Predicting Forward Activation Energies of Elementary Gas-Phase Reactions from Reactant and Product Structures Alone

A Reaction-Difference-Fingerprint and Tree-Ensemble Study on the ωB97X-D3/def2-TZVP Dataset

July 12, 2026

Abstract

Predicting the height of a reaction barrier without first locating its transition state (TS) is a central goal of data-driven chemical kinetics, because the TS search is the dominant computational cost in constructing reaction networks. Using the ωB97X-D3/def2-TZVP subset of the 2020 *Scientific Data* quantum-chemistry reaction dataset of Grambow, Pattanaik and Green [1], we build supervised regression models that predict the forward activation energy (E_a, kcal/mol) using *only* the atom-mapped reactant and product structures — transition-state geometries are never used as model input. Each reaction is encoded as a 2048-bit Morgan count fingerprint (radius 2) [2], and the reaction is represented by a *Reaction Difference Fingerprint*, $\mathrm{FP}_{\mathrm{diff}} = \mathrm{FP}_{\mathrm{products}} - \mathrm{FP}_{\mathrm{reactants}}$, with multimolecular species handled by summing the fingerprints of their components [3]. All 11,961 reactions were featurized with zero RDKit parse failures. We hold out 20% of the reactions (**random seed 42**; 9,568 train / 2,393 test) and tune Random Forest [4], histogram-based gradient boosting [5] and a Ridge baseline by 5-fold cross-validation, comparing against a physics-motivated Bell–Evans–Polanyi (BEP) ΔH baseline. The best model, a Random Forest, achieves a held-out mean absolute error (MAE) of **10.16 kcal/mol**, a root-mean-square error (RMSE) of **14.57 kcal/mol** and a coefficient of determination R^2 = **0.590**; **40.4%** (967/2393) of held-out reactions are predicted within **5 kcal/mol** of the reference barrier. The structural fingerprint model substantially outperforms the thermodynamics-only BEP baseline (MAE 16.32 kcal/mol, $R^2 = 0.233$), quantifying the value added by structural information beyond the reaction enthalpy. We report the full per-reaction predicted-versus-reference barriers for the test set and analyse where the model succeeds and fails, discussing the intrinsic limitations of count-difference fingerprints and concrete routes to closing the gap toward chemical accuracy.

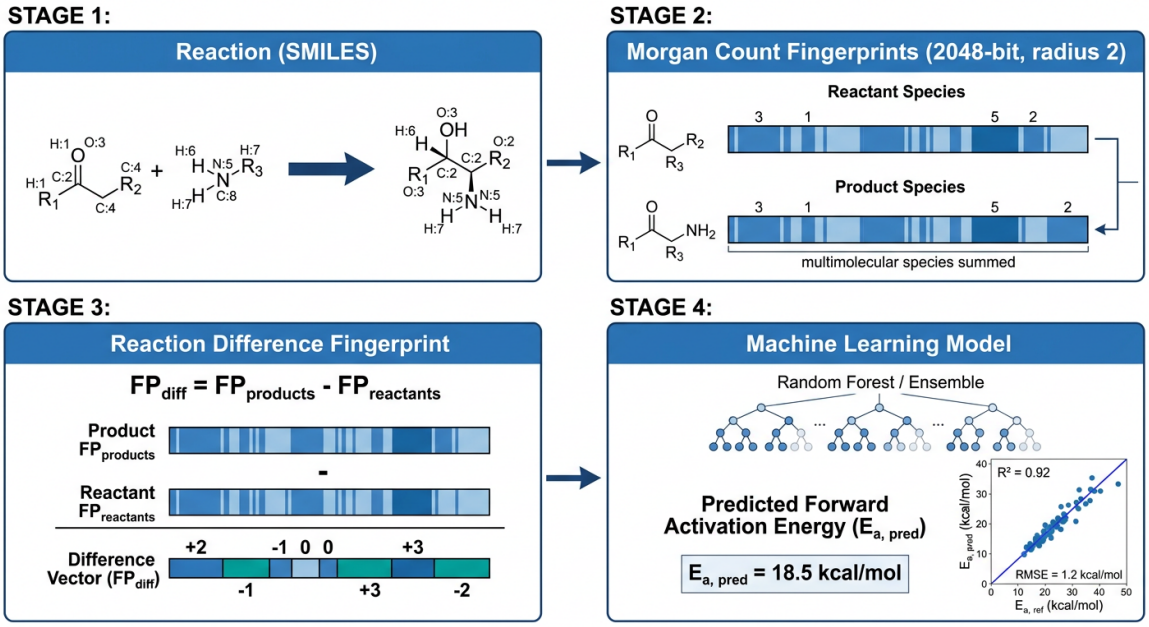

Figure 1: Graphical abstract. The modelling workflow proceeds from an atom-mapped reaction (Stage 1) to per-molecule Morgan count fingerprints (Stage 2, with multimolecular species summed), to a single Reaction Difference Fingerprint that encodes the net change in substructure counts (Stage 3), which is regressed onto the forward activation energy by a tree-ensemble model (Stage 4). No transition-state information enters the model.

1 Introduction

The activation energy E_a — the height of the free-energy or potential-energy barrier that a system must surmount to convert reactants into products — is the single most important quantity governing the rate of an elementary chemical reaction. Through transition-state theory [6], the rate constant depends exponentially on E_a, so an error of even a few kcal/mol translates into orders-of-magnitude uncertainty in predicted rates. Accurate barriers are therefore indispensable for building the detailed kinetic mechanisms that underpin combustion modelling, atmospheric chemistry, pyrolysis and reaction discovery, and automated mechanism generators such as RMG rely on large libraries of barrier estimates to assemble plausible reaction networks [7].

Obtaining E_a from first principles is expensive. It requires locating the transition state (TS) — a first-order saddle point on the potential-energy surface — which is far more costly and failure-prone than optimising the reactant and product minima that flank it. A single TS search may demand many gradient and Hessian evaluations, may converge to the wrong saddle point, and often has to be repeated across conformers. Because the endpoints of a reaction (its reactant and product structures) are comparatively trivial to specify and optimise, a long-standing question is how much of the barrier can be inferred from the endpoints alone. The classical Bell–Evans–Polanyi (BEP) principle and the broader family of linear free-energy relationships formalise the empirical observation that, within a reaction family, more exothermic reactions tend to have lower barriers, i.e. E_a correlates with the reaction enthalpy ΔH. Such relations are powerful but only locally valid, and a single global BEP line explains only a modest fraction of the variance across chemically diverse reactions.

Machine learning (ML) has emerged as a compelling alternative, learning flexible, nonlinear maps from molecular structure to quantum-chemical properties and thereby amortising the cost of expensive electronic-structure calculations across large datasets [8, 9]. The paradigm was established for equilibrium molecular properties by benchmark datasets such as QM9 [10] and

sharpened by delta-learning strategies that predict the correction from a cheap to an expensive level of theory [11]. Extending this idea from molecules to *reactions* requires a representation of the transformation itself. Grambow, Pattanaik and Green demonstrated that a directed message-passing neural network operating on reactant and product graphs could predict activation energies template-free, without any TS input [12]; Heid and Green later generalised this to a condensed-graph-of-reaction representation that predicts activation energies, enthalpies and other reaction properties within a single framework [13]. These graph models build on the directed message-passing architecture introduced for molecular property prediction by Yang et al. [14]. In parallel, higher-accuracy reference data (CCSD(T)-F12 barrier heights in the RDB7 database) have enabled models that approach coupled-cluster accuracy on the same chemical space [15, 16], and continuing work is refining the density-functional reference values themselves [17].

The data that made all of this possible is the 2020 *Scientific Data* descriptor by Grambow, Pattanaik and Green, *"Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry"* [1]. Using automated potential-energy-surface exploration, the authors generated a large, diverse set of elementary gas-phase reactions of small organic molecules (H, C, N, O; reactants drawn from the GDB-7 subset of the GDB-17 chemical universe [18]) and computed optimised geometries, frequencies, activation energies and reaction enthalpies for reactants, products and transition states. The dataset is released at two levels of theory — a broad B97-D3/def2-mSVP exploration of 16,452 reactions and a higher-accuracy ωB97X-D3/def2-TZVP set of 12,001 reactions — and is archived on Zenodo (record 3715478) [19]. It has since become a standard benchmark for reaction-barrier ML.

In this report we study the more accurate ωB97X-D3/def2-TZVP set and deliberately restrict ourselves to a *structure-only, endpoint-only* setting: the model may see the atom-mapped reactant and product SMILES, but never the transition-state geometry. Rather than a deep graph network, we adopt a transparent, computationally light pipeline built on classical cheminformatics fingerprints and tree-ensemble regressors, for three reasons. First, it establishes a strong, easily reproducible baseline against which more elaborate models can be judged. Second, difference fingerprints and gradient-boosted trees remain highly competitive on tabular chemical data and are trivial to deploy. Third, the pipeline exposes cleanly how far a purely topological, count-based description of the net bonding change can carry a barrier prediction. We report MAE, RMSE, the fraction of test reactions predicted within 5 kcal/mol, and the complete per-reaction predicted-versus-reference barriers on a 20% held-out set with a fixed random seed, and we benchmark throughout against a Bell–Evans–Polanyi ΔH baseline that captures the thermodynamic signal alone.

2 Data

We use the ωB97X-D3/def2-TZVP reaction table (`wb97xd3.csv`) from Zenodo record 3715478 [19]. Each row is one elementary reaction described by an integer index, an atom-mapped reactant SMILES (`rsmi`), an atom-mapped product SMILES (`psmi`), the forward activation energy E_a (`ea`, kcal/mol) and the reaction enthalpy ΔH (`dh`, kcal/mol). The reactions are unimolecular on the reactant side but may produce multiple product fragments. After loading, the table contains **11,961** reactions with five columns; this is within tolerance of the commonly cited ∼12,001 figure (a small number of records are absent from the released CSV relative to the raw log-file archive). We found no missing values, no full-row duplicates, and no duplicate (`rsmi`, `psmi`) reaction pairs.

The forward barriers span a wide range (Fig. 2a): mean 83.4 kcal/mol, standard deviation 22.8 kcal/mol, median 81.5 kcal/mol, minimum 10.7 kcal/mol and maximum 204.7 kcal/mol. There are no non-physical negative barriers; four reactions exceed a 3×IQR high-side fence ($>$ 188.7 kcal/mol) and are flagged as extreme but retained, since the dataset is a curated

quantum-chemistry benchmark rather than noisy experimental data. Reaction enthalpies are broadly distributed (mean 35.8 kcal/mol, range [−107.7, 173.7] kcal/mol; Fig. 2b). Consistent with a Bell–Evans–Polanyi trend, E_a and ΔH are positively but only moderately correlated (Pearson $r = 0.464$; Fig. 2c). This correlation is the entire signal available to the thermodynamic BEP baseline, and its modest strength foreshadows the ceiling on any ΔH-only predictor and motivates richer structural features.

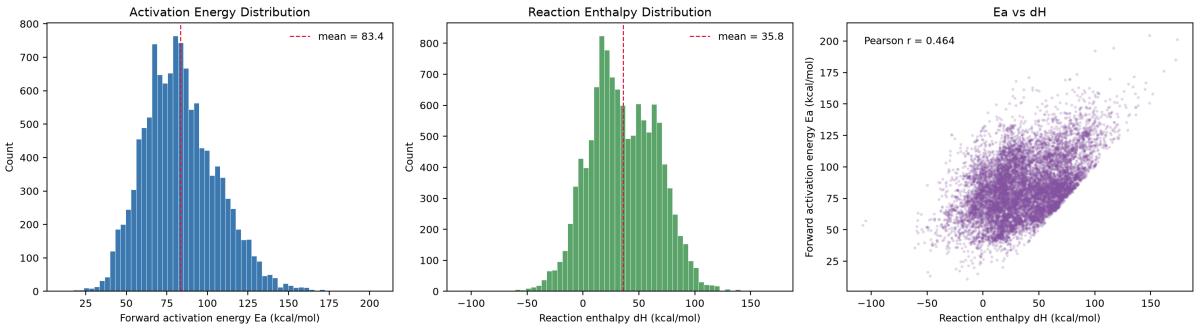

Figure 2: Exploratory analysis of the ωB97X-D3/def2-TZVP set (11,961 reactions). (a) Distribution of forward activation energies E_a (mean 83.4 kcal/mol). (b) Distribution of reaction enthalpies ΔH (mean 35.8 kcal/mol). (c) Joint distribution of E_a versus ΔH: the positive but scattered relationship (Pearson $r = 0.464$) is the Bell–Evans–Polanyi signal that the thermodynamic baseline exploits.

3 Methods

Figure 3 summarises the end-to-end pipeline. All code was run in Python 3.12 with RDKit 2026.03, scikit-learn [20], NumPy, pandas and Matplotlib, using a single fixed random seed (**42**) for the train/test split, cross-validation folds and every stochastic model.

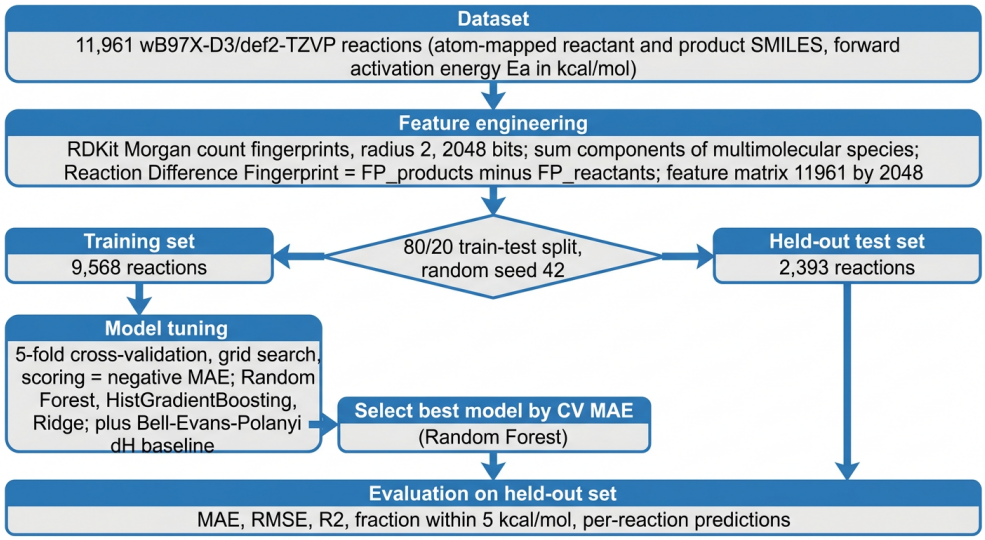

Figure 3: Experimental pipeline. Atom-mapped reactions are converted to Reaction Difference Fingerprints, split 80/20 with seed 42, and used to tune three regressors plus a Bell–Evans–Polanyi baseline by 5-fold cross-validation. The model with the lowest cross-validated MAE (Random Forest) is evaluated once on the untouched 2,393-reaction test set.

3.1 Feature engineering: Reaction Difference Fingerprints

The representation must map a variable-size reaction onto a fixed-length numeric vector while capturing the chemical change that a barrier "feels". We use circular Morgan fingerprints, the open-source realisation [21] of the extended-connectivity fingerprint (ECFP) family [2], which itself descends from Morgan's canonical atom-relabelling algorithm [22]. Our construction proceeds in four steps.

First, every molecule is parsed with RDKit's `MolFromSmiles`. Multimolecular species written as dot-separated SMILES (e.g. $A \to B + C$) are split into components, each of which is fingerprinted independently. Second, each component is encoded as a 2048-bit Morgan *count* fingerprint of radius 2 via `rdFingerprintGenerator.GetMorganGenerator`. We use count (rather than binary) fingerprints deliberately: a barrier depends on how *many* bonds of a given type are broken or formed, information that a binary on/off vector would discard. Radius 2 (equivalent to ECFP4) captures each atom's environment out to two bonds, a standard and robust choice for structure–property modelling. Third, the fingerprints of the components on each side of the reaction are summed to give a single reactant vector and a single product vector, $\mathrm{FP}_{\mathrm{reactants}} = \sum_i \mathrm{FP}(\mathrm{reactant}_i)$ and $\mathrm{FP}_{\mathrm{products}} = \sum_j \mathrm{FP}(\mathrm{product}_j)$. Summation is the natural pooling operation for count fingerprints because it preserves the total substructure inventory of a multi-fragment side. Fourth, the reaction is represented by the *Reaction Difference Fingerprint*

$$\mathrm{FP}_{\mathrm{diff}} \;=\; \mathrm{FP}_{\mathrm{products}} \;-\; \mathrm{FP}_{\mathrm{reactants}}, \tag{1}$$

a signed 2048-dimensional integer vector in which positive entries mark substructures created by the reaction and negative entries mark substructures destroyed. This difference construction is the reaction-fingerprint idea introduced by Schneider et al. [3] and used widely since, for example in the differential reaction fingerprint (DRFP) [23]; here it is instantiated with Morgan count features. Because atom-map indices do not alter the Morgan hashing (which encodes local heavy-atom environments), the atom mapping is used only as a consistency check and not as an information source.

Featurization succeeded for all **11,961** reactions with **zero** parse failures and zero atom-map inconsistencies. The dataset contains 18 reactant-side and 3,557 product-side multimolecular reactions, all handled by the summation rule above. The resulting feature matrix is $X \in \mathbb{Z}^{11961\times 2048}$ (`int32`); across the dataset $\mathrm{FP}_{\mathrm{diff}}$ entries range over $[-7, 6]$, every one of the 2048 columns is used at least once, and a reaction has on average ≈ 23 nonzero difference entries. The regression target is the forward activation energy $y = E_{\mathrm{a}}$.

3.2 Data split and models

We hold out 20% of the reactions as a test set using scikit-learn's `train_test_split` with `random_state=42`, giving **9,568** training and **2,393** test reactions. The test set is untouched during all feature construction, model selection and hyperparameter tuning, and is scored exactly once per model at the end.

On the training set we tune three regressors by 5-fold cross-validation, scoring by negative mean absolute error:

- **Ridge regression** (a linear baseline inside a `StandardScaler` pipeline), grid-searching the regularisation strength α;
- **Random Forest** [4], grid-searching the number of trees, maximum depth, the fraction of features considered per split, and the minimum samples per leaf;
- **Histogram-based Gradient Boosting** [5] (scikit-learn's `HistGradientBoostingRegressor`, the same efficient histogram/leaf-wise design popularised by XGBoost [24] and Light-

GBM [25]), grid-searching the learning rate, maximum number of boosting iterations, tree depth and L_2 regularisation, with early stopping.

We additionally fit a **Bell–Evans–Polanyi (BEP) baseline**: an ordinary least-squares linear regression of E_a on the reaction enthalpy ΔH alone. This baseline embodies the thermodynamic intuition and isolates exactly the incremental value contributed by structural features. After cross-validation, the model with the lowest CV MAE is selected as the primary model; all models (and the BEP baseline) are then evaluated on the held-out test set.

3.3 Evaluation metrics

On the held-out set we report the mean absolute error (MAE) and root-mean-square error (RMSE) in kcal/mol, the coefficient of determination R^2, and — following the accuracy criterion of interest for kinetics — the fraction of test reactions whose predicted barrier lies within 5 kcal/mol of the reference value. We also report the complete list of per-reaction predicted-versus-reference barriers for all 2,393 test reactions (§4.3; full CSV provided as a deliverable).

4 Results

4.1 Held-out test performance

Table 1 collects the held-out metrics for all models. The best cross-validated model is the **Random Forest** (CV MAE = 10.61 kcal/mol; hyperparameters `n_estimators`= 300, `max_depth`=None, `max_features`= 0.3, `min_samples_leaf`= 2). On the untouched 2,393-reaction test set it attains **MAE = 10.16** kcal/mol, **RMSE = 14.57** kcal/mol and $\mathbf{R^2 = 0.590}$, and predicts **40.4**% (967 of 2393) of reactions within 5 kcal/mol of the reference barrier. The median absolute error is only 6.91 kcal/mol — well below the mean — signalling a right-skewed error distribution dominated by a minority of large misses.

Table 1: Held-out test performance (2,393 reactions, seed 42). Lower is better for MAE, RMSE and median absolute error; higher is better for R^2 and the within-5 kcal/mol fraction. The Random Forest, selected by cross-validation, is the best model on every metric.

Model	MAE (kcal/mol)	RMSE (kcal/mol)	R^2	Within 5 kcal/mol (%)	Median |err| (kcal/mol)
Random Forest (best)	**10.16**	**14.57**	**0.590**	**40.4**	**6.91**
Histogram Gradient Boosting	11.83	15.60	0.530	28.8	9.27
Ridge (linear)	15.35	19.47	0.267	20.7	12.93
BEP ΔH baseline	16.32	19.92	0.233	16.0	14.27

Two comparisons stand out (Fig. 4). First, structural information matters: the Random Forest cuts the MAE of the thermodynamics-only BEP baseline from 16.32 to 10.16 kcal/mol (a 38% reduction), raises R^2 from 0.233 to 0.590, and more than doubles the within-5 kcal/mol hit rate from 16.0% to 40.4%. The net change in substructure counts therefore carries substantially more barrier-relevant information than the reaction enthalpy alone. Second, nonlinearity matters: both tree ensembles clearly beat linear Ridge regression on the same features, indicating genuinely nonlinear structure–barrier relationships that a linear model cannot capture. Notably, the linear Ridge model applied to the full difference fingerprint still edges out the BEP baseline, showing that even a linear read-out of structural change is more informative than ΔH.

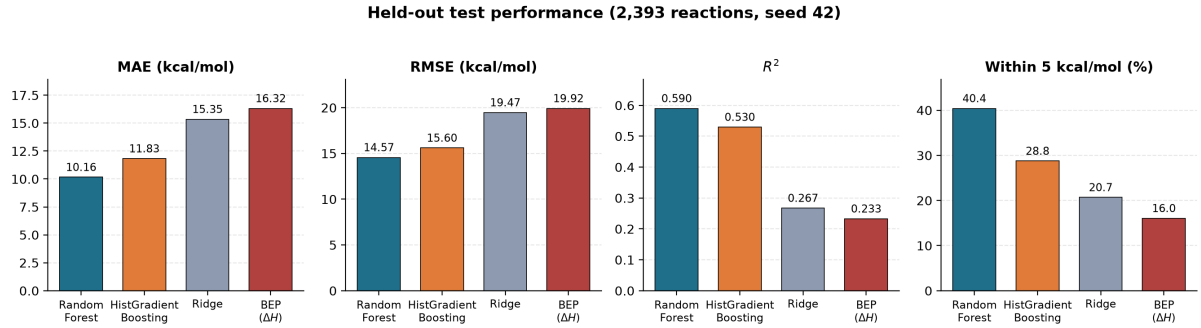

Figure 4: Model comparison on the held-out test set. Across all four metrics — MAE, RMSE, R^2 and the fraction of reactions predicted within 5 kcal/mol — the Random Forest (teal) leads, followed by histogram gradient boosting (orange), Ridge (grey) and the Bell–Evans–Polanyi ΔH baseline (red). Structural difference fingerprints dominate the thermodynamic baseline on every measure.

4.2 Predicted-versus-reference barriers

Figure 5 shows the Random Forest predictions against the reference barriers for all 2,393 test reactions. The predictions cluster along the $y = x$ line with a Pearson correlation of 0.776, and are essentially unbiased overall (mean signed error -0.03 kcal/mol). The model is most reliable in the data-dense core of the distribution ($\sim$60–100 kcal/mol), where the bulk of reactions lie. Two systematic effects are visible: predictions are compressed toward the mean at the extremes — barriers below $\sim$50 kcal/mol tend to be over-predicted and barriers above $\sim$130 kcal/mol under-predicted — which is the expected behaviour of an averaging tree ensemble in sparsely populated regions of feature space. Indeed the predicted barriers span only [44.8, 141.1] kcal/mol whereas the references span [19.9, 192.3] kcal/mol. Reassuringly, the model never predicts a non-physical negative barrier.

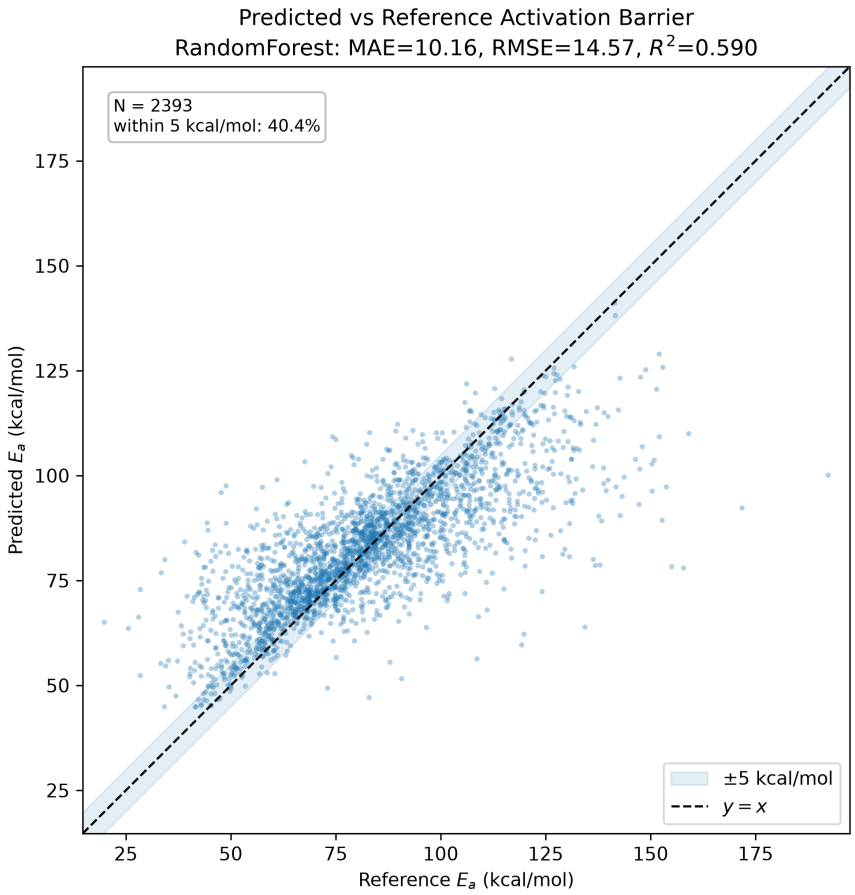

Figure 5: Predicted versus reference forward activation barriers (Random Forest, held-out test set, $N = 2393$). The dashed line is $y = x$ and the shaded band marks ± 5 kcal/mol; 40.4% of points fall inside the band. Predictions are near-unbiased in the central range and regress toward the mean at the extremes.

The shape of the error distribution is quantified in Figure 6. The signed errors (Fig. 6a) are sharply peaked at zero and centred there, but have heavy tails reaching beyond ± 40 kcal/mol; it is these tails that inflate the mean and RMSE relative to the median. The cumulative-accuracy curve (Fig. 6b) makes the practical picture concrete: 22.2% of test reactions are predicted within 2.5 kcal/mol, 40.4% within 5 kcal/mol, 63.2% within 10 kcal/mol, 76.5% within 15 kcal/mol and 85.4% within 20 kcal/mol. In other words, while sub-5 kcal/mol precision is achieved for a large minority, the model places roughly three-quarters of all reactions within 15 kcal/mol of the truth using nothing but reactant and product topology.

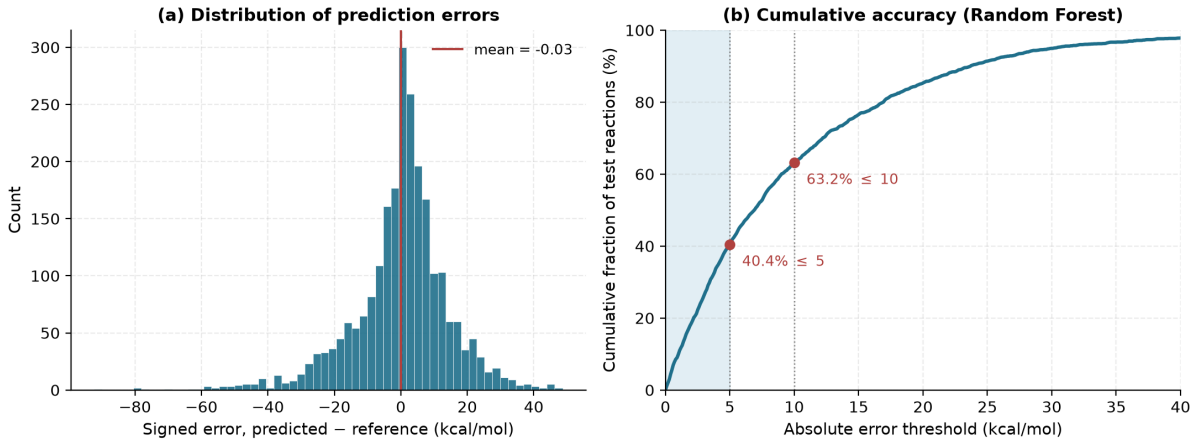

Figure 6: Error structure of the Random Forest predictions. (a) The signed-error histogram is centred at zero (mean −0.03 kcal/mol) with heavy tails. (b) Cumulative fraction of test reactions as a function of the absolute-error threshold; markers highlight 40.4% within 5 kcal/mol and 63.2% within 10 kcal/mol. The shaded region is the ≤ 5 kcal/mol zone.

Where do the large errors concentrate? Figure 7 stratifies the absolute error by reference barrier height. The error is smallest and the within-5 kcal/mol band densest for barriers in the ∼60–100 kcal/mol window that contains most of the training data; MAE degrades sharply for the rare very-low (< 40 kcal/mol) and very-high (> 140 kcal/mol) barriers, where the ensemble has seen few analogous reactions and therefore reverts toward the training mean. The single worst prediction in the test set is an extreme high-barrier reaction (reference 192.3 kcal/mol, predicted 100.1 kcal/mol; error 92.2 kcal/mol), an outlier for which no structurally similar training example exists. This data-density dependence, rather than any particular chemical motif, is the dominant driver of the error tail.

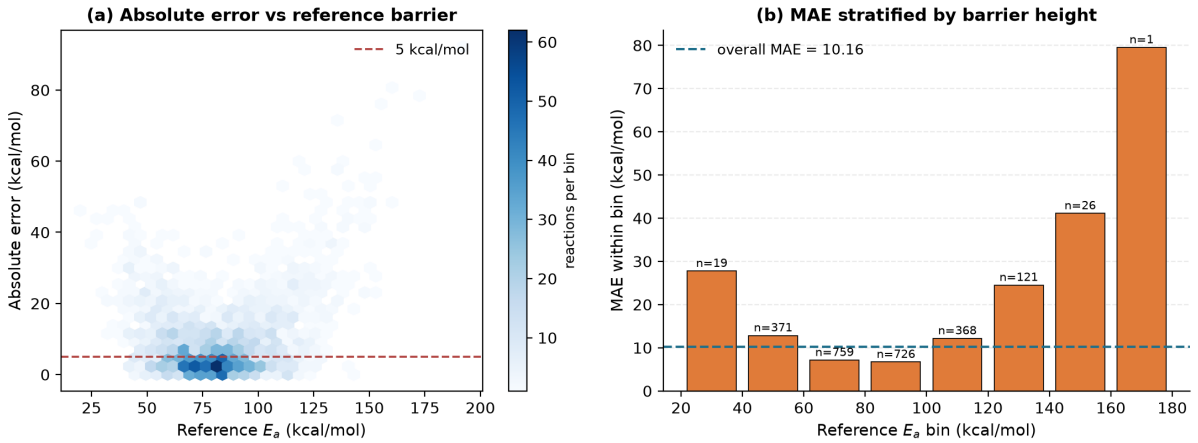

Figure 7: Errors are governed by training-data density. (a) Hexbin of absolute error versus reference barrier; the dashed line marks 5 kcal/mol. Errors are smallest in the densely sampled 60–100 kcal/mol region. (b) MAE within 20 kcal/mol reference-barrier bins (bin counts annotated); accuracy is best where data are abundant and degrades in the sparsely populated high- and low-barrier tails.

4.3 Per-reaction results for the test set

The complete per-reaction predicted-versus-reference barriers for all 2,393 held-out reactions are provided in the accompanying file `test_predictions.csv` (columns: reaction `idx`, reference `ea`, `predicted`, and `abs_error`). Table 2 lists a representative sample of 30 reactions drawn evenly across the reference-barrier range, from the lowest to the highest test barrier, so that both typical

and worst-case behaviour is visible. The sample reproduces the trends above: sub-5 kcal/mol agreement is common in the central range (e.g. reactions with references near 70–100 kcal/mol), while the lowest and highest barriers show the characteristic regression-toward-the-mean over- and under-prediction, culminating in the 192.3 kcal/mol outlier at the foot of the table.

Table 2: Representative per-reaction test predictions (30 of 2,393, ordered by reference barrier). "Signed" is predicted − reference; |err| is the absolute error. The full set is available in `test_predictions.csv`.

Reaction idx	Reference E_a (kcal/mol)	Predicted E_a (kcal/mol)	Signed (kcal/mol)	|err| (kcal/mol)
1110	19.9	65.0	+45.1	45.1
4392	45.7	82.2	+36.5	36.5
9586	50.9	75.2	+24.2	24.2
7681	55.2	65.4	+10.1	10.1
2270	58.0	69.5	+11.5	11.5
5494	60.5	73.2	+12.7	12.7
4250	63.3	93.1	+29.8	29.8
3424	65.8	64.9	−0.9	0.9
4484	68.0	72.8	+4.7	4.7
7692	69.9	71.3	+1.4	1.4
3532	71.7	76.3	+4.7	4.7
3116	74.1	83.2	+9.2	9.2
4352	76.3	81.1	+4.8	4.8
2751	77.8	86.4	+8.5	8.5
6408	80.2	81.4	+1.3	1.3
9243	81.9	87.5	+5.6	5.6
1011	83.5	90.1	+6.5	6.5
11846	85.3	81.5	−3.8	3.8
10504	87.3	88.5	+1.2	1.2
10005	89.8	91.2	+1.4	1.4
373	92.2	100.6	+8.5	8.5
2057	94.2	93.5	−0.7	0.7
9044	97.2	96.5	−0.7	0.7
251	100.6	103.8	+3.2	3.2
3482	105.1	97.8	−7.4	7.4
7488	108.5	96.6	−11.9	11.9
1111	113.1	96.7	−16.4	16.4
3034	118.7	115.6	−3.1	3.1
5173	127.4	122.6	−4.7	4.7
3921	192.3	100.1	−92.2	92.2

5 Discussion

5.1 Interpretation

The central result is that a deliberately simple, endpoint-only model — Morgan count difference fingerprints fed to a Random Forest — predicts forward activation energies of diverse gas-phase organic reactions with a held-out MAE of 10.16 kcal/mol and places 40.4% of reactions within 5 kcal/mol, using no transition-state information whatsoever. The comparison against the Bell–Evans–Polanyi baseline is the most instructive: because the BEP model has access to the exact reaction enthalpy ΔH (itself a quantum-chemical quantity), it is a genuinely competitive physics-based reference, yet the structural model beats it decisively (MAE 10.16 vs 16.32 kcal/mol; R^2 0.590 vs 0.233). This gap is a clean, quantitative statement that the pattern of bonds broken

and formed — the "reaction type" encoded in FP_{diff} — constrains the barrier far more tightly than the thermodynamic driving force alone. It also rationalises why purely thermodynamic barrier estimates are unreliable across chemically heterogeneous reaction sets: a single global BEP line simply cannot span the many reaction families present.

The superiority of the tree ensembles over Ridge regression on identical features confirms that the structure–barrier map is strongly nonlinear. Random Forests and gradient boosting can represent the interactions between co-occurring substructure changes (e.g. a ring opening that is accompanied by a specific heteroatom rearrangement) that a linear model treats additively and independently. The near-zero mean signed error and the tight central cluster in Figure 5 indicate a well-calibrated model in the data-rich regime; the heavy error tails are attributable to sparsity, not bias.

5.2 Limitations of difference fingerprints

Several limitations are intrinsic to the representation and explain the residual error. First, the fingerprint is *two-dimensional and topological*: it encodes which substructures change but not the three-dimensional geometry, strain, steric congestion or through-space electronic effects that can dominate a barrier. Two reactions with identical net bond changes but different conformational or stereoelectronic contexts receive nearly identical features yet may have very different transition states — a fundamental source of irreducible error in any endpoint-only, 2D approach. Second, the difference operation in Eq. 1 discards information: spectator substructures that appear on both sides cancel, and two different reactions can in principle map to the same difference vector, so the representation is not injective. Third, count fingerprints of radius 2 capture only local environments; longer-range effects and the identity of the specific bond being broken (as opposed to the count of bond-types changed) are only implicitly represented. Fourth, tree ensembles cannot extrapolate: their predictions are bounded by the range of training targets, which is precisely why the model compresses predictions toward the mean and fails on the rare very-high and very-low barriers (Fig. 7). Finally, the reference barriers themselves carry method error — the ωB97X-D3/def2-TZVP level is not exact, and recent work shows that its barrier heights deviate from coupled-cluster references in functional-dependent ways [15, 17] — so a fraction of the apparent model error reflects noise and inconsistency in the labels rather than model deficiency.

5.3 Comparison with the literature and context

The accuracy reported here is deliberately a baseline and should be read as such. Graph neural networks trained on this same chemical space reach substantially lower errors: the original template-free D-MPNN of Grambow et al. reported low-single-digit MAEs on related activation-energy data [12], the condensed-graph-of-reaction model of Heid and Green improved robustness across balanced and imbalanced reactions [13], and models trained on the higher-accuracy RDB7 CCSD(T)-F12 data approach coupled-cluster accuracy [15, 16]. Those models learn a task-specific, information-rich embedding of the reaction graph, whereas our fixed difference fingerprint is a generic, hand-crafted descriptor. The value of the present study is therefore not to compete with the state of the art but to quantify what a transparent, fast, fingerprint-based pipeline delivers, and to provide a reproducible reference point — complete with per-reaction predictions and a fixed seed — against which more complex models and richer representations can be measured.

5.4 Future directions

The analysis points to concrete, testable improvements. The most impactful is likely to be adding geometric information: even inexpensive reactant/product 3D descriptors, or ML-generated approximate transition-state geometries, have been shown to sharpen barrier predictions, and generative models for transition-state structures are now practical [26]. Within the endpoint-only setting, several lighter-weight avenues are worth pursuing: (i) concatenating $\mathrm{FP}_{\mathrm{diff}}$ with the reactant and product fingerprints separately (retaining spectator information that the difference cancels), and with cheap global descriptors, so that the thermodynamic and structural signals are combined explicitly rather than implicitly; (ii) substituting the hand-crafted fingerprint with a learned reaction representation such as the condensed graph of reaction [13] or a directed message-passing network [12, 14]; (iii) adding an explicit ΔH feature or a BEP prior as an inductive bias, so that the model reduces to a sensible thermodynamic estimate in data-sparse regions instead of reverting to the global mean; (iv) employing delta-learning to predict the correction from a cheap estimate to the reference barrier [11]; and (v) reporting calibrated per-prediction uncertainties (e.g. from the spread of the forest's trees) so that the unreliable high- and low-barrier extrapolations can be flagged automatically. Training and validating against the higher-accuracy RDB7 labels [15] would additionally reduce the label-noise contribution to the error and give a cleaner measure of representational limits.

6 Conclusion

Using only atom-mapped reactant and product structures from the ωB97X-D3/def2-TZVP subset of the Grambow–Pattanaik–Green dataset [1], and encoding each reaction as a 2048-bit Morgan count Reaction Difference Fingerprint, a cross-validated Random Forest predicts forward activation energies on a 20% held-out set (seed 42) with MAE = 10.16 kcal/mol, RMSE = 14.57 kcal/mol, $R^2 = 0.590$, and 40.4% of reactions within 5 kcal/mol of the reference barrier. The model comfortably outpaces a Bell–Evans–Polanyi thermodynamic baseline, establishing that the topological pattern of bond change is a far stronger predictor of the barrier than the reaction enthalpy alone, while its heavy-tailed errors expose the limits of 2D, count-based descriptors and non-extrapolating ensembles in the sparsely sampled barrier extremes. The complete per-reaction predicted-versus-reference barriers are released with this report as a reproducible, transition-state-free baseline for future barrier-prediction models.

Data and Code Availability

The underlying dataset is openly available from Zenodo record 3715478 (10.5281/zenodo.3715478) [19] and described in Ref. [1]. The engineered features, trained-model evaluation JSON, the full 2,393-row per-reaction prediction table (`test_predictions.csv`) and all figure-generation scripts accompany this report. All experiments use a single fixed random seed (42) for the split, cross-validation and model fitting to ensure exact reproducibility.

Reproducibility Summary

Dataset	ωB97X-D3/def2-TZVP set, 11,961 reactions (Zenodo 3715478)
Features	Morgan count fingerprints (radius 2, 2048 bits); $\mathrm{FP}_{\mathrm{diff}} = \mathrm{FP}_{\mathrm{prod}} - \mathrm{FP}_{\mathrm{react}}$
Split	80/20 train/test, `random_state=42` (9,568 / 2,393)
Model selection	5-fold CV, scoring = negative MAE
Best model	Random Forest (300 trees, max_features = 0.3, min_samples_leaf = 2)
Test MAE / RMSE	10.16 / 14.57 kcal/mol
Test R^2 / within-5 kcal/mol	0.590 / 40.4%

References

[1] Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data*, 7(1):137, 2020. doi: 10.1038/s41597-020-0460-4.

[2] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

[3] Nadine Schneider, Daniel M. Lowe, Roger A. Sayle, and Gregory A. Landrum. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *Journal of Chemical Information and Modeling*, 55(1):39–53, 2015. doi: 10.1021/ci5006614.

[4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[5] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[6] Henry Eyring. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2):107–115, 1935. doi: 10.1063/1.1749604.

[7] Connie W. Gao, Joshua W. Allen, William H. Green, and Richard H. West. Reaction mechanism generator: Automatic construction of chemical kinetic mechanisms. *Computer Physics Communications*, 203:212–225, 2016. doi: 10.1016/j.cpc.2016.02.013.

[8] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018. doi: 10.1038/s41586-018-0337-2.

[9] O. Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry*, 4(7):347–358, 2020. doi: 10.1038/s41570-020-0189-9.

[10] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014. doi: 10.1038/sdata.2014.22.

[11] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015. doi: 10.1021/acs.jctc.5b00099.

[12] Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Deep learning of activation energies. *The Journal of Physical Chemistry Letters*, 11(8):2992–2997, 2020. doi: 10.1021/acs.jpclett.0c00500.

[13] Esther Heid and William H. Green. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *Journal of Chemical Information and Modeling*, 62(9):2101–2110, 2022. doi: 10.1021/acs.jcim.1c00975.

[14] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59 (8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

[15] Kevin A. Spiekermann, Lagnajit Pattanaik, and William H. Green. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. *Scientific Data*, 9(1):417, 2022. doi: 10.1038/s41597-022-01529-6.

[16] Kevin A. Spiekermann, Lagnajit Pattanaik, and William H. Green. Fast predictions of reaction barrier heights: Toward coupled-cluster accuracy. *The Journal of Physical Chemistry A*, 126(25):3976–3986, 2022. doi: 10.1021/acs.jpca.2c02614.

[17] Xiao Liu, Kevin A. Spiekermann, Angiras Menon, William H. Green, and Martin Head-Gordon. Revisiting a large and diverse data set for barrier heights and reaction energies: best practices in density functional theory calculations for chemical kinetics. *Physical Chemistry Chemical Physics*, 27(25):13326–13339, 2025. doi: 10.1039/d5cp01181g.

[18] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012. doi: 10.1021/ci300415d.

[19] Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. `https://doi.org/10.5281/zenodo.3715478`, 2020. Dataset, Zenodo record 3715478.

[20] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] RDKit. RDKit: Open-source cheminformatics. `https://www.rdkit.org`, 2026. Version 2026.03.3; Morgan fingerprint generator (`rdFingerprintGenerator`).

[22] H. L. Morgan. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5 (2):107–113, 1965. doi: 10.1021/c160017a018.

[23] Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital Discovery*, 1(2): 91–97, 2022. doi: 10.1039/D1DD00006C.

[24] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 785–794. Association for Computing Machinery, 2016. doi: 10.1145/2939672.2939785.

[25] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 3146–3154. Curran Associates, Inc., 2017.

[26] Lagnajit Pattanaik, John B. Ingraham, Colin A. Grambow, and William H. Green. Generating transition states of isomerization reactions with deep learning. *Physical Chemistry Chemical Physics*, 22(41):23618–23626, 2020. doi: 10.1039/d0cp04670a.