Learning Conductance: Gaussian Process Regression for Molecular Electronics

Michael Deffner,[*,†,§] Marc Philipp Weise,[†] Haitao Zhang,[†] Maike Mücke,[‡] Jonny Proppe,[‡] Ignacio Franco,[¶] and Carmen Herrmann[*,†,§]

†*Institute of Inorganic and Applied Chemistry, University of Hamburg, Hamburg, Germany*

‡*Institute of Physical Chemistry, Georg-August University, Göttingen, Germany*

¶*Departments of Chemistry and Physics, University of Rochester, Rochester, New York 14627-0216, USA*

§*The Hamburg Centre for Ultrafast Imaging, Hamburg, Germany*

E-mail: michael.deffner@chemie.uni-hamburg.de; carmen.herrmann@chemie.uni-hamburg.de

Abstract

Experimental studies of charge transport through single molecules often rely on break junction setups, where molecular junctions are repeatedly formed and broken while measuring the conductance, leading to a statistical distribution of conductance values. Modeling this experimental situation and the resulting conductance histograms is challenging for theoretical methods, as computations need to capture structural changes in experiments, including the statistics of junction formation and rupture. This type of extensive structural sampling implies that even when evaluating conductance from computationally efficient electronic structure methods, which typically are of reduced accuracy, the evaluation of conductance histograms is too expensive to be a routine task. Highly accurate quantum transport computations are only computationally feasible for a few selected conformations and thus necessarily ignore the

rich conformational space probed in experiments. To overcome these limitations, we investigate the potential of machine learning for modeling conductance histograms, in particular by Gaussian process regression. We show that by selecting specific structural parameters as features, Gaussian process regression can be used to efficiently predict the zero-bias conductance from molecular structures, reducing the computational cost of simulating conductance histograms by an order of magnitude. This enables the efficient calculation of conductance histograms even on the basis of computationally expensive first-principles approaches by effectively reducing the number of necessary charge transport calculations, paving the way towards their routine evaluation.

1 Introduction

In molecular electronics a single molecule bridges the gap between two metallic electrodes. Understanding and exploiting the unique electron transport properties of these molecular junctions offers insights into fundamental physical processes such as quantum interference[1–3] or the behavior of molecules under non-equilibrium conditions[4], helps to improve the performance of solar cells[5–10] and enables new approaches to designing molecular-based devices such as sensors[11–13]. Another intriguing idea is to exploit the spin degree of freedom of molecules to overcome current challenges in the semiconductor industry, such as heat dissipation[14–17]. Studying these systems has implications not only for molecules as electronic building blocks[18,19], but also for the fields of colloidal nanoparticles and nanoparticle arrays[20–23], electrochemistry[24–27] or electrocatalysis[28].

After first discussions dating back to the 1950s, the proposal of a diode based on a single molecule demonstrated the potential of the field of molecular electronics[29,30]. With experimental techniques such as scanning tunneling microscopy or mechanically-controlled break junction (MCBJ) setups, measurements of the charge transport through single molecules, connected by two macroscopic electrodes, have become widely accessible[31]. In break junc-

tion experiments[32,33], nanoscopic electrodes are formed by pulling and eventually breaking a thin gold wire, or by crashing and retracting the STM tip into and from the substrate. When performing this in a solution of molecules of interest (or having molecules deposited on the electrodes beforehand), these molecules can bridge the gap between the two electrodes forming a molecular junction. Once the junction is formed, a bias voltage is applied and the current is measured as the junction is elongated, leading to a so-called conductance trace. Eventually, the junction breaks and the process is repeated 1000s of times to gather a statistically significant data set.

Each individual conductance trace is distinct, as the electrode structures, molecule–electrode binding and molecular conformation vary and fluctuate in and in between experiments[34–36]. Usually, these traces are reported in a conductance histogram, from which the most probable conductance of the molecule in the junction can be identified. These histograms are commonly broad, as the conductance of a molecular junction can vary over several orders of magnitude. The shape of the conductance histogram can potentially be used to identify different configurations of the junction, gain information about the tunneling process or to unveil cooperative effects[37–41].

The shape of the conductance histograms cannot be obtained from calculations of a static junction in a minimum energy conformation, as these calculations do not take into account the conformational variability encountered in experiments. To capture the histograms, it is necessary to perform molecular dynamics (MD) simulations of junction formation and evolution using techniques such as classical force fields[42–46], reactive force fields[47,48], or ab-initio MD techniques[49–54]. Even for such simulations, complete sampling of the experimental situation remains challenging[55]. This is partly because of the significant increase in the number of necessary conductance calculations compared to the static case. In the coherent tunneling regime, such electron transport calculations are usually performed based on electronic structure calculations such as the Landauer approach and the non-equilibrium Green's function formalism[56–58]. Thus, to obtain meaningful histograms it is necessary to perform conduc-

tance calculations for several hundred to thousand snapshots and for relatively large systems, since parts of the gold electrodes have to be included in the calculations in order to correctly describe their interactions with the molecules under study. This comes with a significant computational cost, so, for simulating histograms, one usually has to rely on cheaper and simpler methods of limited accuracy to obtain information about the electron transport, such as (extended) Hückel-based calculations[48,59].

Recently, machine learning (ML) approaches have become an alternative to traditional quantum chemical calculations, which can bring down computational cost by an order of magnitude or more[60–69]. Examples involve the generation of force fields for MD simulations[70–75], prediction of charge transfer integrals[76,77] or even trying to directly solve the Schrödinger equation[78]. While neural networks are attractive in situations, where large datasets are available for training, other methods such as Gaussian process regression (GPR) or kernel ridge regression (KRR) can cope with smaller data sets[61] and have been applied successfully to predict, e.g., interatomic potentials[79]. Others have shown the applicability of GPR for studies of molecular vibrations[80,81] or molecular structure optimization[82–84], the improvement of dispersion corrections[85], and some of us have lately applied GPR to predict exchange spin couplings in transition metal complexes[86]. A valuable feature of GPR is the straightforward accessibility of expected errors on the predictions.

The efficiency of ML methods can be exploited in two ways: To accelerate the computation of conductance histograms with a given (low-accuracy) conductance method, or to enable the construction of conductance histograms with more sophisticated (yet more expensive) approaches to conductance, which so far were reserved for calculations on individual molecular junction structures (such as GW[87]). Here, we focus on the first aspect, since we need the full "traditionally evaluated" conductance histograms as references. Related ML approaches have recently been proposed, which focus on transport through model systems (e.g. representing DNA)[88–90] or atomic wires[91]. The latter demonstrates the application of a neural network in combination with the smooth overlap of atomic positions (SOAP)[92]

descriptor (among others) to encode the structural information for the prediction of conductance values for atomic wires. Such systems are simpler than molecular ones, as the conductance is a multiple of the quantum of conductance $G_0 = \frac{2e^2}{h}$, with e as the elementary charge and h as the Planck constant. Topolnicki *et al.*[93] employ a neural network to predict the conductance of a biphenyl dithiol junction. Their neural network is trained with structural parameters as well as with parameters obtained from electronic structure calculations, with its performance demonstrated by predicting the change of conductance histograms with temperature.

We show that Gaussian process regression can predict the transport properties of a molecular junction and can be used to reliably construct conductance histograms from simulated break junction experiments, yielding a speed-up by one order of magnitude. With problem-tailored descriptors, advantages in speed and performance can be achieved compared with selected general-purpose descriptors. We aim for a method-agnostic approach concerning the MD and transport calculations, which can be applied to small and medium sized data sets, including situations in which the generation of large data sets is computationally unfeasible. While we chose reactive force fields[94,95] and density functional tight–binding (DFTB)[96] calculations for our simulations, any method which performs the structural sampling of a molecular junction and provides transport properties for selected snapshots may provide the basis for our machine learning approach. MD simulations employing reactive force fields have been successfully used in the past to compute trajectories of MCBJ experiments for the same molecule used here or similar systems[47,48,97]. Since electronic structure calculations are the most expensive component in the construction of the conductance histograms, we aim to predict conductance solely on structural information.

2 Simulation of Break Junction Experiments

To simulate the break junction setup, 296 gold atoms are arranged in 24 layers along the fcc(111) direction to form a gold wire. Ten octanedimethylsulfide ($C_8H_{16}(SMe)_2$) molecules are added randomly close to the wire. This system is well characterized by previous experimental and theoretical studies[47,98] and poses a challenge for machine learning approaches, since a huge variety of conformations or electrode-molecule binding configuration can be encountered. Since we try to capture the full evolution of the junction from the initial forming to rupture, we can sample situations where, e.g., multiple molecules form a junction, which is not included in MD simulations based on already-formed molecular junctions. Using reactive force fields as implemented in LAMMPS[94,95], an initial MD simulation is then performed , so that the molecules can adsorb onto the gold wire. Reactive Force fields are required to capture bond breaking and formation, processes inherent to the experiment. Molecules which desorb from the wire are removed from the simulation box.

To simulate the break-junction experiment, three MD simulations are performed while pulling on one end of the gold wire with speed of $1 \times 10^{-4}\,\text{Å}\,\text{fs}^{-1}$, and three more with a speed of $5 \times 10^{-5}\,\text{Å}\,\text{fs}^{-1}$, yielding a total simulation time of 2.4 ns. Structures were dumped every 500 fs resulting in a total of 9600 structures, for which transport calculations were performed.

This extensive structural sampling limits the choice of methods, which is why we employ non-scc DFTB calculations as implemented in DFTB+[96]. Still, the calculation of the conductance for a single structure takes more than six times the time of a calculation of a MD trajectory of 50000 time steps, which needs around 13 minutes on a single core of an Intel Xeon Silver 4110 CPU with a clock speed of 2.10 GHz. The resulting conductance traces and histograms are shown in Figure 1. They show a most probable conductance of around 10^{-5} G_0, which is in line with previous calculations and experiments[47,98].

Several key properties of MCBJ experiments are captured in our simulations. The stochastic nature of successful junction formation is shown, as two out of our six trajec-

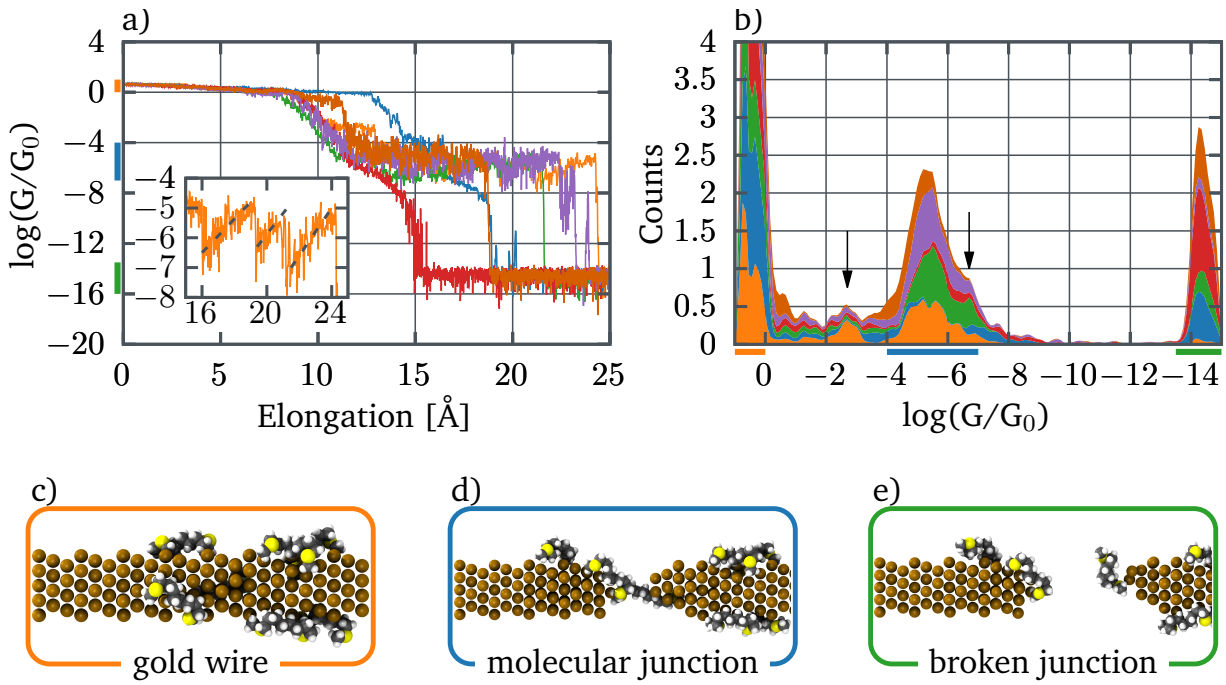

Figure 1: a) Individual conductance traces and b) corresponding stacked histograms for six different MD trajectories. Some peaks with higher or lower conductance than the molecules' main conductance peak are marked with arrows. In two simulations (red and blue curves), no stable junction was formed. c)-d) Representative snapshots of the MD trajectories are shown, representing c) the intact wire, d) a successfully formed molecular junction, and e) a broken junction. The corresponding areas are marked by colored bars at the axes. The inset shows a zoom-in to demonstrate the effect of the gold wire adjusting to the increasing tension by structural rearrangements to release stress.

tories do not result in a stable junction (red and blue line in Figure 1). They both show a small peak between 10^{-6} and $10^{-7}\,G_0$), which stems from a short period of time when a molecule is between the two electrodes, but not properly bound to the electrodes via the anchoring group. One particular trajectory shows a distinct peak at high conductance values between 10^{-2} to 10^{-3} G_0, which is caused by two molecules bridging the junction at the same time. This was attributed before to a distinct shape of a conductance histogram[38], but in our case it leads to two separable peaks. The elongation at which the molecular junction breaks depends on how the gold atoms on the electrode tip are rearranged while pulling and occurs after an elongation of 18 to 25 Å in our simulations. For the successfully created junctions, low-conductance shoulders (10^{-6} to 10^{-7} G_0) can be observed in the histogram. In these shoulders, gauche defects in the molecules decrease the conductance of the system, compared to structures without such effects. These molecules rearrange to anti-conformation, as they are further elongated.

Even though our approach performs structural sampling of a break junction experiment, it does not take into account that experiments average conductance over microseconds. Averaging over all molecular conformations accessible at every point during the elongation. This aspect and its effects on the shape of the resulting conductance histograms has been studied by Li *et al.*[47,48]. Importantly, the methods introduced below can also be used when time-averaged conductances are employed to construct histograms.

3 Molecular Descriptors for Charge Transport Properties

The molecular Cartesian coordinates obtained from the MD simulations have to be converted into a representation suited for a machine learning algorithm[92,99–103]. These representations are called descriptors, as they translate structural information into a so-called feature space. In that way, each molecular structure is represented by a feature vector, with the size of the vector depending on the chosen descriptor.

For different descriptors, the size of the resulting feature vector, the performance for different problems, and the concepts underlying their construction can differ drastically, as discussed in a recent review [103]. To predict the conductance for molecular junctions, we use established and broadly applicable approaches like the ACSF [104], SOAP [92] or F2B [105] descriptors, and construct new, custom descriptors, which aim at establishing structure–property relationships based on our understanding of charge transport through short molecules. As shown in a previous study, the usage of such problem-tailored descriptors can achieve similar or even better performance than established descriptors at decreased computational cost [86].

As ingredients for our custom descriptors, we explore different structural parameters, as detailed in Table 1. The chosen parameters represent information about the local chemical environment of the anchoring groups of the molecules (such as the distance to the closest gold atom), the molecular conformation (by, e.g., measuring the molecular end-to-end distance), and global properties which represent the state of the junction, such as the total length of the system or a histogram of the occurrences of the different atom types along the transport-direction. This histogram counts the number of atoms of a specific type in bins, representing an approximation of an atomic density. In that way, a continuous density of gold atoms represents an intact gold wire, while a gap in the gold density combined with a certain density of carbon atoms in that gap hints at a successfully formed molecular junction (see Figure S3 for an example). The different parameters can be combined to find the best performing ML model while retaining a small feature vector.

The dimensions feature vectors generated by descriptors like SOAP or F2B depend on the chosen settings and types of elements included, not on the system size or number of molecules and electrode atoms. Therefore, they can be used to compare different systems or system sizes. By contrast, the dimension of most of our descriptors scale with the number of molecules (as this relates to the number of anchoring groups/sulfur atoms) included in the simulation, except for the density histogram-based approach. Since we aim to predict the conductance histogram based on MD simulations for a specific molecular junction, this does

Table 1: Structural parameters for constructing custom descriptors for molecular junctions with SMe_2 anchoring groups. N is the number of molecules included in the MD simulation, in our case 7. The size of the feature vector for the density histogram depends on the number of bins per atom, and on the included atom types.

Name	Definition	Dimension
junction length	size of the system in transport direction, measured by the distance between the outermost atoms	1
intra-molecular S–S distances	distances between the terminal sulfur atoms of each molecule	N
Au–S distance	distance of each sulfur atom to the closest gold atom	$2N$
sulfur coordination	number of gold atoms in the vicinity of each sulfur atom within a radius of 3 Å	$2N$
S–S distances	distances between all sulfur atoms	$2N^2 - N$
density histogram	histogram of the atom types along the transport direction. Histograms can differ in bin size, smoothing (denoted by "(smoothed)") or whether hydrogen atoms were included ("+H") (see SI for more details)	∼90-160 (mainly depending on bin size)

not pose any limits to our approach.

The dimension of the feature vector generated by the SOAP descriptor is 2640 for our system, significantly larger than all of the custom descriptors (only surpassed by ACSF with 13328 feature dimensions). The F2B descriptors with 150 dimensions is comparable in size to the custom descriptors.

4 Evaluating Regression Models

We focus on Gaussian process regression (GPR) for predicting the conductance for the molecular junction and use Ridge Regression (RR) as a (regularized) linear baseline model against which we compare GPR performance. Excellent introductions to these methods can be found, e.g., in Refs.[60,62,79,85,106]. GPR performs predictions based on similarities between the feature vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (see Section 3) of different datapoints. A Gaussian Process represents a probability distribution of possible functions which fit to a set of given datapoints, thus providing the opportunity to calculate the mean and associated variance for a prediction. These calculations are not directly performed on the feature vectors, but by a kernel, which provides a measure to obtain similarities between datapoints in a higher dimensional space, into which the features are mapped by the kernel. Ridge Regression on the other hand is a linear model which includes regularization to be able deal with, e.g., highly correlated predictor variables.

Predictions are made employing the SOAP, ACSF, F2B descriptors, and different combinations of the ingredients for our custom descriptors. The coefficient of determination (R^2) as well as the mean absolute error (MAE) are used as measures for the performance of our approach, while plots for the root-mean-squared error (RMSE) and the mean absolute percentage error (MAPE) as well as a discussion of error measures for machine learning can be found in the SI. As a target for the learning algorithm we use the log-conductance instead of the conductance, since otherwise the target would span several orders of magnitudes,

artificially increasing R^2 and decreasing the MAE and RMSE.

As a kernel-based method, the performance of GPR can significantly depend on the choice of the kernel[84,107]. We found the Matérn($\frac{1}{2}$) kernel,

$$k_{\nu=\frac{1}{2}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sigma_f^2 \quad \exp\left(-\frac{|\boldsymbol{x}_1 - \boldsymbol{x}_2|}{l}\right) \tag{1}$$

to deliver a stable and good performance for all our descriptors (see SI). The hyperparameters σ_f^2 (signal variance) and l (length scale) are optimized in the fitting process.

For all features, 25 random splits of the data set into training and test sets are made (a procedure called "random permutations cross-validation" or "shuffle & split"). The obtained values are used to calculate the mean and the standard deviation for MAE and R^2. The results for a training set size of 10 % are summarized in Figure 2. These evaluations are performed for unscaled features as well as for data sets, for which each feature dimension is standardized (i.e., shifted to a mean of zero and divided by its standard deviation).

For all descriptors, the GPR approach outperforms the linear RR model. Very simple custom descriptors with few feature dimensions perform similar or even better than established descriptors such as SOAP. For a training set size of 10 %, SOAP reaches a MAE of $0.45\log(G/G_0)$, while simply taking the distances between all sulfur atoms (d(S-S)) yields a MAE of $0.34\log(G/G_0)$. The performance can be improved by combining different ingredients for the custom descriptors, such as adding the total length to the distance between each sulfur atom and its closest gold atom, d(S-Au): In fact, using only d(S-Au) yields a MAE of $1.4\log(G/G_0)$, while combining this measure with the length of the whole system (in order to capture the elongation of the break junction) further decreases the error to $0.35\log(G/G_0)$. Since the combination of different structural parameters improves the performance of GPR, this represents a modular approach where structural information can be combined to achieve the desired performance.

For a more detailed look, we plot the ML-predicted vs. the target conductance in Figure 3

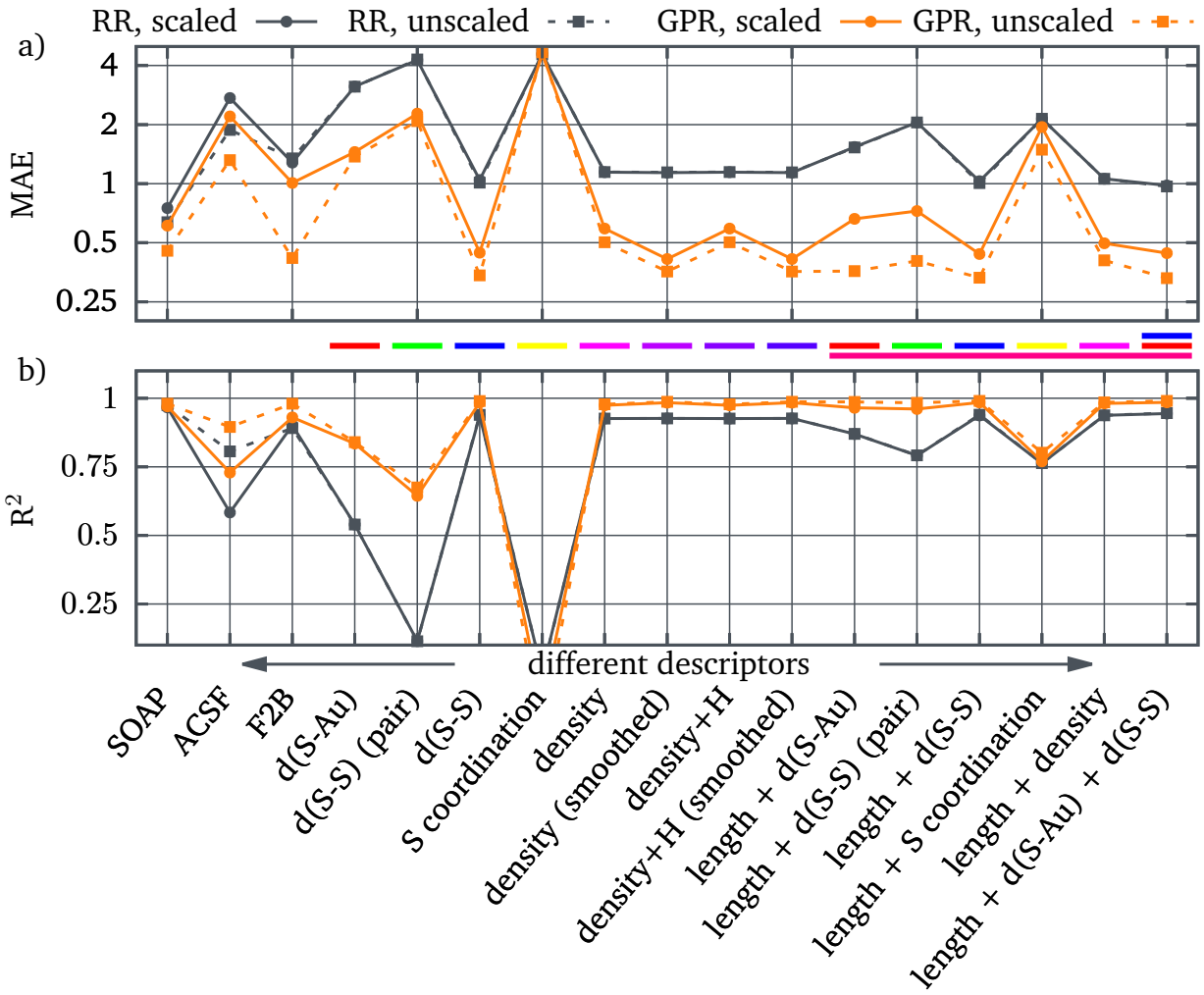

Figure 2: Comparison of the performance of the different descriptors (scaled and unscaled), as well as Ridge Regression (RR) and Gaussian Process Regression (GPR) in terms of a) the mean absolute error and b) coefficient of correlation. A small training set size of 10 % was used. The mean performance and standard deviation was obtained from 25 repetitions of *shuffle and split*. The standard deviation is usually very small, making it only visible for some cases. Colored bars between plots aim to be a guide to the eye for the composition of the different custom descriptors, e.g. the red bar denotes the d(S–Au) descriptors, and the red and the long magenta bar show the usage of the d(S–A) descriptors together with the total length.

for selected descriptors. All chosen descriptors show the correlation between the original and calculated conductance, but differences become especially evident when comparing the RMSE: For the SOAP as well as for the F2B descriptor, the predicted conductance deviates slightly more stronger from the target for structures where the junction is nearly broken or the molecule is detaching ($< 10^{-7}$ G_0). The custom descriptors employing the distances between sulfur atoms (d(S–S)) and d(S–S) combined the distances between the sulfur atoms and the closest gold atom (d(S–Au)) and the length of the total systems (length) show correlation plots very similar to each other (Figure 3), with only slight differences in the MAE and RMSE and a clearly an improved performance compared to SOAP and F2B.

The conductance histograms in Figure 3 constructed from the predictions all resemble the histograms constructed from the targets. Building a histogram involves intrinsic averaging, since different conductance values are grouped together in bins. This affects the final shape of the histogram and clouds minor errors, so that the resulting histograms all show a satisfying agreement with the target.

As discussed above, a significant difference between the descriptors is the size of the feature vector, which affects the fitting times of the GPR. More precisely, it affects the evaluation of the norm inside the kernel (Eq. 1), as it scales linearly with the length of the feature vector. However, it also affects the number of evaluations during the hyperparameter optimization in an unpredictable manner. While the custom and F2B descriptors represent vectors of similar size (size of the feature vector: 100-150), and the resulting computational time for fitting the GPR with a training set size of 10 % is usually around two minutes on an Intel Xeon Silver 4110 CPU with a clock speed of 2.10 GHz, fitting times of the GPR for the SOAP descriptor (size of the feature vector: 2640) are greatly increased by a factor of up to six.

To see whether the number of feature dimensions for, e.g., the SOAP descriptor can be reduced to a similar number as for the other descriptors, we perform Principal Component Analysis (PCA) and evaluate the performance of our approach for different numbers of

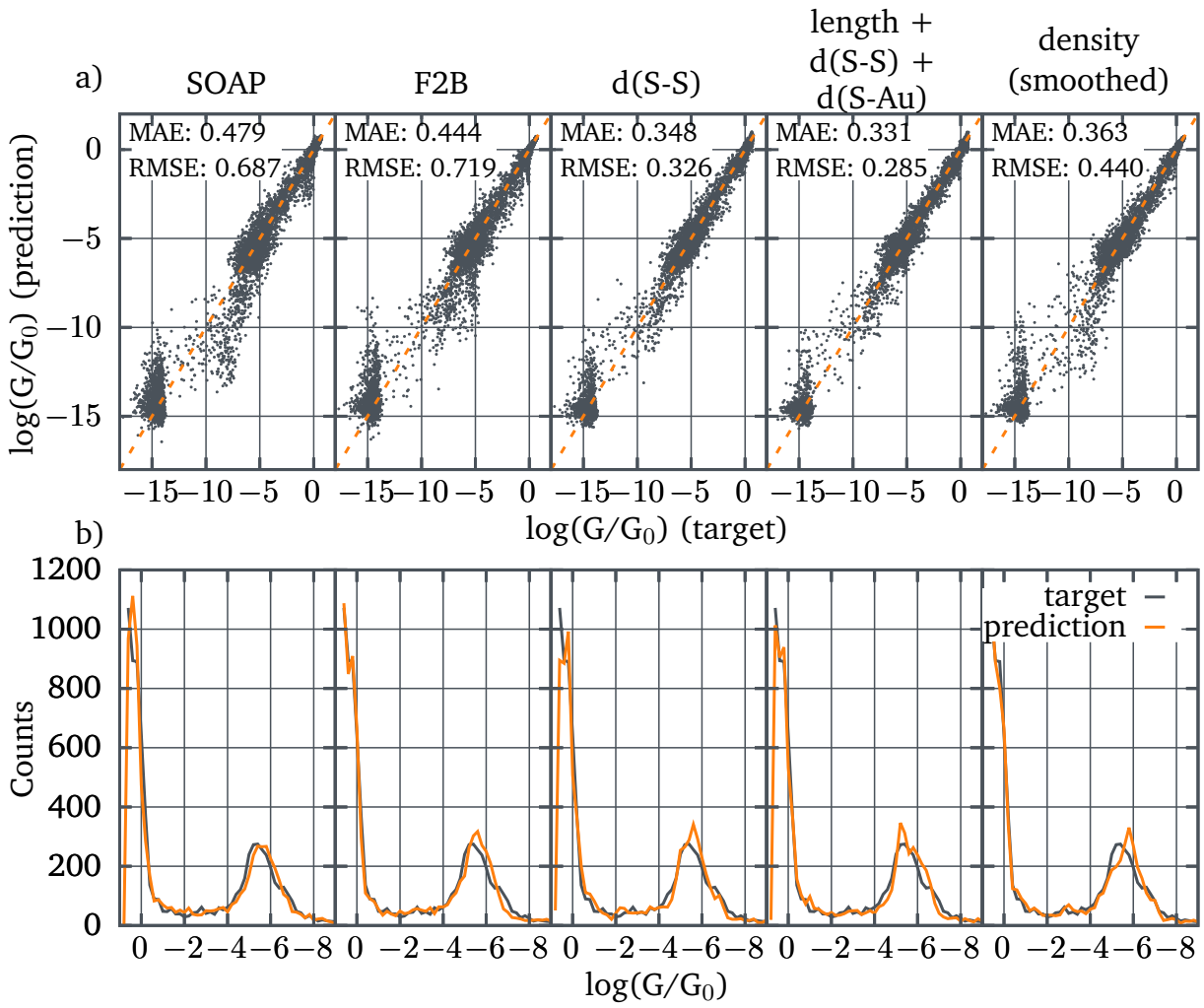

Figure 3: a) Correlation between targets and predicted conductances and b) corresponding conductance histograms. Data is provided for the SOAP & F2B descriptor and for three selected custom descriptors with good performance (unscaled features). The training set size is 10 % throughout. The MAE and RMSE in the upper plots are given for the individual predictions shown here and thus can deviate from the data of Figure 2, which were obtained by shuffle and split. The predicted conductance histograms in the lower row include datapoints from the training set, as this resembles the way the conductance histograms would be created in an actual use case.

Principal Components, with a training set size of 10 %. Figure 4a) shows the MAE and the explained variance for the SOAP, the F2B, the d(S-S), length+d(S-S)+d(S-Au) and the density (smoothed)[1] descriptor for increasing number of PCA dimensions. (The "explained variance" reflects how much of the total variance in the dataset is explained by the selected number of PCA dimensions.)

When the feature dimensions are reduced via PCA to 20 (for SOAP), 10 (for F2B) or 5 (for our custom descriptors), the performance of GPR becomes comparable to the performance of the full feature vectors. In that way, the differences between the descriptors with respect to computational cost become marginal. It is interesting to note how different our two custom descriptors perform: Even though slightly worse in the final performance, only three PCA dimensions are enough for the density-based descriptor to reach the same MAE as for the full feature vector, while for the distance-based one, we need around five dimensions. The PCA of the distance–based descriptor also explains less variance than the other descriptors when only few dimensions are used. Thus, significant feature reduction by PCA is possible here, generating predictions with comparable errors as the original features.

Finally, to show how the predictions improve with increasing training set size, we plot learning curves for selected descriptors. The plots in Figure 4b) clearly demonstrate that, as expected, the performance improves by including more data points into the training set. However, training sets bigger than 10 % yield only minor improvement while increasing computational cost, as the fitting time for GPR scales cubically with the training set size (neglecting the unpredictable timing regarding hyperparameter optimization). Significant differences between the learning curves for the original or PCA-reduced descriptors could not be observed.

[1] density histogram smoothed by a moving average filter

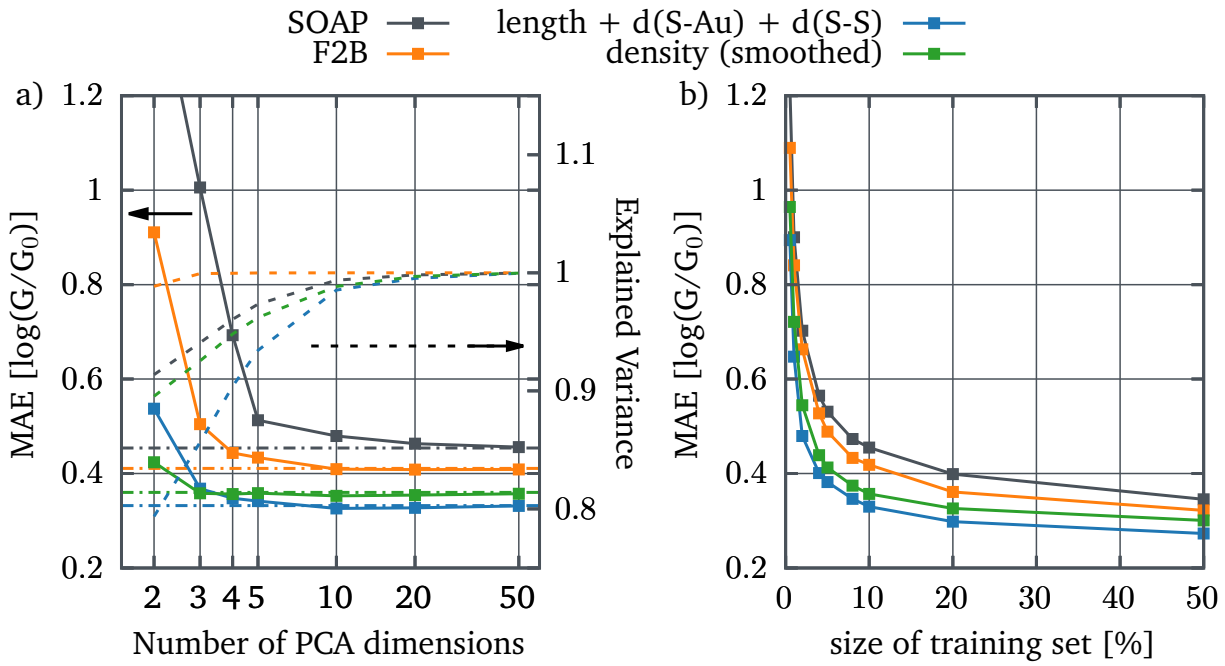

Figure 4: *a)* GPR performance (as measured by the MAE) depending on the number of dimensions of the Principal Component Analysis for the SOAP, F2B, d(S-S), length+d(S-S)+d(S-Au) and the density (smoothed) descriptors. The dashed-dotted horizontal lines show the performance for the original features. The dashed lines give the explained variance (second y-axis). For F2B and the custom descriptors, only 10 PCA dimensions are sufficient to reach a similar performance as for the original features. The SOAP descriptor requires more dimensions ($\geq$ 20), which is still significantly less than the original dimension of the SOAP feature vector (2640). The explained variance does not directly correlate with how close the performance is to the final performance/performance using the original feature vector. *b)* Learning curves for the same descriptors. The shape of the curves is similar for all descriptors; the differences in the performance basically manifest as a shift on the y-axis. As also shown in Figure 3, a training set size of 10 % is reasonable in our case. Increasing the size of the training set gains only minor improvements on the predictions, but comes at a higher computational cost. A training set size of 10 % equals a reduction of necessary electronic structure computations to 10 %, which are the bottleneck for the calculation of our conductance histograms. Learning curves for PCA-reduced versions of the descriptors are shown in the SI.

5 Summary

This study explores the application of Gaussian Process Regression (GPR) for the calculation of conductance histograms based on molecular dynamics simulations of molecular junctions. We show that we can construct such histograms by interpolating between quantum chemical transport calculations for only around 10% of the MD snapshots via GPR. Calculating the conductance of a single structure by quantum chemical methods takes 80 minutes on a single core on our CPUs, while fitting the GPR and predicting the conductance for the remaining data points in conjunction with our custom descriptors is performed within minutes. Given the comparatively small cost of the molecular dynamics simulations and the machine learning, this finally results in a speed-up by one order of magnitude. Predictions based on established molecular descriptors such as SOAP and F2B yield mean absolute errors of about $0.45\log(G/G_0)$ and $0.42\log(G/G_0)$, respectively, but are narrowly outperformed by custom descriptors in terms of speed and performance (down to $0.34\log(G/G_0)$). These custom descriptors aim to capture structural information, which we think are determining the essentials of the conductance of the molecular junction. Reducing the number of feature dimensions via Principal Component Analysis can be used to reduce the feature vector such, that the differences between the different descriptors are negligible in terms of computational efficiency.

Our approach is method-agnostic, so every combination of a (hopefully cheap) method to perform structural sampling and a (potentially expensive) way to calculate electron transport can benefit and be used to construct conductance histograms based on a multitude of structures. For future work, intelligent schemes to cluster the datapoints or, e.g., filter out the ones representing a broken junction may push the boundaries for this approach.

6 Methods

Molecular Dynamics simulations

For all MD simulations, LAMMPS with reactive force fields (ReaxFF) was employed[94,95]. The force field parameters for gold, sulfur, carbon and hydrogen by Bae & Aikens[108] were used. The simulation time step was 0.5 fs, snapshots were generated every 500 steps and the temperature was set to 300 K using an NVT ensemble. Even though all simulations finally stem from the same starting structure, different seeds for the velocities and a equilibration period of before the pulling simulation ensures divergence of the structures. The outermost six layers were frozen in all simulations, due to the requirements of the subsequent transport calculations.

Electron transport calculations

Due to the high number of necessary transport calculations, non-self-consistent electron transport calculations using DFTB+[96] were performed. 386 atoms are included in the central region, and the remaining gold atoms are distributed to the electrodes in six layers each. The `auorg-1-1` parameter set was used[109]. After the calculation of the transmission function in a non-SCC approximation and using the wide band approximation, the zero-bias conductance was evaluated at a Fermi energy of -5 eV by $G = G_0 T(E_F)$.

Feature generation/descriptor

The SOAP and ACSF descriptor was used as implemented in `dscribe` library[110]. For SOAP, the Rcut, $nmax$ and $lmax$ parameter were optimized, further information can be found in the SI. ACSF was employed using the default settings, the symmetry functions were evaluated at the positions of the carbon and sulfur atoms. The custom and F2B descriptors were created using custom python scripts.

Gaussian Process Regression

For the training of Ridge and Gaussian Process Regression as well as the Kernel functions, the `sklearn` library was employed. The training of the GPR was repeated 40 times (using the `gpr_optimizer_restarts` parameter) to ensure the optimization to a global minimum.

Acknowledgement

We would like to thank the high performing computing center of the University of Hamburg for computational resources. We would like to thank Latha Venkataraman and Leopoldo Mejía for helpful discussions. This work is supported by the Cluster of Excellence "Advanced Imaging of Matter" of the Deutsche Forschungsgemeinschaft (DFG) — EXC 2056 — project ID 390715994.

References

(1) Solomon, G. C.; Andrews, D. Q.; Hansen, T.; Goldsmith, R. H.; Wasielewski, M. R.; Duyne, R. P. V.; Ratner, M. A. Understanding quantum interference in coherent molecular conduction. *The Journal of Chemical Physics* **2008**, *129*, 054701.

(2) Markussen, T.; Stadler, R.; Thygesen, K. S. The Relation between Structure and Quantum Interference in Single Molecule Junctions. *Nano Letters* **2010**, *10*, 4260–4265.

(3) Garner, M. H.; Li, H.; Neupane, M.; Zou, Q.; Liu, T.; Su, T. A.; Shangguan, Z.; Paley, D. W.; Ng, F.; Xiao, S.; Nuckolls, C.; Venkataraman, L.; Solomon, G. C. Permethylation Introduces Destructive Quantum Interference in Saturated Silanes. *Journal of the American Chemical Society* **2019**, *141*, 15471–15476.

(4) Thoss, M.; Evers, F. Perspective: Theory of quantum transport in molecular junctions. *The Journal of Chemical Physics* **2018**, *148*, 030901.

(5) Němec, H.; Nienhuys, H.-K.; Perzon, E.; Zhang, F.; Inganäs, O.; Kužel, P.; Sundström, V. Ultrafast conductivity in a low-band-gap polyphenylene and fullerene blend studied by terahertz spectroscopy. *Physical Review B* **2009**, *79*.

(6) Grätzel, M. Recent Advances in Sensitized Mesoscopic Solar Cells. *Accounts of Chemical Research* **2009**, *42*, 1788–1798.

(7) Brabec, C. J.; Gowrisanker, S.; Halls, J. J. M.; Laird, D.; Jia, S.; Williams, S. P. Polymer-Fullerene Bulk-Heterojunction Solar Cells. *Advanced Materials* **2010**, *22*, 3839–3856.

(8) Facchetti, A. π-Conjugated Polymers for Organic Electronics and Photovoltaic Cell Applications†. *Chemistry of Materials* **2011**, *23*, 733–758.

(9) Jin, Z.; Gehrig, D.; Dyer-Smith, C.; Heilweil, E. J.; Laquai, F.; Bonn, M.; Turchinovich, D. Ultrafast Terahertz Photoconductivity of Photovoltaic Polymer–Fullerene Blends: A Comparative Study Correlated with Photovoltaic Device Performance. *The Journal of Physical Chemistry Letters* **2014**, *5*, 3662–3668.

(10) Xue, R.; Zhang, J.; Li, Y.; Li, Y. Organic Solar Cell Materials toward Commercialization. *Small* **2018**, *14*, 1801793.

(11) Schlicke, H.; Rebber, M.; Kunze, S.; Vossmeyer, T. Resistive pressure sensors based on freestanding membranes of gold nanoparticles. *Nanoscale* **2016**, *8*, 183–186.

(12) Wani, I. H.; Jafri, S. H. M.; Warna, J.; Hayat, A.; Li, H.; Shukla, V. A.; Orthaber, A.; Grigoriev, A.; Ahuja, R.; Leifer, K. A sub 20 nm metal-conjugated molecule junction acting as a nitrogen dioxide sensor. *Nanoscale* **2019**, *11*, 6571–6575.

(13) Tao, C.-P.; Jiang, C.-C.; Wang, Y.-H.; Zheng, J.-F.; Shao, Y.; Zhou, X.-S. Single-Molecule Sensing of Interfacial Acid–Base Chemistry. *The Journal of Physical Chemistry Letters* **2020**, *11*, 10023–10028.

(14) Wolf, S. A.; Chtchelkanova, A. Y.; Treger, D. M. Spintronics—A retrospective and perspective. *IBM Journal of Research and Development* **2006**, *50*, 101–110.

(15) Seneor, P.; Bernand-Mantel, A.; Petroff, F. Nanospintronics: when spintronics meets single electron physics. *Journal of Physics: Condensed Matter* **2007**, *19*, 165222.

(16) Bazarnik, M.; Bugenhagen, B.; Elsebach, M.; Sierda, E.; Frank, A.; Prosenc, M. H.; Wiesendanger, R. Toward Tailored All-Spin Molecular Devices. *Nano Letters* **2016**, *16*, 577–582.

(17) Atzori, M.; Sessoli, R. The Second Quantum Revolution: Role and Challenges of Molecular Chemistry. *Journal of the American Chemical Society* **2019**, *141*, 11339–11352.

(18) Reed, M. Molecular-scale electronics. *Proceedings of the IEEE* **1999**, *87*, 652–658.

(19) Vilan, A.; Aswal, D.; Cahen, D. Large-Area, Ensemble Molecular Electronics: Motivation and Challenges. *Chemical Reviews* **2017**, *117*, 4248–4286.

(20) Zabet-Khosousi, A.; Dhirani, A.-A. Charge Transport in Nanoparticle Assemblies. *Chemical Reviews* **2008**, *108*, 4072–4124.

(21) Negre, C. F. A.; Milot, R. L.; Martini, L. A.; Ding, W.; Crabtree, R. H.; Schmuttenmaer, C. A.; Batista, V. S. Efficiency of Interfacial Electron Transfer from Zn-Porphyrin Dyes into TiO_2 - Correlated to the Linker Single Molecule Conductance. *The Journal of Physical Chemistry C* **2013**, *117*, 24462–24470.

(22) Liao, J.; Blok, S.; van der Molen, S. J.; Diefenbach, S.; Holleitner, A. W.; Schönenberger, C.; Vladyka, A.; Calame, M. Ordered nanoparticle arrays interconnected by

molecular linkers: electronic and optoelectronic properties. *Chemical Society Reviews* **2015**, *44*, 999–1014.

(23) Heuer-Jungemann, A.; Feliu, N.; Bakaimi, I.; Hamaly, M.; Alkilany, A.; Chakraborty, I.; Masood, A.; Casula, M. F.; Kostopoulou, A.; Oh, E.; Susumu, K.; Stewart, M. H.; Medintz, I. L.; Stratakis, E.; Parak, W. J.; Kanaras, A. G. The Role of Ligands in the Chemical Synthesis and Applications of Inorganic Nanoparticles. *Chemical Reviews* **2019**, *119*, 4819–4880.

(24) McCreery, R. L. The merger of electrochemistry and molecular electronics. *The Chemical Record* **2011**, *12*, 149–163.

(25) Wierzbinski, E.; Venkatramani, R.; Davis, K. L.; Bezer, S.; Kong, J.; Xing, Y.; Borguet, E.; Achim, C.; Beratan, D. N.; Waldeck, D. H. The Single-Molecule Conductance and Electrochemical Electron-Transfer Rate Are Related by a Power Law. *ACS Nano* **2013**, *7*, 5391–5401.

(26) Bueno, P. R.; Benites, T. A.; Davis, J. J. The Mesoscopic Electrochemistry of Molecular Junctions. *Scientific Reports* **2016**, *6*.

(27) Bueno, P. R. Common Principles of Molecular Electronics and Nanoscale Electrochemistry. *Analytical Chemistry* **2018**, *90*, 7095–7106.

(28) Naaman, R.; Paltiel, Y.; Waldeck, D. H. Chiral Induced Spin Selectivity Gives a New Twist on Spin-Control in Chemistry. *Accounts of Chemical Research* **2020**, *53*, 2659–2667.

(29) Hush, N. S. An Overview of the First Half-Century of Molecular Electronics. *Annals of the New York Academy of Sciences* **2003**, *1006*, 1–20.

(30) Aviram, A.; Ratner, M. A. Molecular rectifiers. *Chemical Physics Letters* **1974**, *29*, 277–283.

(31) Chen, F.; Hihath, J.; Huang, Z.; Li, X.; Tao, N. Measurement of Single-Molecule Conductance. *Annual Review of Physical Chemistry* **2007**, *58*, 535–564.

(32) Xiang, D.; Jeong, H.; Lee, T.; Mayer, D. Mechanically Controllable Break Junctions for Molecular Electronics. *Advanced Materials* **2013**, *25*, 4845–4867.

(33) Gehring, P.; Thijssen, J. M.; van der Zant, H. S. J. Single-molecule quantum-transport phenomena in break junctions. *Nature Reviews Physics* **2019**, *1*, 381–396.

(34) Kiguchi, M.; Takahashi, T.; Kanehara, M.; Teranishi, T.; Murakoshi, K. Effect of End Group Position on the Formation of a Single Porphyrin Molecular Junction. *The Journal of Physical Chemistry C* **2009**, *113*, 9014–9017.

(35) Kaneko, S.; Montes, E.; Suzuki, S.; Fujii, S.; Nishino, T.; Tsukagoshi, K.; Ikeda, K.; Kano, H.; Nakamura, H.; Vázquez, H.; Kiguchi, M. Identifying the molecular adsorption site of a single molecule junction through combined Raman and conductance studies. *Chemical Science* **2019**, *10*, 6261–6269.

(36) Fu, T.; Frommer, K.; Nuckolls, C.; Venkataraman, L. Single-Molecule Junction Formation in Break-Junction Measurements. *The Journal of Physical Chemistry Letters* **2021**, *12*, 10802–10807.

(37) Rivero, S. M.; Arroyo, P. G.; Li, L.; Gunasekaran, S.; Stuyver, T.; Mancheño, M. J.; Alonso, M.; Venkataraman, L.; Segura, J. L.; Casado, J. Single-molecule conductance in a unique cross-conjugated tetra(aminoaryl)ethene. *Chemical Communications* **2021**, *57*, 591–594.

(38) Reuter, M. G.; Hersam, M. C.; Seideman, T.; Ratner, M. A. Signatures of Cooperative Effects and Transport Mechanisms in Conductance Histograms. *Nano Letters* **2012**, *12*, 2243–2248.

(39) Williams, P. D.; Reuter, M. G. Level Alignments and Coupling Strengths in Conductance Histograms: The Information Content of a Single Channel Peak. *The Journal of Physical Chemistry C* **2013**, *117*, 5937–5942.

(40) Quan, R.; Pitler, C. S.; Ratner, M. A.; Reuter, M. G. Quantitative Interpretations of Break Junction Conductance Histograms in Molecular Electron Transport. *ACS Nano* **2015**, *9*, 7704–7713.

(41) Trasobares, J.; Rech, J.; Jonckheere, T.; Martin, T.; Aleveque, O.; Levillain, E.; Diez-Cabanes, V.; Olivier, Y.; Cornil, J.; Nys, J. P.; Sivakumarasamy, R.; Smaali, K.; Leclere, P.; Fujiwara, A.; Théron, D.; Vuillaume, D.; Clément, N. Estimation of π–π Electronic Couplings from Current Measurements. *Nano Letters* **2017**, *17*, 3215–3224.

(42) Pu, Q.; Leng, Y.; Zhao, X.; Cummings, P. T. Molecular Simulation Studies on the Elongation of Gold Nanowires in Benzenedithiol. *The Journal of Physical Chemistry C* **2010**, *114*, 10365–10372.

(43) French, W. R.; Iacovella, C. R.; Cummings, P. T. Large-Scale Atomistic Simulations of Environmental Effects on the Formation and Properties of Molecular Junctions. *ACS Nano* **2012**, *6*, 2779–2789.

(44) 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*, 3654.

(45) French, W. R.; Iacovella, C. R.; Rungger, I.; Souza, A. M.; Sanvito, S.; Cummings, P. T. Structural Origins of Conductance Fluctuations in Gold–Thiolate Molecular Transport Junctions. *The Journal of Physical Chemistry Letters* **2013**, *4*, 887–891.

(46) Mejía, L.; Renaud, N.; Franco, I. Signatures of Conformational Dynamics and Electrode-Molecule Interactions in the Conductance Profile During Pulling of Single-Molecule Junctions. *The Journal of Physical Chemistry Letters* **2018**, *9*, 745–750.

(47) Li, Z.; Franco, I. Molecular Electronics: Toward the Atomistic Modeling of Conductance Histograms. *The Journal of Physical Chemistry C* **2019**, *123*, 9693–9701.

(48) Li, Z.; Mejia, L.; Marrs, J.; Jeong, H.; Hihath, J.; Franco, I. Understanding the Conductance Dispersion of Single-Molecule Junctions. *The Journal of Physical Chemistry C* **2020**, *125*, 3406–3414.

(49) Krüger, D.; Fuchs, H.; Rousseau, R.; Marx, D.; Parrinello, M. Pulling Monatomic Gold Wires with Single Molecules: An Ab-Initio Simulation. *Phys. Rev. Lett.* **2002**, *89*, 186402.

(50) Makk, P.; Visontai, D.; Oroszlány, L.; Manrique, D. Z.; Csonka, S.; Cserti, J.; Lambert, C.; Halbritter, A. Advanced Simulation of Conductance Histograms Validated through Channel-Sensitive Experiments on Indium Nanojunctions. *Phys. Rev. Lett.* **2011**, *107*, 276801.

(51) Szyja, B. M.; Nguyen, H. C.; Kosov, D.; Doltsinis, N. L. Conformation-dependent conductance through a molecular break junction. *Journal of Molecular Modeling* **2013**, *19*, 4173–4180.

(52) Saffarzadeh, A.; Demir, F.; Kirczenow, G. Mechanism of the enhanced conductance of a molecular junction under tensile stress. *Phys. Rev. B* **2014**, *89*, 045431.

(53) Nguyen, H. C.; Szyja, B. M.; Doltsinis, N. L. Electric conductance of a mechanically strained molecular junction from first principles: Crucial role of structural relaxation and conformation sampling. *Phys. Rev. B* **2014**, *90*, 115440.

(54) Hybertsen, M. S. Modeling single molecule junction mechanics as a probe of interface bonding. *The Journal of Chemical Physics* **2017**, *146*, 092323.

(55) Evers, F.; Korytár, R.; Tewari, S.; van Ruitenbeek, J. M. Advances and challenges in single-molecule electron transport. *Reviews of Modern Physics* **2020**, *92*, 035001.

(56) Landauer, R. Spatial Variation of Currents and Fields Due to Localized Scatterers in Metallic Conduction. *IBM Journal of Research and Development* **1957**, *1*, 223–231.

(57) Büttiker, M.; Imry, Y.; Landauer, R.; Pinhas, S. Generalized many-channel conductance formula with application to small rings. *Physical Review B* **1985**, *31*, 6207–6215.

(58) Xue, Y.; Datta, S.; Ratner, M. A. First-principles based matrix Green's function approach to molecular electronic devices: general formalism. *Chemical Physics* **2002**, *281*, 151–170.

(59) Li, Y.; Yu, X.; Zhen, Y.; Dong, H.; Hu, W. Two-Pathway Viewpoint to Interpret Quantum Interference in Molecules Containing Five-Membered Heterocycles: Thienoacenes as Examples. *The Journal of Physical Chemistry C* **2019**, *123*, 15977–15984.

(60) Rasmussen, C. E. *Advanced Lectures on Machine Learning*; Springer Berlin Heidelberg, 2004; pp 63–71.

(61) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *The Journal of Physical Chemistry Letters* **2020**, *11*, 2336–2347.

(62) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer-Verlag New York Inc., 2006.

(63) Murphy, K. P. *Machine Learning*; MIT Press, 2012.

(64) Fabrizio, A.; Meyer, B.; Fabregat, R.; Corminboeuf, C. Quantum Chemistry Meets Machine Learning. *CHIMIA* **2019**, *73*, 983.

(65) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annual Review of Physical Chemistry* **2020**, *71*, 361–390.

(66) Ceriotti, M.; Clementi, C.; von Lilienfeld, O. A. Introduction: Machine Learning at the Atomic Scale. *Chemical Reviews* **2021**, *121*, 9719–9721.

(67) Huang, B.; von Lilienfeld, O. A. Ab Initio Machine Learning in Chemical Compound Space. *Chemical Reviews* **2021**, *121*, 10001–10036.

(68) Duan, C.; Nandy, A.; Kulik, H. J. Machine Learning for the Discovery, Design, and Engineering of Materials. *Annual Review of Chemical and Biomolecular Engineering* **2022**, *13*.

(69) Kulik, H.; Hammerschmidt, T.; Schmidt, J.; Botti, S.; Marques, M. A. L.; Boley, M.; Scheffler, M.; Todorović, M.; Rinke, P.; Oses, C.; Smolyanyuk, A.; Curtarolo, S.; Tkatchenko, A.; Bartok, A.; Manzhos, S.; Ihara, M.; Carrington, T.; Behler, J.; Isayev, O.; Veit, M.; Grisafi, A.; Nigam, J.; Ceriotti, M.; Schütt, K. T.; Westermayr, J.; Gastegger, M.; Maurer, R.; Kalita, B.; Burke, K.; Nagai, R.; Akashi, R.; Sugino, O.; Hermann, J.; Noé, F.; Pilati, S.; Draxl, C.; Kuban, M.; Rigamonti, S.; Scheidgen, M.; Esters, M.; Hicks, D.; Toher, C.; Balachandran, P.; Tamblyn, I.; Whitelam, S.; Bellinger, C.; Ghiringhelli, L. M. Roadmap on Machine Learning in Electronic Structure. *Electronic Structure* **2022**,

(70) Lorenz, S.; Groß, A.; Scheffler, M. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chemical Physics Letters* **2004**, *395*, 210–215.

(71) Agrawal, P. M.; Raff, L. M.; Hagan, M. T.; Komanduri, R. Molecular dynamics investigations of the dissociation of SiO2 on an ab initio potential energy surface obtained using neural network methods. *The Journal of Chemical Physics* **2006**, *124*, 134306.

(72) Ludwig, J.; Vlachos, D. G. Ab initio molecular dynamics of hydrogen dissociation on metal surfaces using neural networks and novelty sampling. *The Journal of Chemical Physics* **2007**, *127*, 154716.

(73) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chemical Science* **2017**, *8*, 6924–6935.

(74) Gebhardt, J.; Kiesel, M.; Riniker, S.; Hansen, N. Combining Molecular Dynamics and Machine Learning to Predict Self-Solvation Free Energies and Limiting Activity Coefficients. *Journal of Chemical Information and Modeling* **2020**, *60*, 5319–5330.

(75) Jurásková, V.; Célerse, F.; Laplaza, R.; Corminboeuf, C. Assessing the persistence of chalcogen bonds in solution with neural network potentials. *The Journal of Chemical Physics* **2022**, *156*, 154112.

(76) Rinderle, M.; Kaiser, W.; Mattoni, A.; Gagliardi, A. Machine-Learned Charge Transfer Integrals for Multiscale Simulations in Organic Thin Films. *The Journal of Physical Chemistry C* **2020**, *124*, 17733–17743.

(77) Brian, D.; Sun, X. Charge-Transfer Landscape Manifesting the Structure–Rate Relationship in the Condensed Phase Via Machine Learning. *The Journal of Physical Chemistry B* **2021**, *125*, 13267–13278.

(78) Manzhos, S. Machine learning for the solution of the Schrödinger equation. *Machine Learning: Science and Technology* **2020**, *1*, 013002.

(79) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chemical Reviews* **2021**, *121*, 10073–10141.

(80) Raimbault, N.; Grisafi, A.; Ceriotti, M.; Rossi, M. Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals. *New Journal of Physics* **2019**, *21*, 105001.

(81) Schmitz, G.; Artiukhin, D. G.; Christiansen, O. Approximate high mode coupling potentials using Gaussian process regression and adaptive density guided sampling. *The Journal of Chemical Physics* **2019**, *150*, 131102.

(82) Denzel, A.; Kästner, J. Gaussian process regression for geometry optimization. *The Journal of Chemical Physics* **2018**, *148*, 094114.

(83) Schmitz, G.; Christiansen, O. Gaussian process regression to accelerate geometry optimizations relying on numerical differentiation. *The Journal of Chemical Physics* **2018**, *148*, 241704.

(84) Meyer, R.; Hauser, A. W. Geometry optimization using Gaussian process regression in internal coordinate systems. *The Journal of Chemical Physics* **2020**, *152*, 084112.

(85) Proppe, J.; Gugler, S.; Reiher, M. Gaussian Process-Based Refinement of Dispersion Corrections. *Journal of Chemical Theory and Computation* **2019**, *15*, 6046–6060.

(86) Bahlke, M. P.; Mogos, N.; Proppe, J.; Herrmann, C. Exchange Spin Coupling from Gaussian Process Regression. *The Journal of Physical Chemistry A* **2020**, *124*, 8708–8723.

(87) Strange, M.; Rostgaard, C.; Häkkinen, H.; Thygesen, K. S. Self-consistent GW calculations of electronic transport in thiol- and amine-linked molecular junctions. *Physical Review B* **2011**, *83*, 115108.

(88) Lopez-Bezanilla, A.; von Lilienfeld, O. A. Modeling electronic quantum transport with machine learning. *Physical Review B* **2014**, *89*, 235411.

(89) Korol, R.; Segal, D. Machine Learning Prediction of DNA Charge Transport. *The Journal of Physical Chemistry B* **2019**, *123*, 2801–2811.

(90) Li, K.; Lu, J.; Zhai, F. Neural networks for modeling electron transport properties of mesoscopic systems. *Physical Review B* **2020**, *102*, 064205.

(91) Bürkle, M.; Perera, U.; Gimbert, F.; Nakamura, H.; Kawata, M.; Asai, Y. Deep-Learning Approach to First-Principles Transport Simulations. *Physical Review Letters* **2021**, *126*, 177701.

(92) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Physical Review B* **2013**, *87*, 184115.

(93) Topolnicki, R.; Kucharczyk, R.; Kamiński, W. Combining Multiscale MD Simulations and Machine Learning Methods to Study Electronic Transport in Molecular Junctions at Finite Temperatures. *The Journal of Physical Chemistry C* **2021**, *125*, 19961–19968.

(94) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics* **1995**, *117*, 1–19.

(95) Aktulga, H.; Fogarty, J.; Pandit, S.; Grama, A. Parallel reactive molecular dynamics: Numerical methods and algorithmic techniques. *Parallel Computing* **2012**, *38*, 245–259.

(96) Hourahine, B.; Aradi, B.; Blum, V.; Bonafé, F.; Buccheri, A.; Camacho, C.; Cevallos, C.; Deshaye, M. Y.; Dumitrică, T.; Dominguez, A.; Ehlert, S.; Elstner, M.; van der Heide, T.; Hermann, J.; Irle, S.; Kranz, J. J.; Köhler, C.; Kowalczyk, T.; Kubař, T.; Lee, I. S.; Lutsker, V.; Maurer, R. J.; Min, S. K.; Mitchell, I.; Negre, C.; Niehaus, T. A.; Niklasson, A. M. N.; Page, A. J.; Pecchia, A.; Penazzi, G.; Persson, M. P.; Řezáč, J.; Sánchez, C. G.; Sternberg, M.; Stöhr, M.; Stuckenberg, F.; Tkatchenko, A.; z. Yu, V. W.; Frauenheim, T. DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *The Journal of Chemical Physics* **2020**, *152*, 124101.

(97) Mejía, L.; Garay-Ruiz, D.; Franco, I. Diels–Alder Reaction in a Molecular Junction. *The Journal of Physical Chemistry C* **2021**, *125*, 14599–14606.

(98) 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. *Journal of the American Chemical Society* **2007**, *129*, 15768–15769.

(99) Todeschini, R. *Handbook of molecular descriptors*; Wiley-VCH: Weinheim New York, 2000.

(100) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Physical Review Letters* **2015**, *114*, 105503.

(101) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *The Journal of Chemical Physics* **2018**, *148*, 241718.

(102) Pozdnyakov, S. N.; Willatt, M. J.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Incompleteness of Atomic Structure Representations. *Physical Review Letters* **2020**, *125*, 166001.

(103) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-Inspired Structural Representations for Molecules and Materials. *Chemical Reviews* **2021**, *121*, 9759–9815.

(104) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics* **2011**, *134*, 074106.

(105) Pronobis, W.; Tkatchenko, A.; Müller, K.-R. Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules. *Journal of Chemical Theory and Computation* **2018**, *14*, 2991–3003.

(106) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer New York, 2009.

(107) Duvenaud, D.; Lloyd, J.; Grosse, R.; Tenenbaum, J.; Zoubin, G. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. Proceedings of the 30th International Conference on Machine Learning. Atlanta, Georgia, USA, 2013; pp 1166–1174.

(108) Bae, G.-T.; Aikens, C. M. Improved ReaxFF Force Field Parameters for Au–S–C–H Systems. *The Journal of Physical Chemistry A* **2013**, *117*, 10438–10446.

(109) Fihey, A.; Hettich, C.; Touzeau, J.; Maurel, F.; Perrier, A.; Köhler, C.; Aradi, B.; Frauenheim, T. SCC-DFTB parameters for simulating hybrid gold-thiolates compounds. *Journal of Computational Chemistry* **2015**, *36*, 2075–2087.

(110) Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications* **2020**, *247*, 106949.