JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

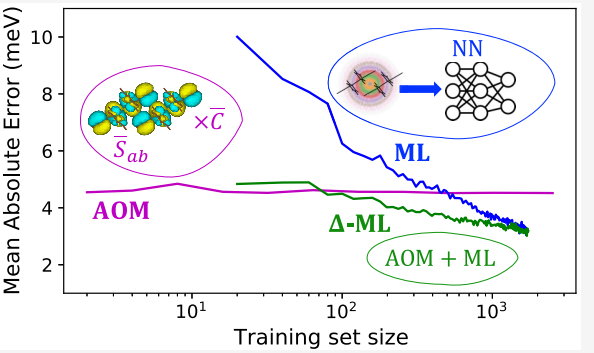
Ultrafast Electronic Coupling Estimators: Neural Networks versus Physics-Based Approaches

Roohollah Hafizi, Jan Elsner, and Jochen Blumberger*

Cite This: https://doi.org/10.1021/acs.jctc.3c00184 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Fast and accurate estimation of electronic coupling matrix elements between molecules is essential for the simulation of charge transfer phenomena in chemistry, materials science, and biology. Here we investigate neural-network-based coupling estimators combined with different protocols for sampling reference data (random, farthest point, and query by committee) and compare their performance to the physics-based analytic overlap method (AOM), introduced previously. We find that neural network approaches can give smaller errors than AOM, in particular smaller maximum errors, while they require an order of magnitude more reference data than AOM, typically one hundred to several hundred training points, down from several thousand required in previous ML works. A Δ-ML approach taking AOM as a baseline is found to give the best overall performance at a relatively small computational overhead of about a factor of 2. Highly flexible π-conjugated organic molecules like non-fullerene acceptors are found to be a particularly challenging case for ML because of the varying (de)localization of the frontier orbitals for different intramolecular geometries sampled along molecular dynamics trajectories. Here the local symmetry functions used in ML are insufficient, and long-range descriptors are expected to give improved performance.



Downloaded via UNIV COLLEGE LONDON on July 5, 2023 at 13:07:32 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

Charge transport simulations in biology and materials science typically begin with the calculation of electronic coupling matrix elements, or transfer integrals.[1−5] There have been significant advances in computing electronic couplings in the last 20 years. Depending on the requirements of the problem at hand, a large number of techniques are now available. The choice of method is dictated by various factors, most importantly by the right balance between accuracy and the associated computational cost. A number of approaches can be employed to accomplish this task: from high accuracy yet expensive ab initio calculations[6−8] to density functional theory (DFT) calculations (e.g., time-dependent DFT,[9] constrained DFT,[10−15] projector operator-based diabatization,[16−18] fragment-orbital DFT,[19,20] and frozen density embedding),[21,22] to fast semiempirical density functional tight binding (DFTB),[7,8,23] to the analytic overlap method (AOM).[24,25]

For a typical simulation of charge carrier transport in soft condensed media (e.g., organic and biological semiconductors) using, e.g., Kinetic Monte Carlo (KMC),[26−29] transient localization theory,[30,31] or non-adiabatic molecular dynamics (NAMD) simulations,[32−35] a very large number of transfer integrals must be evaluated before the simulation of charge mobility is converged. Some time ago, our group introduced the analytic overlap method (AOM), an ultrafast approach for the calculation of electronic coupling matrix elements for electron transfer between π-conjugated molecules. AOM allows one to estimate couplings to a useful degree of accuracy and about 10^5 times faster than with DFT calculations.[24] This method proposes to substitute the computationally expensive calculation of charge transfer integrals by an efficient calculation of the frontier molecular orbital (FMO) overlap integrals, multiplied by a suitable linear scaling coefficient. The FMOs are constructed using an optimized Slater-type orbital (STO) basis set, allowing ultrafast analytical calculations of FMO overlap integrals and electronic couplings for a variety of dimers.[24,36]

While AOM predicts electronic couplings to a useful degree of accuracy for applications in, e.g., KMC or NAMD[32] simulations, they are associated with an error because the relation between overlap and coupling is not strictly linear and the data exhibit a fair amount of scatter (see, e.g., Figure 2). Apart from this, challenging cases for AOM are flexible molecules that may adopt configurations that lead to significant changes in the localization/delocalization of the FMO. In this case the expansion coefficients of the FMOs have to be

Received: February 13, 2023

© XXXX The Authors. Published by American Chemical Society

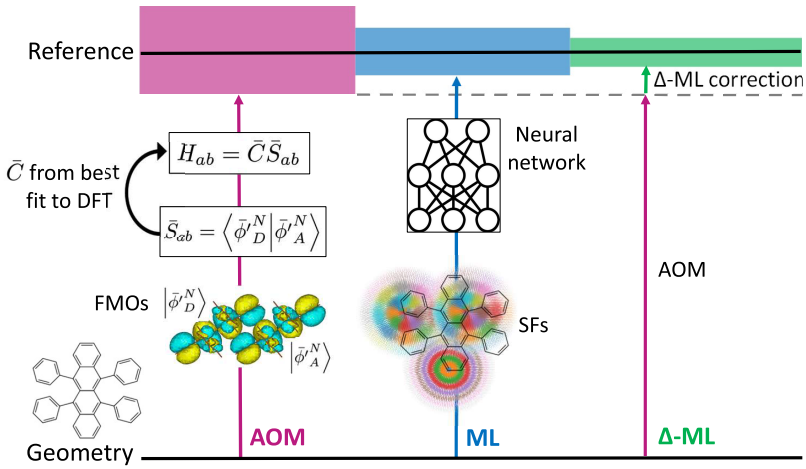


Figure 1. Three models are used in this work to calculate electronic coupling values between dimer molecules: (AOM, purple) The electronic overlap between approximated frontier molecular orbitals is determined, and utilized to estimate electronic couplings between dimer molecules. (ML, blue) The coordinates of dimer atoms are used to calculate the symmetry function and estimate the reference coupling value. (Δ-ML, green) A neural network is employed to correct the AOM model estimations to the reference values. To optimize the number of training points for each model, a few data sampling methods are studied.

reoptimized using expensive DFT calculations, which is not desirable. Besides, there may be distinct dimers in a unit cell with significantly different chemical interactions, necessitating multiple linear scaling constants. One such case, discussed below, is O-IDTBR, a molecule that belongs to the class of non-fullerene acceptors, a promising molecule for the organic photovoltaics industry. The purpose of this study is to explore the potential of atomic neural networks for machine learning of electronic couplings and for error estimation of the physics-based AOM (Δ-ML).

Various machine learning methods have been used to model the electronic coupling between molecular pairs. Musil et al.[37] used the Gaussian process (GP) to predict electronic couplings between rigid molecules based on their relative positions and orientations. A similar approach was taken by Lederer et al.,[38] who employed kernel ridge regression (KRR) to target rigid molecules. In a study by Bag et al.,[39] feature vectors were extracted from DNA via a coarse-grained model, and a neural network was trained to evaluate electronic couplings. Wang et al.[40] and Caylak et al.[41] used Coulomb matrices (CMs) as the molecular descriptor for training GP and deep neural networks, respectively. Also, Miller et al.[42] tried many ML methods, compared their performances, and suggested random forests as the most effective method. There are a number of shortcomings in previous ML models, including either a lack of accuracy in predictions, the necessity to freeze some degrees of freedom of the system, or the requirement for a large number of training (reference) data. We aim in this paper to address all of these issues by "semi-physical" modeling of electronic couplings and to compare our model's performance to that of previous models. The term "semi-physical" refers to the fact that we approximate electronic couplings as the sum of atomic contributions which are modeled by neural networks. At the current stage, ML models predict couplings only between chemically identical molecules, but in arbitrary atomic configurations. In other words, the ML model is not intended to make predictions for molecules other than those for which it has been trained but has the potential to be generalized.

In the following, after a short introduction to the methods, a protocol is developed for sampling reference data points required for the ML model that ensures completeness of sampling with the least number of data points. Figure 1 shows a schematic of the methods used in this work to predict electronic couplings. Two neural network models of electronic couplings of dimers are then trained on (1) DFT reference data points (Figure 1, blue) and (2) the difference between DFT reference data points and AOM (Figure 1, green) in a process called Δ-ML. Results are compared to those from the AOM model (Figure 1, purple). Using a rubrene dimer data set, these models are compared in terms of their performance. We then use the best model to study a challenging molecule for electronic coupling estimation, O-IDTBR. As a point of clarification, throughout the text, the terms "electronic coupling" and "transfer integral" are synonymous.

METHODS

The AOM method assumes a linear relationship between electronic coupling H_{ab} of two diabatic wavefunctions, ψ_a and ψ_b,

$$H_{ab} = \langle \psi_a | H | \psi_b \rangle \tag{1}$$

where H is the electronic Hamiltonian, and the corresponding wavefunction overlap,

$$S_{ab} = \langle \psi_a | \psi_b \rangle \tag{2}$$

such that

$$H_{ab} = CS_{ab} \tag{3}$$

Full calculation of eq 2 is computationally expensive, since it requires explicit diabatic wave functions ψ_a and ψ_b, for example, approximated by Kohn−Sham determinants obtained from constrained density functional theory.[15] Instead, the assumption is made that charge transport is mediated solely by the frontier molecular orbitals (FMOs) of the isolated molecules, ϕ'^{N}_{D} and ϕ'^{N}_{A} (notation as in refs 24 and 25). In the case of hole (electron) transport, these will correspond to the HOMO (LUMO) orbitals of the molecules. To further increase the efficiency of calculations, the FMOs are expressed in a minimum Slater-type orbital (STO) basis. The FMO of a single representative molecule, ϕ'^{N}_{l} (l = D or A), is calculated

https://doi.org/10.1021/acs.jctc.3c00184
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

from an explicit DFT calculation once and is projected onto a minimal STO basis to yield $\overline{\phi}'^{N}_{l}$:

$$|\overline{\phi}'^{N}_{l}\rangle = \sum_{k} c_k|\chi_k\rangle \approx |\phi'^{N}_{l}\rangle \tag{4}$$

where $|\chi_k\rangle$ is the kth STO orbital and c_k is the corresponding expansion coefficient. This allows for an ultrafast analytic calculation of the orbital overlap $\overline{S}_{ab} = \langle\overline{\phi}'^{N}_{D}|\overline{\phi}'^{N}_{A}\rangle$, since the overlap of STO basis functions is known analytically. Further details concerning the DFT calculation and projection to the STO basis can be found in refs 24 and 25. Importantly, orbital expansion coefficients are kept constant for different dimer geometries, while the direction of STO orbitals is updated according to the geometry of the molecule. The validity of this final approximation depends on the extent to which the localization/delocalization of the FMO is preserved for different molecular configurations and will be discussed further below. To account for our representation of the FMO in the STO basis, eq 3 is rewritten as

$$H_{ab} = \overline{C}\overline{S}_{ab} \tag{5}$$

where $\overline{C}$ is obtained from a best fit of $\overline{S}_{ab}$ to H_{ab} reference data computed at the explicit electronic structure level, typically DFT.

Unlike previous attempts at ML of electronic couplings, which map dimer descriptors in the input layer to a single value in the output, our approach approximates the total electronic coupling J_{ij} as the sum of contributions of atoms in monomer i and monomer j:

$$J_{ij} = \sum_{p}^{\substack{\text{all atoms}\\ \text{in dimer}}} J_p \tag{6}$$

This is in analogy to the AOM, where orbital overlap is calculated as a sum of overlap contributions from all atom pairs. Second-generation neural network potentials[43] use a similar approach to approximate energy. This method is included in the open-source code n2p2[44] and is used to train our neural network models. Each element (H, C, etc.) has a network, and atoms of the same element have the same weights and biases. The neural network of each element consists of two hidden layers, each with 20 nodes. Following the notation in ref 45, the neural network architecture is N−20−20−1, where N is the length of the atomic local-environment descriptor. The functional form of a neural network of an element is given by eq 7, and atoms in the same atom types share the same weights and biases. Each element (H, C, etc.) has a separate neural network with the following functional form:

$$J_p = f_1^3\left\{b_1^3 + \sum_{l=1}^{20} a_{l1}^{23}\cdot f_l^2\left[b_l^2 + \sum_{m=1}^{20} a_{ml}^{12}\cdot f_m^1\left(b_m^1 + \sum_{n=1}^{N} a_{nm}^{01}\cdot G_n^{(p)}\right)\right]\right\} \tag{7}$$

in which f's are activation functions (we used $f^1(x) = f^2(x) = \tanh(x)$ and $f^3(x) = x$), a_{kl}^{ij} is the weight connecting node k in layer i to node l in layer j, and b_j^i is the bias attached to node j in layer i. The values of the a_{kl}^{ij} and b_j^i are fitted during the training process. After training the network, the weights and biases remain fixed for all predictions.

At the heart of the function defined in eq 7 lies the N-dimensional structural descriptor vector:

$$G^{(p)} = \{G_n^{(p)}\} = \{\{G_p^2(\eta, R_s)\}, \{G_p^3(\eta, R_s, \zeta, \lambda)\}\},$$
$$n = 1, ..., N \tag{8}$$

The structural descriptors convert each dimer's atomic structure into a rotation-, translation-, and permutation-invariant input for the neural network. Within a cutoff radius of 8 Å, each atom's local environment is described by atom-centered symmetry functions (SFs). The radial environment of each atom is captured using radial symmetry functions:[44,46]

$$G_p^2 = \sum_{q\neq p}^{\substack{\text{all atoms}\\ \text{in dimer}}} e^{-\eta(R_{pq}-R_s)^2} f_C(R_{pq}) \tag{9}$$

where atom p is the central atom for which the symmetry function is calculated, R_s is the shift in the center of the Gaussian peak, η is the width of Gaussians, and $f_C(R_{pq})$ is the cutoff function:

$$f_C(R_{pq}) = \begin{cases} \frac{1}{2}\left[\cos\left(\frac{\pi R_{pq}}{R_c}\right) + 1\right] & \text{for } R_{pq} \le R_C \\ 0 & \text{for } R_{pq} > R_C \end{cases} \tag{10}$$

In order to obtain a better description of the atomic environment, radial symmetry functions are calculated for all element doublets (CC, CH, HC, HH, etc.) We use eight radial symmetry functions for each pair of elements, whose η and R_s parameters are determined using the method introduced by Imbalzano et al.[47] Thus, there are $8 \times N_E$ radial symmetry functions for each atom in a system with N_E elements. For each element triplet, angular functions of the following form are generated to describe the angular environment of each atom:[44,46]

$$G_p^3 = 2^{1-\zeta}\sum_{\substack{q,r\neq p\\ q<r}}^{\substack{\text{all atoms}\\ \text{in dimer}}} (1 + \lambda\cos\theta_{pqr})^{\zeta} e^{-\eta[(R_{pq}-R_s)^2+(R_{pr}-R_s)^2+(R_{qr}-R_s)^2]}$$
$$\times f_C(R_{pq}) f_C(R_{pr}) f_C(R_{qr}) \tag{11}$$

where atom p is the central atom and θ_{pqr} is the angle formed by atoms p, q, and r. Like radial symmetry functions, angular symmetry functions are calculated for all element triplets (CCC, CCH, CHH, etc.) to better describe the environment around the atoms. For each element triplet, there are two R_s values, two ζ values, and two λ values, resulting in eight angular symmetry functions that are automatically selected.[47] As there are $N_E(N_E + 1)/2$ element triplets, each atom will have $8 \times N_E(N_E + 1)/2$ angular descriptors.

A total of 3612 rubrene dimer geometries were taken from an ab initio molecular dynamics trajectory using the optPBE-vdW density functional,[48] a DZVP basis set,[49] and GTH pseudopotentials.[50] Snapshots from four distinct dimer pairs within a supercell of 12 molecules were taken at 50 fs intervals. Further computational details can be found in ref 51. A total of 5770 O-IDTBR dimer pairs were taken from classical molecular dynamics trajectories using a force field parametrized specifically for the family of IDTBR non-fullerene acceptors.[52] All reference DFT electronic couplings were calculated using the projector-operator-based diabatization (POD) method[17] in conjunction with the Perdew−Burke−Ernzerhof (PBE) density functional and a uniform scaling

https://doi.org/10.1021/acs.jctc.3c00184
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

constant of 1.325. This is referred to as the sPOD/PBE method. The scaling factor was obtained from the best fit to ab initio reference values for the HAB79 database of organic dimers.[53]

RESULTS AND DISCUSSION

Optimal Sampling Protocol for AOM Fitting. As a starting point, we present a protocol for fitting the AOM. A representative collection of molecular dimers is sampled from molecular dynamics trajectories to fit the linear scaling relation between AOM overlap and sPOD/PBE electronic coupling (eq 5). Depending on the complexity of the physical system, there may be hundreds to thousands of points sampled.[25,32,51] For crystalline systems, the sampled dimers are arranged in a few clusters that contain fairly similar dimers. Dimers within a cluster may have a wide range of electronic couplings, despite being visually similar. Consequently, chemical intuition cannot easily determine whether the dimer space is undersampled or oversampled. Undersampling of the data limits the validity of the model, whereas oversampling of the data increases the cost of reference electronic coupling calculations. A further consequence of oversampling data is that it may result in inconsistent AOM scaling constants depending on which part of the dimer space has been oversampled. As a means of addressing these issues, we propose a method of systematic sampling that ensures convergence of the scaling factor with a small number of samples.

From a geometrical perspective, the AOM reference data set should represent all possible dimers. It is therefore necessary to use a geometrical descriptor that is capable of accurately capturing the similarity of the structures in order to select the most different structures. In this study, we utilize the average minimum distance (AMD),[54] a geometrical descriptor that has recently been proposed. AMD is rotation-, translation-, and permutation-invariant; it is also stable and computationally efficient. As a test of its quality, we clustered rubrene dimers extracted from an ab initio MD simulation trajectory for a rubrene molecular crystal at room temperature.[51] The result of clustering using the HDBSCAN clustering algorithm[55] are shown in Figure 2.

In HDBSCAN, the distance metric is the Euclidean distance between AMD descriptors. The data set contains 3612 data points arranged in two clusters with size 1806. These two clusters represent the electronic couplings in the a and b crystallographic directions, respectively; they are shown by circles and squares in Figure 2. Using the AMD descriptor, HDBSCAN divides the data set into two clusters indicated by the green and blue colors. It is important to note that no data points remain unclustered and that both clusters are detected correctly without any errors, indicating that AMD is a good descriptor to capture the underlying order of the data.

To sample the reference data points for AOM fitting, AMD is used as a descriptor, and the farthest point sampling (FPS) algorithm[56,57] is applied. This is an iterative optimization strategy that selects the most diverse data points from already-selected data. With a good descriptor, such as AMD, the dimer configurational space will be sampled in a uniform manner such that oversampling and undersampling are prevented. This allows us to sample dimers with the greatest geometric diversity with the smallest number of data points. Figure 3a illustrates the convergence of fitting the AOM's scaling factor, $\bar{C}$, with respect to the size of the training set when data points are sampled randomly (black) or by AMD+FPS (red). As the

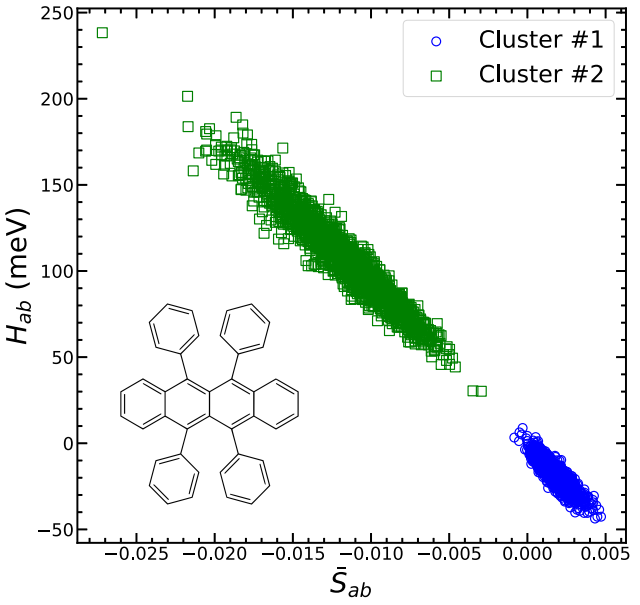


Figure 2. Clustering of the rubrene DFT reference data set using HDBSCAN and AMD descriptor. Two clusters are detected and shown as green squares and blue circles, corresponding to orbital overlap ($\bar{S}_{ab}$) and electronic coupling (H_{ab}) in rubrene dimers along the crystallographic directions a and b, respectively.

number of data points sampled by FPS is increased from 32 to 64, the fitted $\bar{C}$ value differs by less than 1%, and the value can be considered converged for any practical purposes. The $\bar{C}$ value obtained from random sampling depends strongly on the particular sample chosen, especially when training set sizes are small. In Figure 3a we plot the average $\bar{C}$ value obtained from 10 random samples for each training set size, and the root-mean-square deviation of the $\bar{C}$ value across these 10 samples is indicated by error bars. The data show that random sampling can lead to large errors and cannot guarantee a reliable $\bar{C}$ value, while FPS is very robust, even at small sample sizes, and provides an optimal sampling strategy. The completeness of the sampled set is also verified by the convergence of the mean absolute error (MAE), the maximum absolute error (MAX), and the mean unsigned relative error (MURE) in Figure 3b−d. Errors were calculated over 30% of the data that was not used for sampling the training set for fitting $\bar{C}$. The protocol for optimal sampling for AOM is summarized as follows:

1. Run a long-enough MD simulation, on the order of 100 ps to 1 ns.
2. Sample dimers at a frequency which accurately samples the fluctuations of electronic couplings, typically on the order of 100 fs for molecular crystals.
3. Sort the initial data set by FPS using AMD geometrical descriptor.
4. Pick n new data points from the sorted list, calculate the reference ab initio electronic couplings, and include them in the AOM reference data.
5. Fit the scaling constant of AOM, $\bar{C}$.
6. Repeat steps 4 and 5 until the change in $\bar{C}$ is less than 1%.

Using this method, the smallest set of structures with maximum geometrical diversity is collected for AOM, thereby saving a considerable amount of computational time.

Machine Learning Models. While FPS is a good sampling method for selecting data to parametrize the AOM, it is not the

https://doi.org/10.1021/acs.jctc.3c00184
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

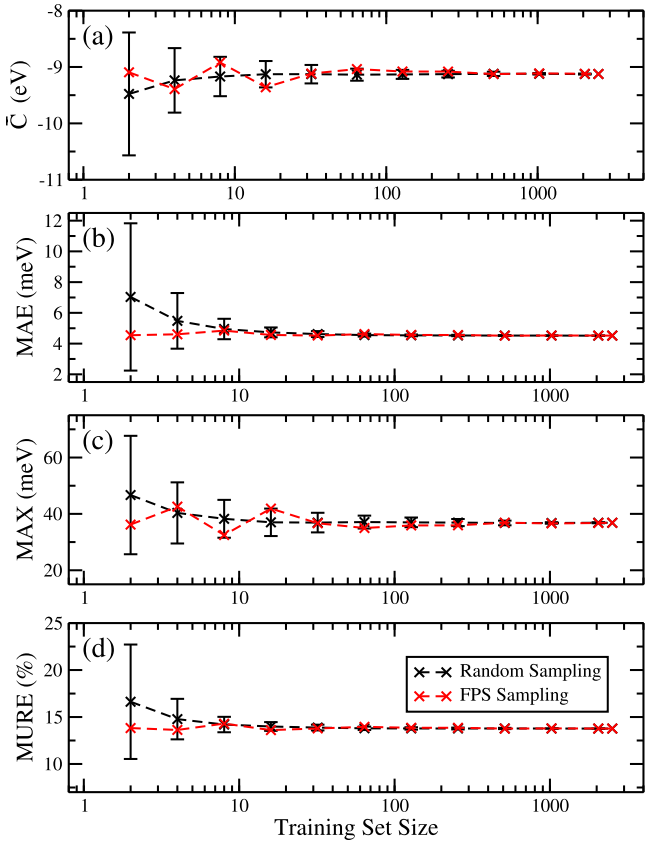


Figure 3. Convergence of AOM scaling constant $\bar{C}$ in rubrene and error metrics with training set size: (a) the scaling constant of the AOM model, (b) the mean absolute error of predictions on the test set, (c) the maximum error of predictions, and (d) the mean unsigned relative error of predictions using either random sampling (black) or the FPS algorithm (red). An error bar indicates one standard deviation based on 10 randomly selected training sets of the corresponding size.

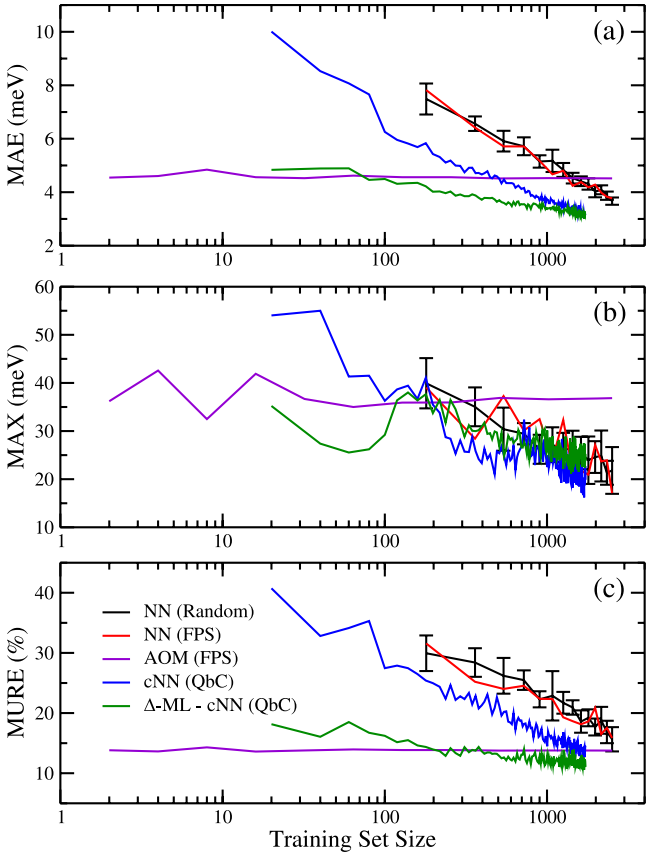


Figure 4. ML of electronic couplings of rubrene dimers. (a) MAE, (b) MAX, and (c) MURE vs training set size when electronic couplings are evaluated by AOM model (purple), a neural network with randomly sampled training set (black), a neural network with FPS-sampled training set (red), a QbC-sampled committee neural network (cNN) (blue), and a QbC-sampled cNN trained on the difference between reference data and AOM values (green).

best algorithm for preparing the training set for neural network (NN) models where larger amounts of data are required compared to the AOM. This was determined by comparing the learning curve of a neural network model for electronic couplings between rubrene dimers when the training set was sampled randomly versus when it was sampled by FPS, as described in the previous section. This comparison is shown in Figure 4 and indicates only a very slight difference between random sampling (black) and FPS sampling (red). FPS sampling was also performed with atomic descriptors based on symmetry functions, and the results were minimally improved compared to random sampling (see Figure S2). In addition to having no advantage over random sampling, FPS sampling also results in larger errors when training sets are small. Consequently, sampling training sets based on geometrical diversity does not necessarily result in improved neural networks. It is evident that both the black and red learning curves show that the NN is capable of achieving a higher level of accuracy than the AOM model (Figure 4, data in purple), in particular with regard to the MAX error, which is reduced by a factor of about 2. However, a rather large reference data set of more than 1000 DFT electronic couplings is needed to outperform AOM. While for rubrene dimers this is computationally manageable, for larger molecules or for systems with a larger number of nearest-neighbor couplings, a more data efficient NN method is desirable.

Active learning methods, in particular committee neural networks (cNN),[58] can often provide greater data efficiency. For this purpose, a committee of N neural networks are trained on slightly differing training sets in order to sample different parts of the hyperdimensional space of neural network weights. The committee is used to make predictions on data that have not been incorporated into the training set. Those data points with the highest level of disagreement, as measured by the standard deviation of the committee's prediction, are added to the training set. This so-called query by committee (QbC) process is repeated iteratively until the disagreement between committee members on the pool of unseen data converges to that of the training set. Detailed information about the approach can be found in the work of Schran et al.[58]

In this study, we utilize a committee of eight neural networks in order to learn the electronic couplings of rubrene dimers; 70% of the data is used for training, and 30% is used for testing. Initially, 20 dimers are selected randomly from the training pool. Each committee member is trained on 80% of this data (16 data points out of 20), such that each committee member sees a slightly different training set. The committee members are applied to the remainder of the training pool, and the 20 structures with the highest disagreement between committee members are added to the committee training pool. This

https://doi.org/10.1021/acs.jctc.3c00184
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

procedure is repeated iteratively until the disagreement between committee members on the training pool and training set data converges. We differ from the work of Schran et al.[58] in that our disagreement score is based on the standard deviation of predicted electronic couplings.

Query by committee takes 2−3 times less data to beat AOM in terms of MAE than random sampling (Figure 4, blue lines). In addition, the cNN method allows us to achieve lower MAE than random sampling: ∼3.2 meV with around 1600 data points selected by the cNN versus ∼3.8 meV when all the training data are included. This indicates that there is hidden redundancy in the data set that biases the model to specific dimer configurations when all of the data are used. The MAX and MURE also decrease faster when QbC is used.

The linear relationship between electronic coupling and orbital overlap shown in Figure 2 indicates that the underlying assumptions of the AOM are a good approximation for this system. However, there is still some scatter. In a Δ-ML approach, we use AOM as a baseline and attempt to learn the difference between DFT electronic coupling and AOM values (scatter or deviation from linear relation) using neural networks. This approach has proven very successful when there is a high correlation between a computationally inexpensive baseline and the target accuracy.[59] AOM provides an excellent baseline for this purpose due to its high correlation with the reference data, R^2 = 0.992, as shown in Figure 2.

A cNN model is trained on the difference between DFT electronic couplings and the AOM value (Δ-ML) using the iterative QbC sampling approach described above. The learning curves are shown in Figure 4 (green lines); about an order of magnitude less data than using standard cNN is required to outperform AOM, a significant improvement. Furthermore, with only 500 dimers used for training, this is the only model that achieves a better MURE (∼11%) than AOM (∼14%). It is worth noting that if the test set includes many reference values close to zero, the MURE will be very large, since small absolute error translates into a large relative error. This is less of a problem for AOM, since the assumed AOM relation eq 5 passes through the origin. We therefore benefit from using AOM as a baseline in order to improve the MURE of predictions.

In the current work, the electronic coupling has been approximated as a sum of atomic contributions, which is the most straightforward approximation if one writes the density in terms of a sum of atomic densities (with any partitioning algorithm). The works of Wang et al.[40] and Caylak et al.[41] are similar in this regard, as both use Coulomb matrices[60] (CMs) as descriptors. CMs describe a system via inverse distances between atoms but do not include higher-order terms. This method is fast to compute and easy to implement, and it allows the reconstruction of an atomistic system using a least-squares approach.[60] Although it uses atomic numbers directly to encode elements, it suffers from discontinuities in the sorted version or from information loss in the diagonalized version since its eigenspectrum is not unique. We instead assign different symmetry functions to different element pairs and triplets depending on their type, both for pairwise radial (G^2) and triplewise angular (G^3) symmetry functions. The use of such descriptors allows for a more accurate description of the chemical environment and thus for a faster and better training of the corresponding model. As opposed to the findings of Musil et al.,[37] we demonstrated that active learning by QbC is a more successful sampling strategy than FPS, at least when neural networks are the ML method of choice. As shown in Figures 4 and S2, the use of FPS did not provide any significant improvement over random sampling when geometrical descriptors, such as AMD and SFs, were used as the input.

A Challenging Case: O-IDTBR. A case study is presented to illustrate the use of the presented machine learning approach to estimate electronic couplings in O-IDTBR. This material belongs to the class of non-fullerene acceptors (NFAs), which have recently attracted significant interest in organic photovoltaics (OPVs). Their contribution to record OPV efficiencies (currently 19%) results from a number of inherent, desirable NFA characteristics, such as synthetically tunable optical properties, improved long-term morphological stability, and high charge carrier mobility (μ).[61,62] In comparison with other NFAs, O-IDTBR exhibits a superior structure/packing motif, resulting in a relatively high electron mobility for this class of materials.[63]

A total of 5770 dimer configurations were taken from classical molecular dynamics trajectories using a force field specifically parametrized for the family of IDTBR NFAs.[52] This data set contains four distinct dimer types, the structures of which are shown in Figure S4. Figure 5 shows electronic

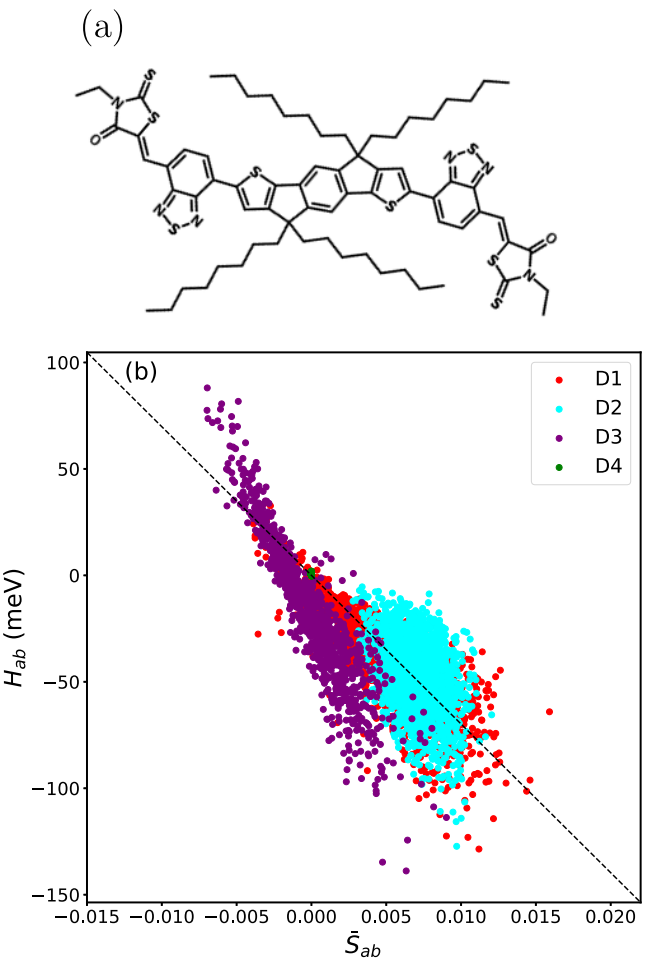


Figure 5. (a) Chemical structure of O-IDTBR. (b) Scatter plot of electronic coupling from DFT calculations (sPOD method) vs orbital overlap from reconstructed FMOs with fixed expansion coefficients. Calculations were carried out for dimers extracted from classical MD simulation of the O-IDTBR crystal. D1 to D4 denote different dimer orientations in the crystal structure.

https://doi.org/10.1021/acs.jctc.3c00184
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

coupling versus orbital overlap for the data set, where overlap values were calculated using the fixed expansion coefficients of the O-IDTBR molecule at the minimum-energy configuration. Different dimer types are labeled by color. Due to a large center-of-mass distance between monomers, overlaps and electronic couplings for D4 dimers are close to zero. Clearly, the deviation from the linear relationship between overlap and coupling is now very significant, and the scatter is much larger than for rubrene. Moreover, it appears that the data would be better described by two slopes, one for dimer geometries labeled D3 and another for the other dimer types. The physical reason for the large scatter is discussed below. We would expect ML models to give a more significant improvement over AOM than for rubrene.

We started by testing the AMD descriptor's ability to assign molecular dimers to one of the four dimer types. We used the HDBSCAN algorithm, AMD descriptors, and Euclidean distances as described previously, results of which are presented in Figure S5. Once again, clustering is successful without any errors or unclustered data points. By randomly assigning 5270 data points to the training set and the remaining 500 data points to the test set, we applied the sampling protocol introduced previously to fit the AOM model. A converged $\bar{C}$ value of −7070 meV was obtained after adding 200 dimers to the training set (see Figure S7). This model results in an MAE of 12.4 meV, a MAX of 94 meV, and a MURE of 62% on the test set (see Figure 6, purple lines).

Figure 5 clearly shows that the correlation between the AOM and reference couplings (R^2 = 0.67) is less strong than for rubrene (R^2 = 0.99). We train cNN models using both direct learning and Δ-learning as described previously. The parameters for the atomic environment descriptors, i.e., symmetry functions, are the same as those for rubrene. O-IDTBR consists of five chemical elements, with each element having 40 radial and 120 angular symmetry functions. Therefore, the dimension of the descriptor is 3 times larger than for rubrene. In order to reduce the computational costs, we first pruned all symmetry functions with values below 10^{-4}. We then trained a simple model on a small subset of the data set and carried out a sensitivity analysis to remove those symmetry functions for which the output layer of the NN is less than 0.4% sensitive to their gradients. Accordingly, we have 111, 127, 119, 85, and 128 symmetry functions for hydrogen, carbon, nitrogen, oxygen, and sulfur atoms, respectively. Using this set of atomic descriptors in the input layer of committees of eight neural networks, we train two cNN models: one is trained on electronic couplings directly, and the other is trained on the correction to the fitted AOM model (Δ-ML).

In the plots of Figure 6, results for the AOM, the direct learning model (cNN), and the Δ-learning model are shown in violet, blue, and green lines, respectively. The AOM results converge fast when the sampling protocol for AOM fitting is employed. Δ-ML outperforms direct learning in terms of both convergence behavior and accuracy metrics. With less than 100 data points, Δ-ML improves on AOM in both the MAE metric (3 meV lower than AOM) and, most significantly, in the MAX error metric (factor of 2 lower than AOM). However, the Δ-ML estimates have larger errors in the MURE metric compared to AOM. As discussed above, this is due to the fact that small absolute errors translate to large relative errors for couplings that are close to zero. However, since the AOM linear scaling relation passes through the origin by definition,

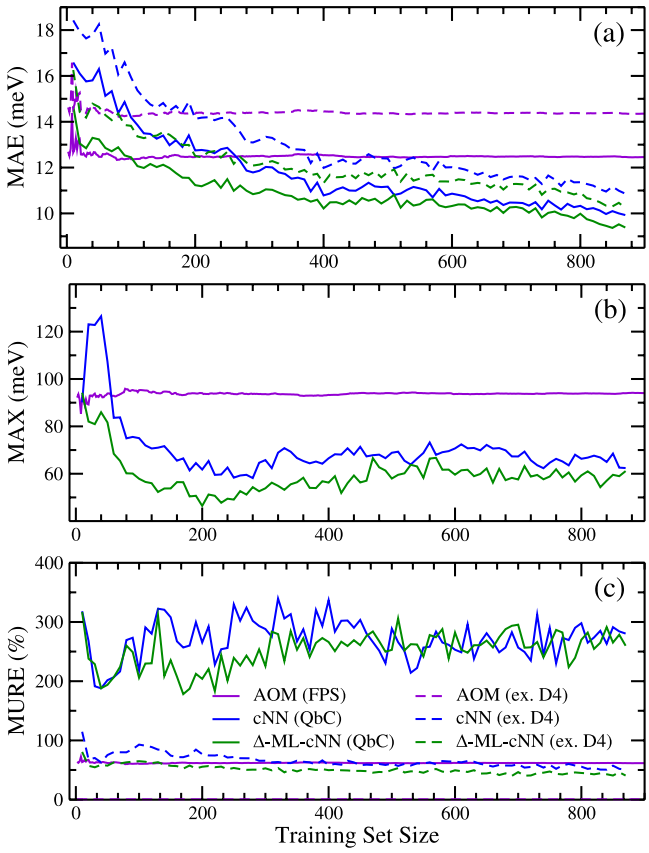


Figure 6. ML of electronic couplings for O-IDTBR. (a) MAE, (b) MAX, and (c) MURE vs training set size when electronic couplings are evaluated by AOM model (purple), a QbC-sampled committee neural network (blue), and a QbC-sampled committee neural network trained on the difference between reference data and AOM values (green). Solid and dashed lines represent test sets that include and exclude D4 dimers.

this problem is somewhat alleviated in this method. We find that the large MURE comes from dimers in the D4 cluster (see Figure S8), which have rather small couplings spread around zero. If these dimers are excluded, the MURE drops to less than 50% for both ML methods (Figure 6c, dashed lines in blue and green).

We now would like to explain the large scatter between overlap and electronic coupling shown in Figure 5. The main reason for its limited accuracy is the use of fixed projection coefficients (same set of c_k's in eq 4 for all geometries) in the AOM model. O-IDTBR exhibits strong thermal fluctuations of the dihedral angles that connect the different molecular units, which results in significant variations in the degree of (de)localization of the FMO along the MD trajectory. Representing the FMOs using fixed expansion coefficients may therefore not be as successful for O-IDTBR as for rubrene. We have illustrated three cases in Figure 7 in which the expansion coefficient of an sp^2 carbon (marked by an arrow) is minimum (0.077), average (0.254), and maximum (0.409). These states correspond to configurations where the FMO (LUMO) of the molecule is preferentially localized on the right-hand side, approximately equally distributed, and preferentially localized on the left-hand side. Obviously, fixed expansion coefficients are no longer a reasonable approximation for this molecule. It results in the very strong scatter shown in Figure 5.

https://doi.org/10.1021/acs.jctc.3c00184
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

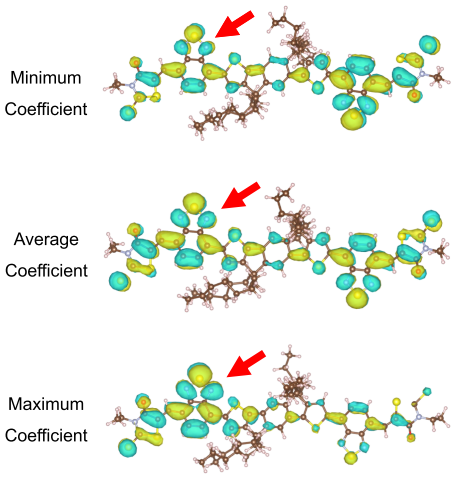


Figure 7. Three cases were the LUMO of an O-IDTBR monomer is localized mostly over the right side (top), both sides (middle), and left side (bottom) of the monomer. Isosurfaces with isovalues of 0.015 are shown and were generated with the VESTA software.[64]

In Figure 8, a subset of 200 randomly selected dimers from Figure 5 are replotted (fixed expansion coefficients, panel a) and compared to the results obtained after reoptimization of the expansion coefficients (panel b). Using the optimized expansion coefficients greatly reduces scatter in the data, resulting in errors that are similar to the ones for rubrene. However, in a charge transport simulation, it is not practical to calculate the DFT FMOs of each monomer at each time step due to the high computational cost involved.

Evidently, ML of the effects of variations in the orbital delocalization on electronic coupling is more challenging than the learning of mainly geometry-based changes as in rubrene. This explains the larger errors for O-IDTBR on all error metrics. The proposed ML models are based on the sum of atomic contributions (eq 6), and the environment around each atom is described by neighbors within the cutoff radius. However, while the local environment of the marked atom is very similar in all three cases of Figure 7, the character of the FMO differs substantially. Changes in FMO may be due to long-range effects not well described by short-range symmetry functions. One would require a cutoff radius of at least 30 Å in order to account for such long-range effects. However, the cost of calculating symmetry functions is the bottleneck for the efficiency of ML models, and such large cutoffs would not be practical.

The development of a model to accurately and efficiently estimate expansion coefficients is a promising avenue for further research, based on the observed improvement in results shown in Figure 8.

Computational Cost. Above we demonstrated that Δ-ML provides a significant improvement to the AOM estimation of electronic couplings. During charge transfer simulations, these estimations are performed millions of times, so their overhead must be kept to a minimum. To benchmark the efficiency of ML models, the calculation time was measured on a single core of an Intel Xeon CPU E5-2650 v4 @ 2.20 GHz. In the case of rubrene, the trained machine learning model requires 40 ms to evaluate electronic coupling for a single dimer, whereas AOM requires 26 ms. This relatively small cost overhead associated with improving accuracy will not be a bottleneck in practical applications. To further improve efficiency, hydrogen atoms

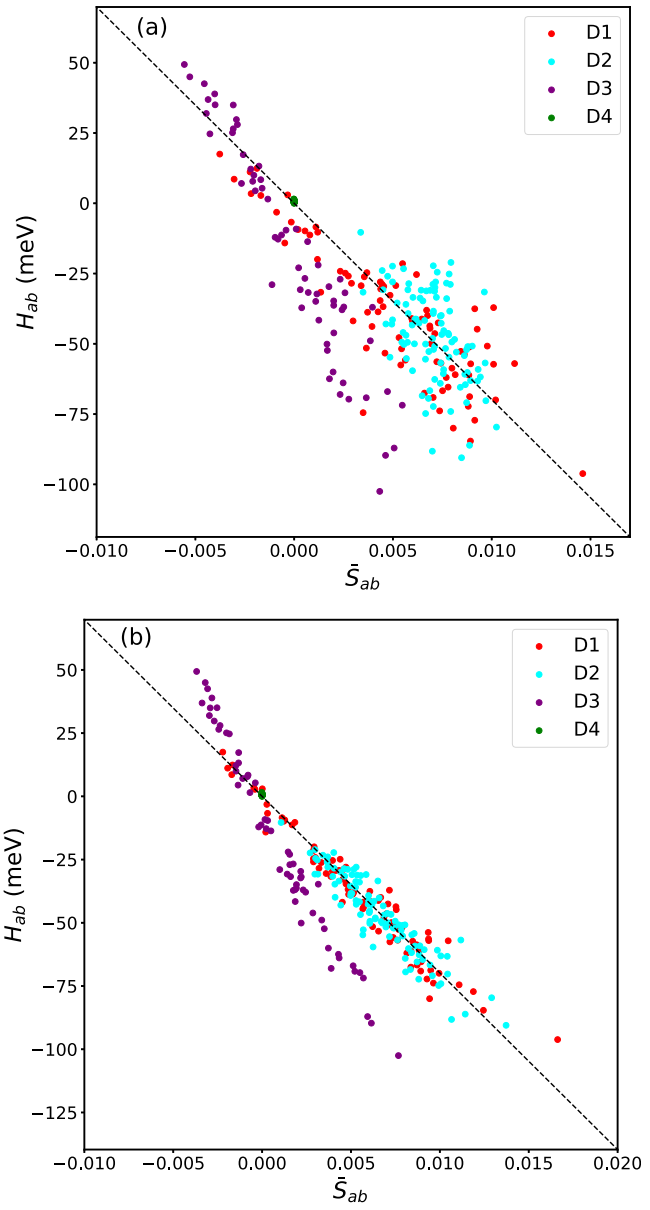


Figure 8. Scatter plots of H_{ab} vs $\bar{S}_{ab}$ for O-IDTBR using (a) fixed expansion coefficients for FMO reconstruction (200 data points taken from Figure 5b) and (b) reoptimized expansion coefficients for each data point. Note the reduction in scatter after reoptimization.

can be removed from the current model, since they have very small contribution to the electronic coupling and play a negligible role in describing the local environment in symmetry functions. The removal of hydrogens from rubrene ($C_{42}H_{28}$) results in 40% fewer atoms being included in the summations in eq 9 and eq 11 when SFs are calculated, resulting in a 3-fold improvement in efficiency, i.e., 14 ms/prediction, with minimal loss in accuracy (see Figure S3). For O-IDTBR dimers, using a single core, the AOM takes ∼190 ms for a single electronic coupling calculation, whereas the Δ-ML model takes ∼130 ms. Again, the cost overhead of improving the AOM model by ML would not be the bottleneck in practical charge transport simulations.

https://doi.org/10.1021/acs.jctc.3c00184
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

CONCLUSIONS

We have presented a set of tools that facilitate ultrafast estimation of electronic couplings between molecules, which is necessary for the simulation of charge transport in organic semiconductors. Initially, a sampling protocol was developed by combining a geometrical descriptor, AMD, and farthest point sampling to sample the reference data for fitting AOM models. Using this protocol ensures convergence of AOM fitting with a small number of reference data calculations, eliminating the need for chemical intuition. Various neural network models and sampling methodologies were examined to model the electronic couplings of rubrene dimers, and it was determined that Δ-ML committee neural networks (cNNs) and query by committee (QbC) sampling provided the most accurate results. By exploiting the physics-motivated AOM model as a baseline, the Δ-ML was trained on the difference between AOM and reference data (sPOD/PBE). This approach allows us to achieve similar or better accuracy than previous efforts[37−42] with training sets that are at least an order of magnitude smaller than those used previously.

With this approach, the accuracy of electronic coupling predictions is improved according to all error metrics employed, at the cost of doubling the computational time. It is important to note, however, that the level of improvement is influenced by the baseline. An AOM model's cost efficiency comes from its use of fixed FMO projection coefficients for the molecule in question. This is found to be a good approximation in rubrene, where the AOM overlaps exhibit a strongly linear correlation with reference electronic couplings. However, in cases where the FMO character of the molecule changes significantly during dynamics, as in O-IDTBR, the AOM will provide a less accurate baseline. Using an additional machine learning model to estimate the expansion coefficients of the FMOs of molecules will provide higher-quality results and serve as a more reliable baseline for Δ-ML. We are currently working along these lines.

Concerning the active learning method, the use of uncertainty as the metric or score, as in QbC, is very popular; however, other scores can also be utilized. The expected model output change (EMOC) is another score recently adopted in the chemical community.[65] With this method, AL will select data points with the largest expected change in model output. Considering that the current study does not compare active learning methods and that QbC performs well, such a comparison needs to be conducted in future studies.

Finally, in terms of generalizability of the model, we expect it to be generalizable for cases such as π-conjugated molecules, where the FMO of a molecule can be approximated by its local geometry. When FMOs have nonlocal dependencies, however, a new model must be trained from scratch. Constructing an AOM+Δ-ML correction is straightforward since with the proposed protocol for fitting AOM models and QbC sampling for training ML models, a small number of reference data point calculations and minimum human intervention are required. In addition, newly visited dimer configurations will be detected during the course of the application by examining the standard deviation of ML predictions. The extrapolation occurrence will be handled automatically by retraining the ML model using such data points.

ASSOCIATED CONTENT

Data Availability Statement

Sampling and prediction codes can be found in https://github.com/blumberger/ml_electronic_coupling.git.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c00184.

Further information about the data sets, clustering, geometry of dimers, and other sampling methods (PDF)

AUTHOR INFORMATION

Corresponding Author

Jochen Blumberger − *Department of Physics and Astronomy and Thomas Young Centre, University College London, London WC1E 6BT, United Kingdom;* orcid.org/0000-0002-1546-6765; Email: j.blumberger@ucl.ac.uk

Authors

Roohollah Hafizi − *Department of Physics and Astronomy and Thomas Young Centre, University College London, London WC1E 6BT, United Kingdom;* orcid.org/0000-0001-6513-4446

Jan Elsner − *Department of Physics and Astronomy and Thomas Young Centre, University College London, London WC1E 6BT, United Kingdom;* orcid.org/0000-0002-3685-3940

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c00184

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Dr. Ljiljana Stojanovic for providing MD-sampled structures for O-IDTBR. R.H. was supported by the European Research Council (ERC) under the European Union Horizon 2020 Research and Innovation Programme (Grant Agreement 682539/SOFTCHARGE). J.E. was supported by a departmental Ph.D. studentship. Through our membership in the UK's HEC Materials Chemistry Consortium [funded by the EPSRC (Grants EP/L000202 and EP/R029431)], this work used the ARCHER2 UK National Supercomputing Service (http://www.archer2.ac.uk) as well as the UK Materials and Molecular Modeling (MMM) Hub [partially funded by the EPSRC (Grant EP/P020194)] for computational resources. The authors also acknowledge UCL for access to the High Performance Computing Facility Kathleen.

REFERENCES

(1) Oberhofer, H.; Reuter, K.; Blumberger, J. Charge transport in molecular materials: An assessment of computational methods. *Chem. Rev.* **2017**, *117*, 10319−10357.

(2) Gryn'ova, G.; Nicolaï, A.; Prlj, A.; Ollitrault, P.; Andrienko, D.; Corminboeuf, C. Charge transport in highly ordered organic nanofibrils: lessons from modelling. *J. Mater. Chem. C* **2017**, *5*, 350−361.

(3) Gryn'ova, G.; Lin, K.-H.; Corminboeuf, C. Read between the molecules: computational insights into organic semiconductors. *J. Am. Chem. Soc.* **2018**, *140*, 16370−16386.

(4) Nematiaram, T.; Ciuchi, S.; Xie, X.; Fratini, S.; Troisi, A. Practical computation of the charge mobility in molecular semiconductors using transient localization theory. *J. Phys. Chem. C* **2019**, *123*, 6989−6997.

(5) Jiang, X.; Burger, B.; Gajdos, F.; Bortolotti, C.; Futera, Z.; Breuer, M.; Blumberger, J. Kinetics of trifurcated electron flow in the decaheme bacterial proteins MtrC and MtrF. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 3425−3430.
(6) Cave, R. J.; Newton, M. D. Calculation of electronic coupling matrix elements for ground and excited state electron transfer reactions: comparison of the generalized Mulliken−Hush and block diagonalization methods. *J. Chem. Phys.* **1997**, *106*, 9213−9226.
(7) Kubas, A.; Hoffmann, F.; Heck, A.; Oberhofer, H.; Elstner, M.; Blumberger, J. Electronic couplings for molecular charge transfer: Benchmarking CDFT, FODFT, and FODFTB against high-level ab initio calculations. *J. Chem. Phys.* **2014**, *140*, 104105.
(8) Kubas, A.; Gajdos, F.; Heck, A.; Oberhofer, H.; Elstner, M.; Blumberger, J. Electronic couplings for molecular charge transfer: benchmarking CDFT, FODFT and FODFTB against high-level ab initio calculations. II. *Phys. Chem. Chem. Phys.* **2015**, *17*, 14342−14354.
(9) Manna, D.; Blumberger, J.; Martin, J. M.; Kronik, L. Prediction of electronic couplings for molecular charge transfer using optimally tuned range-separated hybrid functionals. *Mol. Phys.* **2018**, *116*, 2497−2505.
(10) Wu, Q.; Van Voorhis, T. Extracting electron transfer coupling elements from constrained density functional theory. *J. Chem. Phys.* **2006**, *125*, 164105.
(11) de la Lande, A.; Salahub, D. R. Derivation of interpretative models for long range electron transfer from constrained density functional theory. *J. Mol. Struct.: THEOCHEM* **2010**, *943*, 115−120.
(12) Oberhofer, H.; Blumberger, J. Electronic coupling matrix elements from charge constrained density functional theory calculations using a plane wave basis set. *J. Chem. Phys.* **2010**, *133*, 244105.
(13) McKenna, K. P.; Blumberger, J. Crossover from incoherent to coherent electron tunneling between defects in MgO. *Phys. Rev. B* **2012**, *86*, 245110.
(14) Gillet, N.; Berstis, L.; Wu, X.; Gajdos, F.; Heck, A.; de la Lande, A.; Blumberger, J.; Elstner, M. Electronic coupling calculations for bridge-mediated charge transfer using constrained density functional theory (CDFT) and effective hamiltonian approaches at the density functional theory (DFT) and fragment-orbital density functional tight binding (FODFTB) level. *J. Chem. Theory Comput.* **2016**, *12*, 4793−4805.
(15) Blumberger, J.; McKenna, K. P. Constrained density functional theory applied to electron tunnelling between defects in MgO. *Phys. Chem. Chem. Phys.* **2013**, *15*, 2184−2196.
(16) Kondov, I.; Čížek, M.; Benesch, C.; Wang, H.; Thoss, M. Quantum Dynamics of Photoinduced Electron-Transfer Reactions in Dye−Semiconductor Systems: First-Principles Description and Application to Coumarin 343−TiO_2. *J. Phys. Chem. C* **2007**, *111*, 11970−11981.
(17) Futera, Z.; Blumberger, J. Electronic couplings for charge transfer across molecule/metal and molecule/semiconductor interfaces: Performance of the projector operator-based diabatization approach. *J. Phys. Chem. C* **2017**, *121*, 19677−19689.
(18) Ghan, S.; Kunkel, C.; Reuter, K.; Oberhofer, H. Improved projection-operator diabatization schemes for the calculation of electronic coupling values. *J. Chem. Theory Comput.* **2020**, *16*, 7431−7443.
(19) Senthilkumar, K.; Grozema, F.; Bickelhaupt, F.; Siebbeles, L. Charge transport in columnar stacked triphenylenes: Effects of conformational fluctuations on charge transfer integrals and site energies. *J. Chem. Phys.* **2003**, *119*, 9809−9817.
(20) Oberhofer, H.; Blumberger, J. Insight into the Mechanism of the Ru^{2+}−Ru^{3+} Electron Self-Exchange Reaction from Quantitative Rate Calculations. *Angew. Chem., Int. Ed.* **2010**, *49*, 3631−3634.
(21) Pavanello, M.; Van Voorhis, T.; Visscher, L.; Neugebauer, J. An accurate and linear-scaling method for calculating charge-transfer excitation energies and diabatic couplings. *J. Chem. Phys.* **2013**, *138*, No. 054101.
(22) Ramos, P.; Papadakis, M.; Pavanello, M. Performance of frozen density embedding for modeling hole transfer reactions. *J. Phys. Chem. B* **2015**, *119*, 7541−7557.
(23) Kubar, T.; Elstner, M. Coarse-grained time-dependent density functional simulation of charge transfer in complex systems: application to hole transfer in DNA. *J. Phys. Chem. B* **2010**, *114*, 11221−11240.
(24) Gajdos, F.; Valner, S.; Hoffmann, F.; Spencer, J.; Breuer, M.; Kubas, A.; Dupuis, M.; Blumberger, J. Ultrafast estimation of electronic couplings for electron transfer between π-conjugated organic molecules. *J. Chem. Theory Comput.* **2014**, *10*, 4653−4660.
(25) Ziogos, O. G.; Blumberger, J. Ultrafast estimation of electronic couplings for electron transfer between pi-conjugated organic molecules. II. *J. Chem. Phys.* **2021**, *155*, 244110.
(26) Cornil, J.; Beljonne, D.; Calbert, J.-P.; Brédas, J.-L. Interchain interactions in organic π-conjugated materials: impact on electronic structure, optical response, and charge transport. *Adv. Mater.* **2001**, *13*, 1053−1067.
(27) Brédas, J.-L.; Beljonne, D.; Coropceanu, V.; Cornil, J. Charge-transfer and energy-transfer processes in π-conjugated oligomers and polymers: a molecular picture. *Chem. Rev.* **2004**, *104*, 4971−5004.
(28) Coropceanu, V.; Cornil, J.; da Silva Filho, D. A.; Olivier, Y.; Silbey, R.; Brédas, J.-L. Charge transport in organic semiconductors. *Chem. Rev.* **2007**, *107*, 926−952.
(29) Troisi, A.; Cheung, D. L.; Andrienko, D. Charge transport in semiconductors with multiscale conformational dynamics. *Phys. Rev. Lett.* **2009**, *102*, 116602.
(30) Ciuchi, S.; Fratini, S.; Mayou, D. Transient localization in crystalline organic semiconductors. *Phys. Rev. B* **2011**, *83*, No. 081202.
(31) Fratini, S.; Mayou, D.; Ciuchi, S. The transient localization scenario for charge transport in crystalline organic materials. *Adv. Funct. Mater.* **2016**, *26*, 2292−2315.
(32) Giannini, S.; Carof, A.; Ellis, M.; Yang, H.; Ziogos, O. G.; Ghosh, S.; Blumberger, J. Quantum localization and delocalization of charge carriers in organic semiconducting crystals. *Nat. Commun.* **2019**, *10*, 3843.
(33) Kubař, T.; Elstner, M. A hybrid approach to simulation of electron transfer in complex molecular systems. *J. R. Soc., Interface* **2013**, *10*, 20130415.
(34) Kubař, T.; Elstner, M. Efficient algorithms for the simulation of non-adiabatic electron transfer in complex molecular systems: application to DNA. *Phys. Chem. Chem. Phys.* **2013**, *15*, 5794−5813.
(35) Carof, A.; Giannini, S.; Blumberger, J. Detailed balance, internal consistency, and energy conservation in fragment orbital-based surface hopping. *J. Chem. Phys.* **2017**, *147*, 214113.
(36) Mulliken, R.; Rieke, C.; Orloff, D.; Orloff, H. Formulas and numerical tables for overlap integrals. *J. Chem. Phys.* **1949**, *17*, 1248−1267.
(37) Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M. Machine learning for the structure−energy−property landscapes of molecular crystals. *Chem. Sci.* **2018**, *9*, 1289−1300.
(38) Lederer, J.; Kaiser, W.; Mattoni, A.; Gagliardi, A. Machine learning-based charge transport computation for pentacene. *Adv. Theory Simul.* **2019**, *2*, 1800136.
(39) Bag, S.; Aggarwal, A.; Maiti, P. K. Machine learning prediction of electronic coupling between the guanine bases of DNA. *J. Phys. Chem. A* **2020**, *124*, 7658−7664.
(40) Wang, C.-I.; Braza, M. K. E.; Claudio, G. C.; Nellas, R. B.; Hsu, C.-P. Machine learning for predicting electron transfer coupling. *J. Phys. Chem. A* **2019**, *123*, 7792−7802.
(41) Caylak, O.; Yaman, A.; Baumeier, B. Evolutionary approach to constructing a deep feedforward neural network for prediction of electronic coupling elements in molecular materials. *J. Chem. Theory Comput.* **2019**, *15*, 1777−1784.
(42) Miller, E. D.; Jones, M. L.; Henry, M. M.; Stanfill, B.; Jankowski, E. Machine learning predictions of electronic couplings for charge transport calculations of P3HT. *AIChE J.* **2019**, *65*, No. e16760.

https://doi.org/10.1021/acs.jctc.3c00184
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(43) Behler, J. Four generations of high-dimensional neural network potentials. *Chem. Rev.* **2021**, *121*, 10037−10072.
(44) The latest release version of n2p2 can be found at DOI: 10.5281/zenodo.1344446.
(45) Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930−17955.
(46) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(47) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **2018**, *148*, 241730.
(48) Klimeš, J.; Bowler, D. R.; Michaelides, A. Chemical accuracy for the van der Waals density functional. *J. Phys.: Condens. Matter* **2010**, *22*, No. 022201.
(49) VandeVondele, J.; Hutter, J. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *J. Chem. Phys.* **2007**, *127*, 114105.
(50) Goedecker, S.; Teter, M.; Hutter, J. Separable dual-space Gaussian pseudopotentials. *Phys. Rev. B* **1996**, *54*, 1703.
(51) Elsner, J.; Giannini, S.; Blumberger, J. Mechanoelectric Response of Single-Crystal Rubrene from Ab Initio Molecular Dynamics. *J. Phys. Chem. Lett.* **2021**, *12*, 5857−5863.
(52) Gertsen, A. S.; Sørensen, M. K.; Andreasen, J. W. Nanostructure of organic semiconductor thin films: Molecular dynamics modeling with solvent evaporation. *Phys. Rev. Mater.* **2020**, *4*, No. 075405.
(53) Ziogos, O. G.; Kubas, A.; Futera, Z.; Xie, W.; Elstner, M.; Blumberger, J. HAB79: A new molecular dataset for benchmarking DFT and DFTB electronic couplings against high-level ab initio calculations. *J. Chem. Phys.* **2021**, *155*, 234115.
(54) Widdowson, D.; Mosca, M. M.; Pulido, A.; Cooper, A. I.; Kurlin, V. Average Minimum Distances of Periodic Point Sets—Foundational Invariants for Mapping Periodic Crystals. *MATCH* **2022**, *87*, 529−559.
(55) McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Software* **2017**, *2*, 205.
(56) Rosenkrantz, D. J.; Stearns, R. E.; Lewis, P. M., II. An analysis of several heuristics for the traveling salesman problem. *SIAM J. Comput.* **1977**, *6*, 563−581.
(57) Johnson, M. E.; Moore, L. M.; Ylvisaker, D. Minimax and maximin distance designs. *J. Stat. Planning Inference* **1990**, *26*, 131−148.
(58) Schran, C.; Brezina, K.; Marsalek, O. Committee neural network potentials control generalization errors and enable active learning. *J. Chem. Phys.* **2020**, *153*, 104105.
(59) Egorova, O.; Hafizi, R.; Woods, D. C.; Day, G. M. Multifidelity statistical machine learning for molecular crystal structure prediction. *J. Phys. Chem. A* **2020**, *124*, 8065−8078.
(60) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, No. 058301.
(61) Gao, F. A new acceptor for highly efficient organic solar cells. *Joule* **2019**, *3*, 908−909.
(62) Wadsworth, A.; Moser, M.; Marks, A.; Little, M. S.; Gasparini, N.; Brabec, C. J.; Baran, D.; McCulloch, I. Critical review of the molecular design progress in non-fullerene electron acceptors towards commercially viable organic solar cells. *Chem. Soc. Rev.* **2019**, *48*, 1596−1625.
(63) Bristow, H.; Thorley, K. J.; White, A. J.; Wadsworth, A.; Babics, M.; Hamid, Z.; Zhang, W.; Paterson, A. F.; Kosco, J.; Panidi, J.; et al. Impact of Nonfullerene Acceptor Side Chain Variation on Transistor Mobility. *Adv. Electron. Mater.* **2019**, *5*, 1900344.
(64) Momma, K.; Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Crystallogr.* **2011**, *44*, 1272−1276.
(65) Sivaraman, G.; Jackson, N. E. Coarse-grained density functional theory predictions via deep kernel learning. *J. Chem. Theory Comput.* **2022**, *18*, 1129−1141.

https://doi.org/10.1021/acs.jctc.3c00184
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX