Fang, Y.; Liu, W.; Teat, S. J.; Dey, G.; Shen, Z.; An, L.; Yu, D.; Wang, L.; O'Carroll, D. M.; Li, J., A Systematic Approach to Achieving High Performance Hybrid Lighting Phosphors with Excellent Thermal- and Photostability. *Advanced Functional Materials* **2017,** *27* (3).

97. Wang, S.; Ding, L.; Fan, J.; Wang, Z.; Fang, Y., Bispyrene/surfactant-assembly-based fluorescent sensor array for discriminating lanthanide ions in aqueous solution. *ACS Appl Mater Interfaces* **2014,** *6* (18), 16156-65.

98. Widawsky, J. R.; Darancet, P.; Neaton, J. B.; Venkataraman, L., Simultaneous determination of conductance and thermopower of single molecule junctions. *Nano Lett* **2012,** *12* (1), 354-8.

99. Adak, O.; Rosenthal, E.; Meisner, J.; Andrade, E. F.; Pasupathy, A. N.; Nuckolls, C.; Hybertsen, M. S.; Venkataraman, L., Flicker Noise as a Probe of Electronic Interaction at Metal-Single Molecule Interfaces. *Nano Lett* **2015,** *15* (6), 4143-9.

100. 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* (31), 9971-81.

101. Magyarkuti, A.; Balogh, N.; Balogh, Z.; Venkataraman, L.; Halbritter, A. Unsupervised feature recognition in single molecule break junction data *arXiv e-prints* [Online], 2020. https://ui.adsabs.harvard.edu/abs/2020arXiv200103006M (accessed January 01, 2020).

102. 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* (8), 084111.

103. 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**.

104. Arthur, D.; Vassilvitskii, S., k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics: New Orleans, Louisiana, 2007; pp 1027-1035.

105. Lloyd, S., Least squares quantization in PCM. *IEEE Transactions on Information Theory* **1982,** *28* (2), 129-137.

106. Nair, V.; Hinton, G. E., Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Omnipress: Haifa, Israel, 2010; pp 807-814.

107. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R., Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014,** *15* (1), 1929-1958.

108. Chen, Y.; Fan, H.; Xu, B.; Yan, Z.; Kalantidis, Y.; Rohrbach, M.; Yan, S.; Feng, J. Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks with Octave Convolution *arXiv e-prints* [Online], 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190405049C (accessed April 01, 2019).

109. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift *arXiv e-prints* [Online], 2015. https://ui.adsabs.harvard.edu/abs/2015arXiv150203167I (accessed February 01, 2015).

110. Aradhya, S. V.; Frei, M.; Halbritter, A.; Venkataraman, L., Correlating structure, conductance, and mechanics of silver atomic-scale contacts. *ACS Nano* **2013,** *7* (4), 3706-12.

111. 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* (7105), 904-7.

112. Abadi, M. n.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mane, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viegas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems *arXiv e-prints* [Online], 2016. https://ui.adsabs.harvard.edu/abs/2016arXiv160304467A (accessed March 01, 2016).

113. Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization *arXiv e-prints* [Online], 2014. https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6980K (accessed December 01, 2014).

114. Krogh, A.; Hertz, J. A., A simple weight decay can improve generalization. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, Morgan Kaufmann Publishers Inc.: Denver, Colorado, 1991; pp 950-957.

115. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization *arXiv e-prints* [Online], 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv171105101L (accessed November 01, 2017).

116. Brisendine, J. M.; Refaely-Abramson, S.; Liu, Z. F.; Cui, J.; Ng, F.; Neaton, J. B.; Koder, R. L.; Venkataraman, L., Probing Charge Transport through Peptide Bonds. *J Phys Chem Lett* **2018,** *9* (4), 763-767.

117. Milan, D. C.; Krempe, M.; Ismael, A. K.; Movsisyan, L. D.; Franz, M.; Grace, I.; Brooke, R. J.; Schwarzacher, W.; Higgins, S. J.; Anderson, H. L.; Lambert, C. J.; Tykwinski, R. R.; Nichols, R. J., The single-molecule electrical conductance of a rotaxane-hexayne supramolecular assembly. *Nanoscale* **2017,** *9* (1), 355-361.

118. Wang, L.; Gong, Z. L.; Li, S. Y.; Hong, W.; Zhong, Y. W.; Wang, D.; Wan, L. J., Molecular Conductance through a Quadruple-Hydrogen-Bond-Bridged Supramolecular Junction. *Angew Chem Int Ed Engl* **2016,** *55* (40), 12393-7.

119. Su, T. A.; Neupane, M.; Steigerwald, M. L.; Venkataraman, L.; Nuckolls, C., Chemical principles of single-molecule electronics. *Nature Reviews Materials* **2016,** *1* (3).

120. Xiang, L.; Palma, J. L.; Bruot, C.; Mujica, V.; Ratner, M. A.; Tao, N., Intermediate tunnelling-hopping regime in DNA charge transport. *Nat Chem* **2015,** *7* (3), 221-6.

121. 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* (10).

122. Kamenetska, M.; Koentopp, M.; Whalley, A. C.; Park, Y. S.; Steigerwald, M. L.; Nuckolls, C.; Hybertsen, M. S.; Venkataraman, L., Formation and evolution of single-molecule junctions. *Phys Rev Lett* **2009,** *102* (12), 126803.

123. Quek, S. Y.; Venkataraman, L.; Choi, H. J.; Louie, S. G.; Hybertsen, M. S.; Neaton, J. B., Amine-gold linked single-molecule circuits: experiment and theory. *Nano Lett* **2007,** *7* (11), 3477-82.

124. Huisman, E. H.; Trouwborst, M. L.; Bakker, F. L.; de Boer, B.; van Wees, B. J.; van der Molen, S. J., Stabilizing single atom contacts by molecular bridge formation. *Nano Lett* **2008,** *8* (10), 3381-5.

125. Kaliginedi, V.; Rudnev, A. V.; Moreno-Garcia, P.; Baghernejad, M.; Huang, C.; Hong, W.; Wandlowski, T., Promising anchoring groups for single-molecule conductance measurements. *Phys Chem Chem Phys* **2014,** *16* (43), 23529-39.

126. 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* (51), 15768-9.

127. Hong, W.; Manrique, D. Z.; Moreno-Garcia, 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* (4), 2292-304.

128. 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* (6704), 783-785.

129. McNeely, J.; Miller, N.; Pan, X.; Lawson, B.; Kamenetska, M., Angstrom-Scale Ruler Using Single Molecule Conductance Signatures. *The Journal of Physical Chemistry C* **2020,** *124* (24), 13427-13433.

130. French, W. R.; Iacovella, C. R.; Rungger, I.; Souza, A. M.; Sanvito, S.; Cummings, P. T., Atomistic simulations of highly conductive molecular transport junctions under realistic conditions. *Nanoscale* **2013,** *5* (9), 3654-9.

131. Wang, H.; Leng, Y., Gold/Benzenedithiolate/Gold Molecular Junction: A Driven Dynamics Simulation on Structural Evolution and Breaking Force under Pulling. *The Journal of Physical Chemistry C* **2015,** *119* (27), 15216-15223.
132. Friedman, J. H., Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **2001,** *29* (5).
133. Friedman, J. H.; Meulman, J. J., Multiple additive regression trees with application in epidemiology. *Stat Med* **2003,** *22* (9), 1365-81.
134. He, X.; Pan, J.; Jin, O.; Xu, T.; Liu, B.; Xu, T.; Shi, Y.; Atallah, A.; Herbrich, R.; Bowers, S.; Candela, J. Q., Practical Lessons from Predicting Clicks on Ads at Facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, Association for Computing Machinery: New York, NY, USA, 2014; pp 1-9.
135. Pal, M.; Charan, T. B.; Poriya, A., K-nearest neighbour-based feature selection using hyperspectral data. *Remote Sensing Letters* **2020,** *12* (2), 132-141.
136. Xu, Z.; Huang, G.; Weinberger, K. Q.; Zheng, A. X., Gradient boosted feature selection. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, Association for Computing Machinery: New York, New York, USA, 2014; pp 522-531.
137. Altmann, A.; Tolosi, L.; Sander, O.; Lengauer, T., Permutation importance: a corrected feature importance measure. *Bioinformatics* **2010,** *26* (10), 1340-7.
138. Albanese, D.; Filosi, M.; Visintainer, R.; Riccadonna, S.; Jurman, G.; Furlanello, C., Minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics* **2013,** *29* (3), 407-8.
139. Kinney, J. B.; Atwal, G. S., Equitability, mutual information, and the maximal information coefficient. *Proc Natl Acad Sci U S A* **2014,** *111* (9), 3354-9.
140. Reshef, D. N.; Reshef, Y. A.; Finucane, H. K.; Grossman, S. R.; McVean, G.; Turnbaugh, P. J.; Lander, E. S.; Mitzenmacher, M.; Sabeti, P. C., Detecting novel associations in large data sets. *Science* **2011,** *334* (6062), 1518-24.
141. 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* (3), 1674-1681.
142. Fatemi, V.; Kamenetska, M.; Neaton, J. B.; Venkataraman, L., Environmental control of single-molecule junction transport. *Nano Lett* **2011,** *11* (5), 1988-92.