Charge Transfer Landscape Manifesting Structure-Rate Relationship in the Condensed Phase via Machine Learning

Dominikus Brian[†,‡,¶] and Xiang Sun[*,†,‡,¶,§]

†*Division of Arts and Sciences, NYU Shanghai, 1555 Century Avenue, Shanghai 200122, China*

‡*NYU-ECNU Center for Computational Chemistry at NYU Shanghai, 3663 Zhongshan Road North, Shanghai 200062, China*

¶*Department of Chemistry, New York University, New York, New York 10003, United States*

§*State Key Laboratory of Precision Spectroscopy, East China Normal University, Shanghai 200241, China*

E-mail: xiang.sun@nyu.edu

Abstract

In this work, we develop a machine learning (ML) strategy to map molecular structure to condensed-phase charge transfer (CT) properties including CT rate constants, energy levels, electronic couplings, energy gaps, reorganization energies, and reaction free energies, which are called CT fingerprints. The CT fingerprints of selected landmark structures covering the conformation space of an organic photovoltaic molecule dissolved in an explicit solvent are computed and used to train ML models using kernel ridge regression. The ML models show high predictive power with $R^2 > 0.97$, and both mean absolute error and root mean square error within chemical accuracy. The CT landscape for millions of molecular dynamics sampled structures is thus constructed, which allows for instant prediction of CT rate properties given any conformation of the molecule. We demonstrate some immediate utilities of CT landscape such as calculating the ensemble-averaged CT rate constant and interpreting the effects of molecular structural features on the CT rate. The unprecedented CT landscape will be useful for investigating real-time CT dynamics in nanoscale and mesoscale condensed-phase systems, and the optimal fabrication design for homogeneous and heterogeneous optoelectronic devices.

1. Introduction

Organic semiconductor (OSC) attracts a great deal of attention owing to its immediate and increasingly demanded applications in organic photovoltaics (OPV),[1–5] organic light emitting diodes (OLED),[1,2] molecular electronics,[6] biotechnology,[7] and quantum teleportation.[8] For example, recently discovered OPV solar cells with non-fullerene acceptors have reached a new record of the power conversion efficiency 19.6%.[9] For a given set of functional molecules such as the donor (D) and acceptor (A) in OPV bulk heterojunction solar cells, the nanostructure or morphology has been shown to significantly affect the device charge transport performance.[10,11] The mesoscopic morphology can be tailored through careful material synthesis and advanced fabrication techniques.[12] However, the rational design for attaining optimal device performance is no trivial task, which is typically expensive, time-consuming, and often involves copious

experimental trials and errors. On the microscopic level, the morphology is a direct manifestation of the molecular packing arrangement throughout the bulk of the material. Therefore, establishing a quantitative structure-property relationship (QSPR) would provide insights that guide the exploration of the high-dimensional molecular structural space more efficiently and is effective in accelerating the discovery of high-performance OSC materials.[3,5,10,13]

Unlike the gas-phase D/A molecules or molecular segments, the nanostructure of the D/A interface in the condensed phase could have considerable static and dynamic disorders: the static disorder arises from the time-independent inhomogeneous environments of different D/A pairs, and the dynamic disorder arises from the time-dependent thermally driven conformational fluctuations and is related to the electron-vibration interactions.[14,15] Due to the richness in the molecular conformations in bulk materials, multiscale modeling has been employed to incorporate these disorders in the morphology in studying charge transfer (CT) properties of OSC materials,[16] and we recently showed that different D/A geometries give rise to CT rate constants that can differ by orders of magnitude.[17] The classic Marcus theory is used widely for estimating the CT rate constant in a variety of condensed-phase systems.[18] Traditionally, molecular dynamics (MD) simulations are used to sample molecular structures from a thermal equilibrated bulk material, and then these MD-sampled geometries of D/A molecules will be used for *ab initio* quantum chemistry calculations to obtain energy levels and electronic couplings.[14,15] However, *ab initio* calculation is only feasible for small systems (e.g. up to a few hundred of atoms) and could generate the energy levels and electronic couplings between the excitonic state and CT state for the chosen D/A pair only, and it cannot account for the disordered environment in an atomistic manner. Moreover, the internal reorganization energy computed with the D/A pair does not include the contribution from the environment like the solvent explicitly,[19] whose fluctuations and reorganization are believed to be the most relevant in condensed-phase charge transfer described by the classic Marcus theory.[18] For instance, in the famous aqueous ferrous-ferric electron transfer,[20] there is no internal reorganization energy for the two ions and the reorganization energy comes solely from the surrounding solvent. So it is essential to consider the explicit surrounding molecules when

modeling realistic condensed-phase charge transfer.

Recently, we proposed a systematic way to account for the motion of the environment by performing the all-atom MD simulations for the entire system with explicit environment after *ab initio* calculation of MD-sampled structures.[17,21] The CT rate constant was derived from the linearized semiclassical Fermi's golden rule (LSC FGR), whose Marcus level of approximation is given by[22]

$$k_{D\to A} = \frac{\Gamma^2}{\hbar}\sqrt{\frac{2\pi}{\sigma_U^2}}\exp\left(-\frac{\langle U\rangle^2}{2\sigma_U^2}\right), \tag{1}$$

where Γ is the electronic coupling between the donor (initial) and acceptor (final) electronic states, $\hbar$ is the reduced Planck's constant, $\langle U\rangle$ and σ_U^2 are the ensemble average and corresponding variance of the donor-acceptor energy gap $U(\mathbf{R}) = V_D(\mathbf{R}) - V_A(\mathbf{R})$, respectively. Here, $V_{D/A}(\mathbf{R})$ is the donor or acceptor-state potential energy surface (PES) and $\mathbf{R}$ is the nuclear configuration of the entire system. In fact, the rate constant in Eq. 1 can be expressed using the Marcus parameters as below

$$k_{D\to A}^{\mathrm{M}} = \frac{\Gamma^2}{\hbar}\sqrt{\frac{\pi}{k_B T E_r}}\exp\left(-\frac{(\Delta E + E_r)^2}{4k_B T E_r}\right), \tag{2}$$

where the reorganization energy $E_r = \sigma_U^2/(2k_BT) = -\Delta E - \langle U\rangle$, and ΔE is the donor-to-acceptor reaction free energy which is negative for spontaneous reaction, k_B is the Boltzmann constant, and T is temperature. The activation energy is then $E_a = k_BT\langle U\rangle^2/(2\sigma_U^2)$. It should be noted that $\langle U\rangle < 0$ corresponds to the Marcus normal regime ($-\Delta E < E_r$), whereas $\langle U\rangle > 0$ corresponds to the Marcus inverted regime ($-\Delta E > E_r$).[23] Even if we have implemented the automated calculation of the CT rate constant using Eq. 1 in CTRAMER (Charge-Transfer RAtes from Molecular dynamics, Electronic structure, and Rate theory) package,[24] the computational cost primarily due to the condensed-phase MD simulation is still very high for more than a few D/A geometries.

Machine learning (ML) have recently emerges as a powerful technique for obtaining highly accurate prediction about physical properties with significantly reduced computational cost.[25–27]

In the OPV field alone, many fruitful works have been reported for the prediction of molecular properties such as electronic coupling,[5,16,28–31] reorganization energy,[32,33] energy gap,[34–37] as well as device-level properties such as power conversion efficiency, open-circuit voltage, short-circuit current density, and fill factor.[3,38,39] However, an ML model for directly predicting CT rate in the condensed phase based on bottom-up atomistic description is still missing.

In this work, we aim to construct ML models that maps molecular structure to CT rate properties for condensed-phase systems. To facilitate the discussion, the CT properties for a given conformation of the functional molecule including CT rate constant, electronic coupling, average and variance of the donor-acceptor energy gap, reorganization energy, and reaction free energy are defined as the *CT fingerprint (CTFP)*. So the ML model for a certain CT pathway $D \to A$ is the mapping from OPV molecular conformation, $\mathbf{r}$, to CTFP:

$$\text{Conformation } \mathbf{r} \quad \xrightarrow{\text{ML}} \quad \{k_{D\to A},\, \Gamma,\, \langle U\rangle,\, \sigma_U^2,\, E_r,\, \Delta E\}. \tag{3}$$

It should be noted that $\{\langle U\rangle, \sigma_U^2\}$ and $\{E_r, \Delta E\}$ are equivalent, which means knowing either set would be complete, but we keep $\{E_r, \Delta E\}$ here for straightforward comparison with traditional Marcus parameters. We define a *CT landscape* as a rendering of the CT rate constant $k_{D\to A}$ or other CT properties as a function of molecular conformation, following the similar concept as the energy landscape. Once the ML model is constructed, one can look up the CT properties given a molecular configuration instantly.

We demonstrate our proposed strategy by constructing the CT landscape of a prototypical carotenoid-porphyrin-fullerene molecular triad (CPC_{60}) of the donor-bridge-acceptor type, and it is dissolved in explicit tetrahydrofuran (THF) solvent. The triad has many direct applications such as in artificial light-harvesting,[40] OPV,[41–43] molecular wire,[44] and quantum teleportation.[8] We recently found that different conformations of the triad exhibit significantly different CT rate constants as well as nonequilibrium phenomena in the photoinduced CT process giving rise to time-dependent instantaneous Marcus theory (IMT) CT rate coefficients.[45,46] In the photoinduced CT

processes, two typical pathways are present:[17,19] after the triad is photoexcited to the P-localized excitonic $\pi\pi^*$ state, CP^*C_{60}, there can be a nonradiative electronic transition to the excited P-to-C_{60} CT state, $CP^+C_{60}^-$, which is denoted as CT1: $CPC_{60} \xrightarrow{h\nu} CP^*C_{60}(\pi\pi^*) \rightarrow CP^+C_{60}^-(\mathrm{CT1})$, or transition from the excitonic $\pi\pi^*$ state to the excited C-to-C_{60} charge separated state, $C^+PC_{60}^-$, which is denoted as CT2: $CPC_{60} \xrightarrow{h\nu} CP^*C_{60}(\pi\pi^*) \rightarrow C^+PC_{60}^-(\mathrm{CT2})$. In what follows, we consider the CT rate constants and other CT properties for these two pathways: $\pi\pi^* \rightarrow$ CT1 and $\pi\pi^* \rightarrow$ CT2.

2. Methodology

To construct the CT landscape, we propose a five steps strategy as illustrated in Fig. 1: (1) MD simulations for sampling molecular structures, (2) Extraction of a small set of landmark structures that are representative of the configuration **r**-space, (3) CT rate constant calculation using CTRAMER including the electronic structure and MD simulation to obtain the CTFP database $\{\mathbf{r};\, k_{D\rightarrow A},\, \Gamma,\, \langle U\rangle,\, \sigma_U^2,\, E_r,\, \Delta E\}$ for the landmark structures, (4) ML feature engineering to get molecular structural descriptors, **X**, from the configuration, **r**, and (5) ML model training based on the descriptor features, **X**, and the CTFP reference labels, $\mathbf{y} = \{k_{D\rightarrow A},\, \Gamma,\, \langle U\rangle,\, \sigma_U^2,\, E_r,\, \Delta E\}$, and then use the ML model to construct the CT landscape for millions of structures.

2.1. MD simulations for sampling molecular structures

To obtain a large number of triad conformations, we first performed equilibrium MD simulations in the canonical ensemble (constant NVT). The simulated system is a $100\,\mathring{A} \times 100\,\mathring{A} \times 100\,\mathring{A}$ periodic cubic box containing a single ground-state molecular triad CPC_{60} (207 atoms) and 6741 THF solvent molecules, amounting to a condensed phase system with 87840 atoms, where all molecules are flexible and allowed to move freely throughout the whole simulation. We performed 12 parallel simulations with a duration of 100 ns each using MD time step $\delta t = 2$ fs. A total of 1.2 μs long MD trajectories were sampled every 1 ps, leading to a collection of 1,200,000

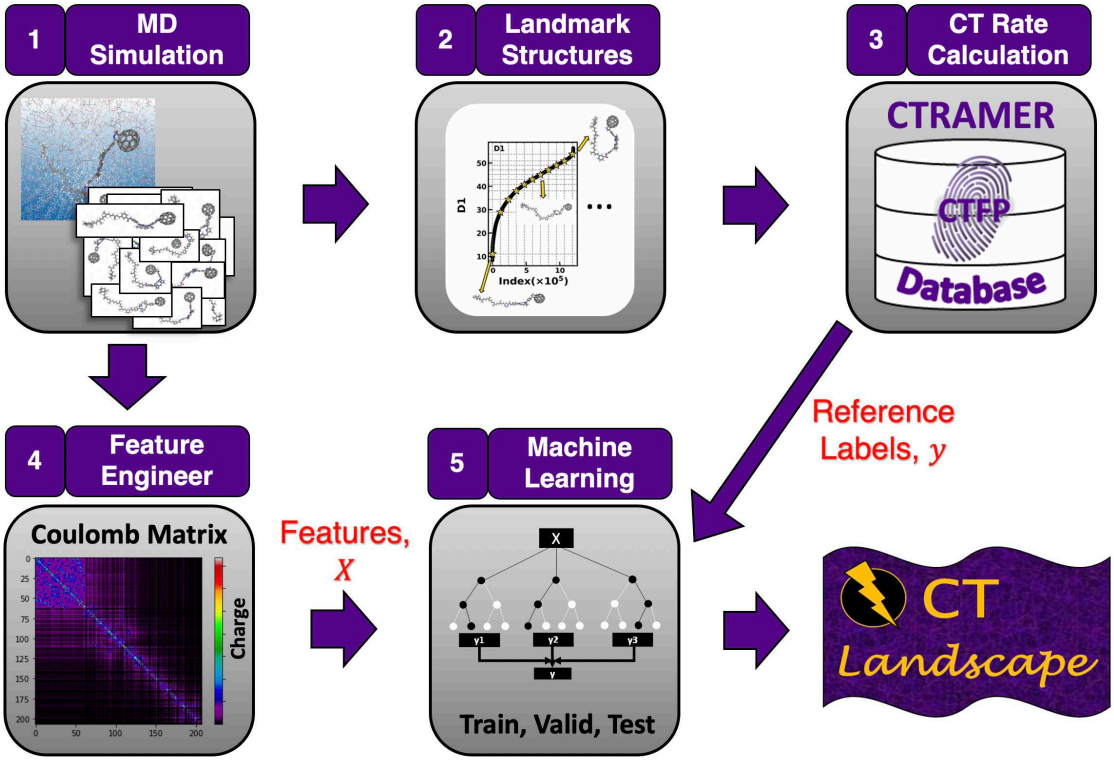


Figure 1: Schematic illustration of the machine learning model construction for charge transfer (CT) landscape of organic photovoltaic molecule in condensed phase. (1) MD simulations for sampling of molecular structures , (2) Extraction of landmark structures, (3) CT rate constant calculation using CTRAMER to obtain CT fingerprint (CTFP) database $\{\mathbf{r}, k_{D\to A}, \Gamma, \langle U\rangle, \sigma_U^2, E_r, \Delta E\}$, (4) ML feature engineering to get descriptors, and (5) Machine learning model training and construct CT landscape.

triad configurations that serve as the conformation database. To ensure that the MD conformation sampling converges, we had performed a convergence analysis for the MD conformation sampling based on the free energy distribution and is detailed in the Supporting Information. The MD simulation was performed using AMBER 2020[47] with GPU acceleration.

2.2. Extraction of landmark structures

Finding a small number of representative structures—landmark structures—to represent the conformation space visited throughout the MD simulation is crucial to reduce the computational cost of the subsequent conformation-specific *ab initio* and MD calculations. Proper selection of landmark structures is important for ML model building since ML works the best when interpolating between the known data points, so the diversity of the selected landmark structures should properly represent that of the original MD-sampled conformation **r**-space. To this end, we select a total of 16 order parameters (OPs), the first 10 of which are geometric descriptors based on chemical knowledge about the triad molecule, and the rest 6 OPs are obtained by performing dimension reduction using 3 different methods and then take the top two leading components. As shown in Fig. 2(a), the first 10 geometric descriptors D1–D10 are various global distances, angles, dihedrals of the constituent fragments of the triad, along with the radius of gyration, solvent accessible surface area, and root mean square deviation (RMSD) from two reference structures of energetic or entropic significance (Conf. I and II in Fig. 2(a)). These descriptors are translational and rotational invariant.

The rest of 6 OPs are from three dimension reduction methods' top two leading components: principal component analysis (PCA)[48] on the raw Cartesian coordinates leading to PC1_XYZ and PC2_XYZ descriptors, PCA on D1–D10 leading to PC1 and PC2, as well as the t-distributed stochastic neighbor embedding (t-SNE)[49] on D1–D10 leading to Z1 and Z2. These 16 OPs are calculated for all structures in the conformation database. Our landmark structure extraction strategy is stochastic sampling around grid points obtained from the equipartition of sequenced order parameter feature spaces.

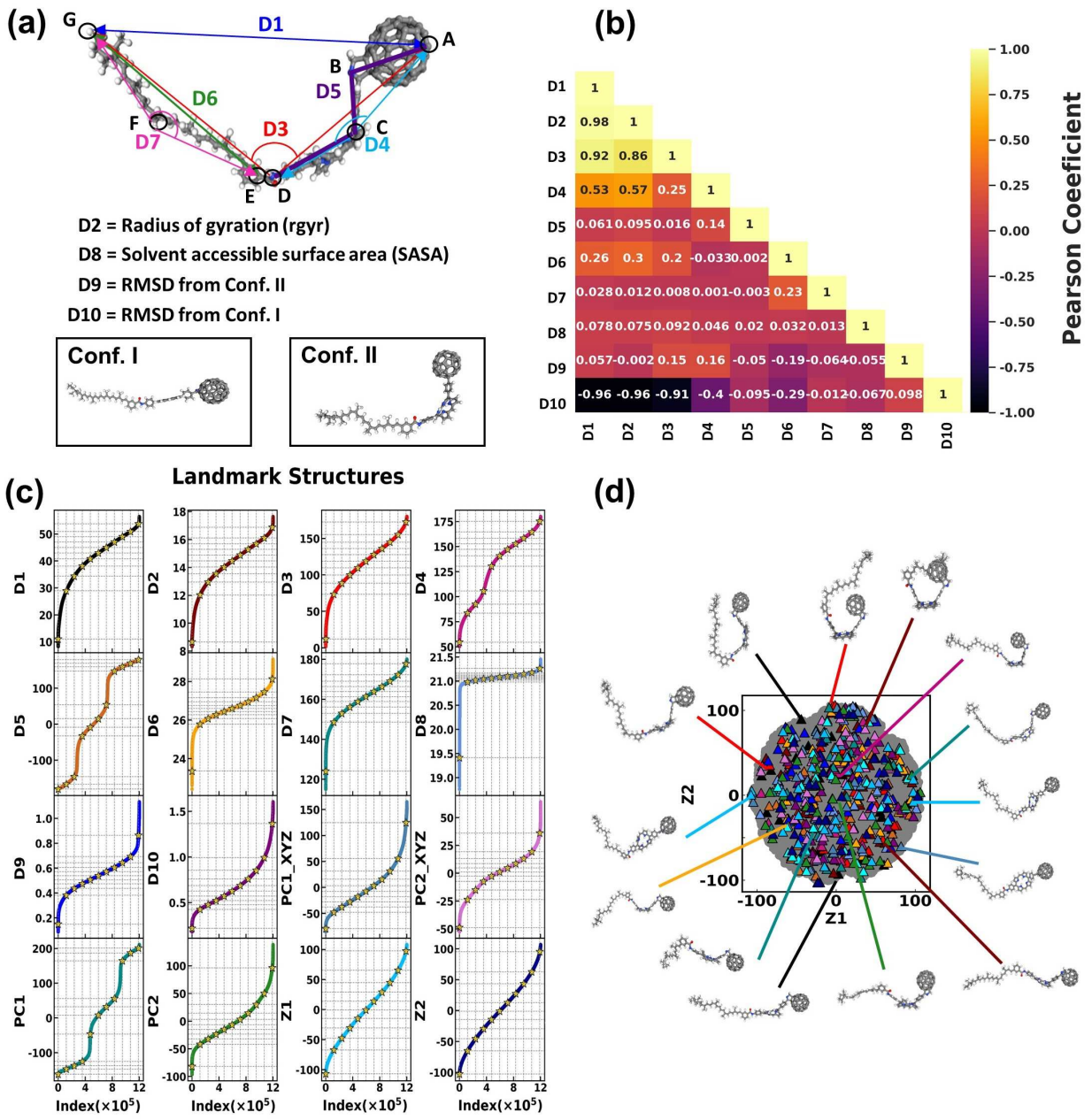


Figure 2: (a) Sketch of geometric descriptors for triad. (b) Pearson correlation between geometric descriptor pairs. (c) Sorted geometric, reduced, and aggregated descriptors for 1.2 million conformations with grid lines and node points used for landmark structure sampling. (d) 528 grid-sampled landmark structures rendered on the first (Z1) and second (Z2) axis of the t-SNE representation, and typical conformations and the reference conformation I and II are highlighted. All the descriptors are in unit of (Å) except for D3, D4, and D7 that are in unit of (°).

2.3. CT rate constant calculation

The CT rate constant calculation implemented in this work was performed using CTRAMER package:[24] (1) For each landmark structure, a gas-phase time-dependent density functional theory (TD-DFT) calculation with the range separated hybrid Baer-Neuhauser-Livshits (BNL)[50] is performed with Q-Chem 4.4[51] and the electronic coupling Γ is obtained with the fragment charge difference (FCD) scheme.[52] (2) Quantitatively analyze the charge transfer character of the 25 lowest excited states and assign the ground, $\pi\pi^*$, CT1, and CT2 states based on pre-designed automated heuristics. The algorithm developed for the automated analysis is described in algorithm 1 given in the Supporting Information. (3) Perform all-atom MD simulation for the triad on the $\pi\pi^*$ state dissolved in 2700 THF solvent molecules, where the partial charges of the triad atoms for different electronic states are obtained from the TD-DFT calculation and all other interaction parameters are the same across different electronic states. For each landmark structure, a total of 100,000 MD snapshots are harvested every 5 fs. The detailed MD procedures and parameters, as well as the trajectory convergence analysis, were further elaborated in the Supporting Information. (4) From the MD trajectories, the average and variance of the donor-acceptor energy gap U, i.e., $\{\langle U\rangle, \sigma_U^2\}$, are obtained, which lead to the Marcus-level CT rate constant via Eq. 1. Thus, we obtain the CTFP $\{k_{D\to A}, \Gamma, \langle U\rangle, \sigma_U^2, E_r, \Delta E\}$ for all landmark structures and the corresponding CT landscape that maps a set of geometric descriptors to CT properties such as $k_{D\to A}$.

It is noted that two kinds of MD simulation are employed: in the first step MD simulation was used to sample different triad conformations, and in the current CT rate constant calculation step, following TD-DFT calculation, the MD simulation of a specific landmark structure was used to obtain the energy gap statistics, i.e., $\langle U\rangle$ and σ_U^2 that depend on the triad conformation.

2.4. Feature engineering

To prepare the input features for ML model, we transform the atomic positions of a triad conformation into a feature vector $\mathbf{X}$, which has desirable translational, orientational, and permutation invariances. In this work, we consider three commonly used feature representations

for constructing ML-model applied on molecular system, namely, smooth overlap of atomic orbitals (SOAP),[53] atom centered symmetry function (ACSF),[54] and coulomb matrix (CM).[55] The conversion from a raw coordinate file to the desired feature representation was performed using the DScribe package.[56] We have tested all the above three feature representations and found only the CM could give promising predictive power when screened using various ML algorithms we considered.

The diagonal CM elements are defined as $C_{ii} = 0.5Z_i^{2.4}$ and the off-diagonal elements are $C_{ij} = Z_iZ_j/|\mathbf{r}_i - \mathbf{r}_j|$ $(i \neq j)$, where Z_i and $\mathbf{r}_i$ are the charge and position of the i-th atom. Note that we leave the CM as is, i.e., no further normalization or sorting is performed, so there is no permutation invariance. This is justified and performs well since we work within the conformation space of a single type of molecule with the same number and order of atoms. We also tested that in our case sorting the CM matrix elements will diminish the model's predictive power since the order of atoms is useful structural information for learning, especially with a small dataset. Actually, if working within a chemical space where different molecules are explored, sorting is necessary to preserve the permutation invariance and will be beneficial to improve model performance. In principle, if provided with big and diverse enough dataset, sorting the CM elements should provide sufficient information for training ML models in conformation space.

2.5. Machine learning for CT landscape

We begin with screening possible combinations of feature representation and ML algorithm for constructing the ML model for the CT landscape. Here, we considered ML algorithms such as kernel ridge regression (KRR), random forest, lasso regression, and ElasticNet. The implementation of the above ML algorithm were performed using the scikit-learn package.[57] After preliminary screening, we decided to focus on the use of KRR algorithm and chose four kernel functions commonly used throughout the literature, viz. the linear, polynomial, Gaussian (or the radial basis function), and Laplacian.[55,58] For validation of the developed models, we use the 5-Fold cross-validation scheme, where only 80% of training data are used in each training iteration

and the other 20 % is used for performance evaluation. This procedure is then repeated five times by randomly shuffling the training and test data sets in each iteration. In each training iteration, a grid search for finding the best model hyperparameters was performed. Details related to the hyperparameters grid search is given in the Supplementary Information.

To evaluate the model performance, we employ three scoring metrics: the mean absolute error $\left(\mathrm{MAE} = \frac{1}{n}\sum_i^n |y_i^{\mathrm{ML}} - y_i|\right)$, the root mean squared error $\left(\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_i^n \left(y_i^{\mathrm{ML}} - y_i\right)^2}\right)$ and the coefficient of determination $\left(R^2 = 1 - \sum_i^n \left(y_i^{\mathrm{ML}} - y_i\right)^2 / \sum_i^n \left(y_i^{\mathrm{ML}} - \langle y \rangle\right)^2\right)$, where y_i^{ML} and y_i are the ML predicted value and the reference value of the i-th structure and n is the total number of instance in the CTFP database. More details for the KRR implementation, the explicit functional form of the kernel function, and the model screening results were supplied in the Supporting Information.

2.6. Computational overhead

We shall briefly describe the computational overhead demanded in each step of the proposed strategy using Intel Xeon Gold 6132 @ 2.60 GHz (28 cores) CPU and Nvidia GeForce RTX 2080 Ti GPU. The first is MD simulation for conformation sampling that costs about 1200 GPU hours in total. The second is the conformation data processing and landmark structures extraction, which costs less than a single CPU core hour and is negligible. The third is the CTRAMER calculation, which includes the TD-DFT calculation that costs about 96 CPU core hours, and the subsequent MD simulation with explicit solvent that costs 1 GPU hour and 100 CPU core hours for each conformation. In total, we have performed CTRAMER calculations for more than 500 conformations that yield an overall cost of about 500 GPU hours and 97000 CPU core hours. The fourth is the feature engineering for all the 1.2 million conformations, which costs less than 2 CPU core hours for each feature representation (SOAP, ACSF, and CM), and is negligible. The fifth is the ML model screening and training, which costs less than 100 CPU core hours for all combinations of feature representations and ML algorithms considered in this work. The last is using the ML model to construct CT Landscape, which only takes a few minutes in a single-core

CPU.

All in all, the computational overhead for the implementation of the 5 steps of constructing CTFP database is about 1700 GPU hours and 97000 CPU core hours. We note that all simulations were performed in a massively parallel high-performance computing (HPC) platform and thus can be implemented in time efficient manner. Importantly, we have also employed a convergence analysis that allows us to make decision that reduced the bottleneck CTRAMER calculation cost by a factor of 100 with minimum trade-off. Details on the convergence analysis is provided in the Supplementary Information.

3. Results and Discussion

3.1. Extraction of landmark structures

In Fig. 2(b), we present the Pearson correlation for D1–D10 OPs (the full correlations between all 16 OPs are given in Supporting Information). We first check the Pearson correlation between all the OPs to ensure they reflect enough diversity for representing the conformation space. It is apparent that most of these OPs are largely uncorrelated and diverse with a notable exception between D10 and $\{$D1, D2, D3$\}$, which shows either highly correlated or anticorrelated, but we still keep D10 for direct chemical insights.

Next, we sort all the structures based on different OPs in ascending order and obtain 16 sorted OP data spaces, where the grid sampling could be performed as shown in Fig. 2(c). In our implementation, the grid sampling has three main parameters that can be tuned: we use grid spacing of $N/11$ with N being the total number of structures, the cutoff of stochastic sampling around grid nodes as $\pm$100 data points, and randomly sample 3 points near each node. Finally, a total of 528 landmark structures were obtained with the stochastic grid sampling approach and their distribution represented on the t-SNE components Z1 and Z2 is shown in Fig. 2(d) with some representative conformations. Projection of all the 528 landmarks structures on each of the individual OP is given in the Supporting Information.

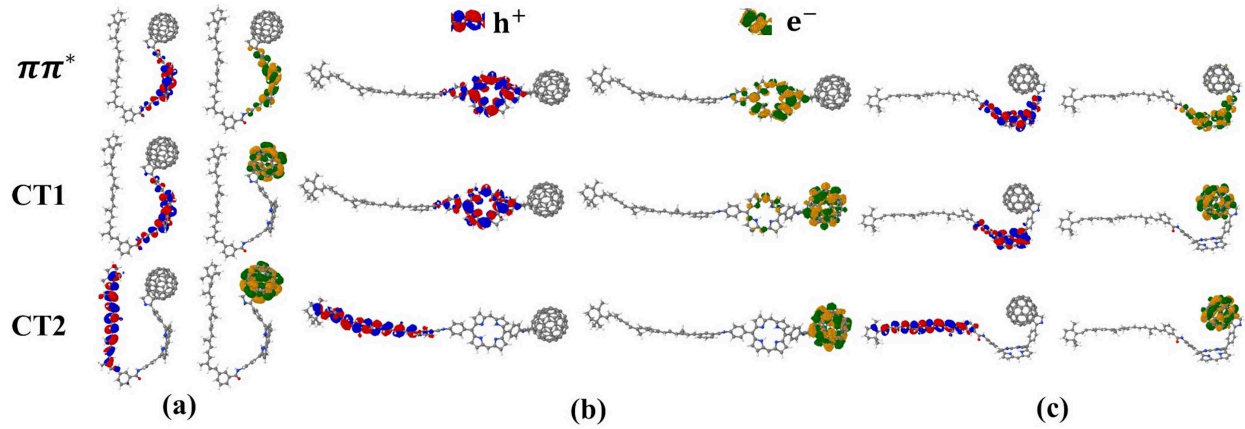
3.2. CTRAMER calculation to obtain CTFP database

With the landmark structures that faithfully represent the triad conformation space, we calculate their CTFP using development version of the CTRAMER package, which nowadays can automatically analyze CT-state characteristics based on the one electron density matrix quantitative analysis.[59,60]

Table 1 shows the CTFP for three example triad conformations, Conf #0 (fold), Conf #32 (linear), and Conf #364 (bent), in addition the natural transition orbital (NTO) diagrams exhibit the localization of hole and electron, which is expected since $\pi\pi^*$ state has a localized excitation on porphyrin, CT1 state has a CT from porphyrin to C_{60}, and CT2 state has a CT from carotenoid to C_{60}. The result of CTRAMER calculation is consistent with the previously reported experimental data[41,43] and numerical calculations,[17,19,46] which show the excitation energies for the CT1 states are within the range of 1.6 – 2.6 eV and for CT2 states to be between 2.1 – 2.7 eV.[61,62] Moreover, the CT rate constant k for $\pi\pi^* \rightarrow$ CT1 transition varies from 8.1×10^9 s^{-1} in the fold Conf #0 to 1.2×10^{14} s^{-1} in the bent Conf #364, as well as for $\pi\pi^* \rightarrow$ CT2 transition k varies from 8.2×10^5 s^{-1} in the linear Conf #32 to 8.0×10^8 s^{-1} in the bent Conf #364. These results clearly show that CT rate constant differs by 5 orders of magnitude in different conformations, which can be traced back to the fact that spatial arrangements of molecular segments affects the electronic structures (Γ, $\langle U \rangle$) and the fluctuations of the surrounding solvents ($\langle U \rangle$, σ_U^2).

It turns out that 495 out of 528 landmark structures have well-defined $\pi\pi^*$, CT1, CT2 excited states. The criterion for defining excitonic and CT states are detailed in the Supporting Information. Figure 3 presents the probability distribution and scatter matrix for the key CT properties of the 495 landmark structures, including logarithm with base 10 of CT rate constant $\log(k/s^{-1})$, Γ, $\langle U \rangle$, and σ_U^2. Several trends can be observed. At first glance, the slope of $\langle U \rangle$ versus $\log(k/s^{-1})$ is negative for the CT1 case and positive for the CT2 case, which seems to indicate anticorrelation and correlation between the two quantities. However, this is a statistical pitfall for misinterpreting without looking at the physics, since the exponential term $\exp\left[-\langle U \rangle^2/(2\sigma_U^2)\right]$ in Eq. 1 dictates that when $\langle U \rangle$ is closer to zero the CT rate is larger. Similarly, larger coupling strength Γ leads to larger

Table 1: Charge transfer properties for three triad conformations. The upper panel shows the natural transition orbital (NTO) diagrams of the localization of hole, h^+ (red-blue) and electron, e^- (green-orange). The bottom table shows the excitation energy for state α, E_α ($\alpha = \pi\pi^*$, CT1, CT2) in (eV), CT fingerprints including CT rate constant, k in (s^{-1}), electronic coupling Γ in (eV), reaction free energy ΔE in (eV), average donor-acceptor energy gap $\langle U \rangle$ in (eV), and its variance σ_U^2 in (eV^2) for both $\pi\pi^* \rightarrow$ CT1 and $\pi\pi^* \rightarrow$ CT2 transitions.



Excitation	(a) Conf #0		(b) Conf #32		(c) Conf #364	
$E_{\pi\pi^*}$	2.58		1.74		2.67	
E_{CT1}	1.87		2.04		1.49	
E_{CT2}	2.01		2.65		2.53	
Transition	$\pi\pi^* \rightarrow$ CT1	$\pi\pi^* \rightarrow$ CT2	$\pi\pi^* \rightarrow$ CT1	$\pi\pi^* \rightarrow$ CT2	$\pi\pi^* \rightarrow$ CT1	$\pi\pi^* \rightarrow$ CT2
k (s^{-1})	8.12×10^{9}	4.41×10^{6}	5.56×10^{13}	8.16×10^{5}	1.21×10^{14}	8.03×10^{8}
Γ (eV)	-2.68×10^{-3}	2.01×10^{-5}	5.47×10^{-2}	-7.20×10^{-5}	6.58×10^{-2}	-4.64×10^{-4}
$\langle U \rangle$ (eV)	0.559	0.2064	-0.124	-0.835	0.000150	-0.440
σ_U^2 (eV^2)	0.0594	0.0726	0.0177	0.0785	0.0186	0.0723
E_r (eV)	1.15	1.40	0.341	1.52	0.360	1.40
ΔE (eV)	-1.71	-1.61	-0.217	-0.681	-0.360	-0.958

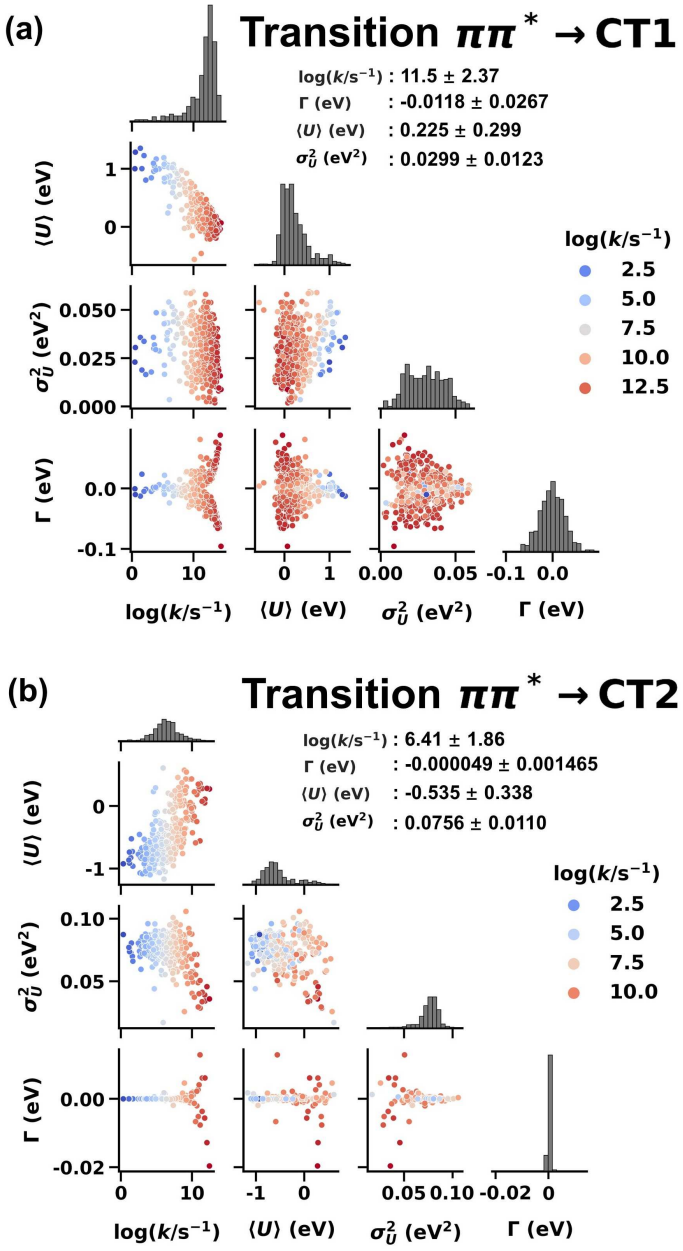


Figure 3: Scatter matrix and probability distribution of the CT properties for 495 landmark structures undergoing (a) $\pi\pi^* \rightarrow$ CT1 and (b) $\pi\pi^* \rightarrow$ CT2 transitions. The data shown are logarithmic base 10 of the CT rate constant, $\log(k/\mathrm{s}^{-1})$, electronic coupling Γ in (eV), average donor-acceptor energy gap $\langle U \rangle$ in (eV), and its variance σ_U^2 in (eV^2), the data are color coded based on the $\log(k/\mathrm{s}^{-1})$ value for each transition.

CT rate is seen. The coupling in $\pi\pi^* \rightarrow$ CT2 transition is typically small and narrowly distributed, which is the main reason for a much smaller CT rate compared with the $\pi\pi^* \rightarrow$ CT1 transition. The variance σ_U^2 or the equivalent reorganization energy does not have a monotonous correlation with CT rate, since it enters both competing terms of the prefactor term and the exponential term in Eq. 1. It is surprising that the span of the CT rate constants can be as large as 12–15 orders of magnitude in these transitions. Therefore, it is absolutely necessary to have conformation-dependent CT rate ML model in order to understand and rationalize their contributions to the macroscopic CT rate.

3.3. Machine Learning for CT landscape

Now, we build ML models to predict the CTFP using the CM features as input. The selected 495 landmark structures' CTFP database was used for training and validation with regularization to avoid overfitting. The challenge here is to predict many CT properties while constrained to work with rather small dataset. Kernel ridge regression (KRR) has been shown to have versatile performance for constructing models using limited amount of training data, though typically still on the order of $\sim 10^4$ instances.[63–65] Here, we made the attempts to implement KRR with only $\sim 10^2$ instances. For the KRR algorithm, we have tested different kernel functions such as linear, polynomial, Gaussian (radial basis function), and Laplacian,[55,58] and found that only the linear and polynomial kernel works well for all CT properties. In this work, we choose to use the third order polynomial kernel function. The validation of the model were performed via the 5-fold cross validation. For each transition, a total of 6 best ML models are constructed for the 6 CTFP properties $\{k_{D\rightarrow A}, \Gamma, \langle U\rangle, \sigma_U^2, E_r, \Delta E\}$.

In Fig. 4, we present the cross-validation results from direct ML modeling for $\log(k/\mathrm{s}^{-1})$, Γ, $\langle U\rangle$, and σ_U^2, as well as the calculated rate constant using the ML predicted $\{\Gamma, \langle U\rangle, \sigma_U^2\}$ using Eq. 1. It is observed that $R^2 > 0.97$ is achieved for all direct ML learned properties (Fig. 4(a–h)) and both MAE and RMSE errors are close within the chemical accuracy of $\sim 1k_BT$ (or ~ 0.03 eV), where k_B is the Boltzmann constant and the temperature $T = 300$ K. The developed direct ML models thus enables us to directly perform a QSPR mapping and reproduce the CT landscape

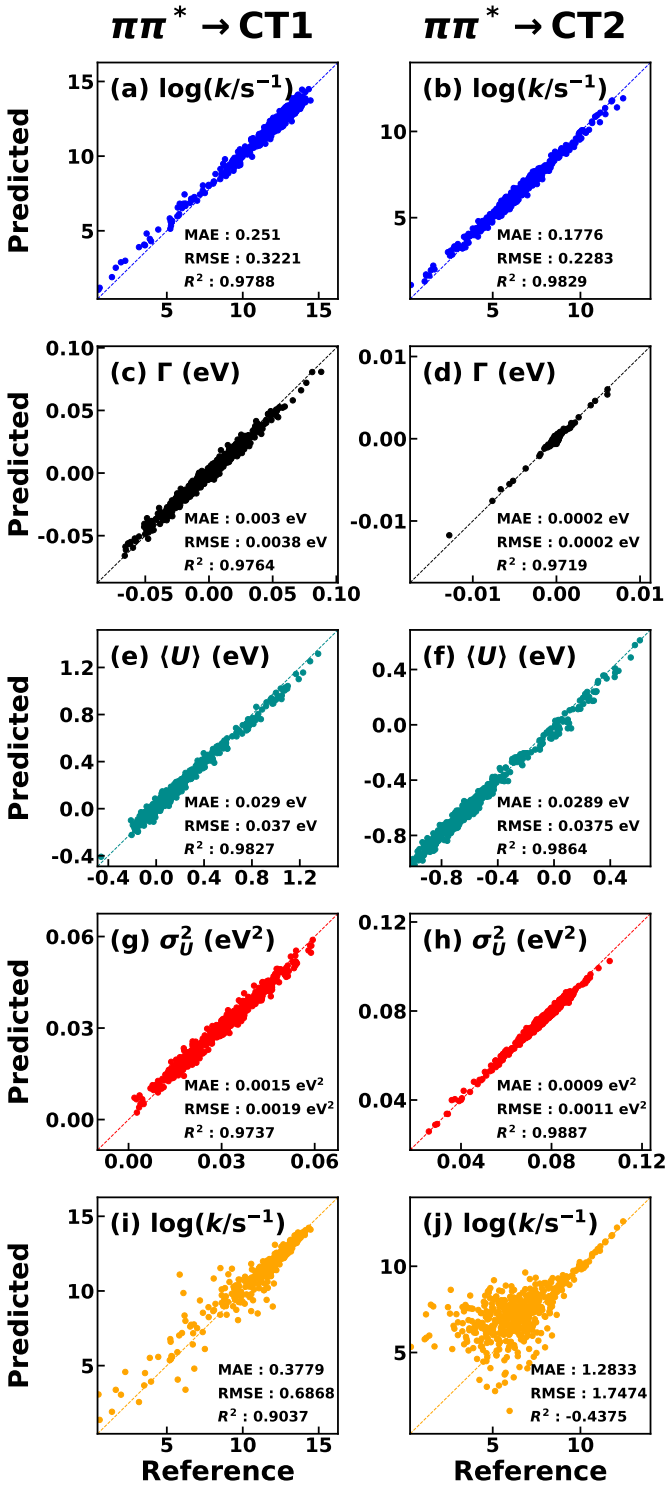


Figure 4: Performance of kernel ridge regression with 5-fold cross validation for transitions $\pi\pi \rightarrow$ CT1 (left panels) and $\pi\pi \rightarrow$ CT2 (right panels). (a,b) directly predicted logarithm of CT rate constant, $\log(k/\mathrm{s}^{-1})$ (blue), (c,d) electronic coupling Γ in (eV) (black), (e,f) average energy gap $\langle U \rangle$ in (eV) (dark cyan), and (g,h) variance of energy gap σ_U^2 in (eV^2) (red), and (i,j) $\log(k/\mathrm{s}^{-1})$ calculated via Eq. 1 (orange) using the predicted parameters, as compared with the CT fingerprint reference values.

of the CTFP database (Fig. 5(a,b)). However, the calculated $\log(k/\mathrm{s}^{-1})$ using Eq. 1 (Fig. 4(i,j)) seems to have noticeably worse performance compared with the directly ML learned $\log(k/\mathrm{s}^{-1})$ (Fig. 4(a,b)). The R^2 score of the $\pi\pi^* \rightarrow$ CT1 transition decreases to 0.9, and for the $\pi\pi^* \rightarrow$ CT2 transition R^2 score is disastrous -0.44. The cause of the failure for the calculated $\log(k/\mathrm{s}^{-1})$ could be ascribed to the fact that Γ for this transition are narrowly distributed close to zero and hence the propagated errors from the $\langle U \rangle$ and σ_U^2 become rather significant. In addition, the failure is also a result of insensitivity of the calculated $k(\mathrm{s}^{-1})$ obtained in the linear scale when used to predict for structure with CT rate many order of magnitude smaller, the error due to this insensitivity become blatantly exposed when the prediction result was projected to the logarithmic space to obtain the calculated $\log(k/\mathrm{s}^{-1})$. This provide useful insight that means if one directly learn the CT rate constant k in the linear scale, only the k with the large order of magnitude will be learned, which will give inaccurate predictions in the case of low k values. Therefore, it is important to train ML model to directly predict the logarithm of the rate constant $\log(k/\mathrm{s}^{-1})$ so as to have high predictive power.

The ML models we obtained using the CM representation with the KRR model is non-trivial. As such, given a training set of the same small size, other combinations of feature representations and models we tested in the model screening process cannot achieve the same level of predictive power for all CTFP properties considered. Moreover, although the combination of CM and KRR has been previously reported for constructing ML model for predicting electronic couplings or energies for gas-phase small molecules with about 10^4 instances,[5,29,55] it is impractical to obtain such a large dataset for the current condensed-phase system that involves hundreds of atoms in the quantum chemistry calculation and tens of thousands of atoms in the MD simulations.

Now, we are ready to construct the CT landscape for the entire conformation space using the ML model for logarithm of CT rate constant $\log(k/\mathrm{s}^{-1})$. This allows us to better explore the mapping between the conformation space and CT properties space. Figure 5 depicts the CT landscapes rendered on the geometric descriptors D1 and D9 for both $\pi\pi^* \rightarrow$ CT1 and $\pi\pi^* \rightarrow$ CT2 transitions. The ML predicted $\log(k/\mathrm{s}^{-1})$ in Fig. 5(b) reproduces the reference $\log(k/\mathrm{s}^{-1})$

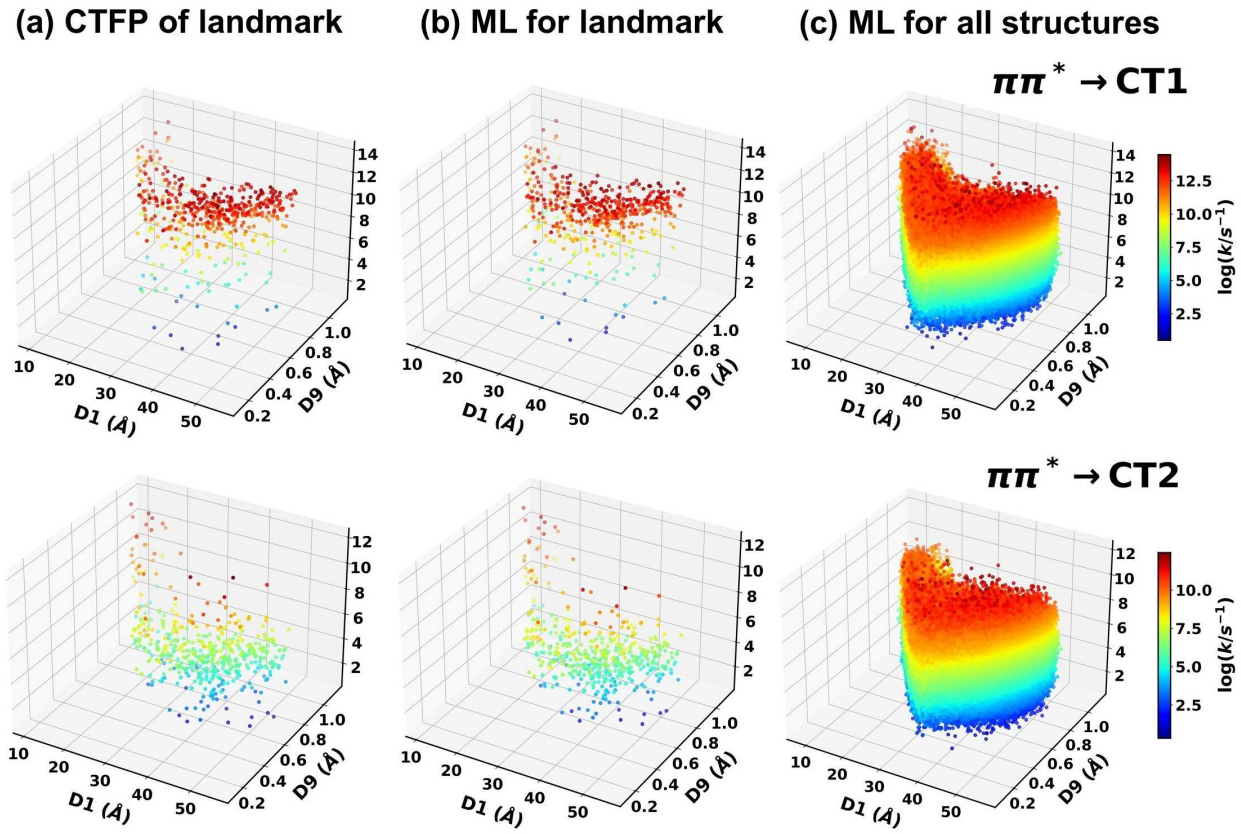


Figure 5: CT landscapes rendered on D1 and D9 feature space with $\log(k/\mathrm{s}^{-1})$ shown on z-axis for (a) the landmark structures in CTFP database, (b) the landmark structures reconstructed with the ML model, (c) the entire 1.2 million conformations predicted by the ML model.

from CTRAMER calculation in Fig. 5(a). Our final result is the CT landscape for the 1.2 million conformations as shown in Fig. 5(c), which for the first time establishes a QSPR relationship between molecular structure and the CT rate constant in realistic condensed-phase system. It is noted that the CT landscape is built on 207×207-dimensional CM feature space, and but plotted in a reduced two-dimensional geometric descriptor space. In general, the CT landscape is based on the full-dimensional conformation space, i.e. $k(\mathbf{X}(\mathbf{r})) = k(\mathbf{r})$, similar to the energy landscape $V(\mathbf{r})$, where the energy is replaced by the rate constant in terms of $\log(k/\mathrm{s}^{-1})$, or other CTFP properties like Γ, $\langle U \rangle$, σ_U^2, etc.

In fact, the CT landscape is informative and can provide the molecular insights of the QSPR in the CT properties of the triad. For example, we show some immediate utilities of the CT landscape based on the CTFP database as follows. In Fig. 6, we present the CT rate constant $\log(k/\mathrm{s}^{-1})$ distribution atop of the 2-dimensional free energy landscape obtained from the conformation database of 1.2 million structures rendered on the PC1_XYZ and PC2_XYZ feature space. We first calculate and extract the Helmholtz free energy, $F_i = -k_B T \ln[\mathrm{Prob.}(\mathbf{D}_i)]$, for the i-th landmark structure, where $\mathrm{Prob.}(\mathbf{D}_i)$ is the probability of finding the structure that is described by the OP pair of the i-th structure $\mathbf{D}_i$, which in this case is $\{$PC1_XYZ, PC2_XYZ$\}$. We then calculate the Boltzmann weight, W_i, as $W_i = e^{-\beta F_i} / \sum_{i=1}^{n} e^{-\beta F_i}$, where n is the number of sampled structures and $\beta = 1/k_B T$ is the inverse temperature with $T = 300$K. The ensemble-averaged CT rate constant is thus given by

$$\langle k_{D\to A} \rangle = \sum_{i=1}^{n} W_i \, k_{D\to A,i}, \tag{4}$$

where $k_{D\to A,i}$ is the CT rate constant for the i-th landmark structure. The ensemble-averaged CT rate constant for the transition $\pi\pi^* \to$ CT1 obtained via the CTFP database of the landmark structures is $(0.82 \pm 0.23) \times 10^{11}$ s^{-1}, and the ensemble average obtained from randomly sampling 770 structures from the ML-based CT landscape of 1.2 million structures is $(5.9 \pm 1.0) \times 10^{11}$ s^{-1}.

Experimental measurements are available for two similar triad molecules dissolved in 2-Methyl-THF at 292 K: the first one corresponds to replacing the porphyrin segment with octaalkylporphyrin,[42] and the second has a longer alkyl chain compared to the first one.[41]

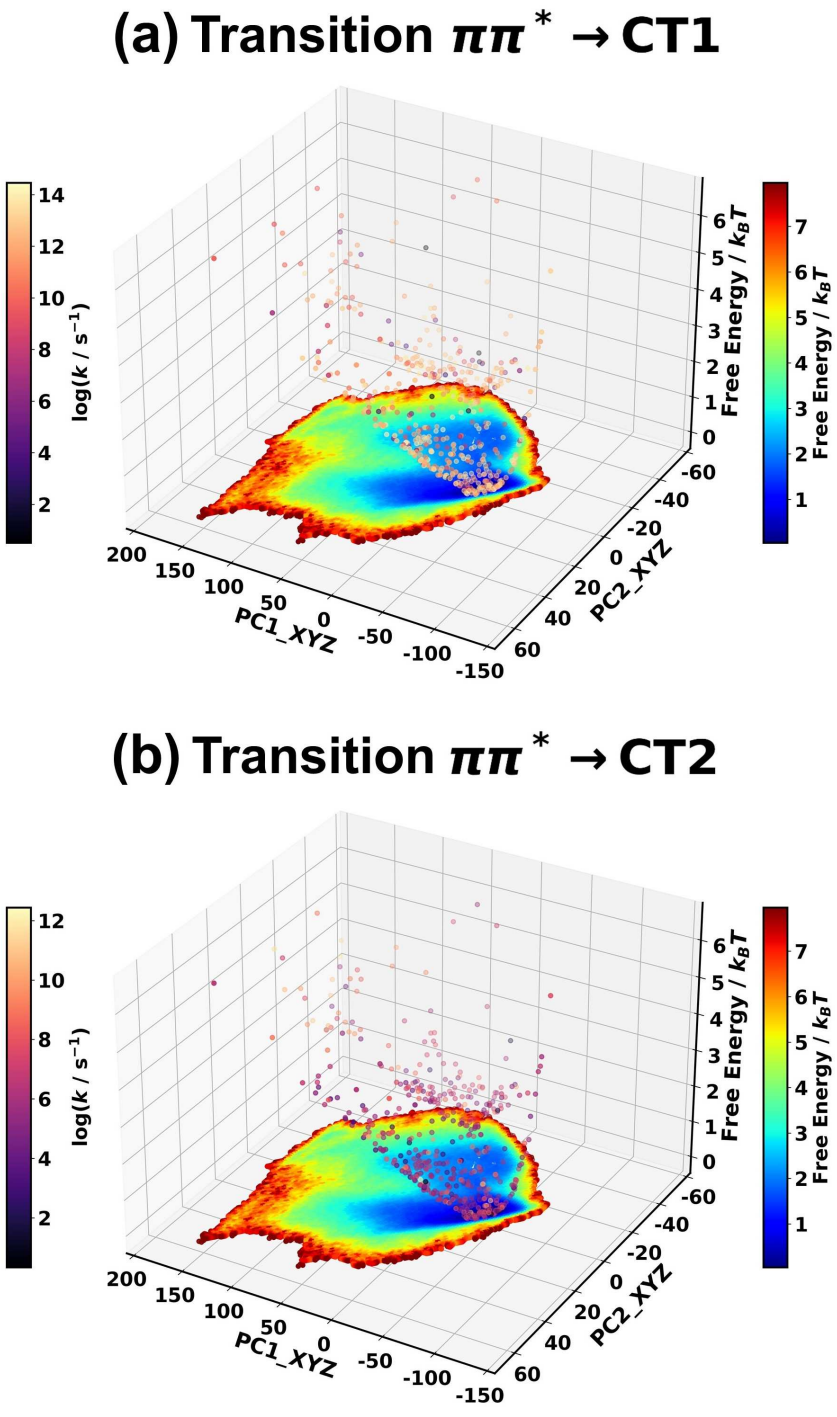


Figure 6: Distribution of the CT rate constant $\log(k/\mathrm{s}^{-1})$ for 495 landmark structures undergoing (a) $\pi\pi^* \rightarrow$ CT1 and (b) $\pi\pi^* \rightarrow$ CT2 transitions rendered atop of the two-dimensional free energy landscape in terms of PC1_XYZ and PC2_XYZ order parameters. The free energy landscape were obtained from the 1.2 million MD sampled configurations.

The experimental values for these two similar triad molecules are 1×10^{11} and 3.3×10^{11}, respectively.[41,42] Thus, the theoretical ensemble averages based on first principles agree with the experimental measurements of similar triad molecules on the order of magnitude. We also note that having the Boltzmann weighting is essential to reflect the physical distribution of conformations, whose ensemble average is measured experimentally.

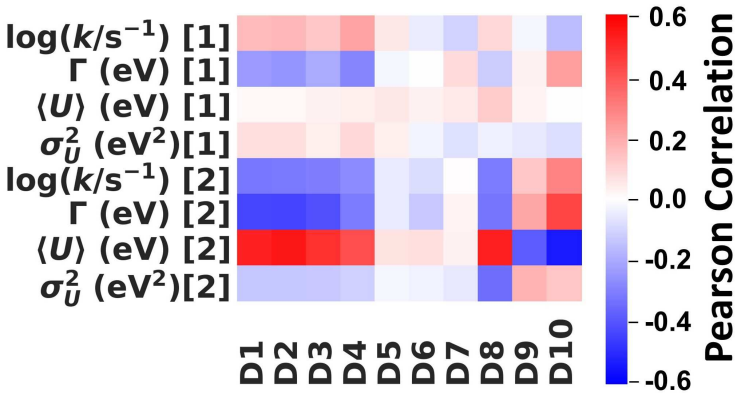


Figure 7: Pearson correlation between CTFP properties undergoing transitions [1]: $\pi\pi^* \rightarrow$ CT1 and [2]: $\pi\pi^* \rightarrow$ CT2 and the geometric descriptors, D1–D10. All the descriptors are in unit of ($\mathring{A}$) with exception of D3, D4, and D7 that are in unit of ($^\circ$).

Another application of the CT landscape is to provide interpretability for understanding the effects of molecular structures on the CT rate. Here, we show a correlation analysis between the CTFP properties ($\log(k/\mathrm{s}^{-1})$, Γ, $\langle U \rangle$, and σ_U^2) and the geometric descriptors (D1–D10) for the CTFP database in Fig. 7. We highlight our observation into the following three points. First, all the geometric descriptors are less correlated with the CT properties of the transition $\pi\pi^* \rightarrow$ CT1 than those of the transition $\pi\pi^* \rightarrow$ CT2. This is expected, since for the same set of structures, the CTFP properties of the $\pi\pi^* \rightarrow$ CT1 transition is more diverse than those in the $\pi\pi^* \rightarrow$ CT2 case (cf. Fig. 3). Second, the descriptors D1–D4 and D8–D10 have strong correlation or anticorrelation with CT properties emphasizing the importance of the relative distance and orientation between the carotenoid and the fullerene. Third, the descriptors D5–D7 that exclusively describe the local shape of the carotenoid was found to be almost uncorrelated to any of the CTFP properties, indicating that the CTFP properties are less sensitive to the the carotenoid local structure. We note that the

quantitative insights obtained via the correlation studies can also be used to serve as an optimization metrics to perform a data-driven search for highly predictive and physically meaningful geometric descriptors.

To better visualize the CT landscape from all possible combinations of geometric descriptors, we implemented an interactive web interface called the *CT landscape explorer*, and the code is available from `https://github.com/xiangsunlab/ct_landscape`. The CT landscape explorer is capable of showing the high-dimensional conformation space in terms of user-selected geometric descriptors in 3 dimensions (like in Fig. 5) and instant displaying the molecular conformation and its CT properties for a chosen data point just a click away (see snapshot in Fig. S9 in the Supporting Information). The entire CTFP database has been integrated into the web interface.

Finally, we envisage that the developed ML modeling strategy is extremely useful for estimating the effects of the static and dynamics disorders to the CT rate in homogeneous or heterogeneous environments, and could also give a time-dependent IMT CT rate coefficient as the molecular structure evolves in time. We will report these immediate applications in our subsequent works. Moreover, the ML-based QSPR models connecting CT properties with the global and local geometric descriptors could be used to reverse search and design the nanostructure or device morphology with desired properties, to achieve high-throughput virtual screening for OPV materials. Optimized data-driven correlation analysis we presented can also be further develop for extracting interpretable molecular insights. Insights gained from both forward and reverse CT landscape exploration could then be use to guide further experimentation and validation.

4. Conclusion

In conclusion, we have developed ML models based on kernel ridge regression that establish the structure–CT rate properties mapping relations for a prototypical OPV molecule in the condensed phase for the first time. It is important to have a good ML algorithm, but the quality and diversity of

the training dataset is the most crucial factor in the model prediction accuracy. We thereby select a small set of landmark structures that are representative of the conformation space using stochastic grid sampling according to their sorted geometrical descriptors. ML models are trained and cross-validated with the landmark structures and their CTFP database computed via *ab initio* calculation and MD simulations with explicit solvent, which includes CT rate constant, electronic coupling, average donor-acceptor energy gap, and its variance, along with the traditional reorganization energy and reaction free energy. In contrast to the previous ML approaches where only the gas-phase CT parameters like electronic coupling and reorganization energy are trained, in this work, we demonstrated that the condensed-phase CT rate constant should be directly trained utilizing the $\log(k/\mathrm{s}^{-1})$ as the target. The developed ML models can accurately predict the CTFP with R^2 score larger than 0.97, and MAE and RMSE within chemical accuracy ($\sim 1\ k_BT$). The CT landscape for the entire conformation space containing 1.2 million MD-sampled structures is constructed with the ML models, allowing extremely efficient exploration of the high-dimensional conformation space by looking up the CT properties for a specific structure instantly. The ML-based CT landscape opens the door to the bottom-up study of charge transport properties on nanoscale or even mesoscale bulk homogenous or heterogeneous OSC materials, where static and dynamic structural disorders and the time-dependent CT phenomena can be addressed. Furthermore, the reverse search for the nanostructure or morphology enables the optimization of fabrication design for optoelectronic devices.

Supporting Information

Convergence analysis of MD conformation sampling; Pearson correlation heat map for all order parameters; landmark structure distribution on order parameter feature spaces; CTRAMER calculation details; algorithm for the automatic CT state analysis; details on ML model building; snapshot of the CT landscape explorer; coordinates for the triad Conf. 364.

Acknowledgement

X.S. acknowledges support from NYU Shanghai, the National Natural Science Foundation of China (No. 21903054), the Hefei National Laboratory for Physical Sciences at the Microscale (No. KF2020008), the Shanghai Sailing Program (No. 19YF1435600), and the Program for Eastern Young Scholar at Shanghai Institutions of Higher Learning. Computing resources were provided by NYU Shanghai HPC.

References

(1) Manzhos, S.; Chueh, C.-C.; Giorgi, G.; Kubo, T.; Saianand, G.; Lüder, J.; Sonar, P.; Ihara, M. Materials Design and Optimization for Next-Generation Solar Cell and Light-Emitting Technologies. *J. Phys. Chem. Lett.* **2021**, *12*, 4638–4657.

(2) Xu, Y.; Cui, Y.; Yao, H.; Zhang, T.; Zhang, J.; Ma, L.; Wang, J.; Wei, Z.; Hou, J. A New Conjugated Polymer that Enables the Integration of Photovoltaic and Light-Emitting Functions in One Device. *Adv. Mater.* **2021**, *33*, 2101090.

(3) Wen, Y.; Liu, Y.; Yan, B.; Gaudin, T.; Ma, J.; Ma, H. Simultaneous Optimization of Donor/Acceptor Pairs and Device Specifications for Nonfullerene Organic Solar Cells Using a QSPR Model with Morphological Descriptors. *J. Phys. Chem. Lett.* **2021**, *12*, 4980–4986.

(4) Häse, F.; Roch, L. M.; Friederich, P.; Aspuru-Guzik, A. Designing and Understanding Light-Harvesting Devices with Machine Learning. *Nat. Comm.* **2020**, *11*.

(5) Wang, C.-I.; Joanito, I.; Lan, C.-F.; Hsu, C.-P. Artificial Neural Networks for Predicting Charge Transfer Coupling. *J. Chem. Phys.* **2020**, *153*, 214113.

(6) Xiang, D.; Wang, X.; Jia, C.; Lee, T.; Guo, X. Molecular-Scale Electronics: From Concept to Function. *Chem. Rev.* **2016**, *116*, 4318–4440.

(7) Borges-González, J.; Kousseff, C. J.; Nielsen, C. B. Organic Semiconductors for Biological Sensing. *J. Mater. Chem. C* **2019**, *7*, 1111–1130.

(8) Kandrashkin, Y. E. Influence of Spin Decoherence on the Yield of Photodriven Quantum Teleportation in Molecular Triads. *J. Phys. Chem. Lett.* **2021**, *12*, 6405–6410.

(9) Wang, J.; Zheng, Z.; Zu, Y.; Wang, Y.; Liu, X.; Zhang, S.; Zhang, M.; Hou, J. A Tandem Organic Photovoltaic Cell with 19.6% Efficiency Enabled by Light Distribution Control. *Adv. Mater.* **2021**, 2102787.

(10) Wang, Y.; Lee, J.; Hou, X.; Labanti, C.; Yan, J.; Mazzolini, E.; Parhar, A.; Nelson, J.; Kim, J.-S.; Li, Z. Recent Progress and Challenges toward Highly Stable Nonfullerene Acceptor-Based Organic Solar Cells. *Adv. Energy Mater.* **2020**, *11*, 2003002.

(11) Hu, Z.; Adachi, T.; Haws, R.; Shuang, B.; Ono, R. J.; Bielawski, C. W.; Landes, C. F.; Rossky, P. J.; Vanden Bout, D. A. Excitonic Energy Migration in Conjugated Polymers: The Critical Role of Interchain Morphology. *J. Am. Chem. Soc.* **2014**, *136*, 16023–16031.

(12) Brian, D.; Eslamian, M. Design and Development of a Coating Device: Multiple-Droplet Drop-Casting (MDDC-Alpha). *Rev. Sci. Instrum.* **2020**, *91*, 033902.

(13) Zhao, Z.-W.; Omar, O. H.; Padula, D.; Geng, Y.; Troisi, A. Computational Identification of Novel Families of Nonfullerene Acceptors by Modification of Known Compounds. *J. Phys. Chem. Lett.* **2021**, *12*, 5009–5015.

(14) Reiser, P.; Konrad, M.; Fediai, A.; Léon, S.; Wenzel, W.; Friederich, P. Analyzing Dynamical Disorder for Charge Transport in Organic Semiconductors via Machine Learning. *J. Chem. Theory Comput.* **2021**, *17*, 3750–3759.

(15) Zheng, Z.; Tummala, N. R.; Wang, T.; Coropceanu, V.; Brédas, J.-L. Charge-Transfer States at Organic–Organic Interfaces: Impact of Static and Dynamic Disorders. *Adv. Energy Mater.* **2019**, *9*, 1803926.

(16) Rinderle, M.; Kaiser, W.; Mattoni, A.; Gagliardi, A. Machine-Learned Charge Transfer Integrals for Multiscale Simulations in Organic Thin Films. *J. Phys. Chem. C* **2020**, *124*, 17733–17743.

(17) Sun, X.; Zhang, P.; Lai, Y.; Williams, K. L.; Cheung, M. S.; Dunietz, B. D.; Geva, E. Computational Study of Charge-Transfer Dynamics in the Carotenoid–Porphyrin–C_{60} Molecular Triad Solvated in Explicit Tetrahydrofuran and Its Spectroscopic Signature. *J. Phys. Chem. C* **2018**, *122*, 11288–11299.

(18) Marcus, R. A. On the Theory of Oxidation–Reduction Reactions Involving Electron Transfer. I. *J. Chem. Phys.* **1956**, *24*, 966–978.

(19) Manna, A. K.; Balamurugan, D.; Cheung, M. S.; Dunietz, B. D. Unraveling the Mechanism of Photoinduced Charge Transfer in Carotenoid–Porphyrin–C_{60} Molecular Triad. *J. Phys. Chem. Lett.* **2015**, *6*, 1231–1237.

(20) Chandler, D. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B. J., Ciccotti, G., Coker, D. F., Eds.; World Scientific: Singapore, 1998; Chapter 2, pp 25–49.

(21) Tinnin, J.; Bhandari, S.; Zhang, P.; Aksu, H.; Maiti, B.; Geva, E.; Dunietz, B. D.; Sun, X.; Cheung, M. S. Molecular-Level Exploration of the Structure-Function Relations Underlying Interfacial Charge Transfer in the Subphthalocyanine/C_{60} Organic Photovoltaic System. *Phys. Rev. Applied* **2020**, *13*, 054075.

(22) Sun, X.; Geva, E. Equilibrium Fermi's Golden Rule Charge Transfer Rate Constants in the Condensed Phase: The Linearized Semiclassical Method vs Classical Marcus Theory. *J. Phys. Chem. A* **2016**, *120*, 2976–2990.

(23) Tong, Z.; Gao, X.; Cheung, M. S.; Dunietz, B. D.; Geva, E.; Sun, X. Charge Transfer Rate Constants for the Carotenoid-Porphyrin-C_{60} Molecular Triad Dissolved in Tetrahydrofuran:

The Spin-Boson Model vs the Linearized Semiclassical Approximation. *J. Chem. Phys.* **2020**, *153*, 044105.

(24) Tinnin, J.; Aksu, H.; Tong, Z.; Zhang, P.; Geva, E.; Dunietz, B. D.; Sun, X.; Cheung, M. S. CTRAMER: An Open-Source Software Package for Correlating Interfacial Charge Transfer Rate Constants with Donor/Acceptor Geometries in Organic Photovoltaic Materials. *J. Chem. Phys.* **2021**, *154*, 214108.

(25) Prezhdo, O. V. Advancing Physical Chemistry with Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 9656–9658.

(26) Ceriotti, M.; Clementi, C.; von Lilienfeld, O. A. Machine Learning Meets Chemical Physics. *J. Chem. Phys.* **2021**, *154*, 160401.

(27) Westermayr, J.; Gastegger, M.; Schütt, K. T.; Maurer, R. J. Perspective on Integrating Machine Learning Into Computational Chemistry and Materials Science. *J. Chem. Phys.* **2021**, *154*, 230903.

(28) Çaylak, O.; Yaman, A.; Baumeier, B. Evolutionary Approach to Constructing a Deep Feedforward Neural Network for Prediction of Electronic Coupling Elements in Molecular Materials. *J. Chem. Theory Comput.* **2019**, *15*, 1777–1784.

(29) Wang, C.-I.; Braza, M. K. E.; Claudio, G. C.; Nellas, R. B.; Hsu, C.-P. Machine Learning for Predicting Electron Transfer Coupling. *J. Phys. Chem. A* **2019**, *123*, 7792–7802.

(30) Bag, S.; Aggarwal, A.; Maiti, P. K. Machine Learning Prediction of Electronic Coupling between the Guanine Bases of DNA. *J. Phys. Chem. A* **2020**, *124*, 7658–7664.

(31) Lederer, J.; Kaiser, W.; Mattoni, A.; Gagliardi, A. Machine Learning–Based Charge Transport Computation for Pentacene. *Adv. Theory Simul.* **2018**, *2*, 1800136.

(32) Misra, M.; Andrienko, D.; Baumeier, B.; Faulon, J.-L.; von Lilienfeld, O. A. Toward

Quantitative Structure–Property Relationships for Charge Transfer Rates of Polycyclic Aromatic Hydrocarbons. *J. Chem. Theory Comput.* **2011**, *7*, 2549–2555.

(33) Atahan-Evrenk, S.; Atalay, F. B. Prediction of Intramolecular Reorganization Energy Using Machine Learning. *J. Phys. Chem. A* **2019**, *123*, 7855–7863.

(34) Liu, Z.; Lin, L.; Jia, Q.; Cheng, Z.; Jiang, Y.; Guo, Y.; Ma, J. Transferable Multilevel Attention Neural Network for Accurate Prediction of Quantum Chemistry Properties via Multitask Learning. *J. Chem. Inf. Model.* **2021**, *61*, 1066–1082.

(35) Lee, C.-K.; Lu, C.; Yu, Y.; Sun, Q.; Hsieh, C.-Y.; Zhang, S.; Liu, Q.; Shi, L. Transfer Learning with Graph Neural Networks for Optoelectronic Properties of Conjugated Oligomers. *J. Chem. Phys.* **2021**, *154*, 024906.

(36) Lu, C.; Liu, Q.; Sun, Q.; Hsieh, C.-Y.; Zhang, S.; Shi, L.; Lee, C.-K. Deep Learning for Optoelectronic Properties of Organic Semiconductors. *J. Phys. Chem. C* **2020**, *124*, 7048–7060.

(37) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(38) Feng, J.; Wang, H.; Ji, Y.; Li, Y. Molecular Design and Performance Improvement in Organic Solar Cells Guided by High-Throughput Screening and Machine Learning. *Nano Select* **2021**,

(39) Sahu, H.; Ma, H. Unraveling Correlations between Molecular Properties and Device Parameters of Organic Solar Cells Using Machine Learning. *J. Phys. Chem. Lett.* **2019**, *10*, 7277–7284.

(40) Rozzi, A. C.; Maria Falke, S.; Spallanzani, N.; Rubio, A.; Molinari, E.; Brida, D.; Maiuri, M.; Cerullo, G.; Schramm, H.; Christoffers, J. et al. Quantum Coherence Controls the Charge Separation in a Prototypical Artificial Light–Harvesting System. *Nat. Comm.* **2013**, *4*, 1602–1607.

(41) Liddell, P. A.; Kuciauskas, D.; Sumida, J. P.; Nash, B.; Nguyen, D.; Moore, A. L.; Moore, T. A.; Gust, D. Photoinduced Charge Separation and Charge Recombination to a Triplet State in a Carotene–Porphyrin–Fullerene Triad. *J. Am. Chem. Soc.* **1997**, *119*, 1400–1405.

(42) Bahr, J. L.; Kuciauskas, D.; Liddell, P. A.; Moore, A. L.; Moore, T. A.; Gust, D. Driving Force and Electronic Coupling Effects on Photoinduced Electron Transfer in a Fullerene-based Molecular Triad. **2000**, *72*, 598.

(43) Liddell, P. A.; Kodis, G.; Moore, A. L.; Moore, T. A.; Gust, D. Photonic Switching of Photoinduced Electron Transfer in a Dithienylethene–Porphyrin–Fullerene Triad Molecule. *J. Am. Chem. Soc.* **2002**, *124*, 7668–7669.

(44) Winters, M. U.; Dahlstedt, E.; Blades, H. E.; Wilson, C. J.; Frampton, M. J.; Anderson, H. L.; Albinsson, B. Probing the Efficiency of Electron Transfer through Porphyrin-Based Molecular Wires. *J. Am. Chem. Soc.* **2007**, *129*, 4291–4297.

(45) Hu, Z.; Tong, Z.; Cheung, M. S.; Dunietz, B. D.; Geva, E.; Sun, X. Photoinduced Charge Transfer Dynamics in the Carotenoid-Porphyrin-C_{60} Triad via the Linearized Semiclassical Nonequilibrium Fermi's Golden Rule. *J. Phys. Chem. B* **2020**, *124*, 9579–9591.

(46) Brian, D.; Sun, X. Linear-Response and Nonlinear-Response Formulations of the Instantaneous Marcus Theory for Nonequilibrium Photoinduced Charge Transfer. *J. Chem. Theory Comput.* **2021**, *17*, 2065–2079.

(47) Case, D. A.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham, III, T. E.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Ghoreishi, D.; Gilson, M. K. et al. AMBER 2020. *University of California, San Francisco* **2020**,

(48) Jolliffe, I. T. *Principal Component Analysis*; Springer New York, 1986; pp 115–128.

(49) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

(50) Baer, R.; Neuhauser, D. Density Functional Theory with Correct Long-Range Asymptotic Behavior. *Phys. Rev. Lett.* **2005**, *94*, 043002.

(51) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X. et al. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113*, 184–215.

(52) Voityuk, A. A.; Rösch, N. Fragment Charge Difference Method for Estimating Donor–Acceptor Electronic Coupling: Application to DNA π-Stacks. *J. Chem. Phys.* **2002**, *117*, 5607–5616.

(53) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87*.

(54) Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

(55) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*.

(56) Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. *Comp. Phys. Comm.* **2020**, *247*, 106949.

(57) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(58) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning

Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.

(59) Plasser, F.; Lischka, H. Analysis of Excitonic and Charge Transfer Interactions from Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2012**, *8*, 2777–2789.

(60) Plasser, F. TheoDORE: A Toolbox for a Detailed and Automated Analysis of Electronic Excited State Computations. *J. Chem. Phys* **2020**, *152*, 084108.

(61) Olguin, M.; Basurto, L.; Zope, R. R.; Baruah, T. The Effect of Structural Changes on Charge Transfer States in a Light-Harvesting Carotenoid-Diaryl-Porphyrin-C60 Molecular Triad. *J. Chem. Phys.* **2014**, *140*, 204309–11.

(62) Baruah, T.; Pederson, M. R. DFT Calculations on Charge-Transfer States of a Carotenoid-Porphyrin-C60 Molecular Triad. *J. Chem. Theory Comput.* **2009**, *5*, 834–843.

(63) Lee, D.; You, D.; Lee, D.; Li, X.; Kim, S. Machine-Learning-Guided Prediction Models of Critical Temperature of Cuprates. *J. Phys. Chem. Lett.* **2021**, *12*, 6211–6217.

(64) Zhang, Y.; Ling, C. A Strategy to Apply Machine Learning to Small Datasets in Materials Science. *npj Comput. Mater.* **2018**, *4*.

(65) Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M. Machine Learning for the Structure–Energy–Property Landscapes of Molecular Crystals. *Chem. Sci.* **2018**, *9*, 1289–1300.

TOC Graphic