Journal of Materials Chemistry C

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5tc04408a

Learning the limits: how data, diversity, and representation control machine-learning predictions of reorganisation energy

Malin Zollner,[a] Yashar Moshfeghi [b] and Tahereh Nematiaram *[a]

Accurate and scalable prediction of hole and electron reorganisation energies (λ_h and λ_e) is a persistent bottleneck in the data-driven design of organic semiconductors, as routine *ab initio* calculations remain impractical for large molecular libraries. This work presents a systematic and interpretable evaluation of how molecular representation, chemical diversity, and dataset size constrain the accuracy and transferability of machine-learning models for predicting λ_h and λ_e. Three complementary datasets are analysed: a chemically diverse benchmark of approximately 5000 molecules with paired λ_h and λ_e values, a thiophene-focused dataset comprising 1486 molecules, and a sequence of progressively augmented datasets extending to nearly 13 000 structures. Fifteen molecular descriptor schemes and twelve learning algorithms, spanning linear, kernel-based, ensemble, and graph-based models, are benchmarked under consistent training and validation protocols. Across broad chemical space, predictive performance is primarily governed by molecular representation, with hybrid descriptors that combine RDKit features and multiple molecular fingerprints consistently outperforming single-source encodings, while graph neural networks underperform in highly diverse regimes. Constraining chemical diversity leads to substantial accuracy gains, particularly for electron reorganisation energies, whereas increasing dataset size improves robustness and generalisation with rapidly diminishing returns beyond modest augmentation. Model interpretation using SHAP analysis reveals stable and physically meaningful design trends across all datasets, showing that rigid, extended π-conjugation, low conformational flexibility, and balanced charge distribution systematically reduce reorganisation energies. These results define realistic performance limits for machine-learning prediction of reorganisation energy and provide concrete guidance on representation choice, dataset design, and molecular optimisation strategies for high-mobility organic electronic materials.

Received 16th December 2025,
Accepted 10th February 2026

DOI: 10.1039/d5tc04408a

rsc.li/materials-c

1 Introduction

Molecular semiconductors have attracted considerable interest owing to their advantages over conventional inorganic counterparts, including mechanical flexibility, low cost, chemical tunability, biocompatibility, and sustainability.[1–3] These attributes enable their integration into a wide range of emerging technologies, such as organic photovoltaics,[4,5] neuromorphic devices,[6,7] and light-emitting diodes.[8,9] The performance of these devices is strongly governed by charge-transport properties; therefore, the discovery and rational design of high-mobility materials remain central to the advancement of organic electronics.[10–14]

Charge transport in molecular semiconductors is inherently complex, and a variety of theoretical frameworks have been developed to elucidate its underlying mechanisms, as reviewed in several recent articles.[15–19] A notable example is a large-scale screening study of the Cambridge Structural Database (CSD),[20] which comprises over one million entries, to estimate charge mobilities across all known molecular semiconductors.[21] This effort not only suggested a plausible upper limit for charge-carrier mobility in these materials but also identified the key parameters that dictate transport, namely reorganisation energy (λ), molecular area, transport two-dimensionality, dynamic disorder, and intermolecular electronic coupling. Among these, λ is distinctive because it can be tuned at the molecular level, whereas the other factors depend critically on crystal packing and solid-state morphology.[22–26]

The reorganisation energy, λ, represents the energetic cost associated with nuclear relaxation accompanying electron or hole transfer in a molecular system. In this work, λ is evaluated

[a] Department of Pure and Applied Chemistry, University of Strathclyde, 295 Cathedral Street, Glasgow G1 1XL, UK.
E-mail: tahereh.nematiaram@strath.ac.uk
[b] Department of Computer and Information Sciences, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, UK

using the four-point scheme originally proposed by Nelsen,[27] which explicitly incorporates structural relaxation on both the initial and final potential energy surfaces. This method is widely adopted because it provides a rigorous description of molecular reorganisation during charge transfer.

It is also worth noting that alternative approaches for estimating reorganisation energies have been reported in the literature. Among these, the vertical gradient method has attracted attention for systems that are relatively rigid, where geometric differences between electronic states are small and the harmonic approximation remains applicable.[28–30] Within this framework, the reorganisation energy is obtained from the gradient of the target electronic state evaluated at the equilibrium geometry of the reference state, together with the mass-weighted Hessian of the reference state. By eliminating the need for an additional geometry optimisation and vibrational frequency calculation, this approach offers a significant reduction in computational expense.

However, the applicability of the vertical gradient approximation is limited. Its accuracy is generally restricted to cases in which charge transfer involves only small-amplitude vibrational displacements. In molecular systems that undergo substantial structural relaxation, conformational reorganisation, or exhibit marked anharmonicity, the approximation can lead to appreciable errors.[30,31] To ensure a uniform and reliable treatment across the chemically diverse systems considered in this study, the four-point scheme is therefore employed throughout. The reorganisation energy is defined as: $\lambda = E(Q') - E(Q) + E'(Q) - E'(Q')$ with E and E' being the energies of the neutral and charged states, respectively.[27] The calculation involves evaluating these energies at two different geometries: Q, which corresponds to the optimised structure of the neutral molecule, and Q', the optimised structure of the charged species. This method applies equally to hole transfer (oxidation) and electron transfer (reduction), where the charged state corresponds to the cation or anion, respectively. Therefore, computing the reorganisation energy λ requires two separate geometry optimisations, one for the neutral state and another for the charged state, making it computationally demanding. This poses a significant challenge for large-scale screening studies, as each calculation requires substantial computational resources.[32,33] As the diversity and complexity of chemical databases continue to grow, this limitation becomes a major bottleneck in the efficient exploration of vast chemical space. Therefore, there is an urgent need for more efficient computational approaches that can accelerate the prediction of λ without compromising accuracy.

Machine learning (ML) has emerged as a powerful strategy for overcoming computational bottlenecks in materials science and molecular modelling.[34] In the context of charge transport, ML enables rapid prediction of key electronic properties, thereby facilitating large-scale screening that would otherwise be impractical using conventional electronic-structure methods. In this work, ML is employed to address the scaling limitations associated with evaluating reorganisation energies across chemically diverse molecular libraries. Whereas a single reorganisation energy computed using the four-point DFT scheme requires on average approximately 2.0×10^4 s of wall time, a trained ML model predicts the same quantity in under one second per molecule. Although the construction of the ML model entails an upfront cost associated with generating quantum-chemical reference data, this expense is incurred only once and can be efficiently amortised over large datasets. As a result, ML enables rapid prescreening of candidate molecules from large compound collections, such as those derived from the ZINC database and related repositories,[35] allowing high-level electronic-structure calculations to be focused on a small subset of the most promising systems. In this way, ML serves as a scalable filtering layer that complements, rather than replaces, accurate quantum-chemical methods.

Early applications of machine learning to organic charge transport have demonstrated promising results.[36,37] However, compared to its widespread use for predicting basic electronic descriptors such as orbital energies or bandgaps, the application of ML to reorganisation energies remains relatively underdeveloped. Existing proof-of-concept studies establish the feasibility of learning λ, but also reveal several outstanding challenges.[38–40] One such challenge is the scarcity of large, chemically diverse, and unbiased datasets of reorganisation energies. Consequently, most reported models are trained on small collections confined to narrow chemical domains, which limits their robustness and transferability.[41,42]

A further limitation concerns the molecular representation strategies used to encode chemical information. Most prior studies rely on a single class of descriptors, typically structural features or molecular fingerprints, without systematically exploring alternative or complementary representations.[42–45] This restricted scope can constrain model expressiveness and adversely affect predictive accuracy. In addition, the existing literature exhibits a pronounced imbalance in its focus on charge-carrier type, with the majority of studies addressing only hole transport through cationic reorganisation energies.[38,39,42,46–49] This bias, rooted in the historically held assumption that electron transport is intrinsically less efficient in organic semiconductors,[50,51] has resulted in comparatively limited data for anionic reorganisation energies and has hindered the development of comprehensive and transferable predictive models.

Here, we address these gaps by developing a comprehensive and interpretable ML framework for predicting both hole and electron reorganisation energies in organic semiconductors. This study advances the field in three ways: (i) we introduce a balanced dataset containing both λ_h and λ_e values across thousands of molecules; (ii) we systematically evaluate the trade-off between chemical diversity and dataset size using three complementary datasets (chemically diverse, chemically focused, and size-augmented); and (iii) we combine extensive benchmarking of molecular representations and algorithms with SHAP-based interpretation[52] to extract physically meaningful design rules. These contributions establish a unified, scalable, and interpretable workflow for accelerated λ prediction, and provide practical guidelines for high-throughput discovery of organic electronic materials.

2 Methodology

The overall ML pipeline developed in this work consists of three main stages: (i) establishing baseline performance across different molecular representations and ML algorithms using a chemically diverse dataset (dataset 1), (ii) evaluating the impact of reduced chemical diversity with a thiophene-focused dataset (dataset 2), and (iii) assessing the effect of dataset size with progressively larger subsets (dataset 3). The following subsections describe the construction of the dataset, feature engineering, model selection, and performance evaluation in detail.

2.1 Dataset

Three datasets with varying chemical scope and size were used.

Dataset 1. Approximately 5000 molecular semiconductors were obtained from ref. 21, in which the hole reorganisation energies (λ_h) had already been computed and deposited in the CSD. To extend this dataset, we calculated the corresponding electron reorganisation energies (λ_e) at the B3LYP/3-21G* level of theory using Gaussian 16,[53] thereby ensuring methodological consistency across charge-carrier types. It is notable that a sample test confirms the consistency between results obtained at the B3LYP/3-21G* and B3LYP/6-31G* levels of theory. Molecular selection criteria originally applied by Padula *et al.*[54] excluded cocrystals, polymers, disordered solids, duplicate entries, and molecules containing more than 100 heavy atoms, while further restricting the HOMO–LUMO gap to the range of 2–4 eV. The resulting dataset is chemically diverse, balanced with respect to hole and electron transport, and provides a robust foundation for machine-learning model development.

Dataset 2. To investigate the effect of reduced chemical diversity, we curated a thiophene-focused dataset containing 1486 unique molecules. These were compiled from three sources: approximately 1200 thiophene-containing structures from Atahan-Evrenk *et al.* (2019),[38] around 250 thiophene monomers from Abarbanel *et al.* (2021),[47] and roughly 90 thiophene-based entries from the CSD identified through substructure searches. After deduplication, all SMILES strings were converted into 3D geometries using RDKit, and both hole (λ_h) and electron (λ_e) reorganisation energies were calculated at the B3LYP/3-21G* level of theory. This dataset emphasizes a narrower chemical space compared to dataset 1, enabling a direct evaluation of how reduced structural diversity influences ML model performance.

Dataset 3. To assess the impact of dataset size on model performance, we began with the ~5000 molecules in dataset 1 as a fixed core. From the larger and chemically diverse dataset of Yang *et al.* (2024),[55] we randomly sampled additional subsets of 500, 1000, 2000, 4000, and 8000 molecules. These were combined with the original core database (*i.e.*, 5000 CSD molecules) to create five extended datasets: dataset 3_500, 3_1000, 3_2000, 3_4000, and 3_8000, respectively. In this way, each dataset is progressively larger but still anchored by the same ~5000-molecule foundation. This controlled expansion allows us to systematically evaluate how increasing dataset size influences prediction accuracy, robustness, and generalisation across broader chemical space. Also, in all steps, to quantify the dataset's chemical diversity, we calculated pairwise Tversky similarity coefficients[56] using the RDKit cheminformatics library. The Tversky coefficient is a generalisation of the more familiar Tanimoto (Jaccard) similarity and it compares two molecular fingerprints by measuring how many features they share relative to the features unique to each molecule. As such, lower similarity values indicate greater structural uniqueness, whereas higher values reflect more closely related molecules.

Before proceeding, we clarify a methodological aspect concerning the calculation of reorganisation energies used throughout this work. It is well established that the choice of exchange–correlation functional represents a significant source of uncertainty in computed reorganisation energies, as different functionals can introduce systematic shifts in the absolute magnitude of λ due to differences in their treatment of exchange, self-interaction error, and charge localisation.[57] In particular, long-range corrected hybrid functionals often yield larger reorganisation energies, whereas conventional hybrid functionals tend to produce smaller values.

In this study, all reference reorganisation energies were computed consistently at the B3LYP level of theory. B3LYP represents a widely adopted and well-characterised compromise between accuracy and computational cost for organic and π-conjugated molecular systems and has been extensively employed in prior studies of charge transport and reorganisation energies.[58–62] Importantly, while B3LYP may not yield quantitatively optimal absolute λ values for every molecular class, it has been shown to provide physically reasonable trends and good qualitative agreement with experimental observations across a broad range of molecular semiconductors.[63–65]

Consequently, the machine-learning models developed here are explicitly trained to reproduce B3LYP-level reorganisation energies rather than to provide functional-independent absolute predictions. Any systematic bias associated with the chosen functional is therefore consistently embedded in both the training data and the target property. This design choice is well aligned with the primary objective of high-throughput screening, where internal consistency, reliable relative trends, and robust molecular ranking within a fixed theoretical framework are of greater importance than absolute accuracy.

Finally, we note that the size and chemical diversity of the present dataset naturally enable future extensions based on multi-fidelity[66] or transfer-learning strategies,[67] in which a smaller subset of reorganisation energies computed with higher-level or alternative functionals could be used to correct systematic offsets and improve absolute accuracy, while preserving the computational scalability of the approach.

2.2 Feature selection

To evaluate the influence of input data quantity and complexity on ML model performance, we examined 15 distinct descriptor sets, summarised in Table 1. These sets encompassed curated physicochemical descriptors, RDKit-calculated descriptors, structural fingerprints, and hybrid combinations.

Table 1 Summary of feature sets used

Set name	Feature length	Features
1.1	13	13 manually selected descriptors
1.2	2059	13 manually selected descriptors & Daylight fingerprints
1.3	141	13 manually selected descriptors & MACCS fingerprints
1.4	2045	13 manually selected descriptors & Morgan fingerprints
1.5	4201	13 manually selected descriptors & Daylight – & MACCS – & Morgan fingerprints
2.1	147	RDKit descriptors
2.2	2193	RDKit descriptors & Daylight fingerprints
2.3	268	RDKit descriptors & MACCS fingerprints
2.4	2167	RDKit descriptors & Morgan fingerprints
2.5	4319	RDKit descriptors & Daylight – & MACCS – & Morgan fingerprints
3.1	2046	Daylight fingerprints
3.2	128	MACCS fingerprints
3.3	2032	Morgan fingerprints
3.4	4188	Daylight & MACCS & Morgan fingerprints
4.1	154	13 manually selected descriptors & RDKit descriptors

Descriptor sets 1.1 to 1.5 consisted of manually selected physicochemical features informed by prior studies on charge transport in organic semiconductors.[36,48] These included molecular weight, bond-related properties such as the number of rotatable or conjugated bonds, atom-related properties such as heteroatom counts and hybridisation states, and ring-related features such as the number of aromatic and non-aromatic rings. These descriptors capture chemically interpretable structural attributes that are directly relevant to charge transport.

Descriptor sets 2.1 to 2.5 were derived from SMILES strings using the RDKit library.[68] This collection represents a broad set of computationally accessible features, including electronic state indices, charge distribution, topological indices, molecular surface areas, and fragment-based representations. For example, the topological polar surface area quantifies the accessible polar surface area, while indices such as SlogP_VSA and SMR_VSA describe contributions of hydrophobicity and refractivity, respectively. These descriptors provide a comprehensive numerical characterisation of molecular properties.

In addition to descriptors, we employed molecular fingerprints, which convert structural information into machine-readable bit vectors suitable for similarity comparisons and predictive modelling. Descriptor sets 3.1 to 3.3 consist of individual fingerprints, including Morgan (circular), Daylight (path-based), and MACCS keys (predefined motifs), while set 3.4 combines all three.[69] Morgan fingerprints capture circular substructures of varying radii around each atom, producing chemically consistent encodings of local environments. Daylight fingerprints encode all possible bond paths up to a defined length, making them particularly useful for substructure searches. MACCS keys rely on a fixed library of structural fragments, offering compact and interpretable representations that are well suited for rapid screening. Each approach offers complementary advantages, and their integration produces a more diverse and informative molecular representation.[38,42,46,47,70,71]

Finally, descriptor set 4.1 combines the curated physicochemical features with RDKit-calculated descriptors to explore whether merging chemically interpretable and automatically generated descriptors improves predictive performance. Altogether, these strategies produced 15 distinct descriptor sets that span both low-dimensional interpretable features and high-dimensional machine-generated encodings.

To mitigate redundancy, we applied pairwise Pearson correlation analysis to all descriptor sets listed in Table 1. Features with absolute correlation coefficients greater than 0.8 were considered highly collinear, and only one representative feature from each correlated pair was retained. The exception was set 1.1, where all descriptors were deliberately retained for their interpretability. It is important to note that all descriptors employed in this study can be computed directly from SMILES strings and therefore do not require any quantum-chemical calculations. This ensures that the workflow remains efficient and scalable, making it well suited for high-throughput screening applications.

2.3 Machine learning model selection and optimisation

To identify the most suitable algorithms for predicting hole and electron reorganisation energies, we evaluated twelve ML models spanning diverse learning paradigms. These included tree-based ensembles, kernel-based methods, neural networks, graph-based architectures, and simple baselines.

Tree-based methods include adaptive boosting (AdaBoost),[72] gradient boosting regression trees (GBRT),[73] light gradient boosting machine (LightGBM),[74] random forest (RF)[75] and extreme gradient boosting (XGBoost).[76] Such models are widely used in molecular property prediction for their ability to capture non-linear relationships and feature interactions.[41,42,47,71] RF reduces overfitting by averaging across multiple decision trees but is computationally demanding. LightGBM improves efficiency through leaf-wise tree growth, making it scalable to large datasets. XGBoost incorporates regularisation to improve generalisation, while GBRT is particularly effective for smaller datasets, albeit at a higher computational cost. AdaBoost improves weak learners through iterative refinement but is sensitive to noisy data. Kernel-based approaches included support vector machines (SVM) with linear and Gaussian kernels.[77] SVMs are well suited for high-dimensional descriptor spaces, though their computational cost grows with dataset size. We also evaluated multi-layer perceptron (MLP) regressors,[78–81] which capture complex non-linear relationships but require careful regularisation to avoid overfitting. For baseline comparison, we included linear regression (LR)[82] and K-nearest neighbours (KNN).[83] LR is computationally efficient and interpretable but poorly suited to non-linear trends, whereas KNN performs well on small datasets but scales poorly with size.

Given the promising results of recent studies,[40,55] we further incorporated graph neural networks (GNNs), which directly operate on molecular graphs where atoms are represented as nodes and bonds as edges. Two widely adopted architectures were implemented using PyTorch geometric,[84] namely the graph convolutional network (GCN),[85] which propagates structural information through spectral convolutions, and the graph

attention network (GAT),[86] which introduces learnable attention weights to prioritise chemically relevant interactions. GNNs capture hierarchical, context-specific features of molecular structures that extend beyond conventional descriptor- or fingerprint-based representations. Finally, we included a deep neural network (DNN) model previously employed by Atahan-Evrenk *et al.*,[38] which demonstrated strong performance for predicting hole reorganisation energies in smaller datasets.

For all models, datasets were split into 80% training and 20% testing. Hyperparameters were optimised using Bayesian optimisation with five-fold cross-validation, minimising the average RMSE across hole (λ_h) and electron (λ_e) reorganisation energies. Key optimised parameters included γ, the regularisation parameter in SVM and XGBoost that controls sample influence; α, the L^2 regularisation term in neural networks that prevents overfitting; and C, the SVM regularisation parameter that balances model complexity and accuracy. A comprehensive overview of the tuned hyperparameters and their ranges is provided in Table S1 of the SI.

The predictive performance of each model–descriptor combination was evaluated using standard regression metrics, namely the coefficient of determination (R^2), the root mean square error (RMSE), and the mean absolute error (MAE). The R^2 value quantifies the fraction of variance in the reorganisation energy explained by the model, while RMSE and MAE characterise the typical magnitude of prediction errors by penalising large deviations and averaging absolute deviations, respectively. These metrics provide a balanced assessment of model performance by capturing both variance reproduction and average deviation from reference values in physically meaningful units. Their use also enables direct comparison with previous machine-learning studies of reorganisation energies in organic semiconductors.[38,40,87]

3 Results and discussion

This section reports the outcomes of our machine learning framework for predicting hole and electron reorganisation energies in molecular semiconductors. Guided by Section 2, we proceed in three steps. First, we evaluate how molecular representation and feature dimensionality affect accuracy on a chemically diverse benchmark (dataset 1). Second, we examine the impact of chemical diversity by retraining the best algorithms on a thiophene-focused set (dataset 2). Third, we assess how increasing dataset size influences accuracy and robustness using progressively augmented subsets (dataset 3). These analyses identify the dominant factors that govern model performance and generalisability.

3.1 Feature impact

We first assess how feature representation and dimensionality influence predictive performance using dataset 1, which provides a diverse benchmark for both electron and hole targets. Fig. 1a shows the distributions of reorganisation energies. Hole values span 0.028–0.864 eV (median 0.162 eV; standard deviation 0.105 eV), and electron values span 0.047–0.890 eV (median 0.217 eV; standard deviation 0.113 eV), confirming substantial electronic variety. Structural diversity, quantified *via* per-molecule average Tversky similarities using MACCS fingerprints (Fig. 1b), ranges from 0.060 to 0.704 (median 0.463; standard deviation 0.103), indicating limited redundancy and broad coverage of chemical space.

All fifteen descriptor sets in Table 1 were evaluated across multiple algorithms with Bayesian hyperparameter optimisation. Full results are reported in Tables S2 and S3 in the SI.

For electrons, the best model is GBRT with feature set 2.5 (R^2 = 0.426, RMSE = 0.081 eV, MAE = 0.053 eV). For holes, the best model is SVM with a Gaussian kernel and feature set 2.5 (R^2 = 0.372, RMSE = 0.082 eV, MAE = 0.050 eV). Linear SVMs perform poorly (maximum R^2 of 0.201 for holes and 0.270 for electrons; RMSE $>$ 0.092 eV and MAE $>$ 0.059 eV), indicating strongly non-linear structure–property relationships. GNNs underperform on dataset 1 (best R^2 = 0.072 for holes and 0.173 for electrons; RMSE = 0.090 and 0.158 eV; MAE = 0.72 and 0.73 eV), suggesting that in broad chemical spaces, information-rich engineered representations remain more effective than end-to-end graph features.

Fig. 2 shows performance as a function of feature length for both electron and hole reorganisation energy prediction in dataset 1. For electrons, median R^2 increases from 0.196 at feature length 13 to 0.361 at 4319, while the best R^2 improves from 0.222 to 0.426. Median RMSE decreases from 0.097 eV to 0.087 eV, and the best RMSE improves from 0.095 eV to 0.081 eV. Consistent with this trend, the median MAE decreases from 0.066 eV to 0.056 eV with the best MAE improving from 0.065 eV

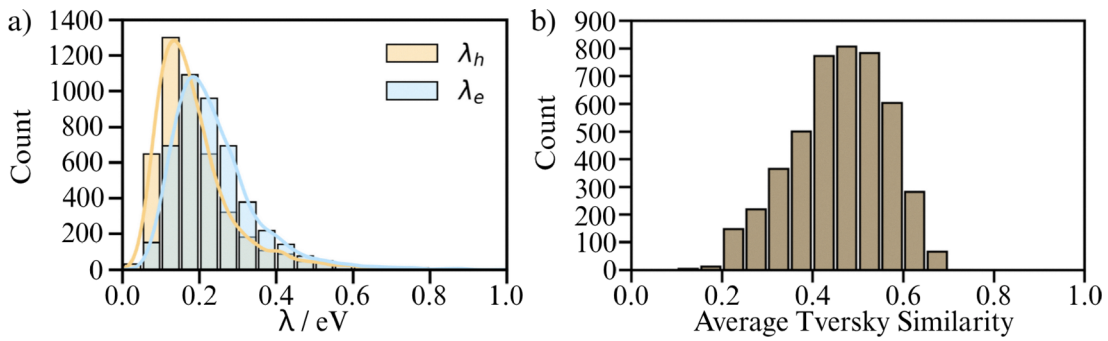

Fig. 1 Distributions of (a) electron (λ_e) and hole (λ_h) reorganisation energies and (b) average Tversky similarity in dataset 1.

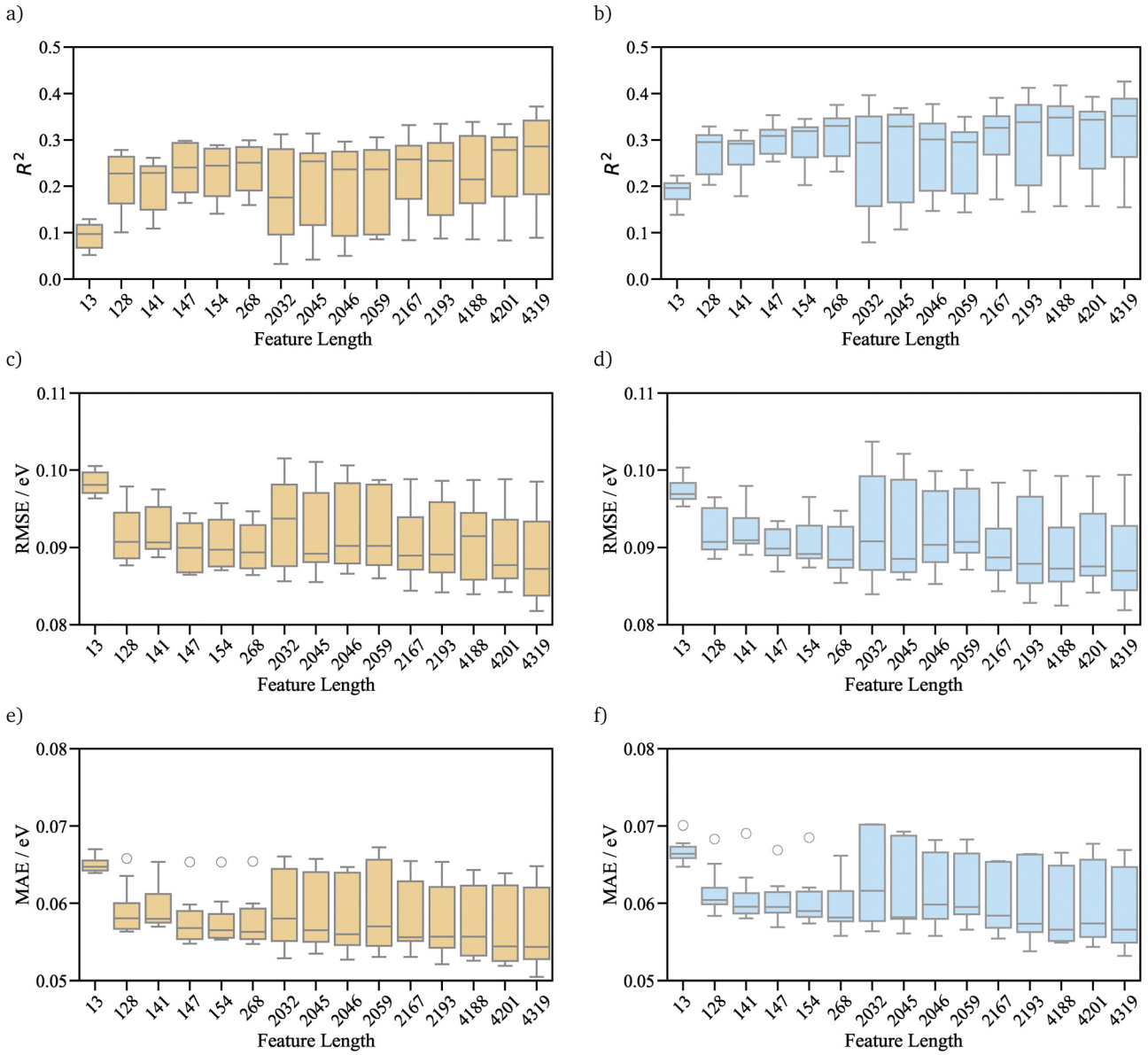

Fig. 2 Performance as a function of feature length for predicting electron (panels a, c, e) and hole (panels b, d, f) reorganisation energies in dataset 1. Panels (a) and (b) report R^2, panels (c) and (d) report RMSE, and panels (e) and (f) report MAE across descriptor sets and learning algorithms.

to 0.053 eV. For holes, median R^2 increases from 0.097 at length 13 to 0.286 at 4319, with the best R^2 reaching 0.129 and 0.372, respectively. Median RMSE decreases from 0.098 eV to 0.087 eV, and the best RMSE improves from 0.096 eV to 0.082 eV. In parallel, the median MAE decreases from 0.064 eV to 0.054 eV and the best MAE from 0.064 eV to 0.050 eV indicating a consistent improvement in absolute prediction accuracy with increasing feature length.

In both cases, models trained on hybrid, high-dimensional encodings that combine multiple fingerprints with RDKit descriptors (*e.g.*, sets 2.5, 3.4, 2.2) consistently outperform compact curated sets or single-source fingerprints. Overall, descriptor richness and complementarity are key to achieving accurate predictions in chemically diverse regimes.

3.2 Diversity impact

We next examined how dataset diversity influences performance by retraining the five best algorithms (see SI) from Section 3.1 (LightGBM, GBRT, RF, SVM with Gaussian kernel, XGBoost) on the thiophene-focused dataset 2 across all descriptor sets. Fig. 3a shows the target distributions; similar to dataset 1, λ_h in dataset 2 range from 0.039 eV to 0.882 eV, with a median of 0.152 eV and a standard deviation of 0.101 eV. λ_e span from 0.090 eV to 0.879 eV, with a median of 0.158 eV and a standard deviation of 0.093 eV. Although the energy ranges are broadly comparable to dataset 1, the structural diversity is markedly lower, as evidenced by higher average Tversky similarities. Values range from 0.361 to 0.816, with a median of 0.745 and a standard deviation of 0.112.

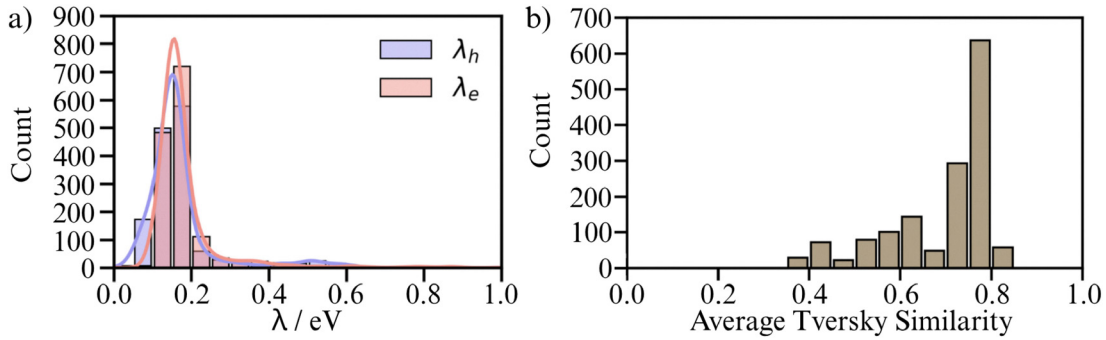

Fig. 3 Distributions of (a) electron (λ_e) and hole (λ_h) reorganisation energies and (b) average Tversky similarity in dataset 2.

This represents a higher median and broader range than dataset 1 (0.463 median; 0.060–0.704 range), indicating increased overall molecular similarity.

Performance improves across the board on dataset 2, particularly for electrons. The best electron model, GBRT with feature set 1.5, reaches R^2 = 0.757 and XGBoost with feature set 4.1 attains the lowest RMSE and MAE of 0.040 eV and 0.016 eV, respectively. For holes, GBRT with feature set 1.5 yields R^2 = 0.508, while the best RMSE is 0.077 eV using XGBoost with feature set 2.4. The minimum MAE for holes is achieved by the RF model with feature set 2.5, reaching 0.031 eV. Fig. 4 summarises the performance improvements relative to dataset 1, with full results in Tables S4 and S5 in the SI. Feature sets that combine hand-picked features with multiple fingerprints remain most effective. Interestingly, in this narrower chemical domain, hand-picked descriptors can outperform RDKit-heavy sets, underscoring that optimal representations depend on the target chemical space.

These results demonstrate that limiting chemical diversity can substantially improve accuracy, even when fewer training samples are available. The associated trade-off, however, is reduced generalisability, which is often essential for the discovery of new chemistries. We therefore explore increasing data quantity as a complementary strategy to enhance performance while retaining chemical diversity.

3.3 Impact of dataset quantity

The previous section demonstrated that reducing chemical diversity can substantially improve predictive performance, though at the expense of generalisability. An alternative strategy is to expand dataset size, thereby increasing both statistical robustness and the coverage of structural motifs. To investigate this, we constructed dataset 3 by systematically augmenting dataset 1 with randomly sampled molecules from the recent Yang *et al.* dataset.[55] This design allowed us to probe whether

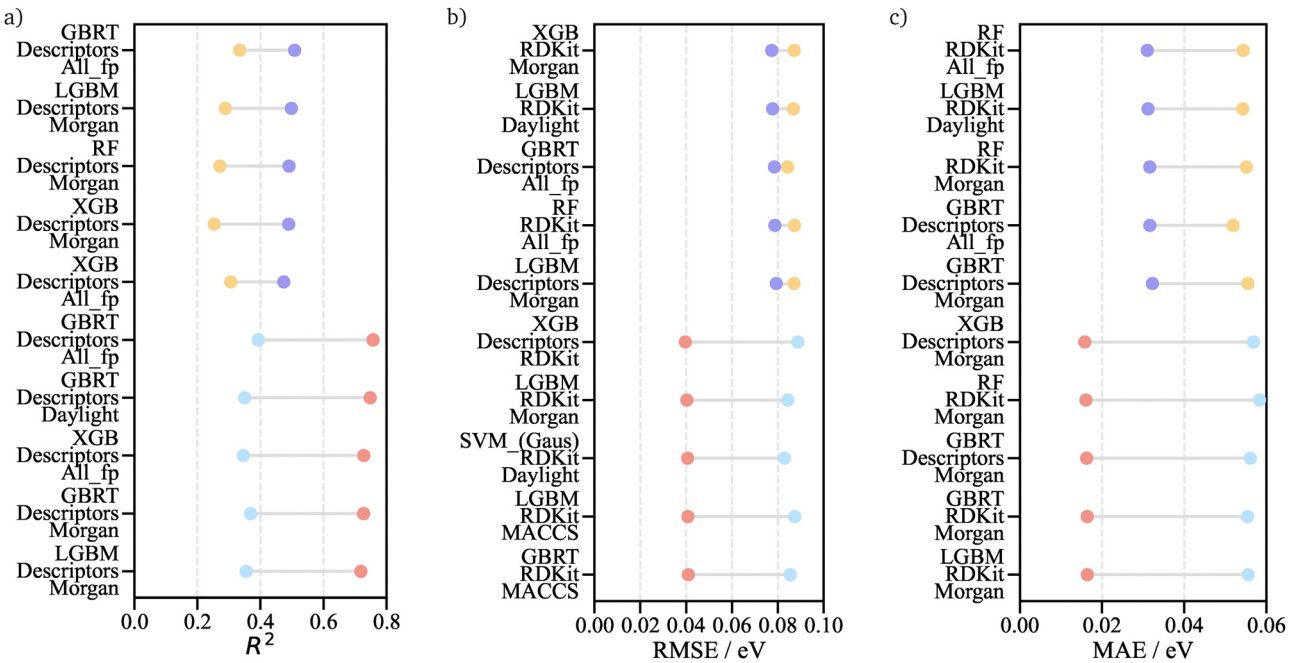

Fig. 4 (a) R^2, (b) RMSE and (c) MAE of the top five models for electrons and holes trained on dataset 2 compared to their counterparts on dataset 1. Yellow: dataset 1 holes; purple: dataset 2 holes; blue: dataset 1 electrons; red: dataset 2 electrons.

larger and more diverse datasets enhance predictive accuracy, and to evaluate whether improvements scale linearly with dataset size.

Fig. 5a illustrates the distribution of hole reorganisation energies across the dataset 3 subsets. Dataset 1 exhibits a pronounced skew toward low values (0.0–0.4 eV), whereas the Yang dataset is biased toward higher energies. By progressively adding random samples from the Yang dataset to dataset 1, we obtained augmented subsets (dataset 3_500, 3_1000, 3_2000, 3_4000, and 3_8000) with increasingly balanced and diverse distributions. The median reorganisation energy rose steadily from 0.162 eV in dataset 1 to 0.392 eV in dataset 3_8000, while the standard deviation increased from 0.105 eV to 0.229 eV, reflecting the growing heterogeneity of the training set. The structural diversity of the datasets is illustrated in Fig. 5b. The similarity values lie between 0.021 and 0.680, with median values for dataset 3_500 to dataset 3_8000 ranging from 0.456 to 0.413. The standard deviation increases from 0.067 to 0.108 across these subsets, indicating slightly broader structural variation. Overall, the dataset is well-suited for isolating the effect of data quantity, as molecular similarity remains consistent and the chemical space retains sufficient diversity.

To benchmark predictive accuracy, we applied the best-performing models and feature sets identified in Section 3.1, namely SVM with a Gaussian kernel, LightGBM, XGBoost, and GBRT, combined with feature sets 1.5, 2.2, 2.5, and 3.4. In all cases, the subset augmented with 8000 additional datapoints (dataset 3_8000) achieved the highest R^2, reaching approximately 0.79. However, improvements were not linear with dataset size. As shown in Fig. 6, the most significant prediction improvement occurred with relatively modest augmentation: adding 500 molecules increased the best model's R^2 from about 0.33 to 0.61. Beyond this point, performance improvement diminished, with the increase from 4000 to 8000 molecules raising R^2 only marginally (from 0.78 to 0.79). RMSE values displayed a complementary pattern, rising from 0.082 eV in dataset 1 to 0.11 eV after augmentation, then stabilising at 0.106 eV for larger subsets. MAE follows a similar saturation behaviour, increasing from an average of 0.062 eV at 500 molecules to 0.069 eV at 1000 and then rising more gradually to 0.075, 0.078 and 0.080 eV at 2000, 4000 and 8000 molecules respectively. This indicates that additional datapoints substantially improved R^2 at small scales and did not further significantly change RMSE or MAE once the dataset reached sufficient size. These results highlight two key insights. First, dataset expansion improves predictive accuracy not only by increasing the number of training examples, but also by reducing the structural and statistical biases inherent to smaller, curated datasets. In particular, the inclusion of molecules with higher reorganisation energies introduces under-represented structural motifs, enabling models to capture a broader range of structure–property relationships. Second, the performance improvement exhibits diminishing returns once the dataset reaches sufficient size and diversity, such that further additions yield only marginal improvements. This suggests that targeted augmentation with under-represented chemistries or structural outliers may be more efficient than indiscriminate expansion.

3.4 SHAP analysis

Predictive accuracy alone does not explain why particular molecules exhibit high or low reorganisation energies. To extract chemically meaningful insight from our best-performing models,

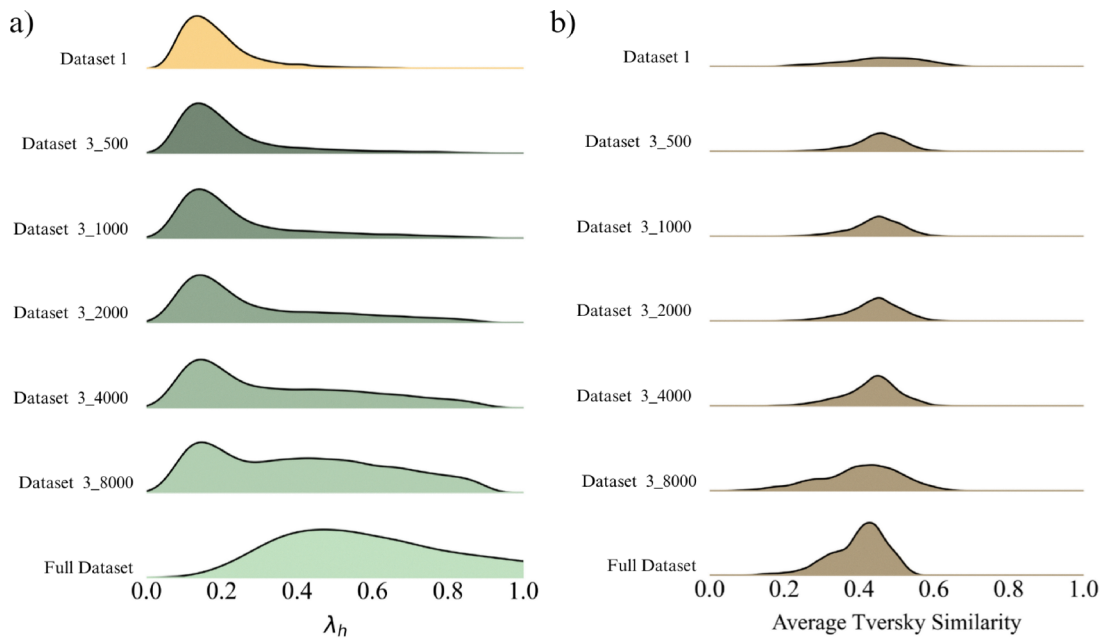

Fig. 5 Distribution of (a) hole reorganisation energies (λ_h) and (b) average Tversky similarity for subsets of dataset 3. Dataset 1 and the full dataset from Yang *et al.* (2024)[55] are shown for reference. Dataset 3_500 corresponds to dataset 1 enhanced with 500 randomly sampled datapoints from the Yang dataset, dataset 3_1000 adds a further 500 datapoints, and so on.

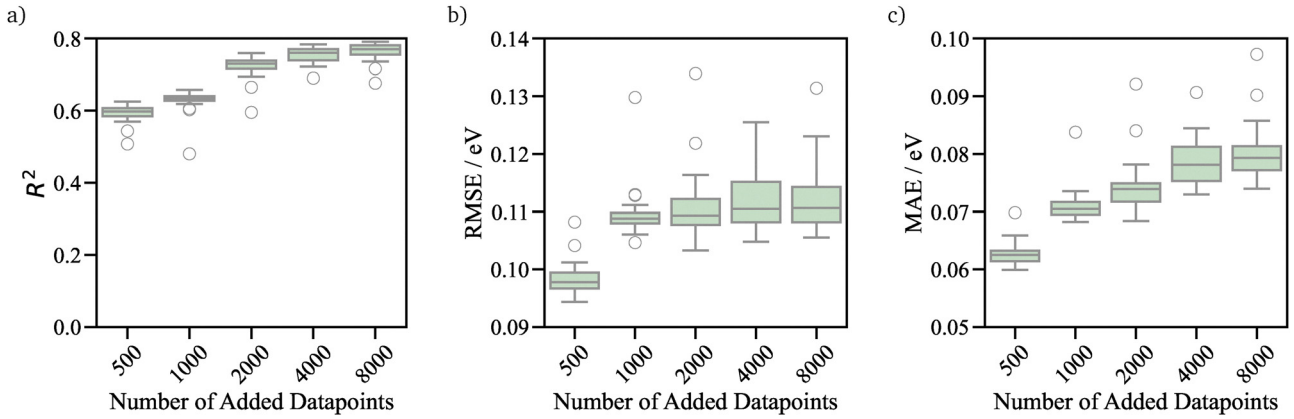

Fig. 6 Change in (a) R^2, (b) RMSE and (c) MAE with increasing dataset size in dataset 3.

we therefore apply *post hoc* interpretability analysis. Specifically, SHAP (SHapley additive exPlanations) is used to quantify the contributions of individual molecular descriptors to the predicted reorganisation energies and to relate these statistically learned effects to established structure–property relationships. Many of the most influential features are RDKit-derived proxy descriptors, including fragment-based logP surface-area bins and empirical partial-charge measures. Accordingly, the interpretations presented below should be understood as statistically inferred structure–property correlations that are broadly consistent with chemical intuition, rather than as direct causal attributions.

The SHAP results are displayed as beeswarm summary plots, in which each point represents an individual molecule in the dataset. For a given descriptor, the horizontal position of a point corresponds to its SHAP value and indicates how that descriptor influences the predicted reorganisation energy for that molecule, *i.e.* points to the right of zero increase the prediction, whereas points to the left decrease it. The overall horizontal extent of the point distribution, therefore, reflects the relative importance of each descriptor across the dataset. Point colour indicates the value of the descriptor for each molecule, ranging from low (blue) to high (red). Colour does not convey importance; rather, it shows whether low or high values of a descriptor are associated with increases or decreases in the predicted reorganisation energy. In this way, the combined position and colour of the points reveal how the model systematically uses each descriptor.

For electron reorganisation energies in dataset 1 (Fig. 7a), the most influential descriptor is SlogP_VSA8. This descriptor quantifies the total van der Waals surface area of atoms whose Wildman–Crippen atomic logP contributions fall within the numerical range defining bin 8. As defined by RDKit, this bin captures atoms with moderately positive hydrophobic contributions arising from their local chemical environments (Fig. S1). High SlogP_VSA8 values, therefore, indicate molecules with substantial weakly polar or non-polar surface area, as encoded by the descriptor representation, and are statistically associated in this dataset with molecular structures exhibiting reduced strongly hydrophilic character. The SHAP analysis reveals that higher values of SlogP_VSA8 are consistently associated, on average, with negative SHAP contributions to λ_e, indicating that molecules encoded as having larger weakly polar surface regions are predicted to exhibit lower electron reorganisation energies.

Several complementary descriptors related to charge proxies contribute in a consistent manner. The minimum absolute

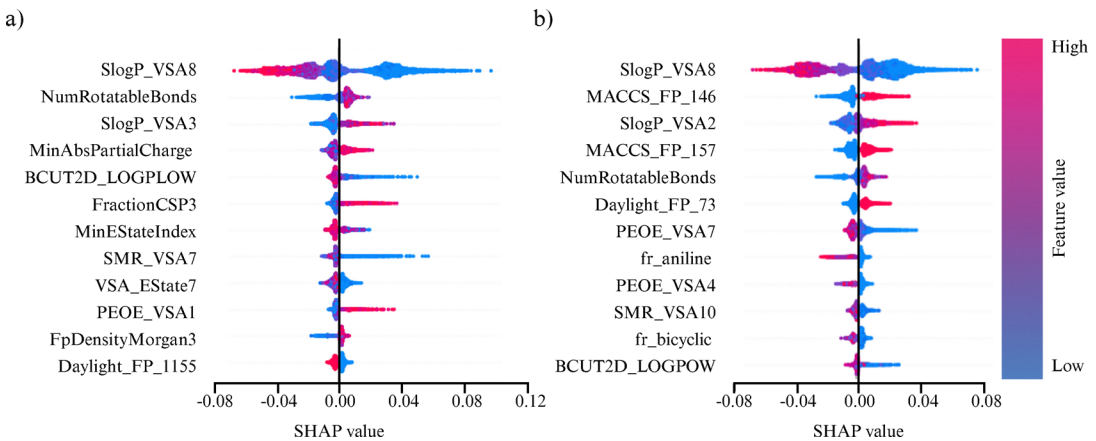

Fig. 7 SHAP analysis for (a) electron and (b) hole reorganisation energy prediction with LightGBM and descriptor set 2.5 for dataset 1.

partial charge, which reflects the smallest magnitude of RDKit-assigned atomic partial charge present within a molecule, shows a positive SHAP association with λ_e. This indicates that molecules in which even the least charged atoms carry larger partial charges, reflecting a more globally polarised charge distribution in the descriptor representation, tend to be assigned higher reorganisation energies by the model. In contrast, BCUT2D_LOGPLOW shows a predominantly negative association with λ_e and higher values tend to contribute negative (or near-zero) SHAP values, whereas lower values more often contribute positively to the predicted λ_e. In addition, PEOE_VSA1, which quantifies the van der Waals surface area associated with atoms bearing low empirical partial charges, also correlates positively with λ_e. Overall, these features indicate that several polarity- and charge-proxy descriptors (*e.g.*, MinAbsPartialCharge and PEOE_VSA1) are associated with increased predicted λ_e, while BCUT2D_LOGPLOW exhibits a distinct, predominantly negative association in this dataset.

The number of rotatable bonds emerges as the second most influential feature and shows a positive association with λ_e. As a descriptor-level measure of molecular flexibility, higher values indicate increased conformational freedom. Within the model, increased flexibility is associated with higher predicted reorganisation energies, consistent with the broader expectation that flexible molecules are more likely to undergo larger geometry changes between charge states.

The third-ranked descriptor, SlogP_VSA3, captures the van der Waals surface area of atoms with atomic logP contributions between −0.2 and 0.0. This range corresponds to weakly polar or near-neutral atomic environments, which frequently occur in aliphatic or non-conjugated regions and in saturated fragments adjacent to heteroatoms (Fig. S2). Increasing SlogP_VSA3 therefore reflects a growing contribution from weakly polar atomic environments in the descriptor representation, and the predominantly positive SHAP association indicates that these environments tend to be linked to increased predicted λ_e. Consistently, FractionCSP3, which measures the proportion of sp^3-hybridised carbons, also correlates positively with λ_e. In contrast, higher values of SMR_VSA7, which quantifies the van der Waals surface area of atoms with moderate molar refractivity (Fig. S3), is associated with reduced predicted λ_e.

Finally, the MinEStateIndex, representing the minimum electrotopological state value within a molecule, exhibits a negative SHAP association with λ_e. This indicates that molecules characterised by lower minimum electrotopological state values in the descriptor representation tend to be assigned lower reorganisation energies by the model. In agreement with this trend, VSA_EState7, which combines van der Waals surface area with electrotopological state information, also correlates negatively with λ_e. Therefore, these descriptors highlight consistent statistical associations between electrotopological state patterns and predicted reorganisation energies.

A comparable SHAP analysis for hole reorganisation energies, λ_h, in dataset 1 (Fig. 7b) reveals broadly similar qualitative trends, although with differences in relative feature importance. SlogP_VSA8 again emerges as the most influential descriptor, indicating that weakly polar atomic environments, as encoded by the descriptor, are consistently associated with lower predicted λ_h values. This pattern is further supported by SlogP_VSA2, which quantifies the van der Waals surface area of atoms with more negative atomic logP contributions and therefore encompasses a range of more polar atomic environments (Fig. S4). Higher SlogP_VSA2 values correlate positively with λ_h, indicating that more polar atomic environments in the descriptor representation are associated with increased hole reorganisation energies.

Descriptors associated with moderately negative empirical partial charge environments, including PEOE_VSA7 and PEOE_VSA4, further refine this picture. Their SHAP contributions indicate that a moderate and distributed degree of negative electrostatic character within the descriptor space is associated with reduced predicted λ_h, whereas stronger localisation is not.

Several additional binary molecular features also influence the predicted hole reorganisation energy, but they do so in opposite directions. The MACCS fingerprint bits MACCS_FP_146 and MACCS_FP_157 are binary indicators that encode the presence or absence of specific predefined substructural patterns within the MACCS fingerprint scheme. In the SHAP analysis, molecules for which these bits are present exhibit predominantly positive SHAP values, indicating that the corresponding encoded substructures are statistically associated with higher predicted λ_h. In contrast, RDKit fragment-count descriptors such as fr_aniline, which encode the presence of an aniline-type functional group, show predominantly negative SHAP values. This indicates that molecules containing such fragments are, on average, assigned lower predicted λ_h by the model. These opposing trends highlight that discrete substructural motifs captured by different molecular representations can contribute to λ_h in qualitatively different ways within the learned model. A related negative association is observed for SMR_VSA10, which quantifies the van der Waals surface area associated with atoms of high molar refractivity (Fig. S5). Higher values of this descriptor correlate with negative SHAP contributions to λ_h, indicating that molecular environments characterised by highly refractive atomic surface area are statistically associated with lower predicted hole reorganisation energies. Consistent with this pattern, fragment-based descriptors corresponding to aromatic motifs, such as fr_aniline, also exhibit negative SHAP contributions, whereas the contribution from fr_bicyclic is weaker and centred closer to zero. These results demonstrate that the model assigns distinct statistical contributions to different discrete molecular features, with some encoded substructures associated with increased predicted λ_h and others associated with reduced values.

Overall, the SHAP analysis indicates that, within the RDKit descriptor and fingerprint representations used here, lower predicted electron and hole reorganisation energies are statistically associated with molecular features characteristic of lower polarity, reduced saturation, and greater structural rigidity. These qualitative trends are consistently observed across reduced-diversity and size-augmented datasets (Fig. S6), with variations in feature ranking primarily reflecting differences in

dataset composition rather than changes in the underlying learned structure–property relationships encoded by the model.

4 Conclusions

This study establishes clear and realistic boundaries for the machine-learning prediction of hole and electron reorganisation energies in organic semiconductors. By systematically disentangling the effects of molecular representation, chemical diversity, and dataset size across multiple complementary datasets, we show that predictive performance is governed less by algorithmic sophistication than by how chemical information is encoded and sampled.

Across chemically diverse molecular space, molecular representation emerges as the primary limiting factor. Hybrid feature sets that combine RDKit-derived descriptors with multiple molecular fingerprints consistently outperform single-source representations, whereas graph neural networks fail to deliver competitive accuracy under conditions of high chemical diversity. This result underscores that, for reorganisation energy prediction, carefully engineered representations remain more effective than end-to-end graph learning when data are heterogeneous and limited in size.

The analysis further reveals a fundamental trade-off between accuracy and transferability that is controlled by dataset composition. Restricting chemical diversity, as demonstrated with thiophene-focused datasets, leads to substantial accuracy gains, particularly for electron reorganisation energies, but at the expense of general applicability. In contrast, expanding dataset size while preserving chemical diversity improves robustness and generalisation, although performance gains saturate rapidly once key structural motifs are sufficiently represented. These findings indicate that indiscriminate data growth is inefficient and that targeted augmentation of under-represented chemistries offers a more effective route to performance improvement.

Importantly, model interpretability confirms that the machine-learning models capture physically meaningful and transferable structure–property relationships. Across all datasets and both charge carriers, rigid and extended π-conjugated architectures consistently minimise reorganisation energies, while molecular flexibility, aliphatic content, and localised charge distributions increase them. Subtle but systematic differences between carriers are observed, with electron transport showing greater sensitivity to charge-distribution descriptors and hole transport being more strongly governed by aromaticity-related features. These trends are fully consistent with established physical understanding of charge transport in molecular semiconductors.

Overall, this work provides a coherent and practical framework for the data-driven prediction of reorganisation energies. It clarifies when machine learning can be expected to succeed, when its limitations are intrinsic rather than technical, and how datasets and representations should be designed to maximise both accuracy and interpretability. As such, the results offer concrete guidance for high-throughput screening and rational molecular optimisation of next-generation organic electronic materials.

Conflicts of interest

There are no conflicts to declare.

Data availability

Supplementary information is available. The supplementary information includes hyperparameter search ranges, full model performance results for all datasets, descriptor structural representations and further SHAP analyses. See DOI: https://doi.org/10.1039/d5tc04408a.

All data underpinning this publication are openly available from https://pureportal.strath.ac.uk/en/datasets/data-for-comprehensive-evaluation-of-strategies-to-improve-machin/.

Notes and references

1 S. R. Forrest, *Nature*, 2004, **428**, 911–918.
2 H. Sirringhaus, *Adv. Mater.*, 2014, **26**, 1319–1335.
3 T. Nematiaram and A. Troisi, *Chem. Mater.*, 2022, **34**, 4050–4061.
4 B. Kippelen and J.-L. Brédas, *Energy Environ. Sci.*, 2009, **2**, 251–261.
5 M. Ernzerhof, M.-A. Bélanger, D. Mayou and T. Nemati Aram, *J. Chem. Phys.*, 2016, **144**, 134102.
6 Y. van De Burgt, A. Melianas, S. T. Keene, G. Malliaras and A. Salleo, *Nat. Electron.*, 2018, **1**, 386–397.
7 P. A. Shaposhnik, S. A. Zapunidi, M. V. Shestakov, E. V. Agina and S. A. Ponomarenko, *Russ. Chem. Rev.*, 2020, **89**, 1483.
8 K. Zhao, Ö. H. Omar, T. Nematiaram, D. Padula and A. Troisi, *J. Mater. Chem. C*, 2021, **9**, 3324–3333.
9 L. R. Blair and T. Nematiaram, *J. Mater. Chem. C*, 2025, **13**, 17769–17779.
10 S. Hutsch, M. Panhans and F. Ortmann, *npj Comput. Mater.*, 2022, **8**, 228.
11 S. Fratini, M. Nikolka, A. Salleo, G. Schweicher and H. Sirringhaus, *Nat. Mater.*, 2020, **19**, 491–502.
12 T. Nematiaram and A. Troisi, *Mater. Horiz.*, 2020, **7**, 2922–2928.
13 N.-E. Lee, J.-J. Zhou, L. A. Agapito and M. Bernardi, *Phys. Rev. B*, 2018, **97**, 115203.
14 J. Ostmeyer, T. Nematiaram, A. Troisi and P. Buividovich, *Phys. Rev. Appl.*, 2024, **22**, L031004.
15 T. Nematiaram, S. Ciuchi, X. Xie, S. Fratini and A. Troisi, *J. Phys. Chem. C*, 2019, **123**, 6989–6997.
16 G. Schweicher, G. Garbay, R. Jouclas, F. Vibert, F. Devaux and Y. H. Geerts, *Adv. Mater.*, 2020, **32**, 1905909.
17 T. Nematiaram and A. Troisi, *J. Chem. Phys.*, 2020, **152**, 190902.
18 S. Fratini, D. Mayou and S. Ciuchi, *Adv. Funct. Mater.*, 2016, **26**, 2292–2315.

19 H. Oberhofer, K. Reuter and J. Blumberger, *Chem. Rev.*, 2017, **117**, 10319–10357.
20 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Struct. Sci.*, 2016, **72**, 171–179.
21 T. Nematiaram, D. Padula, A. Landi and A. Troisi, *Adv. Funct. Mater.*, 2020, **30**, 2001906.
22 D. Vong, T. Nematiaram, M. A. Dettmann, T. L. Murrey, L. S. Cavalcante, S. M. Gurses, D. Radhakrishnan, L. L. Daemen, J. E. Anthony and K. J. Koski, *et al.*, *J. Phys. Chem. Lett.*, 2022, **13**, 5530–5537.
23 S. Giannini, A. Carof, M. Ellis, H. Yang, O. G. Ziogos, S. Ghosh and J. Blumberger, *Nat. Commun.*, 2019, **10**, 3843.
24 V. Dantanarayana, T. Nematiaram, D. Vong, J. E. Anthony, A. Troisi, K. Nguyen Cong, N. Goldman, R. Faller and A. J. Moulé, *J. Chem. Theory Comput.*, 2020, **16**, 3494–3503.
25 E.-G. Kim, V. Coropceanu, N. E. Gruhn, R. S. Sánchez-Carrera, R. Snoeberger, A. J. Matzger and J.-L. Brédas, *J. Am. Chem. Soc.*, 2007, **129**, 13072–13081.
26 J. Kirkpatrick, V. Marcon, J. Nelson, K. Kremer and D. Andrienko, *Phys. Rev. Lett.*, 2007, **98**, 227402.
27 S. F. Nelsen, S. C. Blackstock and Y. Kim, *J. Am. Chem. Soc.*, 1987, **109**, 677–682.
28 W. Cai, C. Zhong, Z.-W. Ma, Z.-Y. Cai, Y. Qiu, Z. Sajid and D.-Y. Wu, *Phys. Chem. Chem. Phys.*, 2024, **26**, 144–152.
29 F. J. A. Ferrer and F. Santoro, *Phys. Chem. Chem. Phys.*, 2012, **14**, 13549–13563.
30 W.-C. Chen and Y.-C. Cheng, *J. Phys. Chem. A*, 2020, **124**, 7644–7657.
31 F. Santoro and D. Jacquemin, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2016, **6**, 460–486.
32 Ö. H. Omar, M. Del Cueto, T. Nematiaram and A. Troisi, *J. Mater. Chem. C*, 2021, **9**, 13557–13583.
33 S. Luo, T. Li, X. Wang, M. Faizan and L. Zhang, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2021, **11**, e1489.
34 M. Ogbaje, V. Bhat and C. Risko, *Annu. Rev. Mater. Res.*, 2025, **55**, 285–306.
35 J. J. Irwin and B. K. Shoichet, *J. Chem. Inf. Model.*, 2005, **45**, 177–182.
36 T. Nematiaram, Z. Lamprou and Y. Moshfeghi, *Chem. Commun.*, 2025, **61**, 3676–3679.
37 J. Lederer, W. Kaiser, A. Mattoni and A. Gagliardi, *Adv. Theory Simul.*, 2019, **2**, 1800136.
38 S. Atahan-Evrenk and F. B. Atalay, *J. Phys. Chem. A*, 2019, **123**, 7855–7863.
39 K. Chen, C. Kunkel, K. Reuter and J. T. Margraf, *Digital Discovery*, 2022, **1**, 147–157.
40 C.-H. Li and D. P. Tabor, *J. Phys. Chem. A*, 2023, **127**, 3484–3489.
41 K. M. Katubi, M. Saqib, M. Maryam, T. Mubashir, M. H. Tahir, M. Sulaman, Z. Alrowaili and M. Al-Buriahi, *Inorg. Chem. Commun.*, 2023, **151**, 110610.
42 C. Shu, G. Mustafa, M. H. Tahir, M. A. El-Tayeb and M. A. Ibrahim, *Dyes Pigm.*, 2024, **231**, 112382.
43 E. Hussain, M. M. Soliman, S. M. El-Bahy, S. Naeem and N. Khan, *Sol. Energy*, 2025, **286**, 113169.
44 T. Ando, N. Shimizu, N. Yamamoto, N. N. Matsuzawa, H. Maeshima and H. Kaneko, *J. Phys. Chem. A*, 2022, **126**, 6336–6347.
45 M. Saqib, M. Sagir, M. H. Tahir, H. O. Elansary and M. Javed, *et al.*, *J. Solid State Chem.*, 2025, **345**, 125250.
46 X. Zhang, G. Ye, C. Wen and Z. Bi, *Comput. Mater. Sci.*, 2023, **228**, 112361.
47 O. D. Abarbanel and G. R. Hutchison, *J. Chem. Phys.*, 2021, **155**, 054106.
48 S. Atahan-Evrenk, *RSC Adv.*, 2018, **8**, 40330–40337.
49 S. K. Pandey and K. Roy, *Mater. Today Commun.*, 2024, **41**, 110430.
50 V. Coropceanu, J. Cornil, D. A. da Silva Filho, Y. Olivier, R. Silbey and J.-L. Brédas, *Chem. Rev.*, 2007, **107**, 926–952.
51 M. Chu, J.-X. Fan, S. Yang, D. Liu, C. F. Ng, H. Dong, A.-M. Ren and Q. Miao, *Adv. Mater.*, 2018, **30**, 1803467.
52 S. M. Lundberg and S.-I. Lee, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 1–10.
53 M. E. Frisch, G. Trucks, H. B. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, G. Petersson, H. Nakatsuji, *et al.*, *Gaussian 16*, 2016.
54 D. Padula, Ö. H. Omar, T. Nematiaram and A. Troisi, *Energy Environ. Sci.*, 2019, **12**, 2412–2416.
55 S. Yang, M. Sun, C. Shi, Y. Liu, Y. Guo, Y. Liu, Z. Lu, Y. Huang and X. Pu, *J. Chem. Theory Comput.*, 2024, **20**, 10633–10648.
56 A. Tversky, *Psychol. Rev.*, 1977, **84**, 327.
57 J. L. Bao, L. Gagliardi and D. G. Truhlar, *J. Phys. Chem. Lett.*, 2018, **9**, 2353–2358.
58 M. Uejima, T. Sato, K. Tanaka and H. Kaji, *Phys. Chem. Chem. Phys.*, 2013, **15**, 14006–14016.
59 N. N. Matsuzawa, H. Arai, M. Sasago, E. Fujii, A. Goldberg, T. J. Mustard, H. S. Kwak, D. J. Giesen, F. Ranalli and M. D. Halls, *J. Phys. Chem. A*, 2020, **124**, 1981–1992.
60 V. Stehr, J. Pfister, R. Fink, B. Engels and C. Deibel, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2011, **83**, 155208.
61 P. Li, Y. Cui, C. Song and H. Zhang, *RSC Adv.*, 2015, **5**, 50212–50222.
62 M. Moral, A. Garzón-Ruiz, M. Castro, J. Canales-Vázquez and J. C. Sancho-Garca, *J. Phys. Chem. C*, 2017, **121**, 28249–28261.
63 P. E. Schwenn, P. L. Burn and B. J. Powell, *Org. Electron.*, 2011, **12**, 394–403.
64 N. W. Mitzel and D. W. Rankin, *Dalton Trans.*, 2003, 3650–3662.
65 S. Fratini, S. Ciuchi, D. Mayou, G. T. De Laissardière and A. Troisi, *Nat. Mater.*, 2017, **16**, 998–1002.
66 V. Vinod, D. Lyu, M. Ruth, P. R. Schreiner, U. Kleinekathöfer and P. Zaspel, *J. Comput. Chem.*, 2025, **46**, e70056.
67 G. J. Moore, O. Bardagot and N. Banerji, *Adv. Theory Simul.*, 2022, **5**, 2100511.
68 G. Landrum, *RDKit: Open-Source Cheminformatics*, https://www.rdkit.org, 2024, Accessed: 2024-03-20.
69 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
70 R. Kawagoe, T. Ando, N. N. Matsuzawa, H. Maeshima and H. Kaneko, *ACS Omega*, 2024, **9**, 48844–48854.

71 M. K. Tufail, S. S. A. Shah, S. Khan, F. Ahmad, L. W. Kiruri, M. S. Abbasi and A. Ahmad, *Chem. Phys. Lett.*, 2024, **834**, 140974.
72 R. E. Schapire, *Empirical inference: festschrift in honor of vladimir N. Vapnik*, Springer, 2013, pp. 37–52.
73 J. H. Friedman, *Comput. Stat. Data Anal.*, 2002, **38**, 367–378.
74 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 3147.
75 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
76 T. Chen and C. Guestrin, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
77 T. Howley and M. G. Madden, *Artif. Intell. Rev.*, 2005, **24**, 379–395.
78 X. Glorot and Y. Bengio, Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249–256.
79 G. E. Hinton, *Machine learning*, Elsevier, 1990, pp. 555–610.
80 K. He, X. Zhang, S. Ren and J. Sun, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, 2015, https://arxiv.org/abs/1502.01852.
81 D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, 2017, https://arxiv.org/abs/1412.6980.
82 M. P. LaValley, *Circulation*, 2008, **117**, 2395–2399.
83 E. Fix and J. L. Hodges, *Int. Stat. Rev.*, 1989, **57**, 238–247.
84 M. Fey and J. E. Lenssen, *arXiv*, 2019, preprint, arXiv:1903.02428, DOI: 10.48550/arXiv.1903.02428.
85 R. V. D. Berg, T. N. Kipf and M. Welling, *arXiv*, 2017, preprint, arXiv:1706.02263, DOI: 10.48550/arXiv.1706.02263.
86 P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, *arXiv*, 2017, preprint, arXiv:1710.10903, DOI: 10.48550/arXiv.1710.10903.
87 Z. Tan, Y. Li, Z. Zhang, T. Penfold, W. Shi, S. Yang and W. Zhang, *New J. Chem.*, 2023, **47**, 9550–9554.