

# Physics-Constrained Machine Learning for Predicting Photovoltaic Performance of 3-Hexylthiophene Decamer:JSM5 Active Layer Integrating Quantum Chemistry and Molecular Dynamics.

Eninges Asmare

[eninges.asmare@wu.edu.et](mailto:eninges.asmare@wu.edu.et)

Wollo University <https://orcid.org/0000-0001-6126-3128>

---

## Research Article

### Keywords:

**Posted Date:** October 27th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11490519/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** The authors declare no competing interests.

---

# Physics-Constrained Machine Learning for Predicting Photovoltaic Performance of 3-Hexylthiophene Decamer:JSM5 Active Layer Integrating Quantum Chemistry and Molecular Dynamics.

**Eninges Asmare**

*Department of Physics, College of Natural Sciences, Wollo University, Ethiopia.*

*eninges.asmare@wu.edu.et*

October 24, 2026

## **Abstract**

This study presents an integrated computational methodology that couples density-functional theory and time-dependent DFT descriptors, all-atom molecular-dynamics (MD) morphology, and physics-constrained machine-learning pipeline applied to 3HT decamer:JSM5 donor-acceptor blend. The oligothiophene series ( $n = 1-10$ ) exhibits near-convergent frontier-orbitals beyond  $n \approx 6$  and monotonically declining reorganization energies (0.251 eV at  $n=6$  to 0.210 eV at  $n=10$ ), justifying the use of the  $n = 10$  oligomer as the finite-length model of the donor polymer. According to NTO analysis, the JSM5 excitation is classified as locally excited with charge-transfer (CT) mixing, whereas the donor excitation is classified as delocalized Frenkel-like. Because morphology descriptors do not exist for the 2,080 literature device training dataset, latent morphology variables, derived from the residuals of the device targets that molecular features fail to explain, are appended to the training table as placeholder inputs, into which the MD-derived descriptors of the new blend could be inserted during prediction. An ensemble of Random Forest, Extra-Trees, Gradient Boosting, and an ARD Matérn Gaussian Process, and a Physics-constrained neural network (PCNN) were trained on dataset with molecular electronic descriptors only and on molecular + latent morphology variables. OPV performance predictions for the 3HT decamer:JSM5 blend show that adding MD-derived descriptors raises the blended (Ensemble/PCNN) PCE from  $\approx 4.0\%$  to  $\approx 4.5\%$ , while the PCNN model alone improves from  $\approx 2.4\%$  to  $\approx 3.7\%$ .

## **1 Introduction**

Over the past decades, the photovoltaic performance of organic solar cells (OSCs) has advanced to exceed a 20% power conversion efficiency (PCE) milestone.<sup>1,2</sup> This performance growth stems largely from the introduction of non-fullerene acceptors<sup>3</sup> that feature tunable energy levels and minimize non-radiative recombination losses<sup>4,5</sup> that enabled simultaneous gains in short-circuit current density ( $J_{SC}$ ), open-circuit voltage ( $V_{OC}$ ), and

fill factor (FF).<sup>6,7</sup> Despite these milestones, the interdependence of key photovoltaic parameters still poses fundamental challenges. Specifically, trying to minimize voltage losses often leads to a simultaneous drop in the fill factor due to field-dependent free-charge generation and geminate recombination at the donor-acceptor interfaces.<sup>8</sup> Consequently, tuning the active layer is essential to improve device performance.<sup>9</sup>

Computational modelings at the quantum mechanical and atomistic levels have become essential complements to experimental characterizations, giving mechanical insights into the electronic and morphological factors that dictate photovoltaic performance. Density functional theory (DFT) and time-dependent DFT (TDDFT) provide essential molecular descriptors - such as frontier orbital energies, reorganization energies, and exciton binding energies - that control charge transfer and separation at the molecular level.<sup>10,11</sup> On the other hand, molecular dynamics (MD) simulations reveal how donor and acceptor materials self-assemble, phase-separate, and form  $\pi$ -stacking within the bulk heterojunction active layer.<sup>12</sup> Recent multiscale frameworks couple all-atom and coarse-grained MD with sequential quantum mechanics/molecular mechanics (QM/MM) calculations. This approach quantifies how structural disorder localizes frontier orbitals and modifies charge-transfer state energetics in polymer films.<sup>13</sup> Furthermore, integrating MD-generated morphologies with Marcus theory<sup>14</sup> enables predicting mesoscopic charge hopping rates and carrier mobilities. This bridges the gap between molecular packing and device-level transport properties.<sup>15,16</sup>

The increasing complexity of organic photovoltaic materials design has catalyzed the adoption of machine learning (ML) approaches,<sup>17-20</sup> which complement earlier empirical design rules that relate frontier-orbital energies to PCE.<sup>21</sup> However, the conventional data-driven models lack physical interpretability and can violate fundamental device constraints such as the detailed balance between  $J_{SC}$ ,  $V_{OC}$ , FF, and PCE.<sup>22-24</sup> Physics constrained ML frames overcome these limitations by embedding domain knowledge directly into model architectures such as physics-informed neural network<sup>25-27</sup> and symbolic regression integrating quantum mechanical descriptors.<sup>28</sup> Recent studies have demonstrated that constraining ML models with physical consistency equations not only enhances predictive accuracy for photovoltaic parameters but also yields interpretable structure-property relationships.<sup>28,29</sup> Such methodologies establish a robust pipeline for translating quantum chemical and morphological data into actionable predictions of device performance.

This study focuses on 3-Hexylthiophene (3HT) decamer:JSM5 active layer, a donor-acceptor combination unexplored computationally. JSM5 is a  $\pi$ -extended,  $\beta$ -side-chain-free Y6-type fused-ring acceptor that has demonstrated 17.3% efficiency when paired with the polymer donor PM6,<sup>30</sup> within the broader family of Y-series and  $\pi$ -extended acceptors that have pushed efficiencies beyond 15-20%.<sup>31,32</sup> The donor, a decameric oligomer of P3HT leverages the electron-rich character and strong  $\pi$ - $\pi$  stacking propensity of thiophene units, which enhance conjugated planarization and reduce packing distances in photovoltaic polymers.<sup>33</sup> By combining DFT/TD-DFT for electronic structure, all-atom MD for morphology, and physics-constrained ML framework this study aims to (i) investigate how supplying MD-simulated morphology descriptors, alongside DFT/TD-DFT descriptors, changes the predicted device performance of this blend and (ii) establish a computational protocol that could be used to predict photovoltaic performance of new materials, see Figure 1.

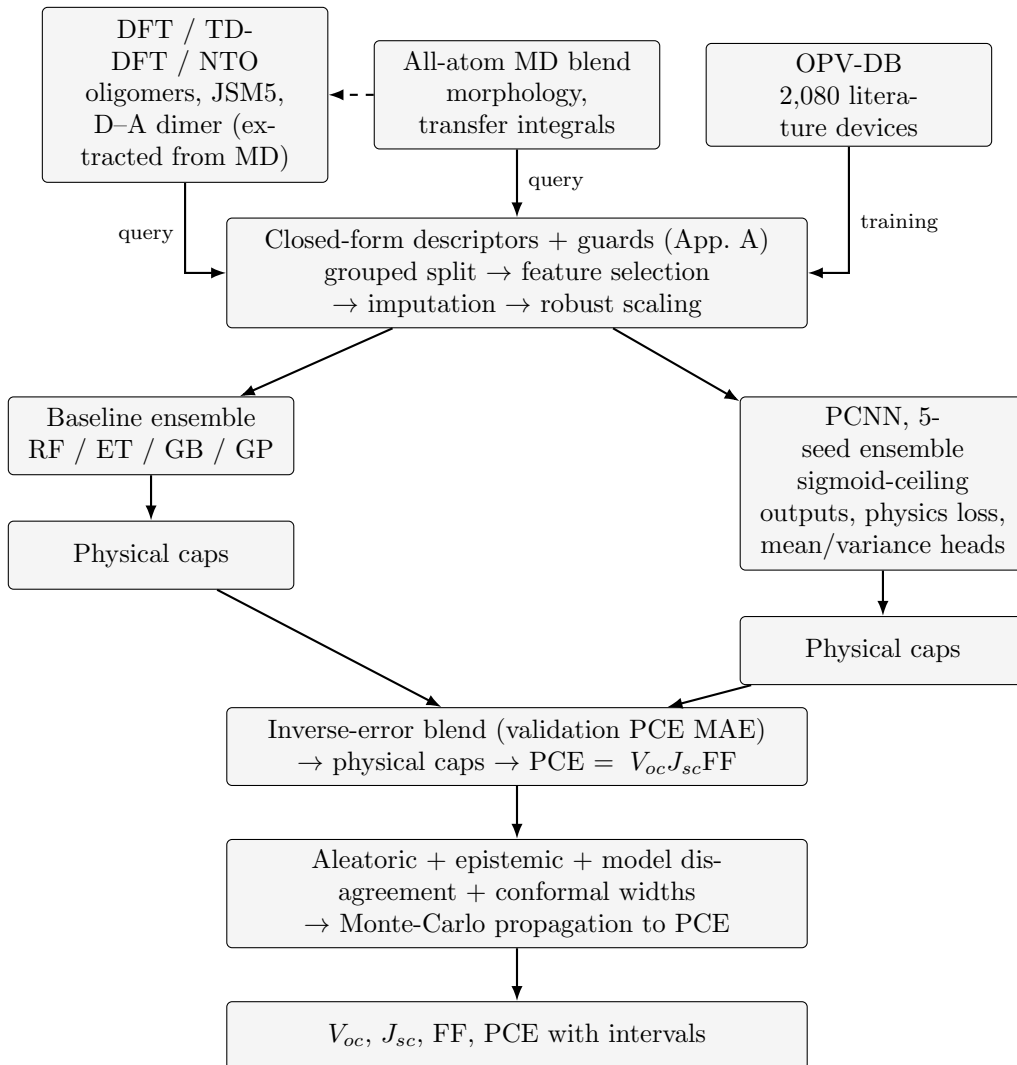


Figure 1: Overview of the multiscale computational workflow. Quantum-chemical and molecular dynamics (MD) calculations provide molecular and aggregate-level descriptors, while literature device data (OPV-DB) are used to train the baseline ensemble and physics-constrained neural network (PCNN). The model predictions are subject to physical constraints and combined using validation-error-based weights to estimate power conversion efficiency (PCE). Predictive uncertainties, including aleatoric and epistemic contributions, model disagreement, and conformal prediction widths, are propagated to PCE through Monte Carlo sampling.

## 2 Computational Methods

### 2.1 Quantum Mechanical Calculations

All ground-state geometry optimizations and electronic structure calculations were performed using the Gaussian 16 (Revision C.01)<sup>34</sup> software package. Ground-state geometries of polythiophene oligomers ( $n = 1 - 10$ ) and the JSM5 molecule were optimized in the gas phase at the B3LYP level of theory. For the polythiophene oligomers, structural optimizations were benchmarked using both the split-valence 6-31G(d, p)<sup>35</sup> and the correlation-consistent cc-pVTZ<sup>36</sup> basis sets. The geometry of JSM5 was optimized utilizing the 6-31G(d, p) basis set. Because the alkyl side chains negligibly affect the electronic properties, they were represented as methyl groups in the calculations in all oligomer and acceptor structures. Frequency calculations were performed at the same level of theory to confirm that all optimized structures represented true local minima on the potential energy surface, evidenced by the absence of imaginary frequencies. The torsional potential energy surface (PES) of 3HT decamer was mapped by performing a relaxed scan of the inter-ring dihedral angle from  $0^\circ$  to  $360^\circ$  at  $30^\circ$  increments. Charge transport properties were evaluated by calculating the internal reorganization energies ( $\lambda$ ) via the standard four-point adiabatic potential energy method.<sup>37,38</sup> Vertical excitation energies, oscillator strengths, and electronic transition characteristics were simulated via TD-DFT using the range-separated hybrid CAM-B3LYP functional.<sup>39</sup> Natural Transition Orbital (NTO)<sup>40</sup> analysis was subsequently performed to visualize and characterize the nature of the optically excited states. Post calculation wavefunction-based analyses (electrostatic potential, MPP/SDP) were carried out with Multiwfn analyzer.<sup>41,42</sup>

### 2.2 Molecular Dynamics Simulations

Classical molecular dynamics (MD) simulations were performed using the LAMMPS<sup>43</sup> package under periodic boundary conditions in all three spatial dimensions. The simulation systems were constructed with Moltemplate.<sup>44</sup> The systems were described using the all-atom OPLS-AA<sup>45-48</sup> force field. Bonded interactions included harmonic bond and angle terms, OPLS dihedrals, and CVFF-form improper dihedrals. Non-bonded interactions were modeled using 12-6 Lennard-Jones plus Coulomb potential with geometric combining rules, a 1-4 scaling factors of 0.5, and 1-2 and 1-3 interactions. Real-space nonbonded interactions were truncated at 12.0 Å, and long-range electrostatics were evaluated using the particle-particle particle-mesh (PPPM)<sup>49</sup> method.

Prior to equilibration, atomic overlaps in the initial configuration were relieved by a soft push-off stage. All nonbonded interactions were temporarily replaced by a soft repulsive potential (cutoff 7.5 Å), while the bonded terms were kept active. The potential's prefactor was ramped linearly from 0 to 60 kcal mol<sup>-1</sup> over 10,000 steps of 0.1 fs, with the maximum atomic displacement limited to 0.01 Å per step. The full force field was then restored, and the structure was relaxed by conjugate-gradient minimization (energy and force tolerances of  $10^{-6}$ ; maximum 5000 iterations and 50,000 force evaluations).

The minimized system was then equilibrated in stages. Velocities were initialized from a Gaussian distribution at 300 K, and a 2.5 ps NVE run (0.25 fs timestep) was performed with a displacement-limited integrator (maximum 0.05 Å per step) to remove residual structural instabilities. The timestep was then increased to 1.0 fs. The system was equilibrated for 50 ps in the NVT ensemble at 300 K using a Nosé-Hoover thermostat<sup>50,51</sup> (damping time 100 fs), followed by 200 ps in the NPT ensemble at 300 K and 1 atm

with isotropic pressure coupling (barostat damping 5000 fs) to relax the cell toward the bulk density. A further 100 ps NVT run at 300 K stabilized the structure after volume relaxation.

Production MD was carried out in the NVT ensemble at 300 K for 500 ps (1.0 fs timestep). Wrapped atomic coordinates with periodic image flags were saved every 5 ps for subsequent structural and electronic-structure analyses.

## 2.3 Performance Prediction

The workflow is implemented as a Python pipeline that parses ground-state DFT, TD-DFT, natural-transition-orbital (NTO) and molecular-dynamics (MD) outputs and combines them with the closed-form photophysical descriptors in Appendix A. The device targets  $V_{oc}$ ,  $J_{sc}$  and FF are learned from a literature compilation (OPV-DB;<sup>52</sup> 2,080 devices after consistency filtering and unit standardization), and the power conversion efficiency is computed as  $PCE = V_{oc}J_{sc}FF$ . Each record contains `sample_id`, donor and acceptor SMILES,  $V_{oc}$ ,  $J_{sc}$ , FF, PCE, device architecture (conventional or inverted), and the donor and acceptor HOMO, LUMO and optical gap. The features omit processing and morphology variables, which largely determine OPV performance; the models therefore may learn little beyond HOMO/LUMO energy-level trends.

**Data splitting** Records are divided 70/15/15 into training, validation, and test sets by grouped shuffling, so that no donor–acceptor (D–A) pair appears in more than one split. Group labels are taken from a configured group column, the donor–acceptor SMILES pair, other material or pair identifier columns, a pattern in `sample_id`, or identical rounded D–A descriptor tuples. The validation set is used for early stopping, blend weights and conformal calibration; the test set is held out for final evaluation.

**Model families** Two complementary model families are trained on the same splits and features:

- i) a *baseline ensemble*, trained separately for each target, of Random Forest,<sup>53</sup> Extra-Trees,<sup>54</sup> Gradient Boosting<sup>55</sup> and a Gaussian Process with an automatic-relevance-determination (ARD) Matérn kernel,<sup>56</sup> combined with fixed weights of 0.15, 0.30, 0.30 and 0.25, respectively; and
- ii) a *Physics-Constrained Neural Network* (PCNN) with mean and log-variance output heads. Each target  $x \in \{V_{oc}, J_{sc}, FF\}$  is bounded by a sample-specific ceiling,

$$x = x_{\min} + (x_{\max} - x_{\min}) \sigma(z), \quad (1)$$

where  $z$  is the raw network output and  $\sigma$  is the logistic function. By default,  $x_{\max} = E_{\text{gap}}^{\text{DA}} \equiv IP_D - EA_A$  for  $V_{oc}$ , and  $x_{\max} = FF_0^{\text{Green}}(V_{oc})$  for FF, so both bounds hold by construction. A ceiling  $J_{sc} \leq m J_{sc}^{\text{AM1.5G}}(E_g)$ , with margin  $m$ , is also available. Each ceiling is disabled automatically if more than 5% of training labels violate it.

The training loss combines a heteroscedastic Gaussian negative log-likelihood with soft physics penalties on  $V_{oc} \leq E_{\text{gap}}^{\text{DA}}$ ,  $J_{sc} \leq J_{sc}^{\text{opt}} \eta_{\text{diss}} \eta_{\text{coll}}$ ,  $J_{sc} \leq J_{sc}^{\text{opt}} \eta_{\text{CS}}$  and the Green fill-factor limit.<sup>57</sup> Samples lacking the efficiency features required by a given penalty are masked from that term rather than imputed. The reported PCNN estimate is the mean over five random seeds.

**Blending and physical caps** The two families are blended using weights inversely proportional to their validation PCE mean absolute error (MAE), clamped to  $[0.25, 0.75]$ . A two-tier guard acts on the relative PCE difference  $\Delta_{\text{PCE}}$  between the families. If  $\Delta_{\text{PCE}} > 0.5$ , the baseline weight is capped at 0.45. If  $\Delta_{\text{PCE}} > 1.0$ , the weight of the better-validated family is raised to between 0.60 and 0.75. The blended prediction is then truncated at two hard physical caps: an energetic cap on  $V_{oc}$ , and an optical cap on  $J_{sc}$ .

**Uncertainty quantification** The total uncertainty combines aleatoric (PCNN variance head), epistemic (Monte-Carlo dropout and between-seed spread),<sup>58,59</sup> model-disagreement and conformal<sup>60</sup> terms. Conformal widths are calibrated on the validation split using PCNN predictions. Because this split also drives early stopping and blend weights, the widths are floored by out-of-fold grouped cross-validation (CV) widths of the baseline ensemble, so that no reported  $\sigma$  is narrower than the baseline’s cross-validated error.

PCE uncertainty is propagated by correlated Monte-Carlo sampling with 20,000 draws, performed in log space for  $V_{oc}$  and  $J_{sc}$  and in logit space for FF. Draws that violate the physical caps are rejected rather than clipped. The resulting intervals are widened by explicit data-quality penalties, applied for an IP/EA scissor correction, a vibronic fallback, an unreliable local-excitation-charge-transfer (LE-CT) coupling, deep-inverted charge separation, tension with a mechanistic ceiling, and conflicting dissociation metrics.

## 2.4 Latent Morphology Representation

Morphological descriptors like molecular orientation, phase separation, domain size, and crystallinity are known to be important in determining organic solar-cell performance.<sup>61</sup> On the other hand, the training dataset has molecular descriptors only. Hence, a different pipeline is used to build a *latent morphology representation* learned from the training devices and predicted from molecular features. The latent variables are derived from the photovoltaic targets, through the part of the targets that molecular features fail to explain. The latent variables are placeholder inputs that allow the trained pipeline to accept MD descriptors for a new blend.

**Ablation design.** The Python pipeline (Section 2.3) is run twice.

- *Without the morphology block (**Without**).* Training uses molecular and electronic descriptors only. The new blend is queried with its own molecular and electronic descriptors.
- *With the morphology block (**With**).* The latent columns are appended to every training device, and the model is fitted on this extended table. The new blend is queried with the same molecular and electronic descriptors as in the **Without** arm, and its MD-derived descriptors of Section 3.3 are included.

**Construction.** (1) *Out-of-fold reference residuals.* A random forest (300 trees, max depth 8, min split 4, min leaf 2) is fit per target on molecular features alone (morphology columns, targets, PCE, and identifiers excluded); five-fold shuffled (not group-aware) out-of-fold predictions give residuals  $r_{ik}$ . (2) *Residual decomposition.* Residuals are standardized, reduced by PCA to  $\min(3, \cdot)$  scores, and, when fewer than four components are

available (always, for three targets), supplemented by the residual magnitude  $m_i = \|\mathbf{z}_i\|_2$  as the fourth axis, assigned in fixed order to  $(\pi\text{-distance}, R_{DA}, S, X_c)$ . (3) *Rank mapping* [Eq. 21]. (4) *Molecular mapper*. A ridge regression with cross-validated penalty ( $\alpha \in [10^{-3}, 10^3]$ ) is fit from standardized molecular features to each  $\tilde{x}_{id}$ , with predictions clipped to  $[\ell_d, u_d]$ .

**Leakage control.** Because the proxies derive from the targets, using them naively as predictors of the same targets would be circular. A nested cross-fit is therefore applied: for each of five outer folds, steps (1)–(4) are repeated on the fitting rows only, and the resulting mapper predicts the held-out rows. Every training device receives out-of-fold latent values produced without its own targets; validation and test devices receive predictions from the final mapper fit on the training partition only. All other preprocessing (feature engineering, selection, imputation, scaling) is likewise fit on the training partition only, and the exported metadata record the target-derived status explicitly.

### 3 Results and Discussion

In order to justify the  $n = 10$  oligomer as a representation of the donor polymer employed in the blend, the polythiophene oligomer series is utilized as a computationally traceable proxy for the 3HT decamer donor throughout, measuring convergence with chain length. The excited state and charge-transport-relevant characteristics of the isolated donor oligomers and the JSM5 acceptor are described after the molecular level. Additionally described is the absorption of the D-A dimer, which probes the blend’s interface. The MD snapshots are used to extract the D-A dimer. Lastly, the machine-learned device performance predictions are discussed, along with their sensitivity to morphology, uncertainty, and the extent of the evidence. The goal of this effort is to demonstrate a protocol that integrates DFT/MD descriptors with ML performance prediction for new donor - acceptor blends.

#### 3.1 Ground-State Molecular Properties

##### HOMO - LUMO (H-L) Energy Levels

Figures 2 and 3, respectively, show front and side views of B3LYP/6-31g(d, p) optimized geometries of the 3HT decamer donor and JSM5 acceptor. The 3HT decamer adopts a near-planar backbone in its front view, with the side view confirming a coplanar arrangement of the thiophene rings that is favorable for intramolecular  $\pi$ -conjugation. The JSM5, on the other hand, has a C-shaped front view. From an electronic perspective, the near-coplanarity of 3HT decamer favors  $p$ -orbital overlap across the backbone, promoting delocalization and effective  $\pi$ -conjugation. For JSM5, while the C-shape limits continuous extendable  $\pi$ -electron delocalization over the entire molecule, the deliberate inclusion of electron-withdrawing terminal groups and a rigid central core preserves local donor-acceptor interactions necessary for modulating the bandgap.

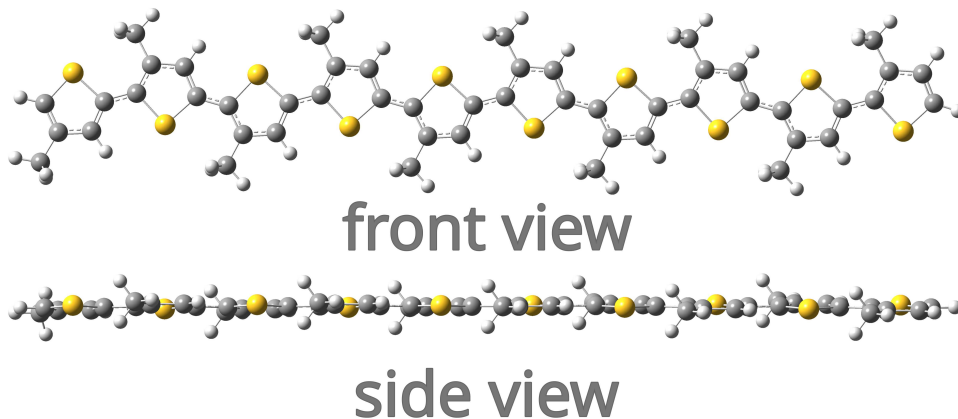


Figure 2: Front and side views of the B3LYP/6-31g(d, p) optimized geometry of the 3HT decamer donor.

The geometry of polythiophene oligomers from 1mer to 10mer were optimized using 6-31g(d, p) and cc-pVTZ basis sets and the corresponding H-L energy levels were extracted as depicted in Figure 4. The basis set effect on the H-L energy levels over the entire oligomers is small, demonstrating robustness of the results. The H-L trend clearly shows the energy gap narrows as the conjugation length increases. The gap-narrowing

rate decreases with conjugation length, relatively stabilizing after 6mer. The H-L gap decreased by 0.88 eV from 3mer to 6mer and by 0.33 eV from 6mer to 10mer, while it decreases by only 0.05 eV from 9mer to 10mer. This suggests that around 6 to 10 units effective conjugation length is nearly reached, which implies additional repeat units contribute progressively less to electronic delocalization.

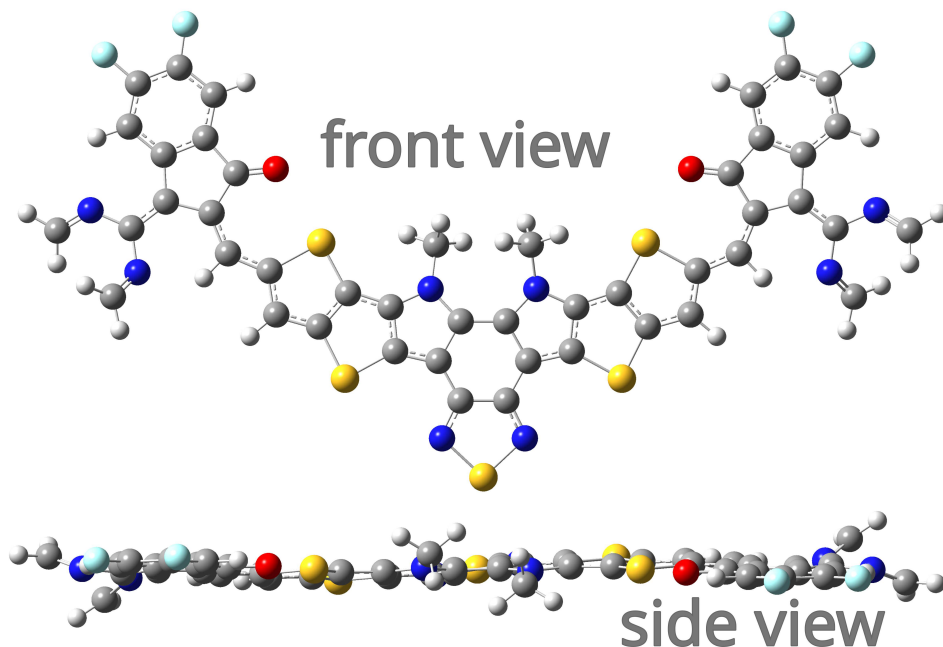


Figure 3: Front and side views of the B3LYP/6-31g(d, p) optimized geometry of the JSM5 acceptor.

Moreover, extrapolating the H-L gap of the oligomers, Figure 5a, to that of infinite repeat units using the modified Kuhn free-electron equation,<sup>62</sup> Equation 2, shows the gap nearly converges to 2.09 eV for 10mer close to the measured gap of P3HT, 1.9 eV to 2.1 eV as reported in the literature.<sup>33</sup> Therefore, the H-L energy gap value of 10mer is valid to be used as a representative donor unit in the device-level prediction.

$$E_g(N) = E_0 \sqrt{1 - 2\beta \cos \left( \frac{\pi}{N+1} \right)} \quad (2)$$

Here  $E_0$  is the effective HOMO-LUMO gap of an uncoupled conjugated unit,  $\beta$  is the dimensionless effective coupling between neighboring units,  $N = 2n$  is the number of linearly conjugated C=C bonds in an oligomer with  $n$  thiophene units, and  $E_g(N)$  is the HOMO-LUMO gap of that oligomer.

Figure 4 also shows the H-L energy level of JSM5 for comparison with polythiophene oligomers. The H-L level of JSM5 compared to polythiophene 10mer, shows the H-L alignment that is potentially suitable for strong electron-withdrawing character and potentially enhanced donor-acceptor charge separation.

### Electrostatic Potential (ESP) Analysis

Figure 6 depicts the ESP map of polythiophene oligomers ( $n = 1$  to  $n = 10$ ), and clearly shows that as conjugation length increases ESP delocalizes along the backbone. This

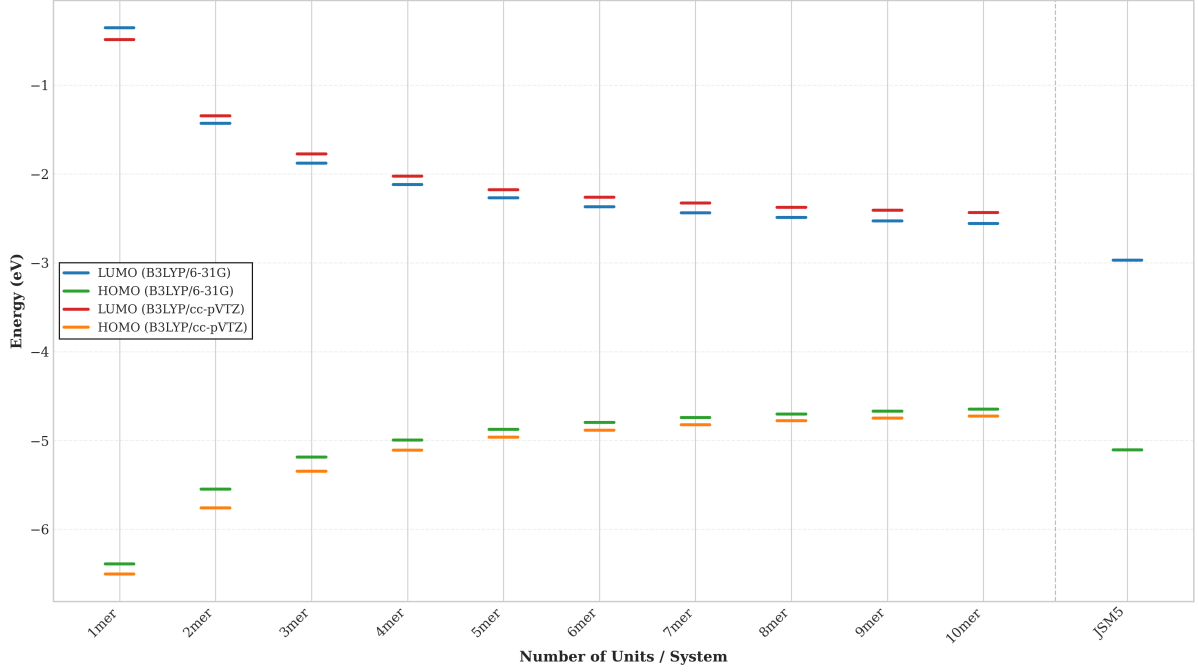


Figure 4: Frontier molecular orbital energies of polythiophene oligomers (1mer - 10mer, illustrating convergence of the band gap in larger oligomers) and JSM5. Results compare 6-31g(d, p) (blue/green) and cc-pVTZ (red/orange) basis sets.

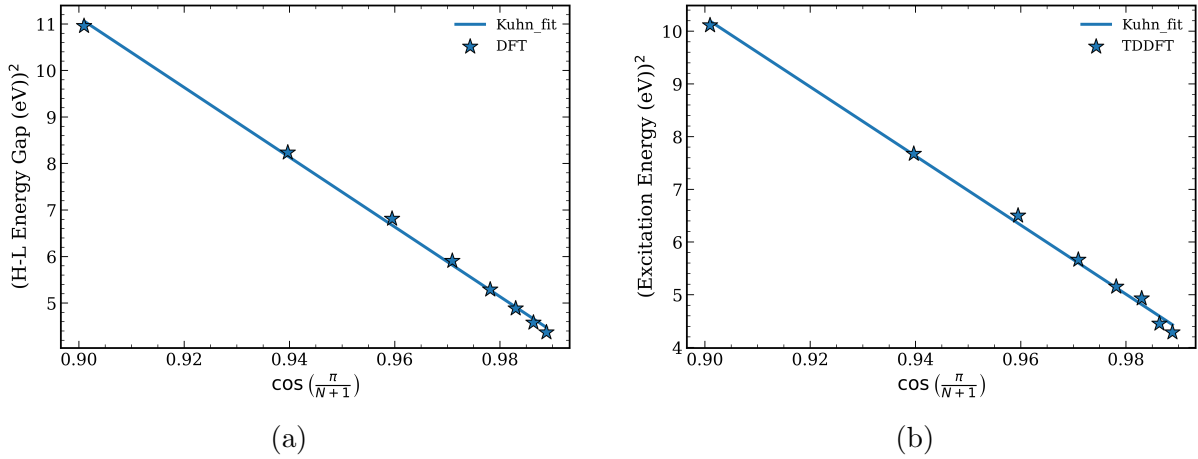


Figure 5: Kuhn free-electron extrapolation (Eq. 2) of (a) the ground-state HOMO–LUMO gap and (b) the lowest TD-DFT singlet excitation energy, both plotted against  $\cos[\frac{\pi}{N+1}]$  for  $n = 3$ –10.

delocalization suggests, as conjugation length increases charge density becomes evenly distributed, polarization becomes less localized, and exciton diffusion should increase which in turn supports enhanced charge transport in longer chains. In fact the ESP map also shows the potential distribution stops changing appreciably beyond about 6 units, supporting the effective conjugation length of about 6-10 units discussed above.

The extended ESP along the backbone is correlated with higher HOMO delocalization, larger integral transfer, and lower reorganization energy. More importantly, the conjugated backbone containing sulfur acts as the dominant electron-rich region indicating high electron density a typical property of conjugated polymers. This is evidenced from the global surface minimum extracted using multiwfn analyser.

### Torsional Potential Energy Surface (PES)

Table 1 reports the molecular planarity parameter (MPP) which reflects the overall deviation of the molecular structure from a fitting plane and the span of deviation from the plane (SDP) that represents the deviation span of the structure relative to the fitting plane.<sup>63</sup> Both MPP and SDP values of a more planar molecular structure is expected to be low. In this regard, the planarity decreases with chain length which is expected as a consequence of chain flexibility. As the MPP and SDP values of JSM5 are greater than the decamer of the donor it also signifies that the acceptor is less planar than the donor.

Table 1: MPP and SDP for the neutral and charged (cationic for donor, anionic for acceptor) optimized geometries of the  $n = 6-10$  of polythiophene donor and JSM5 acceptor.

|                     |     | 6mer    | 7mer    | 8mer    | 9mer    | 10mer   | JSM5    |
|---------------------|-----|---------|---------|---------|---------|---------|---------|
| Neutral Optimized   | MPP | 0.00904 | 0.02251 | 0.06736 | 0.08813 | 0.10782 | 0.26543 |
|                     | SDP | 0.04485 | 0.11132 | 0.33661 | 0.44172 | 0.53823 | 1.92286 |
| Cat/anion Optimized | MPP | 0.04644 | 0.05067 | 0.08468 | 0.10729 | 0.14216 | 0.21849 |
|                     | SDP | 0.15862 | 0.18537 | 0.29896 | 0.38200 | 0.46669 | 1.58314 |

Both metrics increase monotonically with oligomer length in the neutral state, from 0.009 Å/0.045 Å (MPP/SDP) at  $n = 6$  to 0.108 Å/0.538 Å at  $n = 10$ , and are larger still for JSM5 (0.265 Å/1.923 Å). This is the expected consequence of chain flexibility: longer backbones have more rotatable inter-ring bonds and more possible steric clashes between adjacent rings, so some residual non-planarity accumulates even though every individual inter-ring torsion remains close to its planar minimum (see the torsional analysis below). The two observations are not contradictory, because MPP/SDP integrate small local deviations over the whole chain, whereas the PES analysis characterizes each dihedral independently. Together with the frontier-orbital results above, this indicates that repeat units beyond the  $n \approx 6-10$  range add comparatively little electronic delocalization but do add some conformational (planarity) disorder.

Backbone planarity is a prerequisite for the delocalization inferred above, so it was tested directly by relaxing dihedral scans of every inter-ring torsion in the decamer from 0° to 360° with 30° interval, as depicted in Figure 8. Inter-ring torsional angle numbered sequentially along the backbone as defined in Figure 7. The resulting potential energy surface (PES), Figure 8, indicates a global minima near 0°, symmetric maxima near 90° and 270°, and a shallow local feature near 180°. These peaks show that planar conformations maximize  $\pi$  overlap and orthogonal conformations disrupt conjugation. All inter-ring

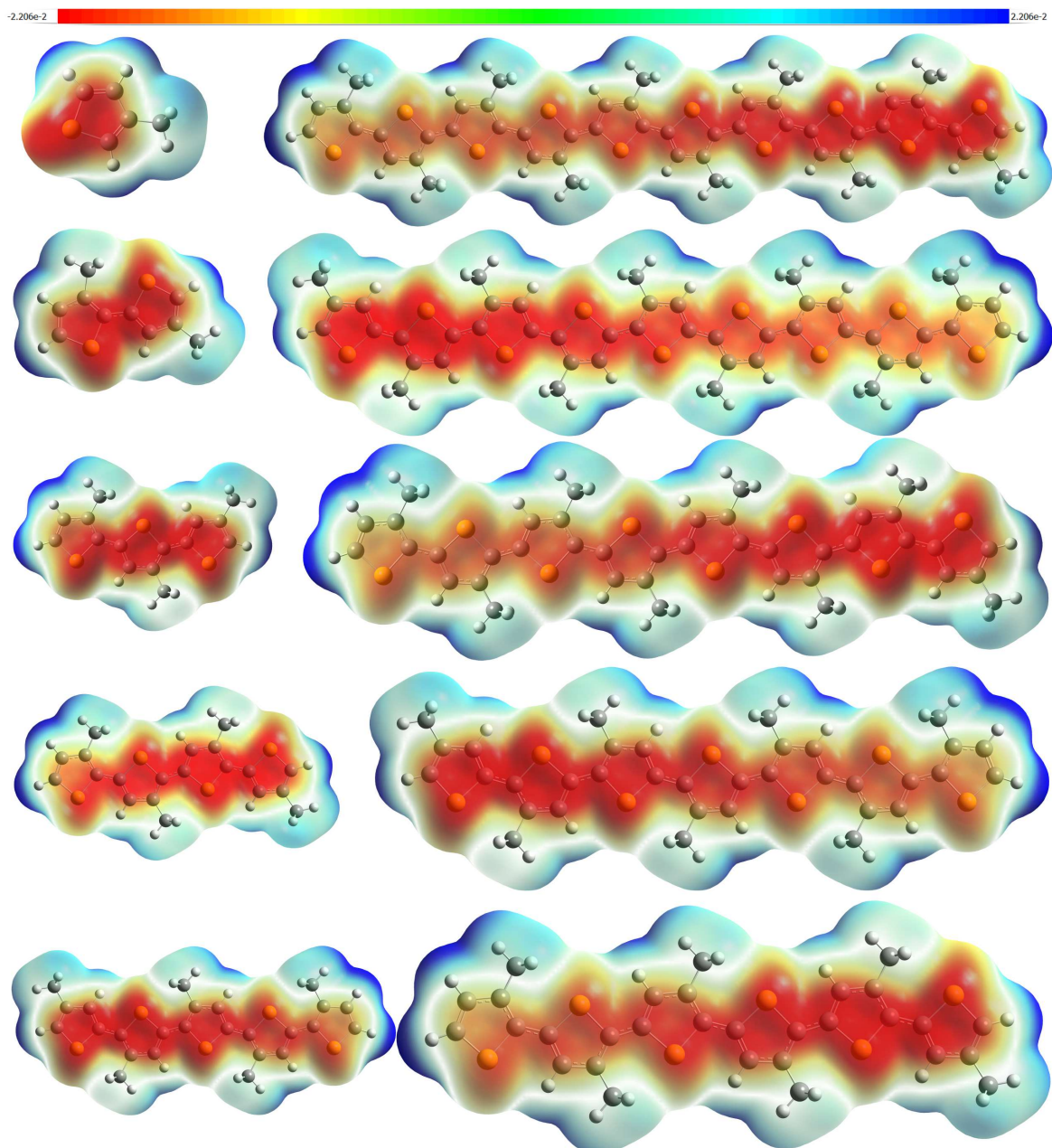


Figure 6: Electrostatic potential distribution on polythiophene oligomers ( $n = 1$  (top left) to 10 (top right)) mapped on the  $\rho = 0.001 \text{ a.u.}$  electron-density isosurface ( $\pm 2.206 \times 10^{-6} \text{ a.u.}$ ), with red regions illustrating higher electron density (more negative potential) and blue regions illustrating lower electron density.

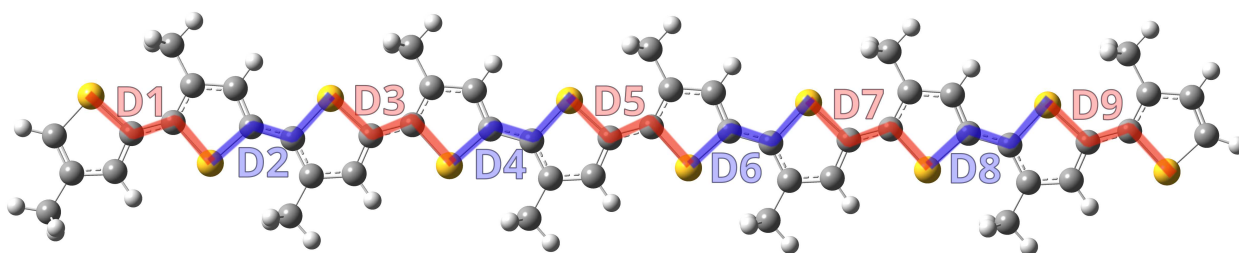


Figure 7: Notation for the nine sequential inter-ring dihedral angles (D1 - D9) scanned along the decamer backbone for PES calculation.

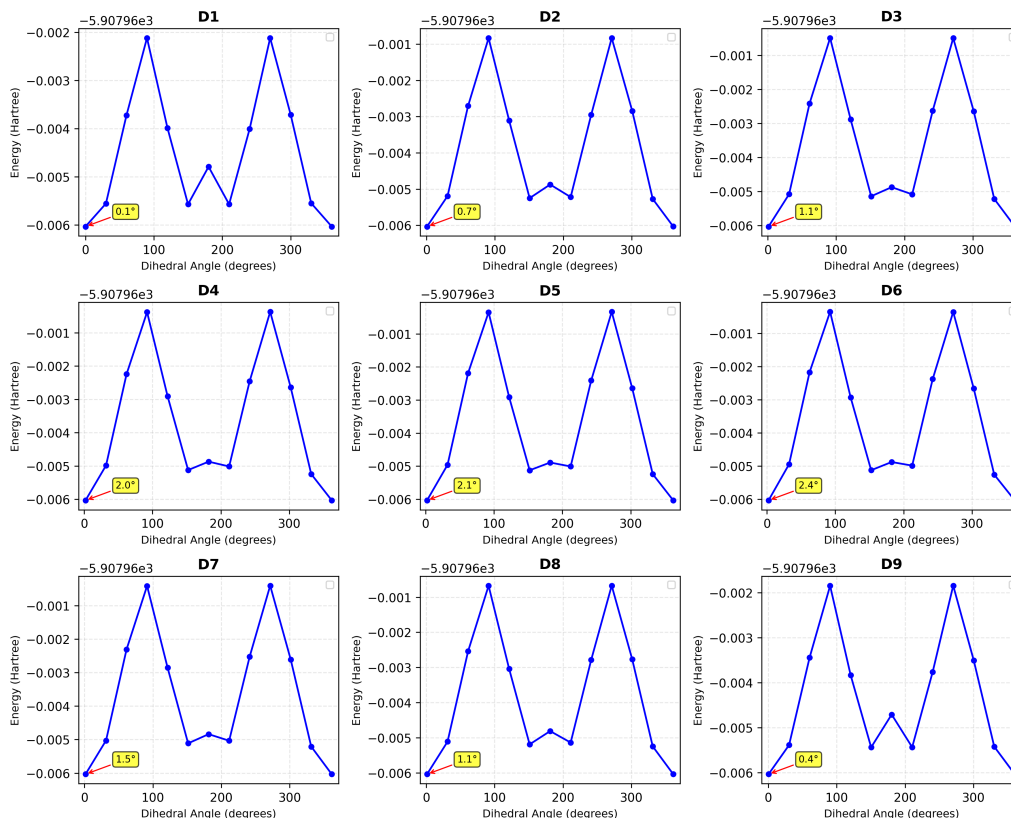


Figure 8: Relaxed torsional PES for the dihedrals D1 - D9 of the decamer, scanned from 0° to 360° in 30° increment (electronic energy reported in Hartree).

torsions relax to values below 3°, indicating an intrinsically planar backbone that facilitates extensive  $\pi$  electron delocalization. On the contrary, the small rotational barrier indicates moderate resistance to twisting, sufficient to preserve planarity at ambient conditions while allowing thermally accessible fluctuations. The low variation among the equilibrium torsions suggests minimal static conformational disorder, which is expected to enhance intrachain electronic delocalization and charge transport. These findings are consistent with the TDDFT and NTO analysis, indicating that the polymer possesses a rigid conjugated framework favorable for photovoltaic and charge-transport applications.

### 3.2 Excited-State and Charge-Transport Properties

Figure 5b the TDDFT of the  $n = 3$  to 10 fitted with Kuhn model supports  $n = 10$

Having established that  $n = 6 - 10$  oligomers approach the converged electronic structure of the donor, TDDFT absorption spectra were computed for  $n = 6 - 10$ , for JSM5, and to probe the interface, for a D-A dimer extracted from the MD blend, Figure 9. The oligomer series shows a systematic red-shift and increasing intensity of the lowest energy ( $\pi - \pi^*$ ) absorption band from  $n = 6 - 10$ , in agreement with the ground-state gap narrowing in Figure 4 providing independent, excited state support for the same chain length trend.

JSM5 features a peculiar absorption profile that includes a low-energy band similar to the  $n = 7$  oligomer characteristic, as well as a higher-energy manifold of transitions near 3.5 - 4.5 eV. The spectroscopic characteristic of the more complicated excited-state landscape anticipated for an acceptor-donor-acceptor NFA architecture is this higher-

energy manifold, which represents excitations localized on or delocalized across the fused core and terminal groups of JSM5. However, this profile does not give a complementary absorption with the donor as expected for a D-A blend. This is because the TDDFT of the small molecule was run for only a single chain and resulted in a far blue shifted profile compared to the experimental absorption of JSM5. This suggests recalculation of TDDFT for a number of chains. (to be calculated)

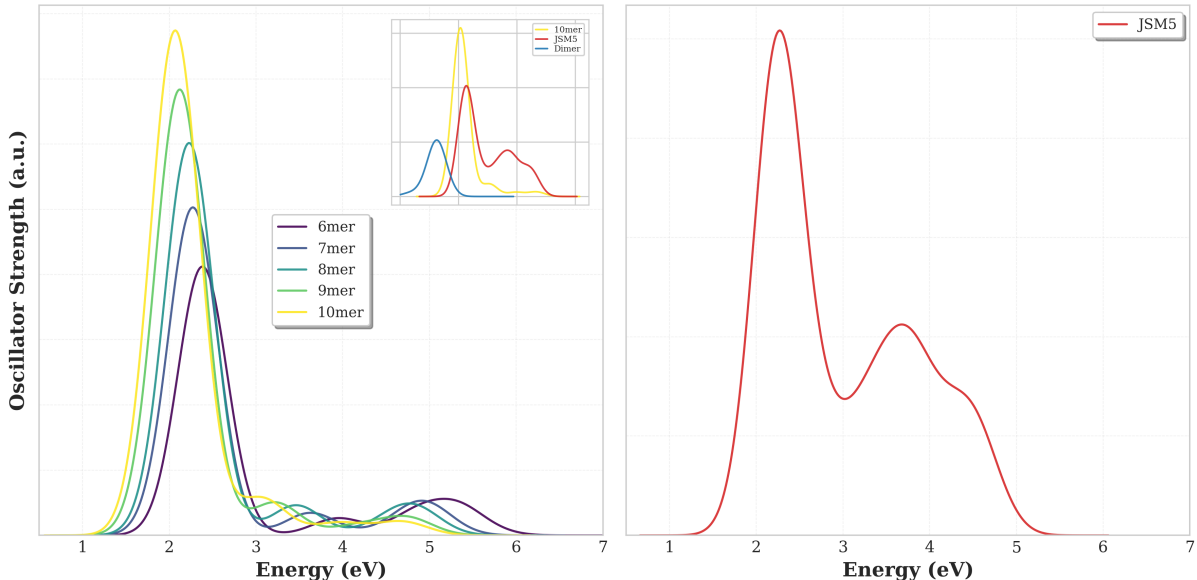


Figure 9: Calculated UV-Vis absorption profiles of polythiophene oligomers, 6mer to 10mer (left) and JSM5 (right). Inset illustrates comparison on oscillator strength and peak absorption energy of 3HT decamer, JSM5, and D-A dimer.

The comparison of the TD-DFT spectra for the 3HT decamer, JSM5, and the D-A dimer, inset of Figure 9, is consistent with the emergence of intermolecular charge-transfer (CT) character in the dimer. The pronounced red shift of the dimer’s lowest absorption band relative 3HT decamer and JSM5 indicates a stabilization of the CT excited state, which is typical when donor and acceptor moieties are electronically coupled. A more reliable assessment of intermolecular CT quality would require natural transition orbitals (NTO) analysis on the dimer (yet to be calculated)

### Reorganization energy ( $\lambda$ )

Table 1 reports two complementary planarity metrics, MPP and SDP, each evaluated for the neutral and the charged optimized geometries. Both metrics increase monotonically with oligomer length in the neutral and charged states. In the neutral state, MPP and SDP rise from 0.009 Å and 0.045 Å, respectively, at  $n = 6$  to 0.108 Å and 0.538 Å at  $n = 10$ . In the charged state, they rise from 0.046 Å and 0.159 Å to 0.142 Å and 0.467 Å over the same range.

The neutral-to-charged comparison in Table 1 (cationic donors, anionic JSM5) shows that while MPP is larger in the cation than in the neutral molecule for every  $n = 6-10$  (by 0.017–0.037 Å), SDP is larger in the cation for  $n = 6-7$  (by 0.11 and 0.07 Å) but smaller for  $n = 8-10$  (by 0.04–0.07 Å). Polaron-induced planarization, a phenomenon common in conducting polymers, is only observed for  $n \geq 8$  only on SDP. The discrepancy on MPP might be due to the methyl side chain. JSM5 shows a clearer effect: both MPP and

SDP decrease upon reduction (MPP  $0.265 \rightarrow 0.218$  Å; SDP  $1.923 \rightarrow 1.583$  Å), i.e., the anion is more planar than the neutral molecule, which is consistent with charge-induced relaxation toward a more conjugated geometry that would stabilize an electron on JSM5.

The reorganization energies of the donor and acceptor were obtained with the four-point scheme using Equation 3. Table 2 reports the internal reorganization energy,  $\lambda$ , computed by the four-point  $\Delta$ SCF method<sup>37,38</sup> (Eq. 3) for hole transport along the oligomer series and for electron transport along the JSM5.  $\lambda$  decreases monotonically from 0.251 eV at  $n = 6$  to 0.210 eV at  $n = 10$ , and JSM5 has the lowest value of the set (0.176 eV). Because the nonadiabatic hopping rate<sup>11,14,64,65</sup> depends exponentially on  $-\lambda/4k_B T$  (Equation 16), this  $\sim 0.04$ – $0.08$  eV reduction is expected to translate into a modest increase in the intrinsic hopping rate at fixed electronic coupling. The trend is directionally consistent with the frontier-orbital and ESP analyses (delocalization improves with chain length over the range studied), and the low  $\lambda_e$  is favorable for electron transport on JSM5.

$$\begin{aligned}\lambda_h &= E_{D,\text{neutral}}^+ - E_{D,\text{neutral}}^0 + E_{D,\text{cation}}^0 - E_{D,\text{cation}}^+ \\ \lambda_e &= E_{A,\text{neutral}}^- - E_{A,\text{neutral}}^0 + E_{A,\text{anion}}^0 - E_{A,\text{anion}}^-\end{aligned}\tag{3}$$

Table 2: Reorganization energy

|                        |       | $\lambda$ (eV) |
|------------------------|-------|----------------|
| polythiophene Oligomer | 6mer  | 0.251          |
|                        | 7mer  | 0.239          |
|                        | 8mer  | 0.229          |
|                        | 9mer  | 0.219          |
|                        | 10mer | 0.210          |
| JSM5                   |       | 0.176          |

### Natural Transition Orbitals (NTO)

Figure 10 shows NTO decomposition<sup>40</sup> of the dominant TD-DFT transition for  $n = 10$  oligomer. Hole and electron NTO pairs co-localize over the same set of thiophene rings along the backbone, with two dominant transition weights ( $\lambda_1 = 0.803$  leading transition and  $\lambda_2 = 0.142$  a secondary, weaker transition). This spatial overlap of hole and electron NTOs is a distinctive characteristic of a locally excited (Frenkel-like), delocalized  $\pi$ – $\pi^*$  state rather than a charge-transfer state, consistent with the chemically homogeneous, non-polar backbone of the oligothiophene donor.

JSM5 presents a qualitatively different picture, see Figure 10. The dominant transition ( $\lambda = 0.845$ ) shows a hole NTO concentrated on the central fused core, while the corresponding electron NTO retains substantial amplitude on the core but also extends measurably onto the peripheral terminal substituents; the secondary transition ( $\lambda = 0.116$ ) shows a similar but more pronounced hole/electron spatial offset. This partial (rather than complete) spatial separation of hole and electron density is consistent with a state that is best described as locally-excited with charge transfer (CT)-admixture: substantially more charge-transfer character than either oligothiophene excitation, but

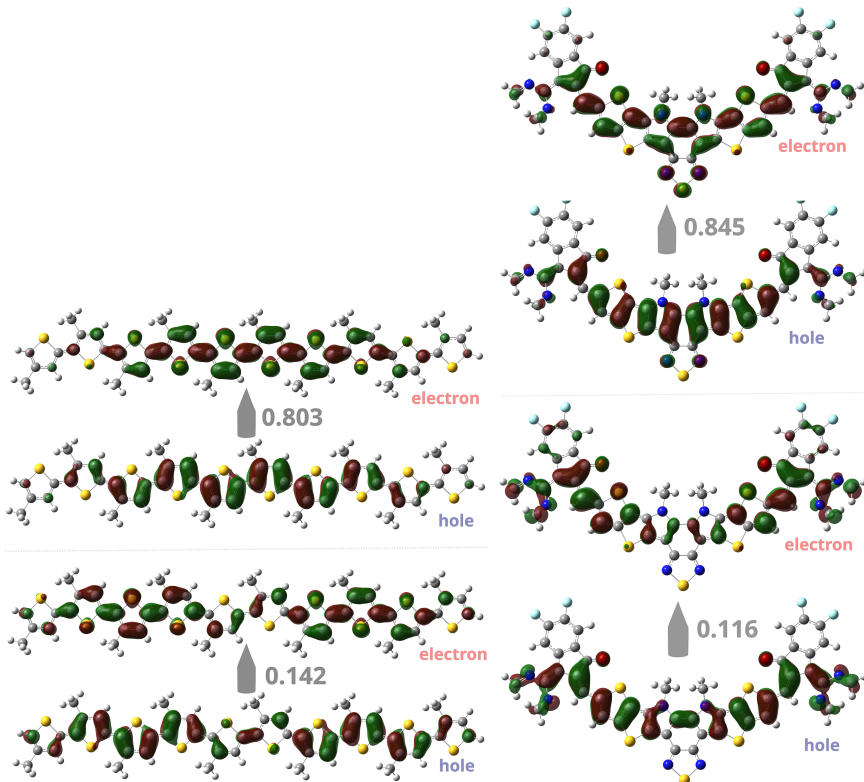


Figure 10: NTO of 3HT decamer and JSM5

not the fully separated hole/electron pair that would be expected for an intermolecular D–A CT state at the blend interface. This distinction matters for the diabatic locally excited (LE)–CT mixing framework of Eq. 15: it indicates that JSM5’s intramolecular excited state already sits partway toward CT character even before intermolecular mixing with the donor is considered, which should lower the effective barrier for the subsequent intermolecular LE→CT step relative to a purely local excitation on either component.

### 3.3 MD-Driven Morphology and Molecular Packing

Figure 11 shows a representative snapshot of the equilibrated production cell (500 ps NVT, 300 K) of the decamer:JSM5 blend and a D–A dimer extracted from it. The cell shows partially interpenetrating donor (blue) and acceptor (green) regions with rare D–A contact, qualitatively signifying poorly intermixed domains. Because charge transport in amorphous blends proceeds by thermally activated hopping between localized sites, transfer-integral statistics extracted from MD morphologies<sup>15</sup> are best interpreted within Marcus-type hopping<sup>14</sup> or, for strongly disordered systems, Miller–Abrahams and Gaussian-disorder pictures.<sup>66–68</sup> From these configurations the morphology descriptors ( $\pi$ -stacking distance,  $R_{DA}$ ,  $S$ ,  $X_c$ , aggregation sizes) and the transfer-integral statistics used by the pipeline are computed.

The descriptors in Table 3 describe a weakly ordered, essentially amorphous blend. The orientational order parameter,  $S = 0.055$ , is close to the isotropic limit ( $S = 0$ ), and the crystalline fraction is low ( $X_c = 0.2$ ). The largest aggregate contains only two molecules a single donor–acceptor network. This is consistent with the qualitative picture of Fig. 11 (rare D–A contact). The  $\pi$ -stacking distance, 3.70 Å, lies in the range usually quoted for thiophene-based donors and Y-type acceptors (roughly 3.5–3.8 Å). The chain

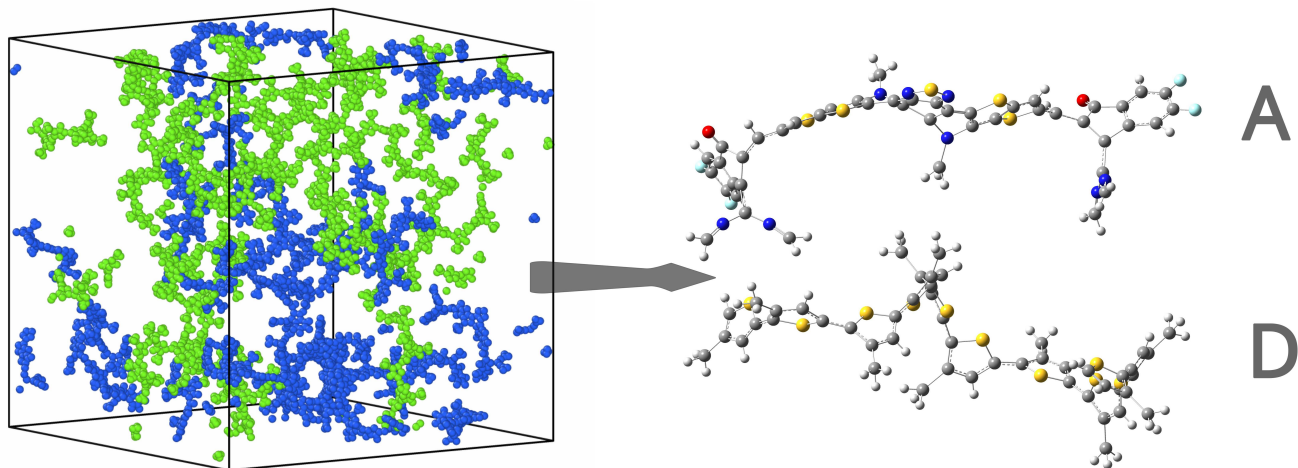


Figure 11: Left: simulation cell of the equilibrated decamer (donor, blue):JSM5 (acceptor, green) blend after 500 ps NVT production at 300 K. Right: a representative D–A dimer extracted from the same configuration, used for coordinate-based transport-descriptor analysis.

Table 3: Average core morphology descriptors extracted from snapshot of the equilibrated production.

| Feature                     | Value    | Likely meaning                              |
|-----------------------------|----------|---------------------------------------------|
| Zip_Rg                      | 11.00 Å  | Chain radius of gyration                    |
| Lp                          | 11.82 Å  | Persistence length                          |
| S                           | 0.055    | Orientational order parameter               |
| pi_distance                 | 3.70 Å   | $\pi$ – $\pi$ nearest-neighbor distance (Å) |
| Xc                          | 0.2      | Crystalline fraction                        |
| agg_size_max / agg_size_min | 2 / 1    | Largest and smallest aggregates observed    |
| J_hole_mean                 | 1.84 meV | Hole transfer integrals                     |
| J_electron_mean             | 1.53 meV | Electron transfer integral                  |
| disorder_hole               | 0.604 eV | Site-energetic disorder                     |
| disorder_electron           | 0.672 eV | Site-energetic disorder                     |

size descriptors are compatible with the near-planar DFT geometries:  $R_g = 11.0$  Å is what a rigid rod of about 38 Å (ten repeat units at  $\approx 3.8$  Å) would give ( $L/\sqrt{12} \approx 11$  Å), while  $L_p = 11.8$  Å is only about a third of that contour length, indicating some bending. The hole and electron transfer integrals (1.84 and 1.53) and the energetic disorder (0.604 and 0.672) are of similar magnitude for the two carriers, which suggests balanced rather than strongly asymmetric transport. A site-energy disorder of 0.6–0.7 eV would be several times larger than values typically reported for amorphous organic semiconductors and three to four times the reorganization energies, which would place transport deep in a trap-limited regime. Table 3 should therefore be read as the as-built, amorphous limit of the blend, not as the morphology of a processed (annealed or solvent-additive-treated) film. **The MD simulation needs recalculation because the side chains on both D and A moities were not included**

### 3.4 Performance Prediction Comparison

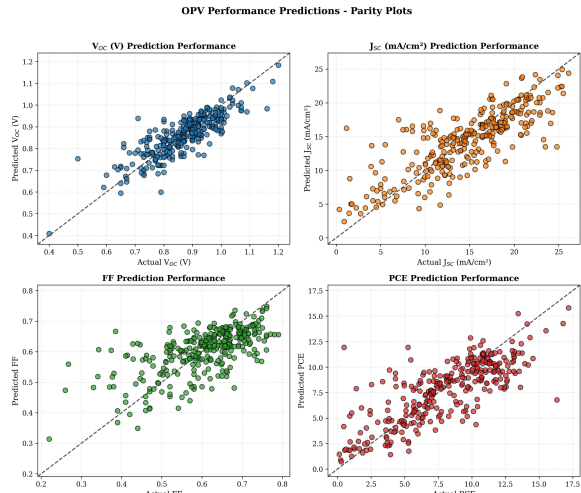
The OPV performance of the decamer:JSM5 blend is evaluated using two ablation arms, **Without** and **With** the morphology block, as described in Section 2.4. To assess the predictive performance and uncertainty of the pipeline, the models were evaluated on a held-out test set of literature-reported devices. The corresponding parity plots and prediction intervals are presented in Figures 12a and 12b, respectively, using predictions from the **Without** morphology-block arm, while the quantitative results are summarized in Table 4. Overall, the models exhibit moderate predictive accuracy, with performance varying across target properties. Open-circuit voltage ( $V_{oc}$ ) is the most accurately predicted property, with test  $R^2$  values of 0.687–0.701, followed by short-circuit current density ( $J_{sc}$ ;  $R^2 = 0.610$ –0.687) and power conversion efficiency (PCE;  $R^2 = 0.542$ –0.595). Fill factor (FF) exhibits the weakest predictive performance ( $R^2 = 0.375$ –0.379), consistent with its sensitivity to morphology, charge-carrier recombination, and processing conditions, which may not be adequately represented by the molecular descriptors alone.

Table 4:  $R^2$  on the grouped held-out test split for the baseline ensemble and the PCNN.

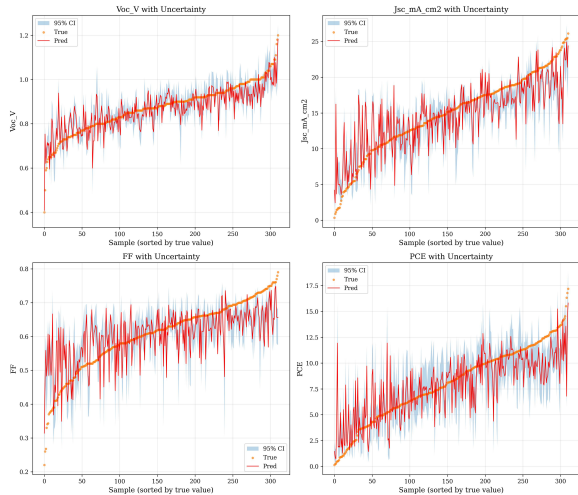
| Metric                          | Baseline Test $R^2$ | PCNN test $R^2$ |
|---------------------------------|---------------------|-----------------|
| $V_{oc}$ (V)                    | 0.701               | 0.687           |
| $J_{sc}$ (mA cm <sup>-2</sup> ) | 0.687               | 0.610           |
| FF                              | 0.375               | 0.379           |
| PCE (%)                         | 0.595               | 0.542           |

The parity plots in Figure 12a further reveal systematic patterns in the prediction errors. Predictions of  $V_{oc}$  generally follow the parity line, although deviations increase at the upper end of the measured range. For example, a device with a measured  $V_{oc}$  of approximately 1.16 V is underpredicted at around 0.98 V. For FF, the predicted values are concentrated within approximately 0.4–0.75, whereas the measured values span a wider range of 0.22–0.79. Consequently, the models overpredict low FF values and underpredict high ones. Similar regression toward the mean is observed for  $J_{sc}$  and PCE, accompanied by several substantial prediction errors. These discrepancies indicate that the models have difficulty capturing extreme values within the heterogeneous literature dataset. Such behavior can arise from limited training dataset features, see Section 2.3. The results suggest that the models capture some of the underlying relationships between

molecular descriptors and device performance but have limited accuracy at the extremes of the experimental distributions. Accordingly, predictive performance should be assessed using both overall statistical metrics and errors across the target-property range, rather than relying on  $R^2$  alone.



(a) Parity plots of predicted versus measured  $V_{oc}$ ,  $J_{sc}$ , FF and PCE on the grouped held-out test split ( $n \approx 310$  devices). The dashed line is  $y = x$ .



(b) Predictions with 95% predictive intervals on the same held-out test split. Devices are sorted by measured value (orange); the red line is the predicted value and the shaded band is the 95% interval.

Figure 12: Learning performance of the models

The prediction intervals in Figure 12b provide additional information about the reliability of individual predictions. For much of the prediction range, the measured values fall within the nominal 95% intervals, particularly for  $V_{oc}$  and PCE. The interval widths increase modestly toward the lower end of the predicted ranges for  $J_{sc}$ , FF, and PCE, while pronounced local increases in width occur for individual devices exhibiting greater disagreement among ensemble predictions. These variations suggest that the estimated uncertainty can help identify predictions for which the model outputs are less consistent. However, several observations at the extremes fall outside the predicted intervals: high-FF and high- $V_{oc}$  measurements exceed the upper bounds, whereas low- $J_{sc}$  and low-PCE measurements fall below the lower bounds when the corresponding predictions overestimate the measured values. This pattern suggests that the intervals may underestimate uncertainty in the tails of the target distributions, where regression toward the mean is most pronounced. If the intervals are derived primarily from ensemble disagreement, they may capture aspects of model uncertainty without fully accounting for experimental variability and other sources of predictive error. The intervals should therefore be interpreted as estimates of model-associated uncertainty rather than as fully calibrated measures of total predictive uncertainty.

The limited learning, seen as regression toward the mean at the extremes of Voc, Jsc, FF, and PCE and as prediction intervals that miss tail observations, most likely arises because the input features (donor and acceptor molecular descriptors, HOMO/LUMO levels, optical gaps, and device type) omit the processing and morphology variables that largely control OPV performance, so the model can capture only energy-level trends and cannot explain the device-to-device variation that drives extreme values. To keep

this limitation explicit rather than claim a fix, the training data are augmented with latent morphology variables derived from the photovoltaic targets themselves that the molecular features fail to explain. These latents are not intended to supply the omitted processing–morphology physics; because they are target-derived, they can only absorb residual outcome variation and may encode the same unexplained device-to-device variation. Their practical role is to provide morphology-compatible input channels and placeholder slots so the trained pipeline can later accept molecular dynamics (MD)-derived descriptors for a new blend. Thus, for an unseen donor–acceptor pair, the learned latents can be replaced with computed morphology descriptors, allowing morphology information to enter prediction through a consistent interface, while the underlying omission of processing variables remains a limitation to be addressed with measured or simulated process–morphology data.

Table 5: Predicted device metrics for the decamer:JSM5 blend from the baseline ensemble, the PCNN, and their blend, **Without** and **With** the morphology block.

| Metric                          | baseline       |             | PCNN           |             | Blend          |             |
|---------------------------------|----------------|-------------|----------------|-------------|----------------|-------------|
|                                 | <b>Without</b> | <b>With</b> | <b>Without</b> | <b>With</b> | <b>Without</b> | <b>With</b> |
| $V_{oc}$ (V)                    | 0.954          | 0.916       | 1.011          | 0.943       | 0.992          | 0.929       |
| $J_{sc}$ (mA cm <sup>-2</sup> ) | 11.295         | 10.062      | 4.667          | 7.711       | 7.650          | 8.909       |
| FF                              | 0.560          | 0.574       | 0.500          | 0.507       | 0.527          | 0.541       |
| PCE (%)                         | 6.038          | 5.288       | 2.385          | 3.685       | 3.998          | 4.478       |

Table 5 gives the predictions for the decamer:JSM5 blend from the two ablation arms and their blend. In the **With** arm the morphology inputs of the new blend are the MD-derived descriptors of Table 3 and DFT, TDDFT and NTO driven descriptor that are used in the **Without** arm. The predictions for the new blend (Table 5) are internally consistent but depend strongly on the model arm. The blended estimate lies between the baseline and PCNN values for every metric. The **With** arm shifts the blended PCE only from 3.998% to 4.478% (+0.48 percentage points, about 12% relative), driven by  $J_{sc}$  (7.650  $\rightarrow$  8.909 mA cm<sup>-2</sup>) and FF (0.527  $\rightarrow$  0.541) and partly offset by a lower  $V_{oc}$  (0.992  $\rightarrow$  0.929 V). Without the MD-derived descriptors the baseline and PCNN differ by 6.63 mA cm<sup>-2</sup> in  $J_{sc}$  and 3.65 percentage points in PCE (6.038 vs. 2.385%). With it, they differ by 2.35 mA cm<sup>-2</sup> and 1.60 percentage points (5.288 vs. 3.685%). The MD-derived descriptors move the two architectures in opposite directions for  $J_{sc}$  (−1.23 for the baseline, +3.04 mA cm<sup>-2</sup> for the PCNN) and for PCE (−0.75 vs. +1.30 percentage points), and in the same direction only for  $V_{oc}$  (decrease) and FF (increase). The reduction in inter-model spread is compatible with the MD-derived descriptors supplying useful information.

**Limitations of the work** Despite capturing some relationships between molecular descriptors and OPV performance, the present framework has some limitations. First, the input features are insufficient to fully represent the factors governing device performance, particularly processing-dependent morphology, charge-carrier transport, and recombination. Second, a potential source of uncertainty arises from the difference between the descriptor sources used for model training and prediction. The training dataset primarily relies on experimentally reported molecular and electronic descriptors, whereas the descriptors for the new decamer blend are obtained from DFT, TDDFT, NTO, and MD

calculations. Differences between experimentally reported and computationally derived descriptors may introduce systematic discrepancies and limit the transferability of the trained models to new donor–acceptor systems. Third, the latent morphology descriptors require further validation and optimization against independently obtained morphology information, either from experimental characterization or from computational approaches such as MD simulations. In the current framework, these target-derived latent variables primarily capture residual variation in device performance rather than explicitly representing the underlying processing–morphology relationships. Consequently, their inclusion should not be interpreted as a complete solution to the omission of morphology-related physics.

## 4 Conclusion

This work presented an end-to-end multiscale study of a 3HT decamer:JSM5 active layer that connects DFT/TD-DFT electronic structure, all-atom MD morphology and a physics-constrained ML framework trained on 2,080 literature devices. Because the literature dataset contains no morphology descriptors, latent morphology variables were appended to the training table. They are derived from the part of the device targets that molecular features fail to explain, and serve solely as input slots into which the MD descriptors of the new blend can be supplied.

At the molecular level, the oligothiophene series approaches a converged electronic structure at  $n \approx 6$ –10: the HOMO–LUMO gap narrows by 0.33 eV from  $n = 6$  to  $n = 10$  but by only 0.05 eV from  $n = 9$  to  $n = 10$ , and the ESP distribution stops changing appreciably beyond about six units. Kuhn extrapolation places the decamer gap at 2.09 eV, within the 1.9–2.1 eV range reported for P3HT, which supports the decamer as the finite-length model of the donor. Relaxed torsional scans show an intrinsically planar backbone, although MPP and SDP grow with chain length, and JSM5 is less planar than the donor. The reorganization energy decreases from 0.251 eV ( $n = 6$ ) to 0.210 eV ( $n = 10$ ), and JSM5 has the lowest value (0.176 eV). The NTO analysis classifies the donor excitation as delocalized and locally excited (Frenkel-like), and the JSM5 excitation as locally excited with appreciable intramolecular CT admixture ( $\lambda = 0.845$ ).

On held-out literature devices (grouped split,  $n \approx 310$ ), the pipeline reproduces the data with moderate accuracy (test  $R^2$  for the baseline ensemble/PCNN:  $V_{oc}$  0.701/0.687,  $J_{sc}$  0.687/0.610, PCE 0.595/0.542, FF 0.375/0.379). Predictions regress toward the mean at the extremes, and the nominal prediction intervals under-cover in the tails, which is attributed to the absence of processing and morphology variables from the training features. The pipeline is therefore a usable but limited reference model for the comparison below; improving its predictive accuracy was not the objective.

For the decamer:JSM5 blend, supplying the MD-derived descriptors shifts the blended PCE prediction from 4.0% to 4.5% (+0.48 percentage points), through  $J_{sc}$  ( $7.65 \rightarrow 8.91$  mA cm<sup>-2</sup>) and FF ( $0.527 \rightarrow 0.541$ ), partly offset by a lower  $V_{oc}$  ( $0.992 \rightarrow 0.929$  V). The two model families respond in opposite directions in  $J_{sc}$  and PCE (baseline PCE:  $6.04 \rightarrow 5.29\%$ ; PCNN PCE:  $2.39 \rightarrow 3.69\%$ ), and their disagreement narrows from 3.65 to 1.60 percentage points. Because the latent variables are target-derived surrogates predicted from molecular features alone, the shift reflects the response of the trained pipeline to MD inputs entered through these slots; it cannot be attributed to a physical morphology effect of the simulated descriptors. The narrower inter-model spread is compatible with, but does not establish, that the descriptors carry useful information.

The influence of MD descriptors on predicted performance will require validation and optimization against independently obtained morphology information, together with sensitivity analysis of the latent-slot construction. Within these limitations, this work provides an electronic-structure characterization of the 3HT decamer blend and a systematic procedure for incorporating MD-derived descriptors into a literature-trained model.

# A Physics-Constrained Descriptor Equations

This section collects, and connects, the closed-form relations that convert quantum-chemical and morphological inputs into the photophysical and device-relevant descriptors used in the ML architecture of Section 2.3. Physical constants are taken from CODATA 2018.<sup>69</sup>

## (a) Absorption and exciton formation

The ionization potentials and electron affinities are obtained from adiabatic  $\Delta$ SCF energies (optimized neutral and optimized ion geometries),<sup>37,38</sup>

$$IP_D = E(D^+) - E(D^0), \quad EA_D = E(D^0) - E(D^-), \quad EA_A = E(A^0) - E(A^-), \quad (4)$$

where  $E(X)$  is the total SCF energy of species  $X$  at its own optimized geometry;  $D^0$ ,  $D^+$ ,  $D^-$  are the neutral, cationic and anionic donor;  $A^0$  and  $A^-$  are the neutral and anionic acceptor;  $IP_D$  is the donor ionization potential; and  $EA_D$ ,  $EA_A$  are the donor and acceptor electron affinities. These establish the energetic basis for the donor transport gap,  $IP_D - EA_D$ , and the interfacial gap,  $E_g^{DA} = IP_D - EA_A$ , used throughout the CT-state and  $V_{oc}$  relations below. (The four-point cycle is used only for the reorganization energies  $\lambda_h$ ,  $\lambda_e$ .) A consistency guard checks the physical requirement  $IP_D - EA_D \geq E_{S_1} + E_b^{\text{floor}}$  on the donor, where  $E_{S_1}$  is the TD-DFT first singlet excitation energy and  $E_b^{\text{floor}} = 0.10$  eV is the minimum admissible binding energy.

The AM1.5G-limited optical short-circuit current is the photon flux integrated above the optical gap,<sup>22</sup>

$$\Phi(E_g) = \int_{\lambda: hc/\lambda \geq E_g} \frac{I_{\text{AM1.5G}}(\lambda)}{hc/\lambda} d\lambda, \quad J_{sc}^{\text{opt}} = 0.1 q \Phi(E_g) \quad [\text{mA cm}^{-2}], \quad (5)$$

where  $E_g$  is the optical gap (eV);  $\lambda$  is the wavelength;  $I_{\text{AM1.5G}}(\lambda)$  is the AM1.5G spectral irradiance ( $\text{W m}^{-2} \text{nm}^{-1}$ );  $hc/\lambda$  is the photon energy, with  $h$  the Planck constant and  $c$  the speed of light;  $\Phi(E_g)$  is the photon flux with energy at or above  $E_g$  ( $\text{photons m}^{-2} \text{s}^{-1}$ );  $q$  is the elementary charge; and  $J_{sc}^{\text{opt}}$  is the resulting photon-limited short-circuit current density. The factor 0.1 converts  $\text{A m}^{-2}$  to  $\text{mA cm}^{-2}$ . The same integral is tabulated once on a 0.005 eV grid between 0.5 and 4.0 eV and interpolated inside the PCNN, so the training-time ceiling and the post-hoc clip use one function and cannot disagree.

The radiative decay rate and lifetime of the singlet exciton follow the Strickler–Berg relation,<sup>70</sup>

$$\omega = \frac{E_{S_1}}{\hbar}, \quad k_r = \frac{e^2 \omega^2 n f}{2\pi \epsilon_0 m_e c^3}, \quad \tau_r = \frac{1}{k_r}, \quad (6)$$

where  $\omega$  is the angular frequency of the  $S_0 \rightarrow S_1$  transition;  $\hbar$  is the reduced Planck constant;  $k_r$  is the radiative decay rate ( $\text{s}^{-1}$ );  $e$  is the elementary charge;  $n$  is the refractive index of the medium;  $f$  is the TD-DFT oscillator strength of the  $S_0 \rightarrow S_1$  transition;  $\epsilon_0$  is the vacuum permittivity;  $m_e$  is the electron mass;  $c$  is the speed of light; and  $\tau_r$  is the radiative lifetime. The TD-DFT  $S_1$  energy and oscillator strength are used directly.

An effective, dielectric-screened exciton binding energy is obtained by applying an Onsager-sphere polarization correction to the *donor* transport (IP–EA) gap,

$$E_{\text{pol}} = 2S_{\text{scale}} \frac{e^2}{8\pi \epsilon_0 r_{\text{sphere}}} \left(1 - \frac{1}{\epsilon_r}\right), \quad E_b^{\text{eff}} = (IP_D - EA_D - E_{\text{pol}}) - E_{S_1} \quad (7)$$

where  $E_{\text{pol}}$  is the polarization (screening) energy of an electron–hole pair, i.e. twice the Born solvation energy of one charge in a dielectric sphere;  $S_{\text{scale}}$  is a dimensionless scaling factor;  $r_{\text{sphere}}$  is the radius of the dielectric sphere;  $\varepsilon_r$  is the relative static permittivity; and  $E_b^{\text{eff}}$  is the effective exciton binding energy.  $E_{\text{pol}}$  is applied once, and only to gas-phase  $\Delta\text{SCF}$  inputs (equivalently, as the Born shifts  $IP \rightarrow IP - E_{\text{pol}}/2$ ,  $EA \rightarrow EA + E_{\text{pol}}/2$ ); for inputs that are already condensed-phase,  $E_{\text{pol}} = 0$ .

For use as a fast ML feature directly from frontier orbitals,

$$E_b = (E_A^{\text{LUMO}} - E_D^{\text{HOMO}} - E_{S_1}), \quad (8)$$

is also computed, where  $E_D^{\text{HOMO}}$  is the donor HOMO energy;  $E_A^{\text{LUMO}}$  is the acceptor LUMO energy (both in eV, so the order LUMO–HOMO gives a positive gap);  $E_{S_1}$  is as above. The singlet–triplet gap,  $\Delta E_{ST} = E_{S_1} - E_{T_1}$  (with  $E_{T_1}$  the first triplet energy), and the Förster radius<sup>71</sup> for exciton (or LE→CT) energy transfer,

$$R_0^6 = \frac{9000(\ln 10) \kappa^2 \Phi_D}{128\pi^5 n^4 N_A} J_{\text{overlap}}, \quad J_{\text{overlap}} \approx \varepsilon_{\text{max}} \lambda_D^4 \frac{\sigma_A}{\sqrt{\sigma_D^2 + \sigma_A^2}} \exp\left[-\frac{(\lambda_D - \lambda_A)^2}{2(\sigma_D^2 + \sigma_A^2)}\right], \quad (9)$$

complete the exciton-formation descriptor set. Here  $R_0$  is the Förster radius (the donor–acceptor separation at which transfer and decay are equally likely);  $\kappa^2$  is the dipole orientation factor (2/3 for random orientation);  $\Phi_D = k_r/(k_r + k_{nr})$  is the donor fluorescence quantum yield, with  $k_{nr}$  the non-radiative decay rate;  $n$  is the refractive index;  $N_A$  is Avogadro’s number; the numerical prefactor 9000ln 10 applies for  $J_{\text{overlap}}$  in  $\text{M}^{-1} \text{cm}^3$  and  $R_0$  in cm;  $J_{\text{overlap}}$  is the donor-emission/acceptor-absorption spectral overlap integral;  $\varepsilon_{\text{max}}$  is the acceptor peak molar extinction coefficient ( $\text{M}^{-1} \text{cm}^{-1}$ );  $\lambda_D$  and  $\lambda_A$  are the donor-emission and acceptor-absorption peak wavelengths; and  $\sigma_D$ ,  $\sigma_A$  are the Gaussian standard deviations of the donor and acceptor line shapes, in the same units as  $\lambda$ .

## (b) Exciton diffusion and dissociation

Following the continuum-electrostatic picture of CT energetics,<sup>72–74</sup> the CT-state energy is evaluated from the donor ionization potential, acceptor electron affinity, and the donor–acceptor contact distance supplied to the pipeline. The implemented Coulomb term is

$$E_{\text{Coulomb}} = \frac{e^2}{4\pi\varepsilon_0\varepsilon_r R_{DA}} = \frac{14.4}{\varepsilon_r R_{DA}[\text{\AA}]} [\text{eV}], \quad (10)$$

where  $E_{\text{Coulomb}}$  is the electron–hole Coulomb binding energy at the CT geometry;  $R_{DA}$  is the donor–acceptor contact distance;  $\varepsilon_r$  is the relative permittivity as above; and 14.4 eV Å is the value of  $e^2/(4\pi\varepsilon_0)$ .

The CT-state energy before LE–CT mixing is

$$E_{CT}^{\text{bare}} = IP_D - EA_A - E_{\text{Coulomb}} - E_{\text{pol}} + E_{\text{image}}, \quad E_{\text{image}} = -\frac{e^2}{16\pi\varepsilon_0\varepsilon_r z_{\text{eff}}} \leq 0. \quad (11)$$

Here  $E_{CT}^{\text{bare}}$  is the unmixed (diabatic) CT energy;  $IP_D - EA_A$  is the interfacial gap;  $E_{\text{pol}}$  is the CT polarization (reaction-field) stabilization;  $E_{\text{image}}$  is the electrode image-charge stabilization (non-positive); and  $z_{\text{eff}} = \min(R_{DA}, d_{\text{electrode}})$ , where  $d_{\text{electrode}}$  is the configured electrode distance.

The LE and CT energies entering the charge-separation relations are the *diabatic* energies: on the TD-DFT/NTD route  $E_{LE}$  and  $E_{CT}$  are the dimer diabats of Eq. 14, and

on the electrostatic fallback  $E_{LE} = E_{S_1}$  and  $E_{CT} = E_{CT}^{\text{bare}}$ . The driving force for interfacial charge separation<sup>75</sup> and the CT $\rightarrow$ S<sub>0</sub> recombination driving force follow directly,

$$\Delta G_{CS} = E_{CT} - E_{LE}, \quad \Delta G_{CR} = -E_{CT}, \quad (12)$$

where  $\Delta G_{CS}$  is the free-energy change of LE $\rightarrow$ CT charge separation;  $\Delta G_{CR}$  is that of CT $\rightarrow$ S<sub>0</sub> charge recombination; and  $E_{LE}$ ,  $E_{CT}$  are the diabatic LE and CT energies defined above. Negative values mean downhill (exergonic) steps. The energetic offset between the LE and CT states, normalized by the thermal energy at 300 K, quantifies how activated the LE $\rightarrow$ CT step is:

$$\Delta E_{CS} = E_{LE} - E_{CT} = -\Delta G_{CS}, \quad \left. \frac{\Delta E_{CS}}{k_B T} \right|_{300 \text{ K}}, \quad (13)$$

where  $\Delta E_{CS}$  is the LE–CT energy offset (positive when CT lies below LE);  $k_B$  is the Boltzmann constant;  $T = 300$  K is the temperature; and  $k_B T = 0.0259$  eV.

The LE–CT coupling  $V_{\text{LE-CT}}$  is treated as an explicit electronic-structure input to the CT-state calculation. For a dimer TD-DFT/NTO route the coupling estimate is

$$V_{\text{LE-CT}} = \Delta E_{12} \sqrt{p(1-p)}, \quad E_{LE} = (1-p)E_1 + pE_2, \quad E_{CT} = pE_1 + (1-p)E_2, \quad (14)$$

where  $V_{\text{LE-CT}}$  is the electronic coupling between the diabatic LE and CT states;  $E_1 < E_2$  are the adiabatic energies of the lower and upper selected states;  $\Delta E_{12} = E_2 - E_1$  is their splitting;  $p$  is the CT fraction of the lower state (from the NTO analysis); and  $E_{LE}$ ,  $E_{CT}$  are the resulting diabatic energies. The routine requires the two selected states to have approximately complementary CT character,  $|p_2 - (1 - p_1)| \leq 0.15$ , where  $p_1$ ,  $p_2$  are the CT fractions of the lower and upper state. This is a two-state TD-DFT/NTO mixing estimate (in the spirit of the generalized Mulliken–Hush scheme<sup>76</sup>), not a fragment-orbital DFT calculation and not a rigorous diabaticization.

Eq. 14 is the inverse of the  $2 \times 2$  LE–CT mixing relation,

$$E_{\pm} = \frac{E_{LE} + E_{CT}}{2} \pm \frac{1}{2} \sqrt{(E_{LE} - E_{CT})^2 + 4V_{\text{LE-CT}}^2}, \quad \chi_{CT} = \frac{1}{2} \left[ 1 + \frac{E_{LE} - E_{CT}}{\sqrt{(E_{LE} - E_{CT})^2 + 4V_{\text{LE-CT}}^2}} \right], \quad (15)$$

where  $E_-$  and  $E_+$  are the lower and upper adiabatic (mixed) energies, and  $\chi_{CT}$  is the CT character (CT population) of the lower adiabatic state, between 0 (pure LE) and 1 (pure CT). The same coupling enters the semiclassical Marcus<sup>14,64,77,78</sup> and Marcus–Levich–Jortner (MLJ)<sup>79</sup> charge-separation rates, with the total reorganization energy  $\lambda = \lambda_{\text{total}} + \lambda_{\text{ext}}$ , where  $\lambda_{\text{total}} = \lambda_h + \lambda_e$  is the four-point inner-sphere sum (hole plus electron) and  $\lambda_{\text{ext}}$  is the Pekar outer-sphere term:

$$k_{\text{LE} \rightarrow \text{CT}}^{\text{Marcus}} = \frac{2\pi}{\hbar} \frac{V_{\text{LE-CT}}^2}{\sqrt{4\pi\lambda k_B T}} \exp \left[ -\frac{(\Delta G_{CS} + \lambda)^2}{4\lambda k_B T} \right], \quad (16)$$

where  $k_{\text{LE} \rightarrow \text{CT}}^{\text{Marcus}}$  is the LE $\rightarrow$ CT charge-separation rate ( $\text{s}^{-1}$ );  $\hbar$  is the reduced Planck constant;  $V_{\text{LE-CT}}$  is the coupling;  $\lambda$  is the total reorganization energy;  $k_B T$  is the thermal energy; and  $\Delta G_{CS}$  is the driving force of Eq. 12.

$$k_{\text{MLJ}} = \frac{2\pi}{\hbar} V_{\text{LE-CT}}^2 \sum_{n=0}^{\infty} \frac{e^{-S} S^n}{n!} \frac{1}{\sqrt{4\pi\lambda_{\text{cl}} k_B T}} \exp \left[ -\frac{(\Delta G_{CS} + \lambda_{\text{cl}} + n\hbar\omega)^2}{4\lambda_{\text{cl}} k_B T} \right], \quad \lambda_q = S\hbar\omega, \quad \lambda_{\text{cl}} = \lambda - \lambda_q. \quad (17)$$

Here  $k_{\text{MLJ}}$  is the MLJ charge-separation rate ( $\text{s}^{-1}$ );  $S$  is the Huang–Rhys factor of the effective high-frequency quantum mode;  $\hbar\omega$  is the energy of that mode (here  $\omega$  is the vibrational angular frequency, not the transition frequency of Eq. 6);  $n$  is the number of vibrational quanta excited in the product state (the Poisson-weighted sum is truncated numerically);  $\lambda_q = S\hbar\omega$  is the quantum (high-frequency) part of the reorganization energy; and  $\lambda_{\text{cl}}$  is the remaining classical part, treated as in Marcus theory. The  $\text{CT} \rightarrow S_0$  recombination rate  $k_{\text{rec}}$  is obtained from the same expressions with  $\Delta G_{CS} \rightarrow \Delta G_{CR}$  and  $V_{\text{LE-CT}} \rightarrow V_{\text{CT-S}_0}$ , the CT-ground-state coupling.

Because Eqs. 16 and 17 assume nonadiabatic electron transfer, the pipeline computes the ratio  $2|V_{\text{LE-CT}}|/\lambda_{\text{total}}$  and the inverted-regime ratio  $(-\Delta G_{CS})/\lambda_{\text{total}}$  (both with the inner-sphere  $\lambda_{\text{total}}$ , i.e. excluding  $\lambda_{\text{ext}}$ ) and *demotes the resulting charge-separation efficiency to diagnostic-only status* when either guard trips (thresholds 0.5 and 1.0, respectively); no physical  $\eta_{CS}$  and no charge-separation-based  $J_{sc}$  ceiling is then reported. Here  $\eta_{CS}$  is the charge-separation efficiency from the Marcus/MLJ kinetics.

Finally, the probability that a bound CT (geminate) pair separates into free carriers, rather than recombining, is estimated with Onsager–Braun field-dependent dissociation theory,<sup>80,81</sup> with  $a = R_{DA}$ :

$$E_b^{CT} = \frac{q^2}{4\pi\epsilon_0\epsilon_r a}, \quad \gamma = \frac{q(\mu_e + \mu_h)}{\epsilon_0\epsilon_r}, \quad b = \frac{q^3 F}{8\pi\epsilon_0\epsilon_r (k_B T)^2}, \quad \phi(b) = \sum_{n=0}^{\infty} \frac{(2b)^n}{n!(n+1)!}, \quad (18)$$

where  $E_b^{CT}$  is the Coulomb binding energy of the CT pair at separation  $a$ ;  $q$  is the elementary charge;  $\gamma$  is the Langevin-type recombination coefficient ( $\text{m}^3 \text{s}^{-1}$ );  $\mu_e$ ,  $\mu_h$  are the electron and hole mobilities;  $b$  is the dimensionless field parameter;  $F$  is the applied electric field; and  $\phi(b)$  is the Onsager field-enhancement factor, with summation index  $n$ .

$$k_{\text{diss}} = \frac{3\gamma}{4\pi a^3} \exp\left(-\frac{E_b^{CT}}{k_B T}\right) \phi(b), \quad \eta_{\text{diss}} = \frac{k_{\text{diss}}}{k_{\text{diss}} + k_{\text{rec}}}, \quad (19)$$

where  $k_{\text{diss}}$  is the CT-pair dissociation rate ( $\text{s}^{-1}$ );  $k_{\text{rec}}$  ( $\text{s}^{-1}$ ) is the CT recombination rate, taken from an explicitly supplied CT recombination rate or, failing that, from the  $\text{CT} \rightarrow S_0$  MLJ/Marcus rate above; and  $\eta_{\text{diss}}$  is the probability that a CT pair dissociates rather than recombines. The mobilities  $\mu_{e,h}$  are the Marcus-derived electron and hole mobilities.

### (c) Device-level constraints and rank mapping

The PCNN fill-factor ceiling is Green’s ideal fill factor,<sup>57</sup>

$$\text{FF}_0^{\text{Green}}(V_{oc}) = \frac{v_{oc} - \ln(v_{oc} + 0.72)}{v_{oc} + 1}, \quad v_{oc} = \frac{qV_{oc}}{n_{\text{id}} k_B T}, \quad (20)$$

where  $n_{\text{id}}$  is the diode ideality factor, and the energetic  $V_{oc}$  ceiling is  $qV_{oc} \leq IP_D - EA_A$ . With  $P_{\text{in}} = 100 \text{ mW cm}^{-2}$ ,  $\text{PCE}(\%) = V_{oc} [\text{V}] J_{sc} [\text{mA cm}^{-2}] \text{FF}$ , which is the form used throughout.

The latent-morphology slots of Section 2.4 are filled by mapping the rank  $r_{id}$  (from 1 to  $N$ ) of the score of training device  $i$  on latent axis  $d$  into that slot’s physical range  $[\ell_d, u_d]$ ,

$$\tilde{x}_{id} = \ell_d + (u_d - \ell_d) \left[ \epsilon + (1 - 2\epsilon) \frac{r_{id} - 0.5}{N} \right], \quad (21)$$

where  $N$  is the number of training devices and  $\epsilon = 0.02$  is a boundary-adjustment factor that keeps  $\tilde{x}_{id}$  within  $[\ell_d + \epsilon(u_d - \ell_d), u_d - \epsilon(u_d - \ell_d)]$ .

## References

- <sup>1</sup> Jiayu Wang, Yuan Xie, Kai Chen, Hongbin Wu, Justin M Hodgkiss, and Xiaowei Zhan. Physical insights into non-fullerene organic photovoltaics. *Nature Reviews Physics*, 6(6):365–381, 2024.
- <sup>2</sup> Yi Lin, Zaifei Ma, and Zheng Tang. Charge carrier recombination and voltage losses in organic solar cells. *Materials Horizons*, 12(19):7714–7748, 2025.
- <sup>3</sup> Cenqi Yan, Stephen Barlow, Zhaohui Wang, He Yan, Alex K.-Y. Jen, Seth R Marder, and Xiaowei Zhan. Non-fullerene acceptors for organic solar cells. *Nature Reviews Materials*, 3(3):18003, 2018.
- <sup>4</sup> Johannes Benduhn, Kristofer Tvingstedt, Fortunato Piersimoni, Sascha Ullbrich, Yeli Fan, Manuel Tropiano, Kathryn A McGarry, Olaf Zeika, Moritz K Riede, Christopher J Douglas, et al. Intrinsic non-radiative voltage losses in fullerene-based organic solar cells. *Nature Energy*, 2(6):17053, 2017.
- <sup>5</sup> Deping Qian, Zilong Zheng, Huifeng Yao, Wolfgang Tress, Thomas R Hopper, Shula Chen, Sunsun Li, Jing Liu, Shangshang Chen, Jiangbin Zhang, et al. Design rules for minimizing voltage losses in high-efficiency organic solar cells. *Nature Materials*, 17(8):703–709, 2018.
- <sup>6</sup> Xingchi Xiao, Malika Chalh, Zhi Rong Loh, Esther Mbina, Tao Xu, Roger C Hiorns, Yujia Li, Maloy Das, Kekeli N’konou, and Leonard WT Ng. Strategies to achieve efficiencies of over 19% for organic solar cells. *Cell Reports Physical Science*, 6(1), 2025.
- <sup>7</sup> Jianqiu Wang, Yafei Wang, Pengqing Bi, Zhihao Chen, Jiawei Qiao, Jiayao Li, Wenxuan Wang, Zhong Zheng, Shaoqing Zhang, Xiaotao Hao, et al. Binary organic solar cells with 19.2% efficiency enabled by solid additive. *Advanced Materials*, 35(25):2301583, 2023.
- <sup>8</sup> Huotian Zhang, Jun Yuan, Tong Wang, Yijie Nai, Nurlan Tokmoldin, Wei Liu, Shanchao Ouyang, Rokas Jasiūnas, Yiting Liu, Yuxuan Li, et al. Overcoming the fill-factor limit of organic solar cells. *Nature Photonics*, pages 1–8, 2026.
- <sup>9</sup> Hosni Meddeb, Maximilian Götz-Köhler, Nils Neugebohrn, Udayan Banik, Norbert Osterthun, Oleg Sergeev, Dennis Berends, Colleen Lattyak, Kai Gehrke, and Martin Vehse. Tunable photovoltaics: adapting solar cell technologies to versatile applications. *Advanced Energy Materials*, 12(28):2200713, 2022.
- <sup>10</sup> Jean-Luc Brédas, David Beljonne, Veaceslav Coropceanu, and Jérôme Cornil. Charge-transfer and energy-transfer processes in  $\pi$ -conjugated oligomers and polymers: a molecular picture. *Chemical reviews*, 104(11):4971–5004, 2004.
- <sup>11</sup> Veaceslav Coropceanu, Jérôme Cornil, Demetrio A da Silva Filho, Yoann Olivier, Robert Silbey, and Jean-Luc Brédas. Charge transport in organic semiconductors. *Chemical reviews*, 107(4):926–952, 2007.
- <sup>12</sup> SY Reddy and Vikram K Kuppaa. Molecular dynamics simulations of organic photovoltaic materials: Structure and dynamics of oligothiophene. *The Journal of Physical Chemistry C*, 116(28):14873–14882, 2012.

- <sup>13</sup> Leandro R Franco, Danilo Valverde, Cleber Marchiori, Ellen Moons, Ergang Wang, and C Moyses Araujo. Multiscale modeling of structural disorder and environmental effects on the ground and excited states properties of a conjugated donor–acceptor polymer in the bulk phase. *Journal of Physics: Energy*, 7(4):045001, 2025.
- <sup>14</sup> Rudolph A Marcus. Electron transfer reactions in chemistry. theory and experiment. *Reviews of Modern Physics*, 65(3):599, 1993.
- <sup>15</sup> Björn Baumeier, James Kirkpatrick, and Denis Andrienko. Density-functional based determination of intermolecular charge transfer properties for large-scale morphologies. *Physical Chemistry Chemical Physics*, 12(36):11103–11113, 2010.
- <sup>16</sup> Hiroki Sato, Syun Kanda, and Hironori Kaji. Elucidation of molecular-level charge transport in an organic amorphous system. *npj Computational Materials*, 11(1):39, 2025.
- <sup>17</sup> Wenbo Sun, Yujie Zheng, Ke Yang, Qi Zhang, Akeel A Shah, Zhou Wu, Yuyang Sun, Liang Feng, Dongyang Chen, Zeyun Xiao, et al. Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science Advances*, 5(11):eaay4275, 2019.
- <sup>18</sup> Daniele Padula, Jack D Simpson, and Alessandro Troisi. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Materials Horizons*, 6(2):343–349, 2019.
- <sup>19</sup> Darya Rasul Ahmed and Fahmi F Muhammadsharif. A review of machine learning in organic solar cells. *Processes*, 13(2):393, 2025.
- <sup>20</sup> Mirza Sanita Haque and Simon Y Foo. Performance evaluation of machine learning algorithms for predicting organic photovoltaic efficiency. *Clean Energy and Sustainability*, 3(4):10016, 2025.
- <sup>21</sup> Markus C Scharber, David Mühlbacher, Markus Koppe, Patrick Denk, Christoph Waldauf, Alan J Heeger, and Christoph J Brabec. Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency. *Advanced Materials*, 18(6):789–794, 2006.
- <sup>22</sup> William Shockley and Hans J Queisser. Detailed balance limit of efficiency of p-n junction solar cells. *Journal of applied physics*, 32(3):510–519, 1961.
- <sup>23</sup> Uwe Rau. Reciprocity relation between photovoltaic quantum efficiency and electroluminescent emission of solar cells. *Physical Review B—Condensed Matter and Materials Physics*, 76(8):085303, 2007.
- <sup>24</sup> Rudranarayan Khatua, Bibhas Das, and Anirban Mondal. Physics-informed machine learning with data-driven equations for predicting organic solar cell performance. *ACS Applied Materials & Interfaces*, 16(42):57467–57480, 2024.
- <sup>25</sup> Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

- <sup>26</sup> George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.
- <sup>27</sup> Mirza Sanita Haque, Monira Khanom Mim, and Simon Y Foo. A physics-guided graph neural network framework for predicting organic solar cell performance parameters. *Algorithms*, 19(6):431, 2026.
- <sup>28</sup> Long-Fei Lv, Cai-Rong Zhang, Cui-Cui Sang, Xiao-Meng Liu, Mei-Ling Zhang, Ji-Jun Gong, Yu-Hong Chen, and Hong-Shan Chen. Integrating deep learning and symbolic regression for molecular design and virtual screening of organic solar cells. *npj Computational Materials*, 12(1):31, 2025.
- <sup>29</sup> Xiumin Liu, Xinyue Zhang, Ye Sheng, Zihe Zhang, Pan Xiong, Xuehai Ju, Junwu Zhu, and Caichao Ye. Advancing organic photovoltaic materials by machine learning-driven design with polymer-unit fingerprints. *npj Computational Materials*, 11(1):107, 2025.
- <sup>30</sup> Jinhui Zhao, Yi Wu, Witold M Bloch, Caroline VI Andersson, Leandro R Franco, Rafael B Ribeiro, Chuangcheng Hong, Wei Zhang, Xun Pan, Joost Kimpel, et al. Side chain elimination enables low-cost fused-ring acceptors and reveals a compact tetrameric structure for high-photocurrent organic solar cells. *Advanced Science*, page e75738, 2026.
- <sup>31</sup> Jun Yuan, Yunqiang Zhang, Liuyang Zhou, Guichuan Zhang, Hin-Lap Yip, Tsz-Ki Lau, Xinhui Lu, Can Zhu, Hongjian Peng, Paul A Johnson, et al. Single-junction organic solar cell with over 15% efficiency using fused-ring acceptor with electron-deficient core. *Joule*, 3(4):1140–1151, 2019.
- <sup>32</sup> Yuandong Sun, Liang Wang, Chuanhang Guo, Jinyi Xiao, Chenhao Liu, Chen Chen, Weiyi Xia, Zirui Gan, Jingchao Cheng, Jinpeng Zhou, et al.  $\pi$ -extended nonfullerene acceptor for compressed molecular packing in organic solar cells to achieve over 20% efficiency. *Journal of the American Chemical Society*, 146(17):12011–12019, 2024.
- <sup>33</sup> Minh Trung Dang, Lionel Hirsch, and Guillaume Wantz. P3ht:pcbm, best seller in polymer photovoltaic research. *Advanced Materials*, 23(31):3597–3602, 2011.
- <sup>34</sup> M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, et al. *Gaussian 16, Revision C.01*. Gaussian, Inc., Wallingford, CT, 2016.
- <sup>35</sup> Praveen C Hariharan and John A Pople. The influence of polarization functions on molecular orbital hydrogenation energies. *Theoretica chimica acta*, 28(3):213–222, 1973.
- <sup>36</sup> Thom H Dunning Jr. Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen. *The Journal of Chemical Physics*, 90(2):1007–1023, 1989.
- <sup>37</sup> Stephen F Nelsen, Silas C Blackstock, and Yaesil Kim. Estimation of inner shell marcus terms for amino nitrogen compounds by molecular orbital calculations. *Journal of the American Chemical Society*, 109(3):677–682, 1987.

- <sup>38</sup> Omar López-Estrada, Humberto G Laguna, Cihuapilli Barraeta-Flores, and Carlos Amador-Bedolla. Reassessment of the four-point approach to the electron-transfer marcus–hush theory. *ACS omega*, 3(2):2130–2140, 2018.
- <sup>39</sup> Takeshi Yanai, David P Tew, and Nicholas C Handy. A new hybrid exchange–correlation functional using the coulomb-attenuating method (cam-b3lyp). *Chemical Physics Letters*, 393(1-3):51–57, 2004.
- <sup>40</sup> Richard L Martin. Natural transition orbitals. *Journal of Chemical Physics*, 118(11):4775–4777, 2003.
- <sup>41</sup> Tian Lu and Feiwu Chen. Multiwfn: A multifunctional wavefunction analyzer. *Journal of computational chemistry*, 33(5):580–592, 2012.
- <sup>42</sup> Tian Lu. A comprehensive electron wavefunction analysis toolbox for chemists, multiwfn. *The Journal of chemical physics*, 161(8), 2024.
- <sup>43</sup> Aidan P Thompson, H Metin Aktulga, Richard Berger, Dan S Bolintineanu, W Michael Brown, Paul S Crozier, Pieter J In’t Veld, Axel Kohlmeyer, Stan G Moore, Trung Dac Nguyen, et al. Lammmps-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer physics communications*, 271:108171, 2022.
- <sup>44</sup> Andrew I Jewett, David Stelter, Jason Lambert, Shyam M Saladi, Otello M Roscioni, Matteo Ricci, Ludovic Autin, Martina Maritan, Saeed M Bashusqeh, Tom Keyes, et al. Moltemplate: A tool for coarse-grained modeling of complex biological matter and soft condensed matter physics. *Journal of molecular biology*, 433(11):166841, 2021.
- <sup>45</sup> William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.
- <sup>46</sup> William L Jorgensen and Julian Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences*, 102(19):6665–6670, 2005.
- <sup>47</sup> Leela S Dodda, Jonah Z Vilseck, Julian Tirado-Rives, and William L Jorgensen. 1.14\* cm1a-lbcc: Localized bond charge corrected cm1a charges for condensed-phase simulations. *The journal of physical chemistry. B*, 121(15):3864, 2017.
- <sup>48</sup> Leela S Dodda, Israel Cabeza de Vaca, Julian Tirado-Rives, and William L Jorgensen. Ligpargen web server: an automatic opls-aa parameter generator for organic ligands. *Nucleic acids research*, 45(W1):W331–W336, 2017.
- <sup>49</sup> Roger W Hockney and James W Eastwood. *Computer Simulation Using Particles*. CRC Press, 1988.
- <sup>50</sup> Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81(1):511–519, 1984.
- <sup>51</sup> William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31(3):1695, 1985.

- <sup>52</sup> Jiangjie Qiu. Opv-db: a literature-mined organic photovoltaic device performance database, jun 2026.
- <sup>53</sup> Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- <sup>54</sup> Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.
- <sup>55</sup> Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.
- <sup>56</sup> Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- <sup>57</sup> Martin A Green. Solar cell fill factors: General graph and empirical expressions. *Solid-State Electronics*, 24(8):788–789, 1981.
- <sup>58</sup> Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.
- <sup>59</sup> Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *PMLR*, pages 1050–1059, 2016.
- <sup>60</sup> Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.
- <sup>61</sup> Qiuju Liang, Mingzhi Duan, Ziqi Geng, Ming Zhang, Wenzheng Xu, Hongjie Geng, Zemin He, and Jiangang Liu. Regulation of molecular orientation in organic solar cells. *Chemical Engineering Journal*, 488:150783, 2024.
- <sup>62</sup> Aditya N Panda, Felix Plasser, Adelia JA Aquino, Irene Burghardt, and Hans Lischka. Electronically excited states in poly (p-phenylenevinylene): Vertical excitations and torsional potentials from high-level ab initio calculations. *The Journal of Physical Chemistry A*, 117(10):2181–2189, 2013.
- <sup>63</sup> Tian Lu. Simple, reliable, and universal metrics of molecular planarity. *Journal of molecular modeling*, 27(9):263, 2021.
- <sup>64</sup> Rudolph A Marcus. On the theory of oxidation-reduction reactions involving electron transfer. i. *The Journal of chemical physics*, 24(5):966–978, 1956.
- <sup>65</sup> Hiroyuki Ishii. Charge transport simulations for organic semiconductors. *Molecular Technology: Energy Innovation*, 1:1–23, 2019.
- <sup>66</sup> Allen Miller and Elihu Abrahams. Impurity conduction at low concentrations. *Physical Review*, 120(3):745, 1960.
- <sup>67</sup> Heinz Bässler. Charge transport in disordered organic photoconductors. a monte carlo simulation study. *Physica Status Solidi B (Basic Research);(Germany)*, 175(1), 1993.
- <sup>68</sup> WF Pasveer, Jeroen Cottaar, C Tanase, Reinder Coehoorn, PA Bobbert, PWM Blom, DM De Leeuw, and MAJ Michels. Unified description of charge-carrier mobilities in disordered semiconducting polymers. *Physical review letters*, 94(20):206601, 2005.

- <sup>69</sup> Eite Tiesinga, Peter J Mohr, David B Newell, and Barry N Taylor. Codata recommended values of the fundamental physical constants: 2018. *Journal of physical and chemical reference data*, 50(3), 2021.
- <sup>70</sup> SJ Strickler and Robert A Berg. Relationship between absorption intensity and fluorescence lifetime of molecules. *The Journal of chemical physics*, 37(4):814–822, 1962.
- <sup>71</sup> Theodor Förster. Zwischenmolekulare energiewanderung und fluoreszenz. *Annalen der Physik*, 437(1-2):55–75, 1948.
- <sup>72</sup> Stijn Verlaak, David Beljonne, David Cheyns, Cedric Rolin, Mathieu Linares, Frédéric Castet, Jérôme Cornil, and Paul Heremans. Electronic structure and geminate pair energetics at organic–organic interfaces: the case of pentacene/c60 heterojunctions. *Advanced Functional Materials*, 19(23):3809–3814, 2009.
- <sup>73</sup> Dirk Veldman, Stefan CJ Meskers, and René AJ Janssen. The energy of charge-transfer states in electron donor–acceptor blends: insight into the energy losses in organic solar cells. *Advanced Functional Materials*, 19(12):1939–1948, 2009.
- <sup>74</sup> Zilong Zheng, Naga Rajesh Tummala, Yao-Tsung Fu, Veaceslav Coropceanu, and Jean-Luc Brédas. Charge-transfer states in organic solar cells: Understanding the impact of polarization, delocalization, and disorder. *ACS applied materials & interfaces*, 9(21):18095–18102, 2017.
- <sup>75</sup> Xiao Liu, Barry P Rand, and Stephen R Forrest. Engineering charge-transfer states for efficient, low-energy-loss organic photovoltaics. *Trends in Chemistry*, 1(9):815–829, 2019.
- <sup>76</sup> Robert J Cave and Marshall D Newton. Generalization of the mulliken-hush treatment for the calculation of electron transfer matrix elements. *Chemical Physics Letters*, 249(1-2):15–19, 1996.
- <sup>77</sup> Joseph E Subotnik, Josh Vura-Weis, Alex J Sodt, and Mark A Ratner. Predicting accurate electronic excitation transfer rates via marcus theory with boys or edmiston-ruedenberg localized diabaticization. *The Journal of Physical Chemistry A*, 114(33):8665–8675, 2010.
- <sup>78</sup> Svante Hedström, Benjamin Rudsteyn, Subhajyoti Chaudhuri, and Victor S Batista. Marcus theory with gaussian and adf—a tutorial. Department of Chemistry, Yale University, New Haven, CT, USA, 2016.
- <sup>79</sup> Joshua Jortner. Temperature dependent activation energy for electron transfer between biological molecules. *The Journal of Chemical Physics*, 64(12):4860–4867, 1976.
- <sup>80</sup> Lars Onsager. Initial recombination of ions. *Physical Review*, 54(8):554, 1938.
- <sup>81</sup> Charles L Braun. Electric field assisted dissociation of charge transfer states as a mechanism of photocarrier production. *The Journal of chemical physics*, 80(9):4157–4161, 1984.