Received: 17 April 2019 | Revised: 5 August 2019 | Accepted: 13 August 2019

DOI: 10.1002/aic.16760

FUTURES ISSUE: THERMODYNAMICS AND MOLECULAR-SCALE PHENOMENA

AIChE JOURNAL

Machine learning predictions of electronic couplings for charge transport calculations of P3HT

Evan D. Miller[1] | Matthew L. Jones[1] | Mike M. Henry[1] | Bryan Stanfill[2] | Eric Jankowski[1]

[1]Micron School of Materials Science and Engineering, Boise State University, Boise, Idaho

[2]Pacific Northwest National Laboratory, Richland, Washington

Correspondence
Eric Jankowski, PhD, Micron School of Materials Science and Engineering, Boise State University, Boise, Idaho 83705, USA.
Email: ericjankowski@boisestate.edu

Funding information
National Science Foundation, Grant/Award Numbers: 1835593, 1653954, 1229709, ACI-1053575

Abstract

The purpose of this work is to lower the computational cost of predicting charge mobilities in organic semiconductors, which will benefit the screening of candidates for inexpensive solar power generation. We characterize efforts to minimize the number of expensive quantum chemical calculations we perform by training machines to predict electronic couplings between monomers of poly-(3-hexylthiophene). We test five machine learning techniques and identify random forests as the most accurate, information-dense, and easy-to-implement approach for this problem, achieving mean-absolute-error of 0.02 [$\times$ 1.6 $\times$ 10^{-19} J], R^2 = 0.986, predicting electronic couplings 390 times faster than quantum chemical calculations, and informing zero-field hole mobilities within 5% of prior work. We discuss strategies for identifying small effective training sets. In sum, we demonstrate an example problem where machine learning techniques provide an effective reduction in computational costs while helping to understand underlying structure–property relationships in a materials system with broad applicability.

KEYWORDS
machine learning, molecular simulation, organic photovoltaics

1 | INTRODUCTION

Finding a needle in a haystack is hard because of all the hay: Inspecting each straight, pointy object drawn from a large haystack rarely reveals needles and it is impractical to inspect all the pointy objects. Searching haystacks is analogous to finding optima in large problem spaces—such as the identification of the best ingredients for high-efficiency, low-cost organic photovoltaics (OPVs) for sustainable power generation, in which, progress is hindered by the experimental and computational expense of enumerating the combination of factors that govern a candidate's viability. Replacing experiments with computer simulations increases the rate of candidate inspection, as computer simulations can be performed at a lower cost and in less time, but does not wholly alleviate the time burden. Here we focus on strategies for further increasing the rate at which candidates can be inspected by lowering the computational cost of connecting OPV structure to its performance.

OPVs are a focus of sustainable energy development because devices with 15% power conversion efficiency (PCE) are theorized as sufficient for one-day energy-pay-back times,[1] which would circumvent economic barriers to widespread deployment. A key difficulty in mass-producing high PCE devices is controlling the self-assembled active-layer morphology (the spontaneously forming microstructure within the electricity generating portion of the device). The majority

of active layers are primarily composed of two components: An electron-donating and an electron-accepting species, and the microstructural order of these two components determine the device's overall efficiency.[2] Recent developments in new OPV ingredients have demonstrated power conversion efficiencies in excess of 15%,[3,4] however mass-produced OPVs still fall below the efficiencies required for widespread commercial viability, and the precise origins of the higher efficiencies are not fully understood. To make OPVs with one-day payback times a reality, a fundamental understanding of how ingredient chemistry and processing determines the active layer morphology and how the resulting features influence PCE is needed.

Here we describe machine learning (ML) efforts toward speeding calculations linking OPV morphology to the mobility of charges through it, which in turn determines the fill factor and PCE [5] of OPV devices. To validate our approach, we focus on the benchmark donor polymer poly-(3-hexylthiophene) (P3HT), which is the archetype for linking the self-assembled morphology to efficiency [6,7] due to its solution processability and history in breakthrough (in 2006, 5% PCE) OPVs.[8] In P3HT devices, faster charge movement (which corresponds to better PCE) can be obtained by creating devices that maximize the degree of crystalline order, [9] which can be accomplished by using high regioregularity [10] and shorter polymer chains.[11,12] Time-of-flight measurements of hole mobility in P3HT experiments range from $\mu = 1 \times 10^{-5}$ to 1×10^{-3} cm^2/Vs.[13-16] Computational work has helped to explain the role of thiophene ring orientation on charge transport,[17] and kinetic Monte Carlo (KMC) simulations of charge transport have predicted mobilities ranging from $\mu = 1 \times 10^{-4}$ to 0.6 cm^2/Vs,[18-22] depending on the degree of ordering in of the P3HT morphologies. These experimental and computational predictions of mobility provide references for validation: Calculated hole mobilities in P3HT should fall between $\mu = 1 \times 10^{-4}$ to 0.6 cm^2/Vs and increase with increasing P3HT crystallinity.

In our own prior work, we predict charge transport through P3HT by first predicting P3HT morphologies at ~350 processing state points [23] (Supporting Information Section 1), then calculating charge mobility through ~100 of these structures[21] using KMC simulations. Doing so requires hopping rates between P3HT chromophores, which we calculate with Marcus semi-classical hopping theory[24] using quantum chemical ZINDO/S [25,26] calculations to obtain the electronic transfer integrals between chromophores (couplings, $J_{i,j}$), which describe the amount of frontier molecular orbital overlap between pairs chromophores. Completely connecting all the neighbors in a representative system requires $\sim 2 \times 10^5$ ZINDO/S calculations per morphology, corresponding to about 26 CPU hours of computation time. We aim to determine the efficacy of using ML to predict $J_{i,j}$ and bypass the numerous, expensive ZINDO/S calculations required to characterize the charge transport properties of a morphology. We take inspiration from recent studies in which ML based on first-principle calculations has been used to accelerate the development of organic light-emitting diodes,[27] OPV candidate compounds,[28] and electronic predictions based on coarse-grained sites.[29] The use of ML to accelerate materials discovery has grown recently due to advances in enabling hardware, algorithms, and open-source libraries.[30-32] The $J_{i,j}$ prediction problem approached here is well-suited to supervised learning algorithms where ample data can inform classification or regression schemes relating inputs features to output properties, especially if discerning these relations would be difficult or tedious for a human.[33-36]

2 | METHODS

We compare two ways of generating electronic transfer integrals ($J_{i,j}$) in P3HT; the control case of quantum chemical ZINDO/S calculations using ORCA,[37] as in Reference 19, and the present test case of machine learning methods trained to predict transfer integrals. Transfer integral generation is required to link morphologies to mobility.

1. Sample OPV morphologies using molecular simulations.
2. Generate transfer integrals between chromophores in each morphologies (with ORCA as in Reference 19 and ML here).
3. Predict charge mobilities from transfer integrals using KMC simulations.

In prior work, we describe combining these steps into the MorphCT [38] software pipeline, the details of said implementation,[19] and applications to P3HT.[21]

To determine charge mobilities with kinetic Monte Carlo (KMC) simulations, morphologies are treated as a weighted network in which each P3HT monomer is considered an electronically active chromophore and charges may hop to neighboring chromophores as defined by neighboring cells from a Voronoi tessellation of thiophene ring centers of mass. We calculate electronic transfer integrals between chromophores using the energy-splitting-in-dimer method (ESD):[39,40]

$$J_{i,j} = \frac{1}{2}\sqrt{\left(E_{\mathrm{HOMO}} - E_{\mathrm{HOMO\text{-}1}}\right)^2 - \left(\Delta E_{i,j}\right)^2}, \quad (1)$$

where the magnitude of the splitting of the highest occupied molecular orbital to a new E_{HOMO} and $E_{\mathrm{HOMO\text{-}1}}$ in the dimer state is compared to the difference in *HOMO* level of the isolated, individual chromophores:

$$\Delta E_{i,j} = E_{\mathrm{HOMO},j} - E_{\mathrm{HOMO},i}. \quad (2)$$

ZINDO/S requires atom positions and types of each chromophore to calculate ($E_{\mathrm{HOMO}} - E_{\mathrm{HOMO\text{-}1}}$) and $\Delta E_{i,j}$.

The rate at which a charge is able to hop from chromophore i to j is given by an adaptation of the semi-classical Marcus theory:[24]

$$k_{i,j} = \frac{|J_{i,j}|^2}{\hbar}\sqrt{\frac{\pi}{\lambda k_B T}}\exp\left(\frac{r_{i,j}}{\alpha}\right)\exp\left(-\frac{\left(\Delta E_{i,j} - \lambda\right)^2}{4\lambda k_B T}\right), \quad (3)$$

where $r_{i,j}$ is the center-of-mass distance between chromophore thiophene rings, $\hbar$ Planck's reduced constant, λ is the reorganization energy, k_B is Boltzmann's constant, and T = 293 K is the temperature of the KMC simulation, which is chosen for room temperature. We

also include an additional exponential term in the hopping rate equation originating from Mott's variable range hopping theory, which is often used in polymers, with α = 0.2 nm here. The material-specific reorganization energy, λ, the energy required to polarize and depolarize a single monomer of P3HT in response to a charge hopping from *i* to *j*, is a constant at 0.306 ($\times$ 1.6 $\times$ 10^{-19} J).[41]

KMC proceeds by stochastically generating a sequence of events and tracking total elapsed time by summing the times associated with the event sequence. In the case of charge motion on P3HT networks considered here, this is implemented by considering the hopping rates $k_{i,j}$ for a located on chromophore *i*, where *j* is the index of a neighboring chromophore. A uniformly distributed random number *x* is generated on [0,1) for each possible hop, and is used to calculate hopping times

$$\tau_{i,j} = \frac{-\ln(x)}{k_{i,j}}, \tag{4}$$

from which the fastest event is selected and performed. Note that this amounts to an importance sampling of possible hops for each event, not a naïve sampling of largest $k_{i,j}$.

By iterating millions of hopping events, a charge's trajectory through the morphology can be followed and the total displacement determined. The systems utilized in this investigation are cubic with a side of length ~15 nm. However, we use periodic boundary conditions to allow the charge to move through the same morphology many times, resulting in a total displacement of hundreds of nanometers. Carriers are permitted to hop for a simulation run time, *t*, at which point the mean squared displacement (MSD) is calculated, and the charge is removed from the system. A new charge is then triggered at a randomized start location and a new trajectory determined. The MSDs are averaged over 10,000 carriers with 1 ns < *t* < 10 ns. The gradient of the MSD as a function of *t* provides the carrier diffusivity, *D*:

$$D = \frac{1}{2n} \cdot \frac{d\mathrm{MSD}}{dt}, \tag{5}$$

where *n* = 3 is the number of dimensions. *D* can then be related to the mobility, μ, through the three-dimensional Einstein–Smoluchowki relation:

$$\mu_0 = \frac{qD}{k_B T}, \tag{6}$$

where *q* is the unit charge. The relation shown in Equation 5 is commonly employed in charge transport studies, and provides an upper-bound for charge carrier diffusivity in the absence of an external driving force. We treat our charges as being isolated, that is, no Coulombic interactions with other charges or electric fields. The mobilities reported, therefore, represent the "best case" zero-field charge mobilities, μ_0, and are analogous to experimental time-of-flight measurements.

2.1 | Machine learning

To predict $J_{i,j}$ using any machine learning approach we select input features that are then related to $J_{i,j}$ calculated by ZINDO/S. Because ZINDO/S requires only atom types and positions, we select nine spatial features that we expect to be predictors of $J_{i,J}$ between P3HT monomers.

1. Whether the monomers are chemically bonded to each other ("Bonded").
2. The distance between their thiophene ring centers of mass (COM–COM).
3. The relative "pitch" between thiophene rings (Figure 1, Y-rot).
4. The relative "roll" between thiophene rings (*X*-rot).
5. The distance between sulfur atoms on the thiophene rings (S–S).
6. The *x*-component of the thiophene ring center separations (*X*-dist).
7. The *y*-component of the thiophene ring center separations (*Y*-dist).
8. The *z*-component of the thiophene ring center separations (*Z*-dist).
9. Energy difference between the chromophores $\Delta E_{i,j}$.

Note that the "yaw" angle about the thiophene's local *z*-axis is missing from this list of features as preliminary work has shown that its effect on the transfer integral is negligible. This is expected as the electron density is delocalized above and below the plane of the thiophene ring, so rotations around the local *z*-axis do not affect the amount of molecular orbital overlap. We aim to limit the chemical specificity of the features used here, and look toward other machine learning techniques that might help automate feature identification in the future.[31] We test ordinary least squares (OLS), support vector machines (SVM), K-nearest neighbors (KNN), artificial neural networks (ANN), and random forests (RF) as machine learning implementations for predicting $J_{i,j}$ from the above nine features. The review article of Reference 31 provides a comprehensive overview of ML techniques in soft matter, and is a recommended starting place for understanding the taxonomy of ML techniques. Briefly, OLS determines coefficients for linear combinations of input features by minimizing error on a training data set; SVM classifies possible outcomes based on hyperplanes dividing the feature space of a training set; KNN uses determines "proximity" in feature-space between elements of a training set and predicts $J_{i,j}$ based on members of clusters that emerge from this grouping. ANN are composed of "layer" matrices that transform inputs into outputs through matrix multiplication, with iterative re-weighting

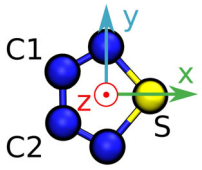

FIGURE 1 Reference thiophene ring and local coordinate axes used to determine relative spatial features between P3HT monomers. The thiophene ring center of mass is used as the origin of the local coordinates. A thiophene ring's rotation about its local *y*-axis relative to another thiophene ring in the reference frame is used to calculate "pitch." A thiophene's rotation about its local *x*-axis relative to the reference ring defines "roll" [Color figure can be viewed at wileyonlinelibrary.com]

matrix elements performed by gradient descent optimization using a training set of known features and $J_{i,j}$. The ANN is implemented in the Python package Tensorflow [42] (version 1.9.0, see Supporting Information Section 2 for ANN details), and all other methods are conducted with the package Scikit-Learn (version 0.19.1) with the default argument values.[43] The code used in this study is available at Reference 44 and the data set at Reference 45.

We explain RFs in more detail, due to their focus in the discussion that follows. RFs are an ensemble technique in which the prediction from many decision trees are combined into an output. A decision tree operates by partitioning the data, based on the features and their values, into progressively smaller subgroups to determine an average outcome ($\bar{y}$) for the group. The decision tree implementation in Scikit-Learn [43] is based on the classification and regression tree (CART) algorithm, which creates a binary split based on a threshold (t_f) for a feature (f) at a "leaf," creating two "branches":

$$d_{f_x} = \begin{cases} \text{Left Branch if } f_x < t_f \\ \text{Right Branch if } f_x \geq t_f \end{cases}, \tag{7}$$

in which d_{fx} signifies the branch decision for sample x. The threshold t_f is determined by minimizing the cost function:

$$C(d_f) = \frac{n_{left}}{N} E_{left}(d_f) + \frac{n_{right}}{N} E_{right}(d_f), \tag{8}$$

where n_{left} and n_{right} are the number of samples on each branch (based on the decision d_f), N is the total number samples on the leaf, and $E_{left,\ right}(d_f)$ is the error from assigning the samples to the left and right branches. This error is measured as the mean-squared error:

$$E(d_f) = \frac{1}{n_m} \sum_{i}^{n_m} (y_i - \bar{y})^2, \tag{9}$$

where y_i is the true output and n_m is the number of samples in the left or right branch. This process is repeated further with additional cut-offs, thereby growing the tree and partitioning the data into smaller and smaller partitions, reducing the error on each leaf, until a stopping criteria (such as a maximum depth) is met. RFs avoid over-fitting by providing each tree with a different subset of the total training data, then taking the ensemble average over each tree "voting" on the outcome.

Here we draw training set chromophore pairs from one "disordered," one "semi-crystalline," and one "crystalline" morphology from prior work.[21] The degree of crystallinity is reported using ψ' as in Reference 21, with "disordered," "semi-crystalline," and "crystalline" corresponding to ψ' of 0.17, 0.25, and 0.33, respectively. ψ' is a quantification of fraction of thiophene rings composed into "large" clusters and the deviations in the aliphatic bond lengths. A description of the origins and implementation of ψ' is included in Supporting Information Section 5 and References 21 and 23. Each morphology is composed of 15,000 P3HT repeat units, giving about 230,000 chromophore pairs (as defined by the Voronoi tessellation around thiophene centers). The ML techniques are trained against some or all of these 700,000 chromophore pairs and their associated ZINDO/S calculations of $J_{i,j}$. The ML techniques are tested against 6.48 million chromophore pairs from 9 additional "disordered," 9 "semi-crystalline," and 9 "crystalline" morphologies.

3 | RESULTS AND DISCUSSION

In this section, we first summarize the accuracy of five machine learning techniques for correlating our nine chosen structural features with $J_{i,j}$ calculated using ZINDO/S. We show that Random Forests are the optimal choice here for their ease of implementation and accuracy. We then evaluate the KMC charge mobility calculations from the RF-predicted $J_{i,j}$. We discuss the time saved through using RFs in place of ZINDO/S. Finally, we determine which features matter most for $J_{i,j}$ and investigate the relationship between the training set population and RF's prediction capabilities to understand the minimal information needed for accurate RF training.

3.1 | Comparison of ML techniques

Prediction accuracies of OLS, KNN, SVM, ANN, and RF techniques are shown in Figure 2. We orient the reader to two regions in each accuracy plot: There is a cluster of bonded chromophore pairs with $0.6 < J_{i,j} < 1.1$ and a cluster of nonbonded pairs with $J_{i,j} < 0.5$. The more test pairs that are not on the diagonal line indicating perfect agreement between predicted and actual $J_{i,j}$, and the further their distance from the diagonal line of agreement, the worse the method. The poor predictive capabilities of OLS (Figure 2a), despite the surprisingly high R^2 = 0.96, suggests nonlinear relationships between features determines $J_{i,j}$. SVM accurately predicts bonded $J_{i,j}$ but fails when the chromophores are nonbonded (yellow region near [Actual = 0, Predicted = 0.4]). This results in a large number of $J_{i,j} \sim 0.4$ [$\times 1.6 \times 10^{-19}$ J] predictions for hops that should have zero coupling, leading to a low R^2 value and high mean-absolute-error (MAE). KNN provides predictions that are more accurate than OLS and SVMs and with better predictions of nonbonded pairs, but with over-prediction of bonded interactions, which can be seen as a "tail" extending above the perfect match diagonal around (Actual = 0.6, Predicted = $0.8 \times 1.6 \times 10^{-19}$ J). Both the RF and the ANN outperform the aforementioned techniques, with RF slightly outperforming ANN. Because the ANN has a larger number of hyperparameters to tune (number of hidden layers, neurons per layer, activation function type, optimization method [See Supporting Information]) and is less accurate than RF, we focus on RFs henceforth.

3.2 | Mobility predictions

The predicted $J_{i,j}$'s from the random forest closely track the actual values, with an R^2 value of 0.986 and a MAE of 0.020 ($\times 1.6 \times 10^{-19}$ J), though there exist outliers (Figure 2e). For example, the predicted average nonbonded $J_{i,j}$ value is slightly higher (0.0015 [$\times 1.6 \times 10^{-19}$ J]) than the actual mean (<0.001 [$\times 1.6 \times 10^{-19}$ J]) (see

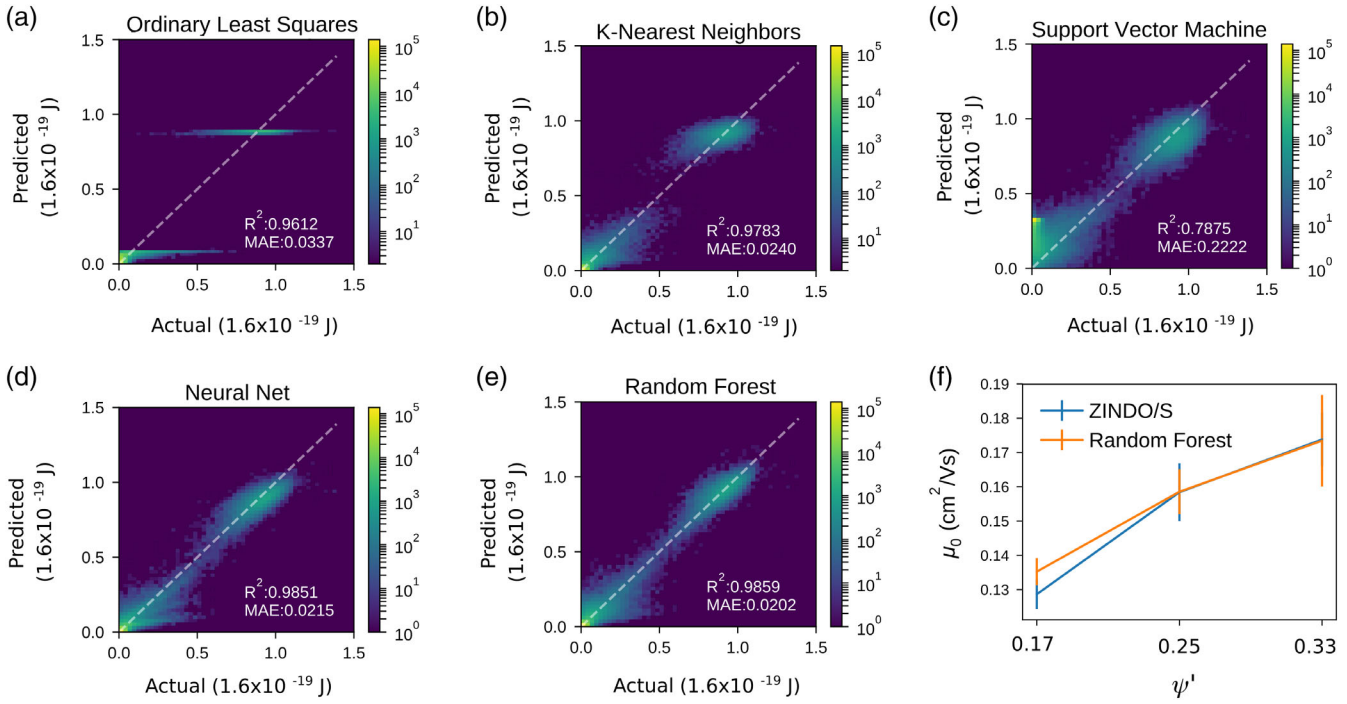

FIGURE 2 Accuracy of predictions of ZINDO/S $J_{i,j}$ from (a) OLS, (b) KNN, (c) SVM, (d) ANN, and (e) RF. The *x*-axes of each plot describe $J_{i,j}$ calculated with ZINDO/S and the *y*-axis corresponds to the predicted $J_{i,j}$ for a ML technique, with each chromophore pair from the training set occupying one pixel on these axes. The number of chromophore pairs at a particular location is represented by the purple-to-yellow color bar. (f) The mobilities from RF $J_{i,j}$ are commensurate with those using ZINDO/S $J_{i,j}$. In the disordered morphology case, the RF-informed mobilities are ~5% higher than ZINDO/S-informed mobilities. Error bars show the standard error of the mobility calculations [Color figure can be viewed at wileyonlinelibrary.com]

Supporting Information Section 3). With the ultimate goal of determining the efficacy of ML in predicting overall charge carrier mobilities through a morphology, we test the significance of these deviations by using predicted $J_{i,j}$ values in KMC simulations to calculate the final hole mobility for the system (Figure 2f). The mobilities calculated from the RF predictions are slightly higher than those determined with ZINDO/S for the disordered system. We hypothesize this over-prediction stems from our features incompletely describing structural perturbations that occur more frequently in disordered systems. For example, it is known that the dihedral angle between two chromophores will affect the charge transport along the chain,[17] so trying out explicit dihedral angle features rather than the present combinations of rotations may provide marginal accuracy gains. Despite the small over-prediction of disordered P3HT mobility, the resultant mobilities are close (within 5% of ZINDO/S-informed mobilities), and follow the expected trend of increasing mobility with increasing crystallinity. These agreements are encouraging, as mobilities can vary by several orders-of-magnitude for different chemistries and processing conditions, and suggest that RF-predicted transfer integrals are an effective replacement for the relatively expensive ZINDO/S calculations.

3.3 | Performance benefit

To quantify the computational burden alleviated by using random forests we consider representative times for training the RF, generating $J_{i,j}$ with ZINDO/S for one morphology, and the frequency of calculating $J_{i,j}$ for multiple morphologies. Applying a trained RF to a representative system of ~200,000 chromophore pairs (with unknown energy levels and transfer integrals) requires 4 min on an Intel Haswell CPU, compared to ~26 CPU hours using Intel Xeon CPUs with ZINDO/S calculations. This factor of 390× speedup for a single simulation snapshot is multiplied in ensemble sampling studies: It is gained for each of the independent samples in an equilibrated simulation trajectory. This transferability of RFs trained across disordered, semi-crystalline and crystalline P3HT demonstrates that a single RF can be used to accurately infer ensemble charge mobilities across hundreds of state points, each with hundreds of morphology snapshots. Using RFs, therefore, *enables* such screening studies, replacing 1.08×10^4 CPU-days of ZINDO/S calculations with 28 CPU-days of RF lookups.

3.4 | RF training requirements

We consider here the minimal training set (the fewest ZINDO/S calculations) needed for accurate RF prediction of $J_{i,j}$, helping to gauge what "plenty of data to train against" means for the present problem. We evaluate the performance of several RFs, calibrated with different sizes of training data. In each case, the number of samples was selected randomly from the complete database of ~700,000 samples. Figure 3a shows that R^2 and MAE converge exponentially to high and low values, respectively, with as few as 100 training samples. The fast convergence is due to the algorithm quickly learning that bonded

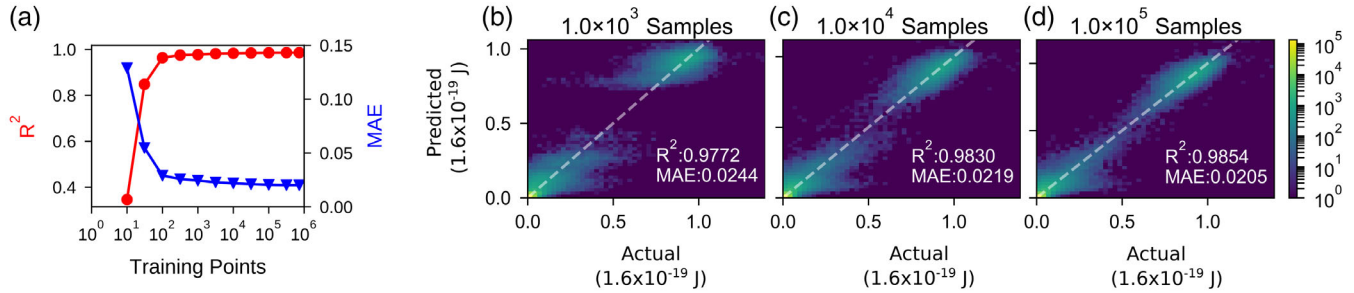

FIGURE 3 (a) Dependence of the R^2 and MAE on number of training examples shows that prediction accuracy converges around tens-of-thousands of pairs. (b–d) Despite relatively "good" R^2 and MAE values, significant deviations from the diagonal of perfect prediction are seen below ~100,000 training samples [Color figure can be viewed at wileyonlinelibrary.com]

chromophores typically result in high $J_{i,j}$ (>0.7 [$\times$ 1.6 $\times$ 10^{-19} J]) and nonbonded chromophores resulting in low $J_{i,j}$ (<0.3 [$\times$ 1.6 $\times$ 10^{-19} J]).

Although convergence to a fairly accurate prediction ($R^2 \sim 0.977$) is quickly achieved based on bonded/nonbonded chromophores, it can be seen in Figure 3 that with 1×10^3 samples, the distribution between bonded/nonbonded transfer integrals is bimodal, with high nonbonded $J_{i,j}$ and low bonded $J_{i,j}$ that occur in the range (0.4, 0.7) ($\times$ 1.6 $\times$ 10^{-19} J) being missed. When 1×10^4 samples are used, the (0.4, 0.7) ($\times$ 1.6 $\times$ 10^{-19} J) gap begins to fill in (Figure 3c), but it is not until 1×10^5 samples are used that the high/low nonbonded/bonded are correctly captured by the RF (Figure 3d). Extracting and training on these features from a simulation takes a negligible amount of time (~2 min for extracting, 14 s for training on 1×10^5 samples). The most expensive part of the process will be conducting the ZINDO/S calculations to train on these 1×10^5 samples (~13 hr).

3.5 | Feature comparison

We compare the relative importance of the nine features we currently use in predicting $J_{i,j}$, relying on the RF's advantage of feature transparency. Specifically, we use permutation importance, which compares the accuracy of the RF (R^2 value) on a validation set with true values and when the features' values have been shuffled. The importance is then the difference in R^2 caused by permuting that feature. The permutation mechanism is more computationally expensive than the mean decrease in impurity (or Gini importance) which is built into Scikit-Learn's RF algorithm but is more reliable. We note that the X, Y, and Z displacements are permuted in aggregate, that is, in testing the X, Y, and Z importances, all three columns are permuted at the same time so that their importance relative the COM-COM feature can be better distinguished. The calculated feature importances, normalized to sum to one, are shown in Figure 4. By far, the most important feature in predicting $J_{i,j}$ is whether or not two chromophores are directly bonded to each other. This is due to charges being delocalized over neighboring chromophores, which result in very high $J_{i,j}$ values. When the "bonded" feature is missing, many low, bonded $J_{i,j}$ are over-predicted and high nonbonded $J_{i,j}$ are under-predicted.

In Figure 5, we summarize the prediction accuracies of RFs trained, but with select features omitted from the training sets. The

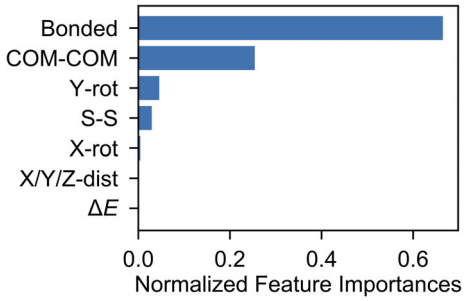

FIGURE 4 The feature importances for the RF algorithm. The X, Y, and Z distances are all combined into one feature importance [Color figure can be viewed at wileyonlinelibrary.com]

biggest deviation from champion accuracy (R^2 = 0.9858) is observed when the bonded feature is omitted, as expected. Removing the COM–COM feature results in an over-prediction of the "bonded" $J_{i,j}$ values–transfer integrals in the 0.8–1.0 ($\times$ 1.6 $\times$ 10^{-19} J) region are shifted closer to 1.0 ($\times$ 1.6 $\times$ 10^{-19} J) (Figure 5b). The importance of having close chromophores is somewhat unsurprising as the transfer integrals decrease rapidly as the two chromophores move away from each other.[17,39,46,47] We note that the COM-COM feature is directly dependent on the X, Y, and Z displacements as it is the square-root of the squared-sums of the X, Y, and Z offsets. Although it is a composite feature, explicitly training on the COM-COM distance is very important for predicting the $J_{i,j}$. The individual *X*-, *Y*-, and *Z*-dist features have negligible feature importance, even when permuted in aggregate (Figure 4). This is likely to be due to the small size and relative symmetry of the thiophene ring, and the nonlinear relationship between the individual features and the aggregate COM-COM feature. If larger or asymmetric chromophores were used, such as a coronene or a perylene derivative, the displacements along the different axes are likely to dominate and increase relative feature importance (see Figure 5c).[47]

Relative rotation around the *Y*-axis ("pitch") is the third most important feature, and is more important than rotation around the *X*-axis ("roll"; Figure 5d). This is likely because rotations around Y move the sulfur atom in the ring, as opposed to rotations around X in which the sulfur is stationary. The importance of the relative sulfur positions is further highlighted by the S–S distance being the fourth most

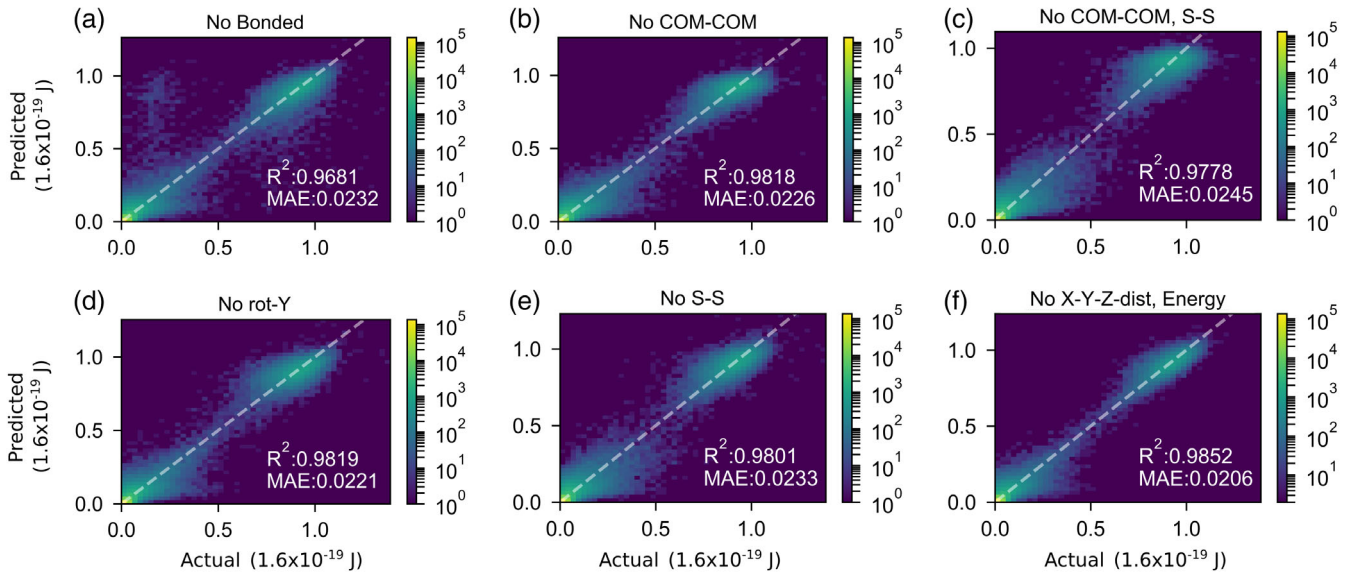

FIGURE 5 (a) Removing the bonded feature results in a high number of outliers as both bonded and nonbonded $J_{i,j}$ values are under and over predicted. (b) Removing the COM–COM constraint results in a flattening and broadening of the "bonded" $J_{i,j}$ distribution. (c) If both the COM–COM and S–S distances are removed (and therefore only the displacements along the X, Y, and Z axes are considered) the distribution of $J_{i,j}$ is much more split between "bonded" and "nonbonded." (d, e) Removing the rotation around Y and the S–S distance create more noise. (f) The X, Y, and Z displacements and the $\Delta E_{i,j}$ can all be omitted in training and result in high accuracies [Color figure can be viewed at wileyonlinelibrary.com]

important feature, and this feature is responsible for obtaining correct predictions for high nonbonded $J_{i,j}$ and low bonded $J_{i,j}$ (Figure 5e). This indicates that in order to have high $J_{i,j}$, electronegative atoms within the chromophores must be proximal in order to act as bridges between the two chromophores.

The $\Delta E_{i,j}$ feature in this experiment is unimportant for predicting $J_{i,j}$. This unimportance is not surprising as the MD simulations represent the thiophene ring with a rigid-body, which means the relative positions of all the atoms in the ring are fixed throughout the simulation. With this model, differences in energies can only arise based on conformational differences of the aliphatic tails. The effect of these tails on energy is likely to be small, and many studies omit the tails as a way to reduce computational burden and still obtain correct results. Consequently, if $\Delta E_{i,j}$ is small compared to the HOMO and HOMO-1 splitting in Equation 1, it becomes negligible for $J_{i,j}$. If flexible thiophene rings were used, the importance of the $\Delta E_{i,j}$ feature would increase (although thiophene ring perturbations are still likely to be small because of the aromatic structure of the ring). Despite the insignificance of $\Delta E_{i,j}$ in predicting $J_{i,j}$, we do not argue that $\Delta E_{i,j}$ will be unimportant for predicting mobility values as Equation 3 explicitly considers $\Delta E_{i,j}$ within an exponential and it will likely still have non-negligible effects on the hopping rate. Here, we show that omitting the X–Y–Z displacements and $\Delta E_{i,j}$ features entirely has a negligible effect on the accuracy of only our $J_{i,j}$ predictions (Figure 5f).

3.6 | Curating a training set

Here we consider the possibility of curating a "universal" training set of chromophore pairs that inform an RF with predictive capabilities for P3HT morphologies with disparate degrees of order. This experiment is motivated by (a) the above observation that only 10^4–10^5 sufficient for the present work, and (b) knowing that ML methods excel when there is an abundance of training data. So, is it possible to curate a minimal set of chromophore pairs that will work on the present morphologies, be transferable to other morphologies with different distributions of chromophore positions, and be straightforward to create? If it is possible, then generating libraries of chromophore positions could be a general strategy for speeding the calculation of mobilities in new materials: Quantum chemical calculations on monomers can be performed once and used in novel blends of materials, and transfer integrals usable for many morphologies can be calculated before the first MD simulation is performed, saving time. To curate the training data, we duplicate a chromophore (parent) to create a child chromophore, resulting in all ΔE values being 0. The child chromophore is then moved along each axis ($\leq$ 0.5 nm) and rotated around the x- and y-axes ($\leq \pi$) resulting in 1×10^4 training pairs. The child movement and rotation is done in two ways: At distinct steps, for example, steps of 0, 0.1, 0.2 nm and uniformly distributed over the range (shown in Figure 6). For each offset, we apply the constraint that the COM–COM distance must be greater than 0.3 nm, as COM–COM distances shorter than this are unphysical. With this uniform sampling of positions and orientations, close packings and large separations observed in simulations are underrepresented (Figure 6a), as are aligned and anti-aligned orientations of thiophene rings (Figure 6b). We expect that the undersampling of pi-stacked configurations will most negatively impact accuracy, as $J_{i,j}$ is negligible for large separations. This data curation generates COM–COM and S–S distributions similar in shape around 0.5 nm, though missing pairs separated at larger distances that are observed in simulations (Figure 6c and d).

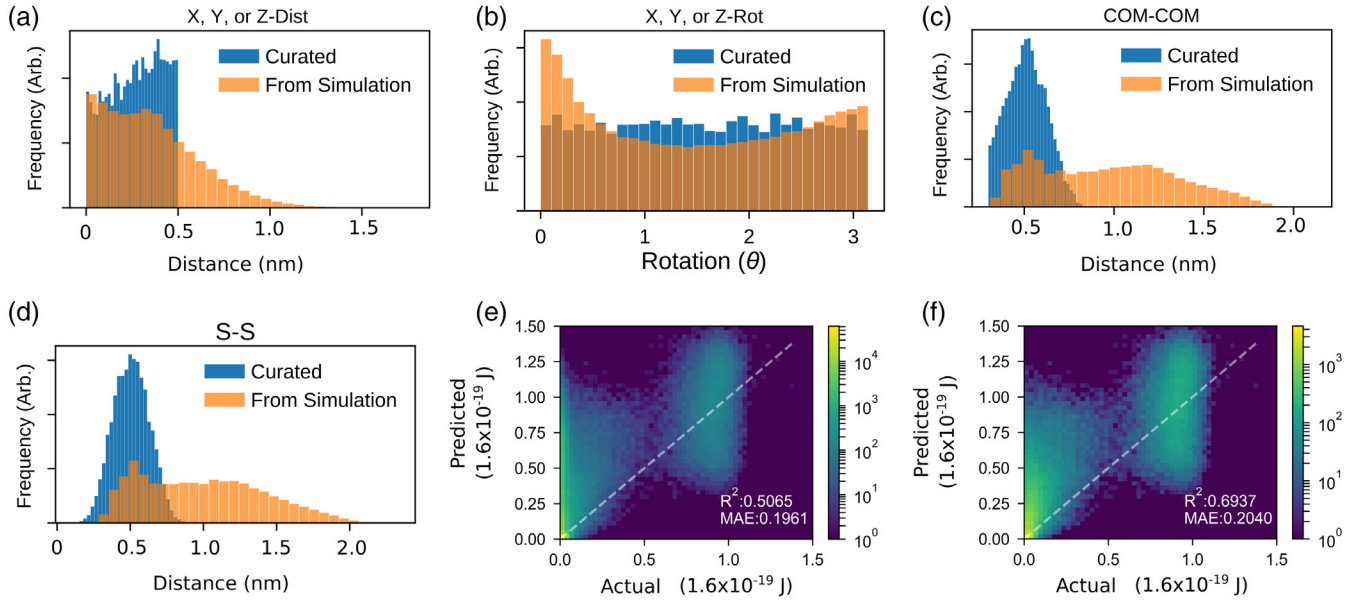

FIGURE 6 The normalized distributions of the training features in the curated set (blue) and in the simulation (orange). (a) Displacements and (b) rotations are determined based on a uniform distribution along each axis; 0 ≤ distance ≤0.5 nm for displacements along the axes and 0 ≤ rotation ≤ π for rotations and show that a uniform distribution fails to capture more energetically favorable close configurations and similar alignments. (c) COM–COM and (d) S–S distances are then calculated based on the displacements along the various axes. The curated training set does a poor job of predicting $J_{i,j}$ whether (e) the curated set is used to evaluate all chromophore pairs or (f) the chromophore pairs that lie within the range of the curated set [Color figure can be viewed at wileyonlinelibrary.com]

Though these larger spacings are prevalent in the simulated structures, we find they contribute negligibly to charge transport.

We train the RF using this curated training set and validate it against the simulation produced $J_{i,j}$. As is seen in Figure 6e, the RF trained on the curated set does a poor job of predicting $J_{i,j}$. The largest error in the predictions arises from the over-prediction of the low (≤0.2 [$\times$ 1.6 $\times$ 10^{-19} J]) $J_{i,j}$ in the system. This error can be reduced somewhat by considering only chromophore pairs that lie within the range of the curated dataset (within 0.5 nm along each axis). This restriction of the validation data improves the R^2 value (0.5 → 0.7) while the MAE decreases slightly (both ~0.2 [$\times$ 1.6 $\times$ 10^{-19} J]), however, will come at the cost of missing long-range pairs or inflating/diluting the training set with pairs that are likely to be negligibly small. Despite the small improvement, these curated data provide low predictive utility (Figure 6f). This failure of the curated set serves as a reminder that equilibrium simulations efficiently perform importance samplings of configurations, and that a uniform sampling of configurations in a similar range is an insufficient proxy for those configurations that matter most. Related, if training samples are selected from only a single simulation snapshot, it is best here to select them from crystalline morphologies because the relative absence of high $J_{i,j}$ in other morphologies disproportionately lowers the RF prediction accuracy (Supporting Information Section 4).

4 | CONCLUSION

The expensive quantum chemical calculation of electronic couplings ($J_{i,j}$) between P3HT chromophores need not be repeated if a representative training set of chromophores is used to train a machine to infer the couplings from chromophore features. We have shown that artificial neural networks and random forests are sufficiently predictive of $J_{i,j}$, resulting in expected charge mobilities. Here, random forests are recommended over artificial neural networks because we begin with a physical intuition for the features salient to $J_{i,j}$, so the RF ability to transparently rank feature importances and the ease of implementing RFs in Scikit-Learn give benefits at no added cost. We show that $J_{i,j}$ is obtained ~390× faster when the RF is used to look up ZINDO/S calculations, and we identify chromophore bonding, distance, "pitch," and sulfur-separation between chromophores to be the strongest predictors. Two conclusions arose from our investigations into minimal training sets: (a) The failure to accurately predict $J_{i,j}$ from a training set curated on chromophore separations and rotations informed by the ranked feature importance highlights the importance of drawing training data from a thermodynamic simulation method in which importance samplings of configurations are performed, and (b) Training sets as small as 1 $\times$ 10^5 chromophore pairs are sufficient to generate $J_{i,j}$ and resultant mobilities in agreement with prior work.[21] In sum, this work demonstrates one example of where significant computational speedups can be gained in exchange for a small amount of machine learning tuning. In future work, we look toward identifying other bottlenecks where RFs and ANNs will provide similar speedups, toward the automatic identification of molecular descriptors that allow the prediction of $\Delta E_{i,j}$, and extending this work to additional chemistries.

ACKNOWLEDGMENTS

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation

grant number ACI-1053575.[48] This material is based upon work supported by the National Science Foundation under Grant No. 1229709, 1653954, and 1835593. We would like to acknowledge high-performance computing support of the R2 compute cluster (DOI: 10.18122/B2S41H) provided by Boise State University's Research Computing Department. This research made use of the resources of the High-Performance Computing Center at Idaho National Laboratory, which is supported by the Office of Nuclear Energy of the U.S. Department of Energy and the Nuclear Science User Facilities under Contract No. DE AC07-05ID14517.

CONFLICT OF INTEREST

The authors declare no competing financial interest.

ORCID

Eric Jankowski https://orcid.org/0000-0002-3267-1410

REFERENCES

1. Espinosa N, Hösel M, Angmo D, Krebs FC. Solar cells with one-day energy payback for the factories of the future. *Energ Environ Sci*. 2012;5:5117.
2. Vandewal K, Himmelberger S, Salleo A. Structural factors that affect the performance of organic bulk heterojunction solar cells. *Macromolecules*. 2013;46:6379-6387.
3. Jun Y, Yunqiang Z, Liuyang Z, et al. Single-junction organic solar cell with over 15% efficiency using fused-ring acceptor with electron-deficient Core. *Joule*. 2019;3:1-12.
4. Lingxian M, Yamin Z, Wan X, et al. Organic and solution-processed tandem solar cells with 17.3% efficiency. *Science*. 2018;2612: eaat2612.
5. Xueping Y, Bhoj G, Iordania C, et al. Impact of nonfullerene molecular architecture on charge generation, transport, and morphology in PTB7-Th-based organic solar cells. *Adv Funct Mater*. 2018;28:1-9.
6. Dang MT, Hirsch L, Wantz G. P3HT:PCBM, best seller in polymer photovoltaic research. *Adv Mater*. 2011;23:3597-3602.
7. Mazzio KA, Luscombe CK. The future of organic photovoltaics. *Chem Soc Rev*. 2015;44:78-90.
8. Ma W, Yang C, Gong X, Lee K, Heeger AJ. Thermally stable, efficient polymer solar cells with Nanoscale control of the interpenetrating network morphology. *Adv Funct Mater*. 2005;15:1617-1622.
9. Chang J-F, Sun B, Breiby DW, et al. Enhanced mobility of poly (3-hexylthiophene) transistors by spin-coating from high-boiling-point solvents. *Chem Mater*. 2004;16:4772-4776.
10. Sirringhaus H, Brown PJ, Friend RH, et al. Two-dimensional charge transport in self-organized, high-mobility conjugated polymers. *Nature*. 1999;401:685-688.
11. Kline RJ, McGehee MD, Kadnikova EN, Liu J, Fréchet JMJ. Controlling the field-effect mobility of Regioregular Polythiophene by changing the molecular weight. *Adv Mater*. 2003;15:1519-1522.
12. Brinkmann M, Rannou P. Effect of molecular weight on the structure and morphology of oriented thin films of Regioregular poly (3-hexylthiophene) grown by directional epitaxial solidification. *Adv Funct Mater*. 2007;17:101-108.
13. Pandey SS, Takashima W, Nagamatsu S, Endo T, Rikukawa M, Kaneto K. Regioregularity vs regiorandomness: effect on photocarrier transport in poly(3-hexylthiophene). *Jpn J Appl Phys*. 2000;39: L94-L97.
14. Kim Y, Cook S, Tuladhar SM, et al. A strong regioregularity effect in self-organizing conjugated polymer films and high-efficiency polythiophene:fullerene solar cells. *Nat Mater*. 2006;5:197-203.
15. Ballantyne AM, Chen L, Dane J, et al. The effect of poly (3-hexylthiophene) molecular weight on charge transport and the performance of polymer:fullerene solar cells. *Adv Funct Mater*. 2008;18: 2373-2380.
16. Mauer R, Kastler M, Laquai F. The impact of polymer regioregularity on charge transport and efficiency of P3HT:PCBM photovoltaic devices. *Adv Funct Mater*. 2010;20:2085-2092.
17. Lan Y, Huang C. A theoretical study of the charge transfer behavior of the highly regioregular Poly-3-hexylthiophene in the ordered state. *J Phys Chem B*. 2008;112:14857-14862.
18. Jones ML, Huang DM, Chakrabarti B, Groves C. Relating molecular morphology to charge mobility in semicrystalline conjugated polymers. *J Phys Chem C*. 2016;120:4240-4250.
19. Jones ML, Jankowski E. Computationally connecting organic photovoltaic performance to atomistic arrangements and bulk morphology. *Mol Simulat*. 2017;43:1-18.
20. Van E, Jones ML, Jankowski E, Wodo O. Using graphs to quantify energetic and structural order in semicrystalline oligothiophene thin films. *Mol Syst Des Eng*. 2018;1:273-277.
21. Miller ED, Jones ML, Jankowski E. Tying together multiscale calculations for charge transport in P3HT: structural descriptors, morphology, and tie-chains. *Polymers*. 2018;10:1358.
22. Greco C, Melnyk A, Kremer K, Andrienko D, KCh D. Generic model for lamellar self-assembly in conjugated polymers: linking mesoscopic morphology and charge transport in P3HT. *Macromolecules*. 2019;52: 968-981.
23. Miller ED, Jones ML, Henry MM, Chery P, Miller K, Jankowski E. Optimization and validation of efficient models for predicting polythiophene self-assembly. *Polymers*. 2018;10:1305.
24. Marcus RA. On the theory of oxidation-reduction reactions involving electron transfer. *J Chem Phys*. 1956;24:966.
25. Ridley J, Zerner M. An intermediate neglect of differential overlap technique for spectroscopy: pyrrole and the azines. *Theor Chim Acta*. 1973;32:111-134.
26. Kirkpatrick J. An approximate method for calculating transfer integrals based on the ZINDO Hamiltonian. *Int J Quantum Chem*. 2008; 108:51-56.
27. Gómez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel TD, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat Mater*. 2016;15:1120-1127.
28. Pyzer-Knapp EO, Li K, Aspuru-Guzik A. Learning from the Harvard clean energy project: the use of neural networks to accelerate materials discovery. *Adv Funct Mater*. 2015;25:6495-6502.
29. Jackson NE, Bowen AS, Antony LW, Webb MA, Vishwanath V, de Pablo JJ. Electronic structure at coarse-grained resolutions from supervised machine learning. *Sci Adv*. 2019;5:1-38.
30. Mueller T, Kusne A, Ramprasad R. Machine learning in materials science: recent Progress and emerging applications. In: Parrill AL, Lipkowitz KB, eds. *Reviews in computational chemistry*. Hoboken, NJ: John Wiley & Sons; 2016:186-273.
31. Ferguson AL. Machine learning and data science in soft materials engineering. *J Phys Condens Matter*. 2018;30:043002.
32. Venkatasubramanian V. The promise of artificial intelligence in chemical engineering: is it here, finally? *AIChE J*. 2019;65:466-478.
33. Pilania G, Wang C, Jiang X, Rajasekaran S, Ramprasad R. Accelerating materials property predictions using machine learning. *Sci Rep*. 2013; 3:1-6.
34. Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. *NPJ Comput Mater*. 2017;3:1-27.

35. Liu Y, Zhao T, Ju W, Shi S. Materials discovery and design using machine learning. *J Materiomics*. 2017;3:159-177.
36. Leibowitz MH, Miller ED, Henry MM, Jankowski E. Application of artificial neural networks to identify equilibration in computer simulations. *J Phys*. 2017;921:012013.
37. Neese F. The ORCA program system. *Wiley Interdiscip Rev Comput Mol Sci*. 2012;2:73-78.
38. Jones M. L., Henry, M.M. matty-jones/MorphCT: MorphCT v3.0. Available from: https://zenodo.org/record/1243843#.XV6pVkdS_IU; 2018. doi: https://doi.org/10.5281/zenodo.1243843.
39. Deng W-Q, Goddard WA. Predictions of hole mobilities in oligoacene organic semiconductors from quantum mechanical calculations. *J Phys Chem B*. 2004;108:8614-8621.
40. 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.
41. Johansson E, Larsson S. Electronic structure and mechanism for conductivity in thiophene oligomers and regioregular polymer. *Synth Met*. 2004;144:183-191.
42. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Available from: https://arxiv.org/abs/1603.04467; 2016.
43. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2012;12:2825-2830.
44. Miller ED, Jones ML, Jankowski E. Machine learning predictions of electronic couplings for charge transport calculations of P3HT. *AIChE J*. 2019. https://doi.org/10.5281/zenodo.2635495
45. Miller ED, Jones ML, Jankowski E. Machine learning for structure-performance relationships in organic semiconducting devices. *Comput Mater Eng Lab*. 2018;3. https://doi.org/10.18122/cme_lab/3/boisestate
46. Bredas JL, Calbert JP, da Silva Filho DA, Cornil J. Organic semiconductors: a theoretical characterization of the basic parameters governing charge transport. *Proc Natl Acad Sci USA*. 2002;99:5804-5809.
47. Coropceanu V, Cornil J, da Silva FDA, Olivier Y, Silbey R, Brédas J-L. Charge transport in organic semiconductors. *Chem Rev*. 2007;107:926-952.
48. Towns J, Cockerill T, Dahan M, et al. XSEDE: accelerating scientific discovery. *Comput Sci Eng*. 2014;16:62-74.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Miller ED, Jones ML, Henry MM, Stanfill B, Jankowski E. Machine learning predictions of electronic couplings for charge transport calculations of P3HT. *AIChE J*. 2019;e16760. https://doi.org/10.1002/aic.16760