Reaction Chemistry & Engineering

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6re00049e

Automated analysis of DFT output files for molecular descriptor extraction and reactivity modeling

Yu-Chien Huang,[a] Dennis Chung-Yang Huang *[bc] and Yun-Cheng Tsai*[a]

Received 11th February 2026,
Accepted 24th April 2026

DOI: 10.1039/d6re00049e

rsc.li/reaction-engineering

Understanding the relationship between molecular structure and chemical reactivity or properties is fundamental to rational molecular design. Linear free energy relationships (LFERs), particularly Hammett analysis, have long served as powerful tools in organic chemistry. Recently, these approaches have been enhanced by the incorporation of computationally derived parameters, enabling broader applicability across diverse molecules and reactions. To facilitate and scale this process, we present *DFTDescriptorPipeline*, a fully automated workflow for extracting quantum chemical descriptors from Gaussian log files and constructing structure–property/reactivity relationships using multivariate linear regression (MLR) models. We validate the workflow across four case studies, including photoswitchable molecules and catalytic reactions. In each case, the models provide interpretable results, demonstrating the versatility of this approach and relevance to a wide range of chemical contexts. We anticipate that this platform will serve as a generalizable framework for integrating quantum chemical calculations into data-driven molecular design.

Introduction

Bridging molecular structures with their properties or reactivities has long been a central goal in chemistry. In organic chemistry, linear free energy relationships (LFERs) provide a quantitative framework for understanding how structural variations influence reaction rates or selectivity.[1] A classic example is Hammett analysis, in which the acidity of substituted benzoic acids, expressed as the substituent constant (σ value), serves as a surrogate parameter for the electronic properties of arenes (Fig. 1a).[2]

The strength of this approach lies in its ability to infer transition-state energetics (activation barriers) from readily measurable ground-state characteristics. As one of the most influential tools in physical organic chemistry, Hammett analysis has been widely applied to correlate the electron-donating or -withdrawing nature of aromatic substituents with chemical reactivity, selectivity, and other properties.[3]

However, traditional substituent constants are experimentally derived and often unavailable for structurally complex arenes. Moreover, Hammett analysis does not account for steric effects, which can critically influence the catalytic reactivity and selectivity.[4,5] To address these limitations, Sigman and coworkers introduced a novel approach in 2016 that leverages computationally derived parameters (Fig. 1b)[6,7] Specifically, they employed density functional theory (DFT) to calculate electronic descriptors (DFT featurization) and incorporated Sterimol values to quantify steric effects.[8,9] This method is computationally inexpensive, requiring calculations only for the varying substituents and no transition-state modeling. Using this approach, they constructed multivariate linear regression (MLR) models that successfully correlated reactant structures with the selectivity of catalytic transformations. More recently, one of us adapted this methodology to model the thermal half-lives of *N*-aryl-substituted indigo photoswitches, demonstrating the utility of MLR models in establishing structure–property relationships for functional molecules.[10]

This modern form of Hammett analysis consists of three key steps: 1) parameter extraction from DFT output files, 2) descriptor tabulation, and 3) construction of MLR models (Fig. 1c). In previous studies, these steps were carried out manually, limiting scalability to larger datasets. Automation of steps 1 and 2 has been reported by Doyle (*Auto-QChem*),[11] Paton (*AQME*),[12] and Sigman (*Get_Properties*)[13] groups. Although these tools have shown great value in data

[a] *Department of Technology Application and Human Resource Development, National Taiwan Normal University, Taipei 106, Taiwan. E-mail: pecu610@gmail.com*
[b] *Department of Chemistry, Oklahoma State University, 107 Physical Sciences I, Stillwater, Oklahoma 74078, USA. E-mail: dcyhuang@okstate.edu*
[c] *Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21, Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0021, Japan*

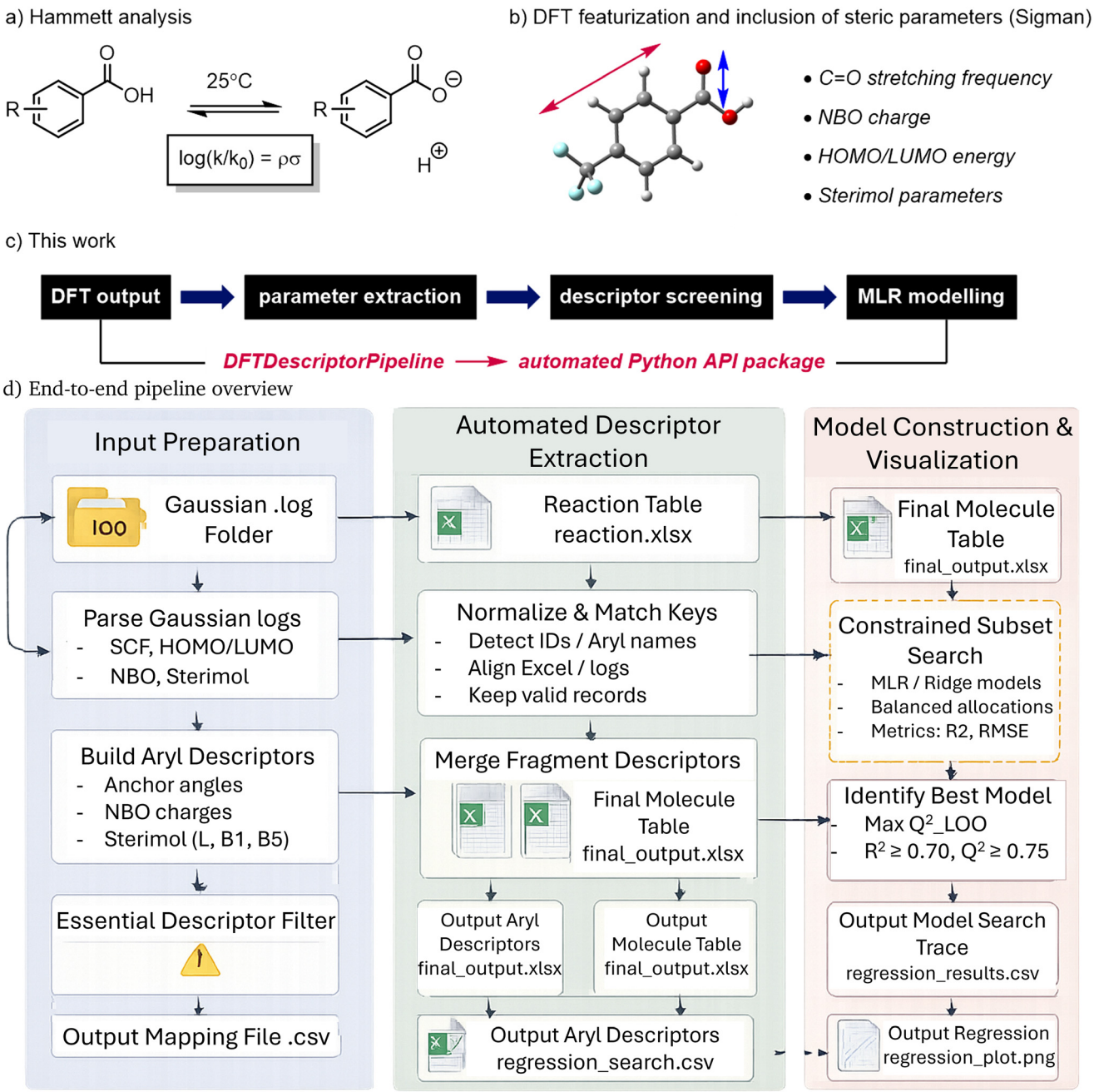


Fig. 1 Integrated overview of the proposed *DFTDescriptorPipeline*. (a) Classical Hammett analysis linking substituent constants to reactivity. (b) DFT-based featurization incorporating steric and electronic parameters. (c) The present work: *DFTDescriptorPipeline* automates descriptor extraction, screening, and multivariate linear regression modeling. (d) End-to-end pipeline overview of *DFTDescriptorPipeline*: input preparation, automated descriptor extraction, substituent matching, model construction, and visualization.

chemistry by connecting computation and molecular parametrization to downstream tasks, a workflow that further integrates LFER model construction remains elusive.[14,15]

We envision a platform that enables users to upload DFT output files along with experimental data, automatically extract relevant molecular descriptors, and identify optimal MLR models. This tool should ideally require minimal coding interventions, thereby lowering the barrier to adoption for general scientists. Herein, we report the development of *DFTDescriptorPipeline*, an open-source Python package designed for automated parameter extraction and reaction modeling, and demonstrate its utility through four case studies. It is expected that this platform will add to the toolbox for connecting data science and physical organic chemistry.[16,17]

Parametrization

We developed a fully automated parametrization workflow, available at https://github.com/peculab/DFTDescriptorPipeline, that extracts quantum-chemical descriptors from Gaussian output (.log) files and aligns them with experimental identifiers. The workflow consists of three stages.

1. Extract electronic descriptors from Gaussian log files, including HOMO/LUMO energies, dipole moment, isotropic polarizability, and NBO-derived features defined on anchor atoms.

2. Quantify steric effects using Sterimol parameters (L, B_1, B_5) computed along the C1–C2 axis.

3. Aggregate all features into a unified pandas table, using substituent-aware prefixes to keep variables consistent for downstream modeling.

This end-to-end process ensures reproducible and scalable descriptor generation across diverse reaction systems. A schematic overview of the complete pipeline is provided in Fig. 1d.

Detailed extraction procedures, anchor-based NBO descriptors, Sterimol geometry processing, and the final feature aggregation framework are described in section *Descriptor extraction*.

Anchor indices and Sterimol (implementation details)

Anchor atoms (a–g) are inferred from the NBO summary by locating the O–H bond(s) of the acid, the carbonyl carbon (C1), and the adjacent aryl carbon (C2) (Fig. 2). Steric descriptors are computed after removing atoms a, b, and d from the final Standard orientation geometry; C1–C2 defines the Sterimol axis. Bondi radii are used with a hydrogen adjustment (H = 1.09 Å). The resulting L, B_1, and B_5 values are stored as Ar_Ster_L, Ar_Ster_B1, Ar_Ster_B5.

Descriptor extraction

For each molecule, the pipeline systematically parses Gaussian log files to obtain a comprehensive set of quantum chemical descriptors relevant to structure–reactivity modeling. The extraction is implemented in a module, which integrates regular-expression parsing, numerical computation, and error handling to ensure robustness across varied log file formats (https://github.com/peculab/DFTDescriptorPipeline).

Frontier orbital energies (HOMO/LUMO)

Frontier orbital energies are extracted from the final converged self-consistent field (SCF) results, specifically the highest occupied (HOMO) and lowest unoccupied (LUMO) molecular orbital energies reported at the end of the electronic-structure calculation. These descriptors reflect the electron-donating and electron-accepting tendencies of substituents.

Dipole moment and isotropic polarizability

The field-independent dipole magnitude is parsed from the "dipole moment" block in the Gaussian output and reported in Debye units. This descriptor suggests overall charge distribution within the molecule.

Isotropic polarizability is computed as the arithmetic mean of the tensor components from the "exact polarizability" block in the Gaussian output, reflecting the molecular responsiveness to external electric fields.

NBO analysis and anchor atoms

This module locates the "natural bond orbitals (summary)" block and identifies the O–H bonds of the carboxylic acid moiety, which serve as reference points for defining seven characteristic atoms (a–g) (Fig. 2a). Atom c (C1) corresponds to the carbonyl carbon of the carboxylic acid, and atom e (C2) to the connecting carbon on the aromatic ring, while a, b, d, f, and g represent adjacent atoms and aryl extensions. These anchor atoms provide consistent geometric references for extracting:

1. Occupancies and orbital energies of the C1–O and C1–C2 bonds,

2. Atomic NBO charges on the C1, C2, and O atoms within the carboxyl group,

3. IR vibrational frequency and intensity of the C═O stretch (in the range of 1800–1900 cm^{-1}),

4. C1–C2 bond length, computed from the final "Standard orientation" geometry.

All parsing routines are tolerant to format variations and automatically report missing atomic indices.

Sterimol parameter computation

To capture steric effects, Sterimol parameters (L, B_1, B_5) are automatically computed using the morfeus package.[9] Prior to computation, atoms a, b, and d are excluded from the geometry, and the C1–C2 bond is designated as the reference X-axis (Fig. 2b). The structured procedure is as follows:

1. **Anchor atom selection:** indices defining the substituent framework (a–g) are identified from NBO analysis. Atoms c (C1) and e (C2) define the substituent attachment axis.

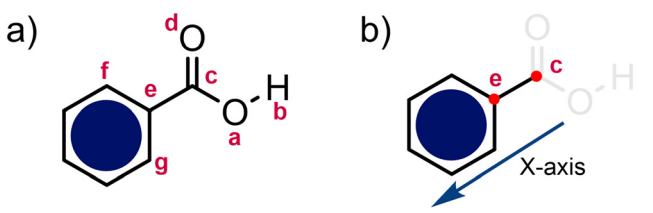


Fig. 2 (a) Definition of the seven anchor atoms (a–g) used for natural bond orbital (NBO) analysis. Atom c (C1) denotes the carbonyl carbon, and atom e (C2) denotes the aryl carbon directly bonded to C1. (b) Definition of the C1–C2 axis used to compute Sterimol parameters (L, B_1, and B_5). Atomic boundaries are defined using Bondi radii.

2. **Geometry filtering:** atoms a, b, and d are removed from the final "standard orientation" geometry, and the filtered coordinates are written to an intermediate .xyz file.

3. **Axis definition and descriptor computation:** using the C1–C2 bond as the principal *X*-axis, the Sterimol parameters are computed as: L (maximum substituent length), B_1 (minimum width), and B_5 (maximum width). Atomic radii are assigned using Bondi radii, with adjusted values for hydrogen.

4. **Error handling:** molecules lacking valid anchor atoms or yielding invalid geometries are skipped. Missing values are set to none, and all exceptions are logged for transparency.

5. **Integration into feature table:** the resulting Sterimol parameters are appended to the descriptor dataframe with substituent-specific prefixes (*e.g.*, Ar1_Ster_L, Ar2_Ster_B5).

Feature table aggregation

All extracted descriptors (electronic, vibrational, geometric, and steric) are consolidated into a unified tabular representation that serves as the direct input to downstream regression analyses. The aggregation procedure is designed to preserve provenance (*i.e.*, which substituent slot a descriptor originates from), and maintain robustness against partial extraction failures. Concretely, the workflow proceeds as follows:

1. **Feature-table assembly:** for each unique substituent identifier (*e.g.*, Ar, Ar1, Ar2), the extracted descriptors are compiled into a single dataframe, one row per substituent. Column names are systematically prefixed by the corresponding substituent slot to retain traceability after merging at the molecule level.

2. **Completeness screening:** a code-defined set of essential descriptors (including NBO charges, HOMO/LUMO energies, IR vibrational frequency/intensity of C═O stretch, and Sterimol parameters) is checked for missing values. Entries lacking any critical fields are excluded to prevent ill-posed regression fits and to improve the stability of cross-validated performance.

3. **Non-blocking error reporting:** extraction anomalies (*e.g.*, missing anchor atoms, failed parsing blocks, or geometry-related exceptions) are recorded in a structured diagnostic report. This design enables *post hoc* inspection and debugging without interrupting the end-to-end pipeline execution.

4. **Interoperable export:** the finalized feature table is exported in common tabular formats suitable for reproducible statistical analysis and for integration with external modeling toolchains.

Modeling

A complete list of descriptors extracted through this workflow is available in SI as well as the project repository (https://github.com/peculab/DFTDescriptorPipeline). The modeling module of *DFTDescriptorPipeline* automatically correlates the cleaned descriptor table with experimental reactivity metrics, such as $\ln k_{obs}$ and $\Delta\Delta G^{\ddagger}$, using multivariate linear regression (MLR). Implemented entirely in open-source Python, the workflow ensures full reproducibility from descriptor extraction to model construction and publication-ready plots.

Data cleaning and effective dataset size

Before model fitting, rows with missing entries in any descriptor or target column are removed programmatically. The resulting effective dataset size (n_{eff}) can therefore be smaller than the raw spreadsheet count (n_{raw}), as only molecules with complete descriptor vectors are used for regression. All reported R^2, Q^2_{LOO}, and RMSE values correspond to this filtered dataset.

Descriptor grouping

Each descriptor column carries an explicit aryl prefix (Ar1_, Ar2_, …), allowing substituent-specific grouping. Features are collected into groups based on these prefixes to ensure that each substituent contributes at least one descriptor. The algorithm constructs balanced models by enforcing per-group bounds (default 1–3 features per group) while varying the total descriptor count up to the user-defined maximum.

Regression search logic

For each feasible feature allocation, all possible combinations are enumerated and a linear model is fitted to the target variable. Candidate models are screened and ranked through:

(i) **In-sample fit.** Models with $R^2 < 0.70$ are discarded.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

(ii) **Cross-validation.** Leave-one-out cross-validation (LOOCV):

$$Q^2_{LOO} = 1 - \frac{\sum_i \left(y_i - \hat{y}_{i,-i}\right)^2}{\sum_i (y_i - \bar{y})^2}.$$

(iii) **Error metric.** Root-mean-square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}.$$

Model selection

Models are ranked primarily by Q^2_{LOO} and secondarily by RMSE. Among models of comparable performance, the algorithm favors those with fewer descriptors and balanced group representation. The final model M_{best} takes the form

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j,$$

Algorithm 1. Code-consistent automated extraction and subset model search (concise but detailed)

Require: Reaction table $\mathcal{X}$ (aryl slots `Ar1`, `Ar2`, …, target y); Gaussian logs $\mathcal{D} = \{\ell_a\}$ keyed by aryl ID a; join mode J; $k_{\max}$; group bounds $[m_g, M_g]$; required descriptor set $\mathcal{E}$; acceptance filter $\mathcal{F}$ (e.g., $R^2 \geq 0.70$).
Ensure: Merged table $\mathcal{T}$; ranked records $\mathcal{R}$; best subset $S^\star$ and coefficients $\hat{\beta}$.

Normalize keys and define valid aryl set.
Normalize slot IDs in $\mathcal{X}$ (trim/case/unify delimiters).
$\mathcal{A} \leftarrow$ unique aryl IDs appearing in required slots; $\mathcal{A} \leftarrow \{a \in \mathcal{A} : a \in \mathrm{keys}(\mathcal{D})\}$.
Aryl-level descriptor extraction.
for all $a \in \mathcal{A}$ **do**
  Parse ℓ_a and compute descriptor vector $\mathbf{f}_a$ (electronic/structural/steric).
  If parsing fails or $\mathbf{f}_a[\mathcal{E}]$ incomplete: **skip** a.
  Store $(a, \mathbf{f}_a)$ in $\mathcal{F}_{Ar}$.
end for
Drop constant/all-missing descriptor columns in $\mathcal{F}_{Ar}$.
Prefix-merge to reaction level.
$\mathcal{T} \leftarrow \mathcal{X}$.
for all required slots s under join mode J **do**
  Prefix descriptors: $d \mapsto$ `s_`d (e.g., `Ar1_*`); left-join on $\mathcal{T}[s] = a$.
end for

Extract modeling arrays $(\mathbf{y}, \mathbf{X})$ from $\mathcal{T}$; remove rows with missing y and rows failing missingness policy for predictors; remove constant/duplicate columns.
Define grouped constraints.
Partition predictors by slot-prefix into groups $\{\mathcal{G}_g\}$.
Feasible subset predicate: $\mathcal{C}(S) : |S| = k \ \wedge \ \forall g,\ m_g \leq |S \cap \mathcal{G}_g| \leq M_g$.
Enumerate constrained subsets and evaluate via closed-form LOO.
$\mathcal{R} \leftarrow \emptyset$.
for $k = 1$ **to** $k_{\max}$ **do**
  Enumerate feasible $S \subseteq \mathcal{P}$ satisfying $\mathcal{C}(S)$ (via group allocations k_g and within-group combinations).
  for all feasible subsets S **do**
    Fit OLS (with intercept) on $\mathbf{X}_S$ to obtain $\hat{\beta}_S$ and residuals $\mathbf{e}$.
    Compute R^2 and RMSE on training data.
    Compute closed-form LOO:
      Hat diagonal h_{ii} from $\mathbf{H} = \mathbf{X}_S(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{X}_S^\top$,
      $e^{(-i)} = \frac{e_i}{1-h_{ii}}$, $\mathrm{RMSE}_{\mathrm{LOO}} = \sqrt{\frac{1}{n}\sum_i (e^{(-i)})^2}$,
      $Q^2_{\mathrm{LOO}} = 1 - \frac{\sum_i (e^{(-i)})^2}{\sum_i (y_i - \bar{y})^2}$.
    If model fails $\mathcal{F}$: **continue**.
    Append record $(k, S, R^2, Q^2_{\mathrm{LOO}}, \mathrm{RMSE}_{\mathrm{LOO}}, \hat{\beta}_S, n)$ to $\mathcal{R}$.
  end for
end for
Select and report best model.
Save $\mathcal{R}$ (ranked by Q^2_{LOO}) to disk.
$S^\star \leftarrow \arg\max_{\mathcal{R}} Q^2_{\mathrm{LOO}}$ (ties: min $\mathrm{RMSE}_{\mathrm{LOO}}$; then smaller k).
Refit OLS on full data using $S^\star$ to obtain final $\hat{\beta}$; report equation and diagnostics.

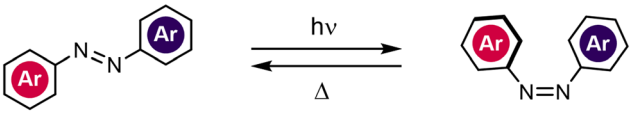
where x_j are standardized descriptors and β_j the fitted coefficients.

Visualization

Each final model is visualized through a parity plot (predicted *vs.* experimental) annotated with R^2, Q^2_{LOO}, RMSE, and n_{eff}, providing an at-a-glance validation of model reliability.

Case studies

To demonstrate a consistent regression-style reporting across datasets with different coupling schemes (single- *vs.* pair-join) and different targets, we applied the proposed workflow to four case studies: (a) thermal back-reaction of azoarene photoswitches, (b) redox-relay Heck coupling with boronic acids, (c) thermal back-isomerization of *N*-aryl-*N*′-alkylindigo



Scheme 1 Thermal back-reaction of azoarene photoswitches.

photoswitches, (d) thermal back-isomerization of *N*,*N*′-diarylindigo photoswitches.

Thermal back-reaction of azoarene photoswitches

Azoarenes are widely used photoswitches, and their thermal back-reaction rates are primarily governed by substituent electronic effects and steric constraints (Scheme 1). In this dataset, we adopt a **pair-join** scheme (pair: Ar1 + Ar2), where each reaction entry is constructed by pairing two aromatic fragments, Ar_1 and Ar_2.

Redox-relay Heck coupling with boronic acids

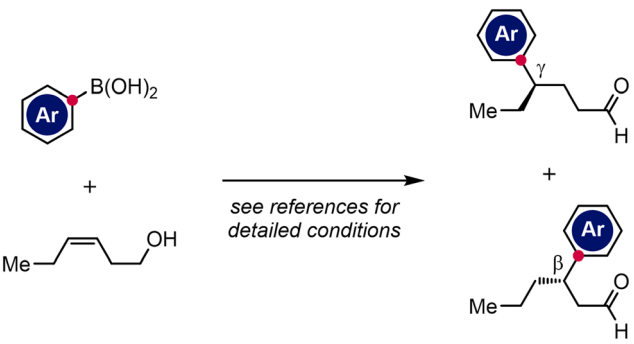
To assess transferability, we next examined redox-relay Heck couplings between boronic acids and allylic alcohols (Scheme 2).[6] This Heck–boronic acids dataset adopts a **single-join** scheme (single: Ar), in which each reaction entry is associated with a single aromatic fragment.

Thermal back-isomerization of *N*-aryl-*N*′-alkylindigo photoswitches

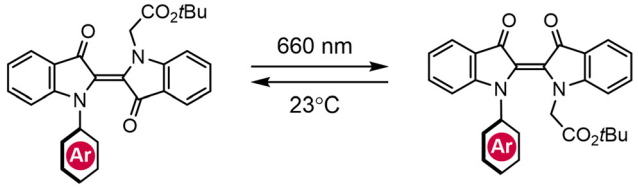
Indigos have recently emerged as a red-light-responsive class of photoswitches, with thermal back-isomerization rates that are strongly modulated by *N*-substituent effects.[18] We therefore applied the proposed workflow to the *N*-aryl-*N*′-alkylindigo dataset (Scheme 3).

Thermal back-isomerization of *N*,*N*′-diarylindigo photoswitches

Finally, we examined bis-aryl indigo derivatives in which two aryl substituents vary independently (Scheme 4). This indigo diaryl dataset adopts a **pair-join** scheme (pair: Ar1 + Ar2) and targets $\ln(k_{obs})$.



Scheme 2 Redox-relay Heck coupling with boronic acids.



Scheme 3 Thermal back-isomerization of *N*-aryl-*N*′-alkylindigo photoswitches.

Results and discussion

For each case study, we identify the best model produced by the constrained subset search, ranked by the LOOCV-based predictive coefficient Q^2_{LOO}. We explicitly indicate whether the selected model satisfies the predefined pass criteria; when it does not, we still report the top-performing model(s) to support transparency and enable direct comparison across datasets. Table 1 provides a quantitative overview of all four case studies, including the join mode, target, key descriptors, and performance metrics. To keep the main text focused, Fig. 3–6 then visualize only the single best-performing model for each case (selected by Q^2_{LOO}).

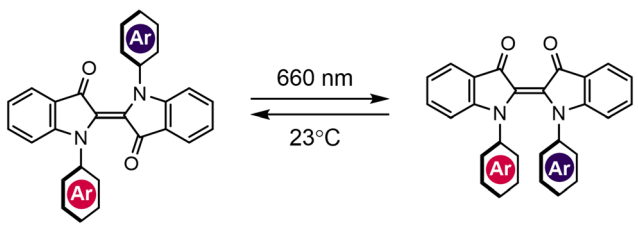
All models are linear regressions selected using the same constrained-subset protocol and evaluated using leave-one-out cross-validation. For each dataset, we examine descriptor set sizes k = 3, 4, 5 and report the best-performing subset by Q^2_{LOO} together with R^2 and RMSE, as well as an interpretable closed-form regression equation. Complete results—including all k = 3, 4, 5 runs and the top-five models (with regression formulas) for each case—are provided in the associated content of the project repository (https://github.com/peculab/DFTDescriptorPipeline).

Furthermore, compilation of descriptor correlations, feature importance, as well as the pipeline runtime of each case study can be found in SI as well as the project repository.

Thermal back-reaction of azoarene photoswitches

Across k = 3, 4, 5, the best LOOCV generalization within the OLS search range was obtained at k = 4 (R^2 = 0.618, Q^2 = 0.468, RMSE = 2.74; Fig. 3). For interpretability, we express the best k = 4 OLS model as:

$$\ln(k_{obs}) = -214.89 - 88.08x_1 + 88.10x_2 - 142.71x_3 - 142.71x_4 \quad (1)$$



Scheme 4 Thermal back-isomerization of *N*,*N*′-diarylindigo photoswitches.

Table 1 Summary of the four case studies and their best-performing models, including method, join mode, target property, data size, selected descriptor subsets, and predictive performance metrics

Case study	Method/join mode	Target	Data size	Key descriptors (best, normalized)	Performance
Azoarene	Ordinary least squares (OLS) pair: Ar1 + Ar2	$\ln(k_{obs})$	25/25 k = 4	ar2_scf_energy_hartree ar2_gibbs_free_energy_hartree ar1_nbo_q_a ar1_Ar_NBO_OH	R^2 = 0.618 Q^2 = 0.468 RMSE = 2.74
Azoarene	Support vector regression (SVR) with a radial basis function (RBF) kernel pair: Ar1 + Ar2	$\ln(k_{obs})$	25/25 k = 5	ar2_homo_hartree ar1_bond_length_c_a ar2_gap_hartree ar2_bond_length_c_e ar1_nbo_q_a	R^2 = 0.815 Q^2 = 0.486 RMSE = 1.907
Heck–boronic-acids	Ordinary least squares (OLS) single: Ar	$\Delta\Delta G$	17/17 k = 5	bond_length_c_e q_e enthalpy_hartree gibbs_free_energy_hartree homo_hartree	R^2 = 0.831 Q^2 = 0.641 RMSE = 0.274
N-Aryl-*N*′-alkylindigo	Ordinary least squares (OLS) single: Ar	$\ln(k_{obs})_{MeCN}$	21/21 k = 5	Ar_v_C═O Ar_NBO_–O LUMO L_C1–C2 Ar_NBO_C2_x	R^2 = 0.932 Q^2 = 0.841 RMSE = 0.141
N,*N*′-Diarylindigo	Ordinary least squares (OLS) pair: Ar1 + Ar2	$\ln(k_{obs})$	12/12 k = 4	ar1_nbo_q_b ar2_bond_length_c_a ar2_bond_length_c_d Ar2_v_C═O	R^2 = 0.989 Q^2 = 0.976 RMSE = 0.191

where x_1 and x_2 denote the SCF total energy and Gibbs free energy of Ar_2, respectively, and x_3 and x_4 denote Ar_1-side NBO charge descriptors at the mapped anchor position. Overall, the selected subset suggests that frontier energetic stabilization on Ar_2 and localized electronic distribution on Ar_1 jointly influence the thermal back-reaction rate in the azoarene family.

The limited performance of the linear model may reflect the greater mechanistic complexity of azoarene thermal back-reactions, which is not fully captured by the MLR framework.

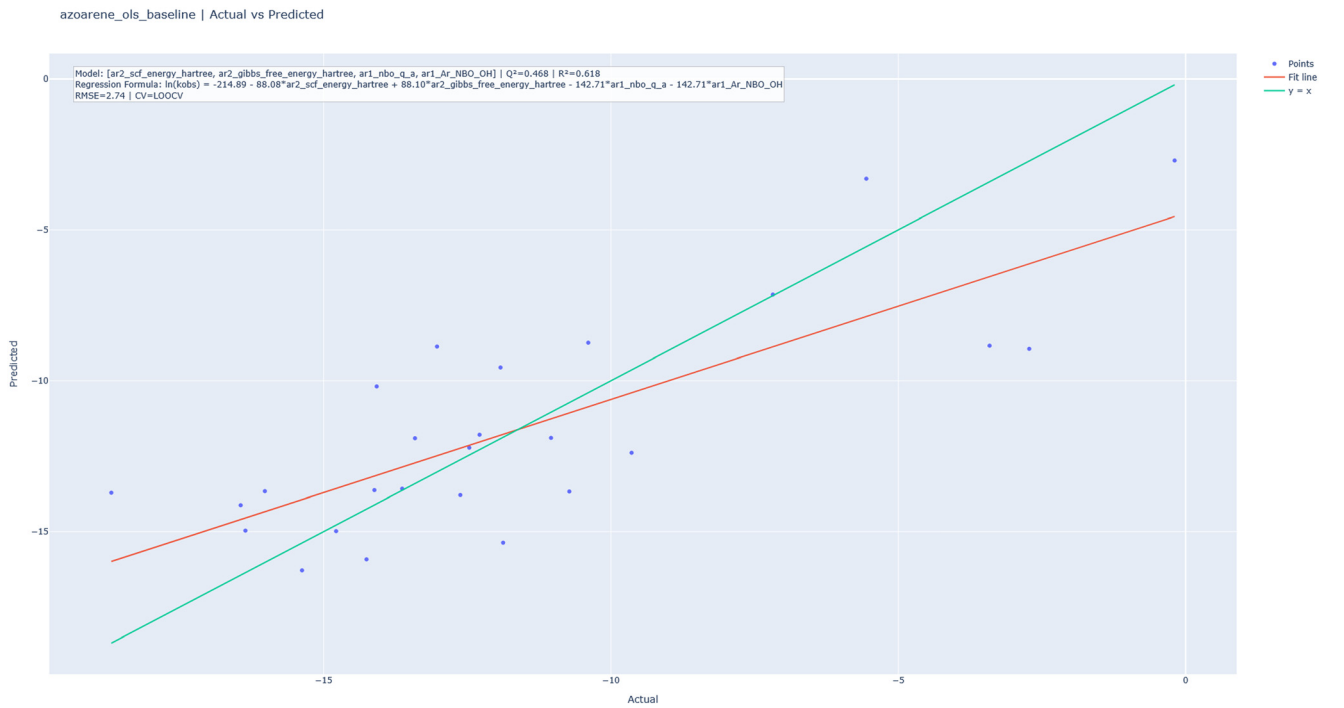


Fig. 3 Actual *vs.* predicted plot for the best-performing LOOCV model: thermal back-reaction of azoarene photoswitches. The selected model maximizes Q^2_{LOO} under the constrained subset search.

To further examine this possibility, we applied support vector regression (SVR) with a radial basis function (RBF) kernel, resulting in improved R^2 and Q^2 values, as summarized in Table 1.

Redox-relay Heck coupling with boronic acids

Across k = 3, 4, 5, the best LOOCV performance was obtained at k = 5 (R^2 = 0.831, Q^2 = 0.641, RMSE = 0.274, n = 17; Fig. 4). Although this model does not pass the pre-defined Q^2 threshold, it still provides the strongest predictive performance among the tested subsets for this case study. The best k = 5 model is:

$$\Delta\Delta G = -78.17 + 53.06x_1 - 1.19x_2 - 47.60x_3 + 47.60x_4 - 13.06x_5 \quad (2)$$

where x_1, …, x_5 correspond to bond_length_c_e, q_e, enthalpy_hartree, gibbs_free_energy_hartree, and homo_hartree, respectively. Unlike the previous charge-only solution, the updated best subset combines a local geometric descriptor, a site-specific electronic descriptor, two thermodynamic energy terms, and one frontier-orbital descriptor. This suggests that $\Delta\Delta G$ in the Heck–boronic-acids dataset is better explained by a mixed representation of local structure and global electronic/energetic effects, rather than by atomic charges alone.

Thermal back-isomerization of *N*-aryl-*N*′-alkylindigo photoswitches

This indigo aryl–alkyl dataset uses a **single-join** scheme (single: Ar) and targets the rate of thermal back-reaction in acetonitrile, denoted as $\ln(k_{obs})_{MeCN}$. Across k = 3, 4, 5, the best LOOCV performance was again obtained at k = 5 (R^2 = 0.932, Q^2 0.841, RMSE = 0.141, n = 21; Fig. 5), which remains above the pass threshold. The best k = 5 model is:

$$\ln(k_{obs})_{MeCN} = -131.17 - 0.09x_1 - 99.30x_2 + 10.27x_3 + 152.37x_4 - 8.64x_5 \quad (3)$$

where x_1, …, x_5 denote Ar_v_C═O, Ar_NBO_–O, LUMO, L_C1–C2, and Ar_NBO_C2_x, respectively. The resulting model indicates that the thermal back-reaction rate in this dataset is highly correlated to a combination of carbonyl-related vibrational and charge descriptors, frontier-orbital energy, local bond-length variation, and aryl-site electronic distribution.

Thermal back-isomerization of *N*,*N*′-diarylindigo photoswitches

In contrast to the previous two lower-performing case studies, this dataset yields models that clearly satisfy the pass criteria. Across k = 3, 4, 5, the best LOOCV performance was obtained at k = 4 (R^2 = 0.989, Q^2 = 0.976, RMSE = 0.191, n = 12; Fig. 6), indicating excellent predictive accuracy and strong generalization. The best k = 4 model is:

$$\ln(k_{obs}) = -3691.23 - 650.61x_1 + 235.59x_2 + 2451.33x_3 + 0.41x_4 \quad (4)$$

where x_1 denotes the NBO charge at site b on Ar_1 (ar1_nbo_q_b), x_2 denotes the bond length between sites c and

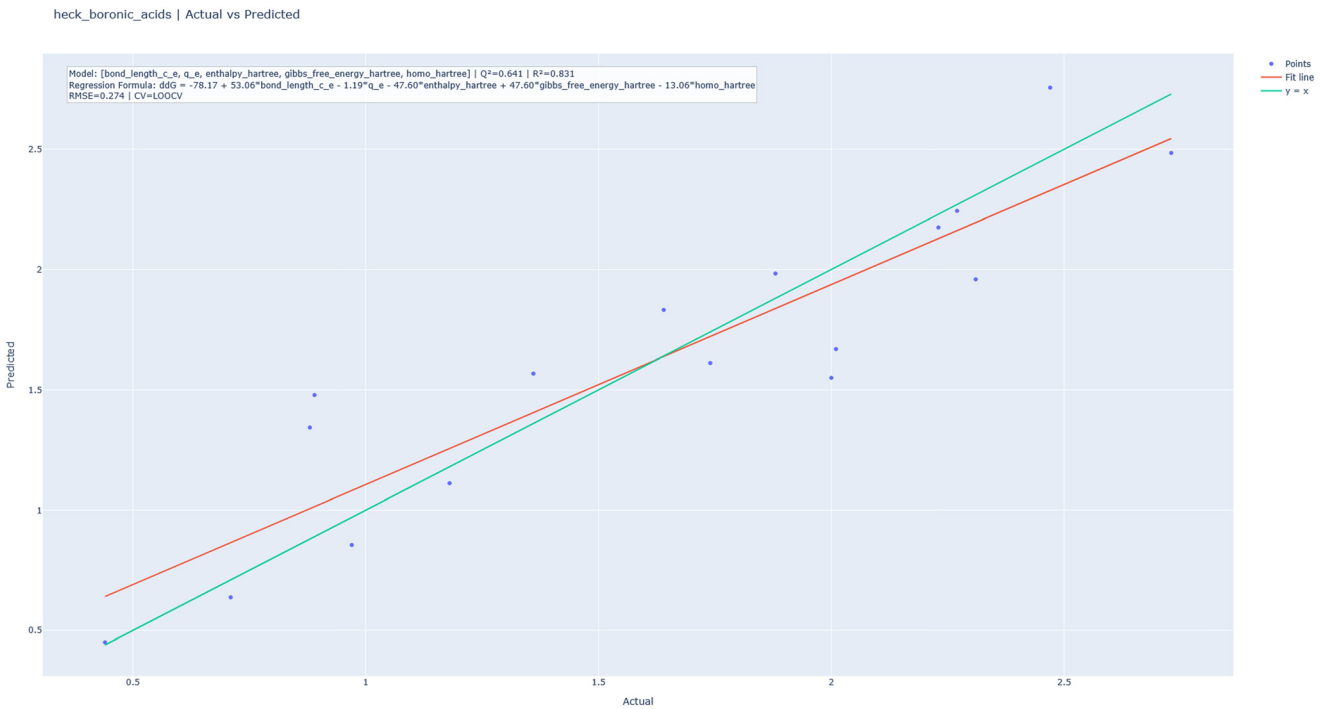


Fig. 4 Actual *vs.* predicted plot for the best-performing LOOCV model: redox-relay Heck coupling with boronic acids. The selected model maximizes Q^2_{LOO} under the constrained subset search.

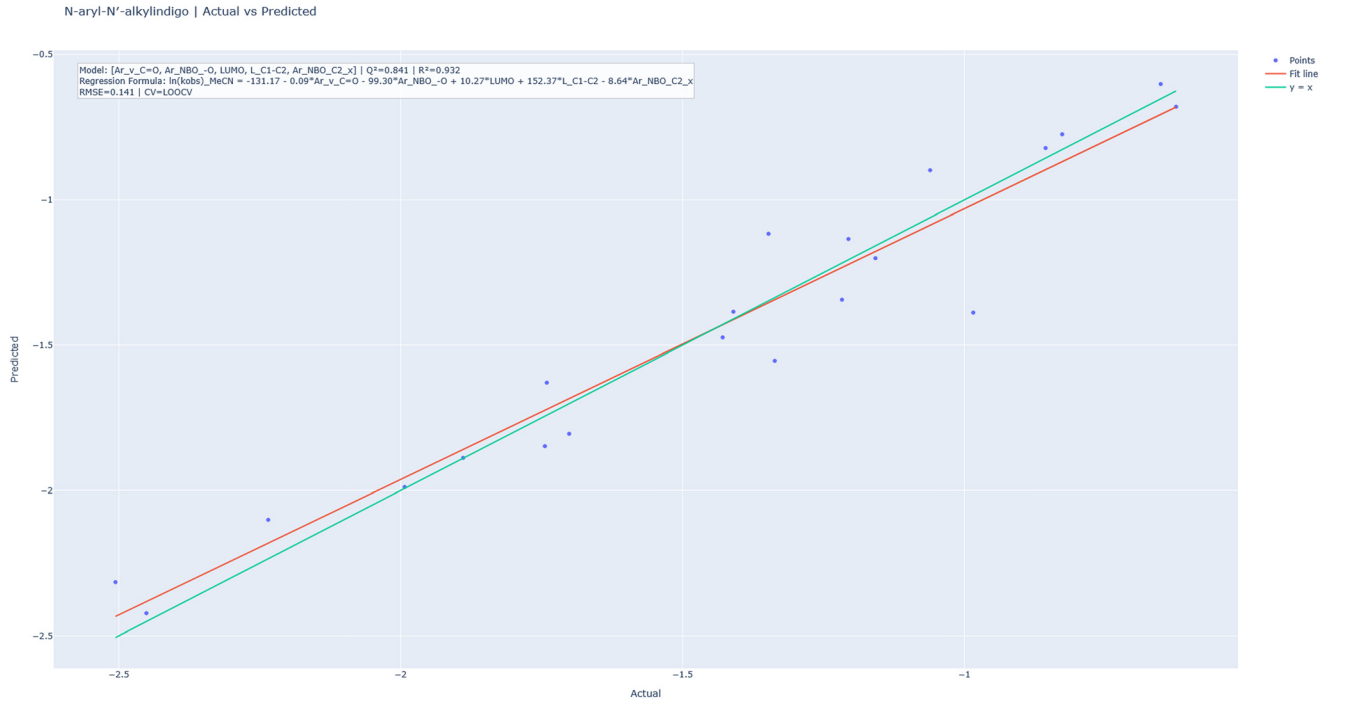


Fig. 5 Actual *vs.* predicted plot for the best-performing LOOCV model: thermal back-isomerization of *N*-aryl-*N*′-alkylindigo photoswitches. The selected model maximizes Q^2_{LOO} under the constrained subset search.

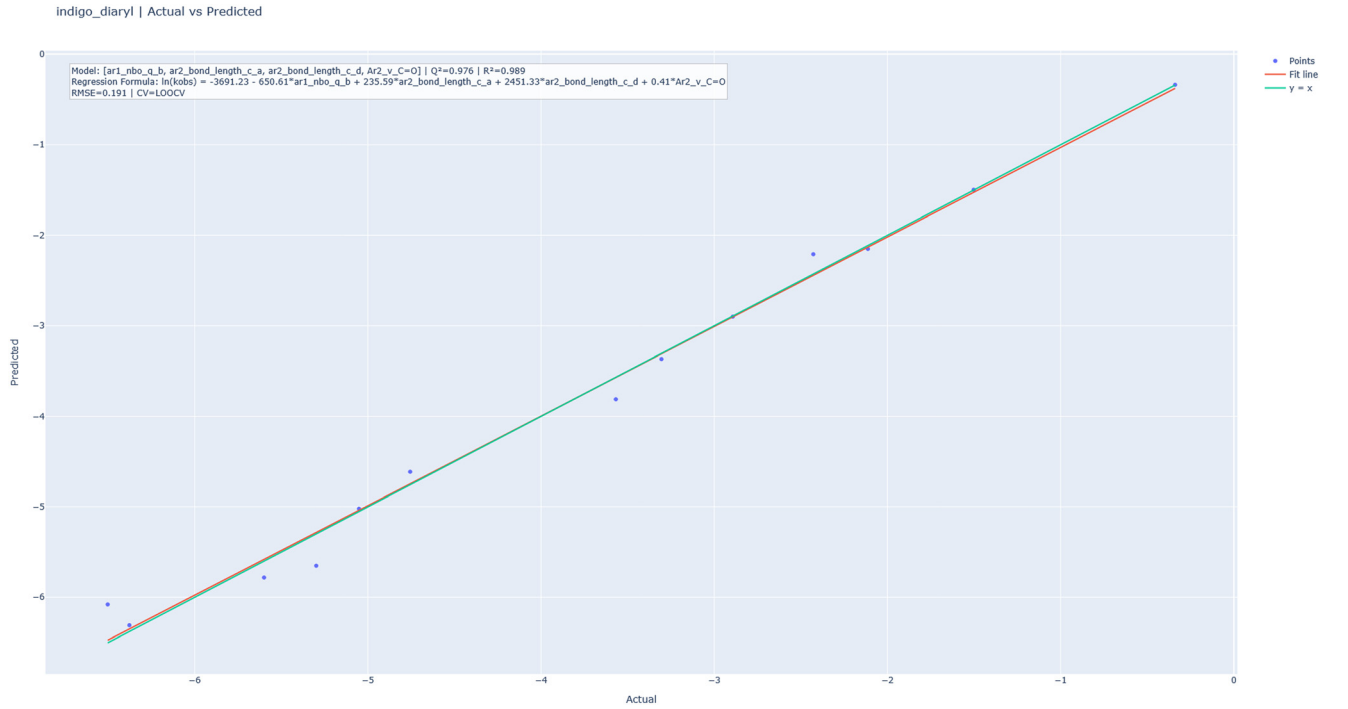


Fig. 6 Actual *vs.* predicted plot for the best-performing LOOCV model: thermal back-isomerization of *N*,*N*′-diarylindigo photoswitches. The selected model maximizes Q^2_{LOO} under the constrained subset search.

a on Ar_2 (ar2_bond_length_c_a), x_3 denotes the bond length between sites c and d on Ar_2 (ar2_bond_length_c_d), and x_4 denotes the carbonyl-related vibrational descriptor on Ar_2 (Ar2_v_C═O). This updated subset highlights a chemically interpretable combination of one localized electronic descriptor and three Ar_2-side structural/vibrational descriptors, suggesting

that both substituent-resolved charge distribution and carbonyl-related geometric signatures are important parameters in reflecting thermal back-isomerization behavior.

Conclusions

The *DFTDescriptorPipeline* provides a convenient workflow for establishing structure–property/reactivity relationships. By integrating fully automated descriptor extraction, flexible feature engineering, and multivariate linear regression into a unified open-source platform, this work accelerates the translation of quantum-chemical calculations into actionable, data-driven insights.

A key strength of the pipeline is its extensibility: each module is independently testable, and new descriptor types can be incorporated with minimal code modification. This modular design enables rapid adaptation to new molecular classes, reaction families, or emerging physical descriptors, helping the tool remain relevant as computational methods and chemical knowledge advance. The platform is also suitable for larger datasets, with descriptor extraction expected to scale approximately linearly with dataset size, while model-building cost is driven mainly by the size of the descriptor search space.

Broad adoption and continual growth of *DFTDescriptorPipeline* as a community-driven platform for data-centric chemistry can thus be expected. Future extensions include support for richer statistical/ML models, automated uncertainty quantification, and more direct feedback loops between quantum calculations and experiments. Interfacing with structure-input tools and quantum-chemistry execution platforms is expected to further increase scalability and usability. Additionally, although the current pipeline is limited to Gaussian output files, it can be extended to other DFT software packages through modest modifications to the parser layer. Compared with existing tools, our platform provides a comparable set of global and local electronic descriptors, as well as steric descriptors, and additional user-required parameters can be incorporated with minimal code modification. In contrast to descriptor-extraction-only workflows, our pipeline unifies descriptor extraction, MLR model construction, and model selection within a single workflow.

We anticipate that, by supporting high-throughput and fully reproducible modeling with minimal user intervention, the pipeline will lower the barrier for non-specialists to apply quantum-chemical data in experimental and industrial settings. Its design also supports integration with laboratory automation, database mining, and downstream machine-learning pipelines, making it a practical foundation for future closed-loop discovery platforms in catalysis, materials science, and medicinal chemistry.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data for this article, including code examples and the job/log files for the four case studies, are openly available in the project repository: https://github.com/peculab/DFTDescriptorPipeline.

Supplementary information (SI): containing further details of modelling and case studies as well as a step-by-step user guide. See DOI: https://doi.org/10.1039/d6re00049e

Acknowledgements

D. C.-Y. H. acknowledges support from WPI-ICReDD (Hokkaido University), Oklahoma State University, and JSPS (JP23K13734). Y.-C. H. and Y.-C. T. acknowledge financial support from National Taiwan Normal University (NTNU). The authors used ChatGPT (GPT-5.2) solely for English-language editing (grammar checking and sentence polishing). The tool was not used to generate, analyze, or interpret any experimental results, data, or figures. The study design, methodology, manuscript structure, and all scientific content are original and were developed by the authors, who take full responsibility for the final manuscript.

Notes and references

1 P. R. Wells, *Chem. Rev.*, 1963, **63**, 171–219.
2 L. P. Hammett, *J. Am. Chem. Soc.*, 1937, **59**, 96–103.
3 H. H. Jaffé, *Chem. Rev.*, 1953, **53**, 191–261.
4 R. W. Taft Jr., *J. Am. Chem. Soc.*, 1952, **74**, 3120–3128.
5 R. W. Taft Jr., *J. Am. Chem. Soc.*, 1953, **75**, 4538–4539.
6 C. B. Santiago, A. Milo and M. S. Sigman, *J. Am. Chem. Soc.*, 2016, **138**, 13424–13430.
7 C. B. Santiago, J.-Y. Guo and M. S. Sigman, *Chem. Sci.*, 2018, **9**, 2398–2412.
8 A. Verloop, *Drug Design*, Academic Press, New York, 1976, vol. III.
9 A. V. Brethomé, S. P. Fletcher and R. S. Paton, *ACS Catal.*, 2019, **9**, 2313–2323.
10 A. K. Jaiswal, P. Saha, J. Jiang, K. Suzuki, A. Jasny, B. M. Schmidt, S. Maeda, S. Hecht and C.-Y. D. Huang, *J. Am. Chem. Soc.*, 2024, **146**, 21367–21376.
11 A. M. Zuranski, J. Y. Wang, B. J. Shields and A. G. Doyle, *React. Chem. Eng.*, 2022, **7**, 1276–1284.
12 J. V. Alegre-Requena, S. V. S. Sowndarya, R. Pérez-Soto, T. M. Alturaifi and R. S. Paton, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2023, **13**, e1663.
13 B. C. Haas and M. A. Hardy, SigmanGroup/Get_Properties: Get_Properties_v1.0.3, *Zenodo*, 2024.
14 G. Luchini and R. S. Paton, *ACS Phys. Chem. Au*, 2024, **4**, 259–267.
15 O. S. Lee, M. C. Gather and E. Zysman-Colman, *Digital Discovery*, 2024, **3**, 1695–1713.
16 J. M. Crawford, C. Kingston, F. D. Toste and M. S. Sigman, *Acc. Chem. Res.*, 2021, **54**, 3136–3148.
17 W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle and E. V. Anslyn, *ACS Cent. Sci.*, 2021, **7**, 1622–1637.
18 C.-Y. Huang and S. Hecht, *Chem. – Eur. J.*, 2023, **29**, e202300981.

