Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: C. Pieters and A. Grinberg Dana, *Digital Discovery*, 2026, DOI: 10.1039/D6DD00113K.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D6DD00113K

Learning Rates: Predicting Rate Coefficients for Hydrogen Abstraction Reactions

Calvin Pieters† and Alon Grinberg Dana*,†,‡,¶

†*Wolfson Department of Chemical Engineering, Technion – Israel Institute of Technology, Haifa 3200003, Israel*

‡*Grand Technion Energy Program (GTEP), Technion – Israel Institute of Technology, Haifa 3200003, Israel*

¶*Boeing-Technion SAF Innovation Center, Technion – Israel Institute of Technology, Haifa 3200003, Israel*

E-mail: alon@technion.ac.il

View Article Online
DOI: 10.1039/D6DD00113K

Abstract

Accelerating the discovery of complex chemical systems, from sustainable aviation fuels to atmospheric models, requires the rapid determination of thousands of elementary rate coefficients, a task fundamentally bottlenecked by traditional, low-throughput transition-state searching. Here we develop a high-throughput digital pipeline and a reaction-aware geometric message-passing framework for predicting the three parameters of the modified Arrhenius equation directly from molecular structure. A dataset of ∼1,800 hydrogen-abstraction reactions was generated using automated workflows and high-level electronic-structure calculations. By incorporating reactive-atom distance (RAD) features – a novel data representation that solves the "spatial blindness" of standard molecular graphs – the model achieves a cross-validated median error of 0.285 dex (∼1.9x) in $k(T)$ across 300–3000 K. While accuracy is modestly lower in heteroatom-rich environments, the framework robustly captures the underlying structural trends and directly yields the complete Arrhenius parameter triplet, ensuring a rigorous, continuous temperature dependence across the entire evaluated range. These results establish reaction-aware representation learning as a scalable strategy to replace weeks of quantum chemical compute with near-instantaneous inference, providing a clear path for the data-driven acceleration of kinetic modeling.

View Article Online
DOI: 10.1039/D6DD00113K

1 Introduction

Understanding and accurately predicting the kinetics of chemical reactions is crucial to modeling complex chemical processes, such as fuel combustion and atmospheric chemistry.[1–5] Among the diverse classes of reactions, hydrogen-abstraction reactions—where a radical species (e.g. H· or ·OH) abstracts a hydrogen atom from a donor molecule, whether organic or inorganic (e.g. C_2H_6, H_2S, N_2H_2)—are particularly influential.[6–10] In combustion systems, these reactions govern chain propagation and radical interconversion,[11,12] significantly affecting ignition timing,[11–13] flame speeds[14–16] and pollutant formation.[17] Accurate prediction on a large-scale of hydrogen-abstraction rate coefficients is therefore critical, as these values are key inputs for quantitative models of combustion and atmospheric chemistry, often determining whether simulations yield reliable or misleading outcomes.

Determining the rate coefficients with high confidence is not trivial. These reactions often involve transient radicals[18] and are important at extreme conditions,[19–21] complicating direct experimental kinetic measurements and limiting the available data to narrow temperature ranges. While experimental determination of rate coefficients remains the gold standard, particularly for validating critical reactions, it is often impractical to characterize all important individual steps in complex kinetic networks. The time, cost, and difficulty of radical reaction experiments—especially those involving short-lived intermediates like ·OH or H· radicals—restrict experimental efforts to a limited set of major reactions.[22–24]

As experimental determination of rate coefficients remains limited by cost and scope,[25] quantum chemical methods combined with transition state theory (TST) have become the primary means of determining reaction kinetics.[2,26–28] In principle, this enables the construction of kinetic models purely from *ab initio* parameters, providing predictive insight even in the absence of experimental data. In practice, however, the generation of accurate rate coefficients across large reaction networks remains computationally demanding. Modern tools such as CREST,[29] EStokTP,[30] and ARC[31] have automated many steps—conformer generation, transition-state (TS) searching, and refinement—but the exploration of the vast TS

View Article Online
DOI: 10.1039/D6DD00113K

conformational space continues to be a major bottleneck. Even for reactions with only a few degrees of freedom (e.g., two key dihedral angles adjacent to the reaction zone), identifying the correct TS conformer can require dozens of high-level quantum calculations and still often demands expert supervision, making the process time-consuming, especially when scaled. TS searches typically involve iterative optimization or heuristic guidance and must be validated via intrinsic reaction coordinate (IRC) calculations.[32] Vibrational frequency analysis is then used to confirm a single imaginary mode and to compute partition functions for TST rate evaluation. For flexible molecules, multiple candidate TS geometries arising from distinct rotamers may each require separate optimization and validation, adding to the computational burden. To refine energetic accuracy, high-level single-point energy calculations such as CCSD(T) are usually performed on lower-level optimized geometries, but their steep scaling, $\sim O(N^7)$, confines their use to small systems or a limited subset of benchmark reactions. Consequently, in detailed kinetic mechanisms—such as those used in combustion—only a small fraction of reactions are computed at the highest levels of theory, while most are estimated using more approximate quantum chemical or empirical approaches.[2,33,34] This imbalance introduces substantial uncertainty, as missing or inaccurate rate coefficients can propagate unpredictably through reaction networks.[35,36] Thus, while *ab initio* kinetics provide a rigorous theoretical foundation, their broad application remains constrained by computational cost, scalability, and the complexity of exploring the TS landscape.

The low-throughput nature of the quantum computations involved creates an iterative bottleneck that prevents the rapid exploration of the thousands of individual rate coefficients required to model complex chemical systems, such as sustainable aviation fuels (SAF) or atmospheric oxidation pathways. To meet these demands, the field requires a transition from manual, case-by-case mechanistic studies toward scalable, data-driven kinetic modeling that can instantly generalize across diverse chemical spaces.

To bridge this gap, chemists have turned to systematic data-driven and knowledge-based methods to fill in missing kinetic data. For decades, rate rules and correlation formulas

View Article Online
DOI: 10.1039/D6DD00113K

have served as the workhorses of reaction rate estimation. Early empirical relationships such as the Evans–Polanyi correlation approximate activation energies as linear functions of reaction enthalpy.[37,38] Another foundational method is group additivity, pioneered by Sidney Benson, which estimates thermochemical properties—and later kinetic parameters—by summing contributions from molecular substructures.[39,40] Notably, Saeys et al. developed an *ab initio* group additivity scheme for hydrogen-abstraction reactions,[41] while Sumathi et al. applied group-based rules to estimate rate coefficients for alkane abstractions by H· and ·CH_3 radicals.[42]

Building on these ideas, the Reaction Mechanism Generator (RMG)[43] utilizes a rule-based rate estimation approach, whereby a new reaction is mapped onto the most specific applicable node in a hierarchical kinetic tree; if no exact match exists, RMG falls back to parent nodes or averages over sibling rules. The rules are organized in decision trees whose branching reflects increasing structural specificity. This scheme enables rapid and interpretable estimation of missing rate coefficients at negligible computational cost.[43,44] However, its accuracy ultimately depends on the coverage and granularity of the underlying training reactions and the expressiveness of the group definitions; extrapolation beyond the represented chemical space, or into environments where important non-local interactions are not explicitly encoded, can lead to unreliable estimates.[43,45] To address these challenges, RMG incorporates automated tree generation methods that construct and update rate-rule hierarchies directly from training reactions using the Subgraph Isomorphic Decision Tree (SIDT) algorithm.[45,46] This automation improves scalability and consistency across reaction families, though the estimator's performance remains limited by sparse data coverage and the local nature of the group-based representation. These ongoing limitations motivate a shift toward more flexible, data-driven approaches capable of learning reactivity patterns directly from the molecular structure.

Machine learning (ML) has rapidly emerged as a powerful tool to overcome the limitations of traditional kinetics calculations, enabling fast and accurate predictions of reaction param-

View Article Online
DOI: 10.1039/D6DD00113K

eters directly from molecular structure. Recent studies have demonstrated that graph-based neural networks (NNs), including message passing neural networks (MPNNs), can predict activation energies with errors as low as 2 kcal mol^{-1} to 3 kcal mol^{-1}, rivaling DFT-level accuracy.[47,48] For example, Grambow et al. applied a variant of the MPNN model, called DMPNN (Directed Message Passing Neural Network) to over 12,000 reactions and achieved remarkably accurate activation energy predictions across multiple reaction types.[47] More recently, Spiekermann et al. showed that the same DMPNN trained on ∼24,000 reactions could predict barrier heights to within ∼2.6 kcal mol^{-1} of coupled-cluster reference values - outperforming even high-quality DFT - while being many orders of magnitude faster to evaluate.[49] These successes demonstrate the game-changing potential of ML: once properly trained, such models can generalize to new reactions and instantly return chemical parameters that would otherwise require expensive, time-consuming quantum chemical calculations.

Recent studies have also focused on improving ML performance in data-scarce regimes, a common challenge in chemistry. Chang et al.[48] demonstrated that incorporating global molecular descriptors, such as overall charge or electron count, into graph neural networks (GNNs) can modestly improve barrier height predictions (though often loosely referred to as activation energies in the ML literature) when high-level training data are limited. They also explored strategies such as Δ-learning, where the model learns the difference between a low-level and a high-level calculation, achieving high accuracy with far few expensive computations.

Hydrogen-abstraction reactions have become a prime testing ground for these approaches because of their central role in combustion, atmospheric oxidation, and radical chemistry, coupled with the difficulty of obtaining accurate rate coefficients experimentally. While the Spiekermann et al.[49] work contains many hydrogen-abstraction reactions, it focuses on barrier heights (Ea), leaving the Arrhenius pre-exponential factor (A) and temperature exponent (n) in the modified (three-parameter) Arrhenius expression largely unexplored. A few recent studies have begun to address this gap, particularly for specific hydrogen-

View Article Online
DOI: 10.1039/D6DD00113K

abstraction families.

Yu et al.[8] developed a deep NN model to predict rate coefficients for hydrogen-abstraction reactions between alkyl esters and atomic hydrogen, successfully reproducing temperature-dependent rate coefficients from high-level calculations. In a separate study,[50] Yu et al. proposed a hybrid ML approach for $\cdot CH_3$ + alkane abstractions, combining gradient boosting decision trees (XGBoost) with a feed-forward NN to predict A, n, and Ea from bond dissociation energies and steric/electronic descriptors. Using six optimized descriptors, this hybrid model reduced the cross-validated RMSE from 0.239, Feedforward Neural Network (FNN) alone, to 0.152 and the average deviation from 74% to 42% on the prediction set, and achieved residual standard errors of $0.07 - 0.16$.

Xia et al.[51] extended the hybrid XGB–FNN framework to $\cdot OH/HO_2\cdot$ + alkane abstractions, training on thousands of TST-derived rate coefficients with descriptors including bond dissociation energies, steric factors, and molecular symmetry numbers. For $\cdot OH$ + alkane hydrogen abstraction, this hybrid model achieved median leave-one-out deviations mostly below 100% with training, validation and prediction set errors of 46.2%, 45.4%, and 89.1% respectively - meeting the "accurate" threshold by combustion-modeling standards.[52] Performance was weaker for $HO_2\cdot$ + alkane reactions, with corresponding deviations of 148–191%, reflecting the smaller, less diverse training set and the need to extrapolate to larger, more complex molecules. Importantly, the $\cdot OH$ model reproduced well-known site-specific reactivity patterns (primary $<$ secondary $<$ tertiary barrier heights), consistent with prior *ab initio* studies, whereas the $HO_2\cdot$ model predicted more uniform reactivity across sites. Both models generalized reasonably to unmeasured alkanes up to $C_{16}H_{34}$ and produced temperature trends that fit smoothly to a modified three-parameter Arrhenius form, indicating physically plausible extrapolation beyond the training domain.

Despite this progress, four persistent challenges limit the deployment of ML-predicted H-abstraction rate coefficients in large-scale kinetic models: (i) severe data scarcity, with most datasets restricted to a single radical and/or substrate class; (ii) incomplete prediction of the

View Article Online
DOI: 10.1039/D6DD00113K

Arrhenius triplet, with many studies focusing on Ea alone; (iii) limited generalization across radical–substrate combinations, with accuracy dropping sharply outside the training domain; (iv) Lack of physics-constrained rate prediction frameworks enforcing Arrhenius consistency. Addressing these challenges is essential for producing physically reliable, broadly applicable rate predictions that can replace costly high-level calculations in combustion and atmospheric chemistry.

Here, we present a graph neural network (GNN) framework for predicting the full Arrhenius triplet (A, n, E_a) for hydrogen-abstraction reactions, trained on a new database of ∼1,800 diverse reactions. TS geometries were optimized at the ωB97X-D/def2-TZVP level of theory, followed by single-point energy refinements using DLPNO-CCSD(T)-F12/cc-pVTZ-F12 to obtain high-accuracy barrier heights. The resulting dataset spans a broad range of radicals and substrates, including carbon-, oxygen-, nitrogen-, and sulfur-centered radicals and both saturated and unsaturated donor substrates.

While topological representations provide a strong foundation, standard 2D molecular graphs often fail to capture the subtle electronic and spatial effects governing TS structures and energetics. To bridge this gap, recent studies have successfully integrated quantum-mechanical (QM) descriptors such as on-the-fly predicted atomic charges, bond orders, and Fukui indices as explicit node or edge features in GNNs to improve barrier prediction and regioselectivity.[53–55] Furthermore, alternative reaction representations, such as Condensed Graphs of Reaction (CGR) and unified reaction-level graphs have been shown to consistently outperform reactant-only baselines by explicitly encoding the mechanistic transformation into the graph structure.[56,57] Concurrently, the rise of 3D-aware architectures, which incorporate explicit reactant/product coordinates or extract learned descriptors from generated TS geometries, has demonstrated the critical importance of spatial information for activation energy prediction.[58,59] Building on these methodological foundations, the present approach provides a highly targeted, reaction-aware geometric representation that captures essential TS physics without the substantial computational overhead or noise sensitivity

View Article Online
DOI: 10.1039/D6DD00113K

associated with full 3D equivariant message passing or on-the-fly coordinate generation.

The model is based on the Directed Message Passing Neural Network (DMPNN) architecture, incorporating a physics-informed Arrhenius output head that predicts (A, n, E_a) and deterministically computes $\ln k(T)$ over a sampled temperature range during training, thereby enforcing internal curve-fit consistency. We also explore the Communicative MPNN (CMPNN)[60] as an alternative encoder, enabling a direct comparison of different message-passing strategies. By combining molecular graph representations with a multi-target kinetic loss, our framework predicts the complete Arrhenius parameter set and captures continuous structure–reactivity relationships across diverse radical–substrate families.

2 Methods

2.1 Data Set Construction

We generated a database of hydrogen-abstraction reactions (see general form in Eq. 1) spanning diverse donor/acceptor chemistries. Approximately 1,800 unique reactions were computed using a composite electronic-structure protocol. All high-throughput workflows were automated with the Automated Rate Calculator (ARC)[31] tool. Crucially, the pipeline computes site-specific rate coefficients, and the dataset contains numerous instances of competing abstraction channels by the same reactants.

For each reaction, ARC identified the lowest-energy conformers of all reactant and product species using a search algorithm that uses random RDKit[61] MMFF94s[62] conformers, and performed geometry optimizations and harmonic vibrational frequency calculations at the ωB97X-D/Def2-TZVP level of theory. All Density-Functional Theory (DFT) calculations were carried out using Gaussian 09.[63] These vibrational frequencies were used to compute zero-point energy correction.

View Article Online
DOI: 10.1039/D6DD00113K

$$R_1H + R_2\cdot \rightleftharpoons R_1\cdot + R_2H \quad (1)$$

Transition-state (TS) guesses were generated using two independent workflows, ARC heuristics and AutoTST.[64] Conformational sampling was performed with CREST[29,65–69] while constraining reactive-region distances. All TS candidates were optimized at the ωB97X-D/Def2-TZVP level and verified by a single imaginary frequency corresponding to hydrogen transfer. Details of the ARC hydrogen-abstraction heuristic are provided in Section **??**.

For all optimized reactants, products, and transition states (TSs), high-level single-point energies were computed using the DLPNO-CCSD(T)-F12 method[70] with the cc-pVTZ-F12 orbital basis set,[71] as implemented in ORCA (version 5.4).[72,73] The RIJCOSX[74] approximation was used to accelerate the Hartree-Fock exchange and Coulomb integrals. For density fitting, the aug-cc-pVTZ/C auxiliary basis[75] (for correlation) and the def2/J auxiliary basis[76] (for Coulomb fitting) were employed. The cc-pVTZ-F12-CABS complementary auxiliary basis[77] was used for the resolution of the identity in the F12 integrals and for the CABS singles correction to the Hartree-Fock energy.

Confirmed TS geometries, together with all optimized species, DFT vibrational frequencies, and DLPNO-CCSD(T)-F12 single-point energies, were subsequently processed using Arkane.[78] Arkane combines the high-level electronic energies with DFT-derived zero-point and thermal corrections to compute thermochemical properties and transition-state-theory (TST) rate coefficients. Internal rotors were not treated explicitly; for practical high-throughput workflow robustness, all species were modeled within the rigid-rotor harmonic-oscillator (RRHO) approximation. Quantum tunneling effects were incorporated using the Eckart correction.[79] External symmetry numbers were determined by Arkane.[78]

TST rate coefficients $k(T)$ were computed and fitted to the modified Arrhenius expression to obtain the pre-exponential factor A, temperature exponent n, and activation energy

View Article Online
DOI: 10.1039/D6DD00113K

E_a via nonlinear least-squares regression over the 300–3000 K temperature range. The resulting dataset, including optimized geometries, vibrational frequencies, high-level electronic energies, and fitted Arrhenius parameters, is publicly available on Zenodo[80]

2.2 Molecular Representations

We progressively enriched molecular graph representations with additional chemical and geometric information, enabling controlled ablation studies that isolate the contributions of electronic descriptors, global molecular conformation, and reaction-centered geometry. All representations share a common graph backbone and differ only in the auxiliary features provided to nodes and edges.

Each hydrogen-abstraction reaction was represented as a two-component molecular graph: the hydrogen donor (R_1H) and the acceptor (R_2H), both taken here in the H-atom bearing form (see Eq. 1). Optimized 3D geometries from ARC were stored in Structure Data Files (SDF format), ensuring consistent atom and bond indexing. Baseline atom and bond features followed the Chemprop[81] D–MPNN featurization scheme, implemented using RDKit.[61] Our training and evaluation pipeline builds on the Chemprop codebase and extends it to multi-component reaction inputs and Arrhenius-parameter prediction. Atom features included atomic number, number of directly bonded neighbors, formal charge, chirality, hydrogen count, hybridization, aromaticity, and normalized atomic mass. Bond features comprised null-bond indicators, bond order, conjugation, ring membership, and stereochemistry. These features served as inputs to the MPNN.

Beyond the standard Chemprop atom features, we considered an atom-augmented representation that incorporated per-atom electronic descriptors derived from the quantum chemical computations. Specifically, we included the Mulliken atomic charge, the atomic polar tensor charge, and the Mulliken spin-density magnitude for open-shell species, which together provided a local, geometry-independent description of charge distribution and radical character.

View Article Online
DOI: 10.1039/D6DD00113K

In addition, atoms were annotated with binary role indicators encoding the known reaction center for hydrogen abstraction. The donor heavy atom, acceptor heavy atom, and transferring hydrogen were explicitly identified; the hydrogen was defined as H_{donor} in the reactants and as H_{acceptor} in the products. First-neighbor atoms of the donor and acceptor centers were also labeled. These role flags were concatenated into the atom feature vector. By explicitly flagging the breaking and forming bonds, this geometry-free baseline topological representation fully specifies the reaction mapping. It functions similarly to the atom-mapped CGR approach used in reaction property prediction models.[82] Consequently, performance gains observed in our geometry-enhanced configurations can be attributed to the inclusion of 3D spatial information rather than to the mere identification of the reaction center.

2.2.1 Local Features

Standard molecular graphs are essentially 'spatially blind' to the critical 3D structural information of the TS, such as bond-breaking and bond-forming distances. We solve this by developing Reactive-Atom Distance (RAD) descriptors, a novel data representation that explicitly injects reaction-aware geometric context into the GNN. This allows the model to perceive the local structural environment of the reaction center while retaining the scalability of a message-passing architecture. We incorporate 3D structural information through two complementary geometric channels: RAD and edge-level continuous geometric features. All geometric quantities are computed from the DFT-optimized ground-state reactant structures (R1H and R2H), ensuring consistency between geometric descriptors and the kinetic labels.

Seven feature configurations were evaluated, organised along two orthogonal design axes — RAD treatment (none / full-graph / path-restricted) and edge-level geometry (off / on) — together with a Baseline reference. (i) Baseline: standard DMPNN representation with atom and bond features but no continuous geometric information. (ii) Atom: extends the Baseline with per-atom physicochemical descriptors and role-anchored binary indicators. (iii)

View Article Online
DOI: 10.1039/D6DD00113K

RA: adds full-graph reaction-anchored RAD geometric descriptors (radial, angular, dihedral) computed for every atom reachable from H_{ref}. (iv) Path: same as RA but with RAD geometric quantities defined only along simple covalent paths between H_{ref} and the focus atom. (v) Geom: omits RAD descriptors and instead provides edge-level continuous geometric features (bond distances, bond angles, dihedral angles) derived from the optimized reactant structures. (vi) RA+Geom: the union — full-graph RAD plus edge-level continuous geometry, with RAD values and existence indicators masked beyond four graph hops from H_{ref}. (vii) Path+Geom: path-restricted RAD plus edge-level continuous geometry, again with RAD masked beyond four graph hops. Table 1 summarises these configurations.

Table 1: Overview of feature configurations evaluated. RAD scope refers to whether reaction-anchored geometric descriptors are computed for all atoms reachable from H_{ref} (full graph) or only for atoms along simple covalent paths to H_{ref} (path-restricted). When edge-level geometry is enabled, RAD geometric values and existence indicators are masked beyond four graph hops from H_{ref}.

Config.	Mode	Atom-augmented	Node-level RAD	RAD scope	Edge-level geom.
(i)	Baseline	–	–	–	–
(ii)	Atom	✓	–	–	–
(iii)	RA	✓	✓	full graph	–
(iv)	Path	✓	✓	path-restricted	–
(v)	Geom	✓	–	–	✓
(vi)	RA+Geom	✓	✓	full graph (4-hop masked)	✓
(vii)	Path+Geom	✓	✓	path-restricted	✓

The RAD descriptors encode local abstraction geometry in a reaction-centered reference frame. The reactive hydrogen, H_{ref}, is identified via reaction mapping. For each atom v in the molecular graph, geometric quantities are computed relative to H_{ref}. Pivot_1 is the heavy atom covalently bonded to H_{ref}, and Pivot_2 is a heavy neighbor of Pivot_1 selected by minimal graph distance to H_{ref}. For each focus atom v, we compute the radial distance

$$r = \|H_{\text{ref}} - v\|,$$

View Article Online
DOI: 10.1039/D6DD00113K

the bond angle

$$\theta = \angle(H_{\text{ref}}, \text{Pivot}_1, v),$$

and the dihedral angle

$$\tau = \angle(H_{\text{ref}}, \text{Pivot}_1, \text{Pivot}_2, v).$$

Bond angles θ are retained in their raw angular form. Dihedral angles τ are represented using sine and cosine components to preserve periodicity and avoid discontinuities at $\pm\pi$.

In the path-restricted RAD variant, geometric quantities are defined only along simple covalent paths between H_{ref} and v. Let P denote the simple path between these atoms. The radial distance r is retained only if $|P| = 2$, the bond angle θ only if $|P| = 3$, and the dihedral τ only if $|P| = 4$, using intermediate vertices in canonical order as geometric pivots. Quantities that do not satisfy these path-length constraints are set to zero and accompanied by binary existence indicators.

To assess the spatial extent of reactive geometry, we introduce a locality-masked RAD variant. Let $d_{\text{graph}}(H_{\text{ref}}, v)$ denote the shortest-path bond distance between H_{ref} and atom v. For atoms satisfying

$$d_{\text{graph}}(H_{\text{ref}}, v) > h,$$

with $h = 4$ bonds in our experiments, RAD quantities and their existence indicators are set to zero. Base atom features remain unchanged. In the global variant, RAD descriptors are retained wherever structurally defined.

In the edge-geometric configurations, continuous geometric features are constructed for each directed bond $i \to j$ from the 3D reactant geometry (Fig. 1). The Euclidean bond length d_{ij} is expanded using a radial basis function (RBF) representation,

$$\phi_k(d_{ij}) = \exp\left[-\gamma(d_{ij} - \mu_k)^2\right],$$

where $\{\mu_k\}$ are fixed centers spanning a predefined distance range.

View Article Online
DOI: 10.1039/D6DD00113K

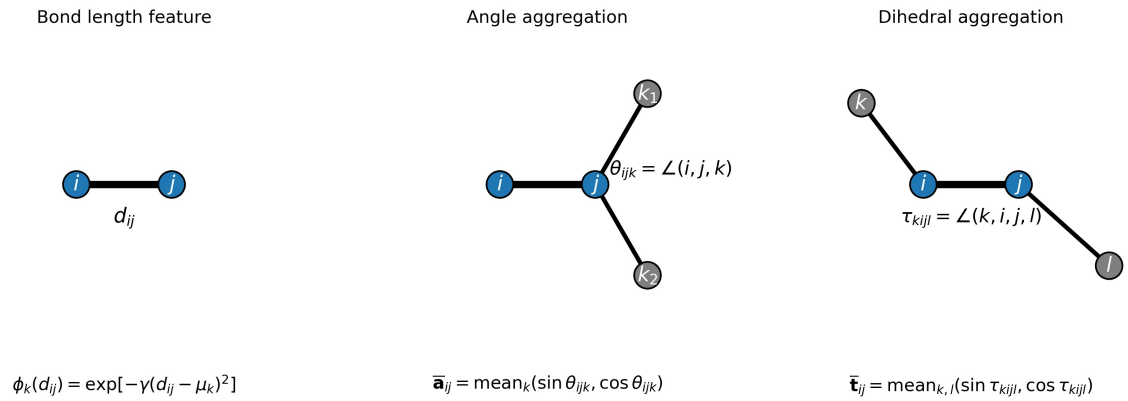

Figure 1: Construction of edge-level geometric features for a directed bond $i \to j$. Bond length is expanded via an RBF basis, while angular and torsional information are encoded using sine–cosine representations and aggregated by mean pooling over neighboring atoms. The resulting components are concatenated to form the directed-edge feature vector.

Local angular structure is incorporated by aggregating bond angles $\theta_{ijk} = \angle(i,j,k)$ around the terminal atom j using mean-pooled sine–cosine encodings over neighboring atoms k. Torsional structure about the central bond (i,j) is incorporated analogously by mean pooling sine-cosine encodings of dihedral angles $\tau_{kijl} = \angle(k,i,j,l)$ over valid (k,l) combinations. A detailed mathematical specification of the neighbor sets and pooling operations is provided in Section **??**.

2.2.2 Global Features

Global molecular descriptors were evaluated as an optional augmentation to the pooled graph embedding produced by the underlying MPNN architecture (DMPNN or CMPNN). When enabled, global descriptors were concatenated to the graph embedding immediately prior to the prediction head.

Two classes of global representations were considered. First, circular Morgan fingerprints (ECFP) were computed with radius $r = 2$ and length 2048 using RDKit, evaluated in both binary (bit) and count (occurrence) variants.[83] Second, we used 200-dimensional RDKit2D descriptors generated via Descriptastorus.[84] The RDKit2D features were normalized using empirical cumulative distribution functions to ensure consistent scaling across heterogeneous

View Article Online
DOI: 10.1039/D6DD00113K

descriptors. This design enabled a direct evaluation of whether whole-molecule context provides complementary information beyond local reactive and geometric descriptors.

2.3 Model Architectures

We evaluated two message-passing encoders for molecular graph representation learning: DMPNN,[85] which propagates messages along directed bonds to prevent backtracking, and CMPNN,[60] which extends DMPNN by enabling atom–bond communication during message updates. Each encoder was tested in two configurations: a dual-encoder mode, where donor and acceptor molecules are processed by separate networks, and a shared-encoder mode with weight sharing between components.

From the resulting donor and acceptor embeddings, denoted h_d and h_a, respectively, we constructed bidirectional composite vectors $h_{d\to a} = [h_d; h_a]$ and $h_{a\to d} = [h_a; h_d]$. Because the hydrogen-abstraction reaction family is formally reversible yet chemically asymmetric with respect to radical localization and bond polarity, we evaluated both order-aware and order-invariant treatments of the donor–acceptor pairing. Both DMPNN and CMPNN encoders were evaluated under each order treatment.

2.3.1 Order-aware and Order-invariant Configurations

In the order-aware configuration, donor and acceptor embeddings are treated as distinct inputs reflecting the directional nature of hydrogen-abstraction reactions. Each ordered concatenation is passed to a separate multilayer perceptron head, allowing the model to learn asymmetric representations for the forward and reverse directions. By contrast, the order-invariant configuration enforces symmetry with respect to donor–acceptor ordering. The two directional concatenations are symmetrically averaged to yield a representation invariant to donor–acceptor ordering. Schematic illustrations of these architectures are provided in Figures 2 and 3.

Each head predicted the three modified Arrhenius parameters (A_{10}, n, E_a), where $A_{10} =$

View Article Online
DOI: 10.1039/D6DD00113K

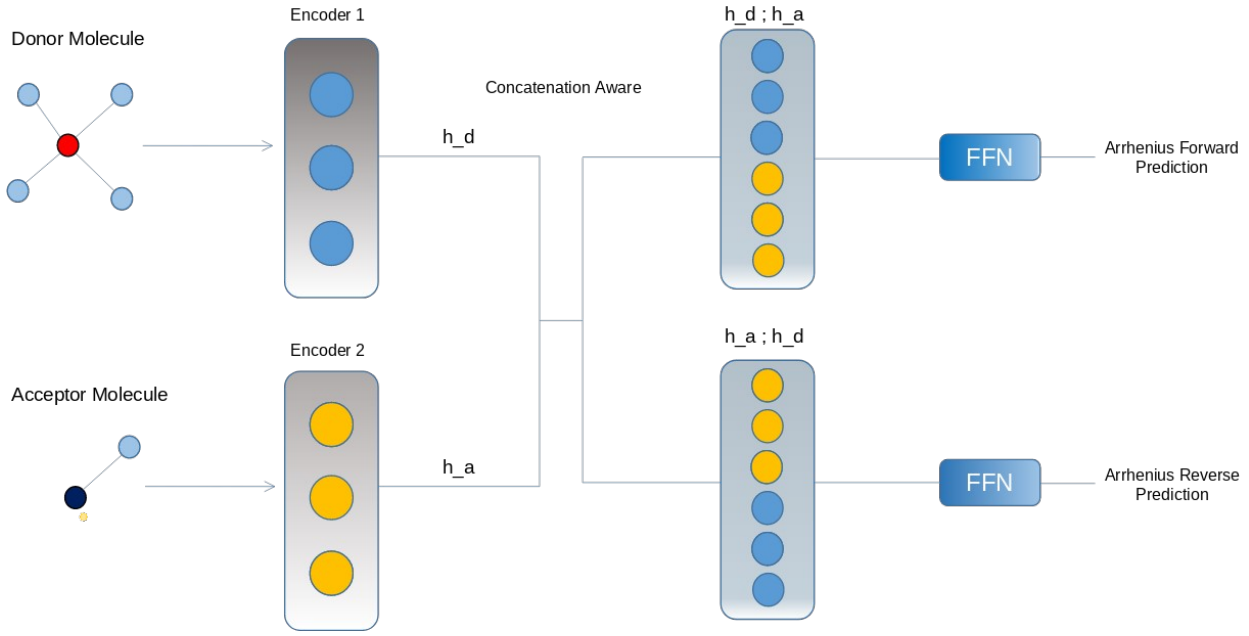

Figure 2: Order-aware architecture for hydrogen-abstraction reactions. Donor and acceptor embeddings are concatenated in forward ($[h_d; h_a]$) and reverse ($[h_a; h_d]$) order and passed to direction-specific prediction heads to predict (A_{10}, n, E_a).

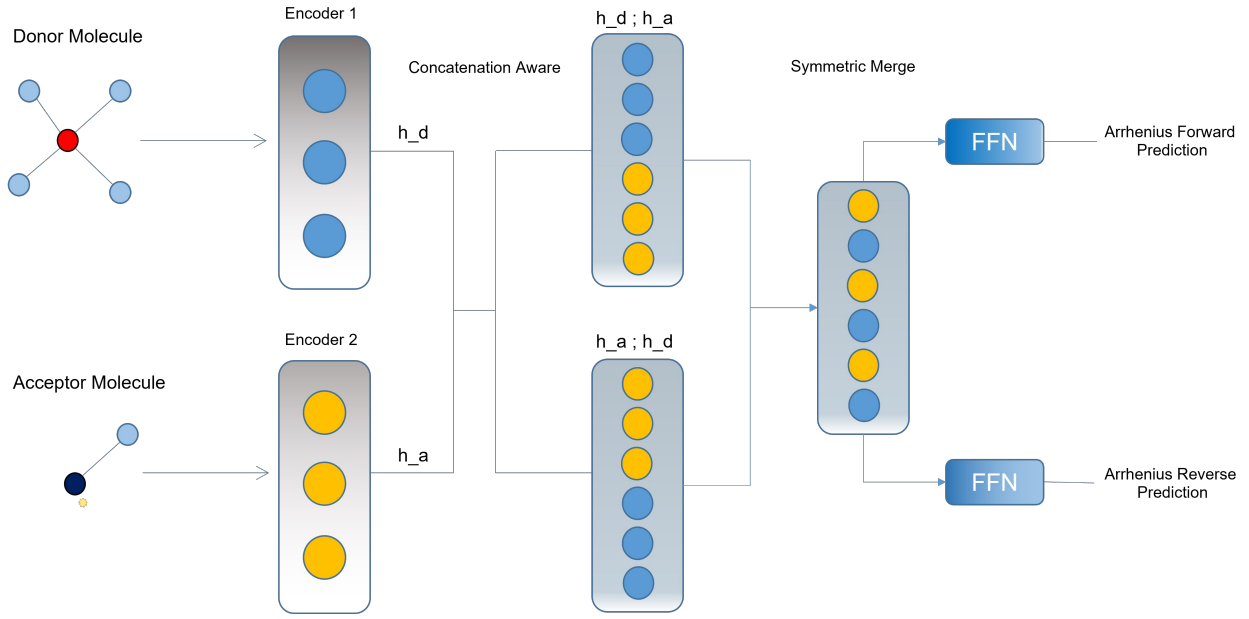

Figure 3: Order-invariant architecture. Directional concatenations are symmetrically averaged before Arrhenius prediction.

View Article Online
DOI: 10.1039/D6DD00113K

$\log_{10} A$. These outputs were passed through a differentiable Arrhenius layer to compute temperature-dependent rate coefficients $\ln k(T)$ over a fixed temperature grid. Specifically,

$$\widehat{\ln k(T)} = \widehat{A}_{10} \ln 10 + \widehat{n} \ln T - \frac{\widehat{E}_a}{RT}. \tag{2}$$

The model therefore predicts Arrhenius parameters rather than rate coefficients directly; temperature-dependent rates are obtained deterministically from the predicted parameters. This design constrains predictions to the modified Arrhenius functional form, embedding physical structure directly into the learning objective.

2.3.2 Training Objective

The model predicts forward and reverse Arrhenius triplets $\widehat{\mathbf{y}} = (\widehat{A}_{10}, \widehat{n}, \widehat{E}_a)$. These parameters are passed through a differentiable Arrhenius layer to obtain temperature-dependent rate coefficients. Because $A_{10} = \log_{10} A$ while rate coefficients are modeled in natural-log space, conversion of the prefactor introduces a factor of $\ln 10$.

Rates are evaluated on a fixed temperature grid from 300 to 3000 K in 100 K increments:

$$\widehat{\ln k(T)} = \widehat{A}_{10} \ln 10 + \widehat{n} \ln T - \frac{\widehat{E}_a}{RT}.$$

This formulation enforces cross-temperature consistency between the predicted Arrhenius parameters and the resulting rate trajectory, reducing compensating errors among (A_{10}, n, E_a).

The loss combines parameter regression and trajectory fidelity terms. Because training is performed in $\ln k$ space, additive errors correspond to multiplicative deviations in rate constants. Parameter errors are evaluated using the Huber[86] (smooth L_1) loss,

$$\mathcal{L}_{\text{param}} = \omega_{A_{10}} \mathcal{L}_{\text{Huber}}(\widehat{A}_{10}, A_{10}) + \omega_n \mathcal{L}_{\text{Huber}}(\widehat{n}, n) + \omega_{E_a} \mathcal{L}_{\text{Huber}}(\widehat{E}_a, E_a),$$

View Article Online
DOI: 10.1039/D6DD00113K

while trajectory error over the temperature grid is measured using mean squared error (MSE),

$$\mathcal{L}_{\ln k} = \omega_{\ln k}\mathrm{MSE}\big(\widehat{\ln k(T)},\, \ln k(T)\big).$$

Here $\omega_{A_{10}}, \omega_n, \omega_{E_a}$, and $\omega_{\ln k}$ are scalar weighting coefficients balancing parameter and trajectory contributions to the total loss.

2.4 Evaluation Metric

We report mean absolute error (MAE) in the logarithmic rate coefficient,

$$\mathrm{MAE}_{\ln k} = \frac{1}{N}\sum_i \big|\ln k_i^{\mathrm{pred}} - \ln k_i^{\mathrm{true}}\big|,$$

because kinetic accuracy is most naturally multiplicative: a constant deviation in $\ln k$ corresponds to a fixed factor error in k, independent of magnitude. Since all reactions in the dataset are reversible, metrics are computed separately for the forward and reverse directions and their average is reported,

$$\mathrm{MAE}_{\ln k,\mathrm{avg}} = \tfrac{1}{2}\big(\mathrm{MAE}_{\ln k,\mathrm{fwd}} + \mathrm{MAE}_{\ln k,\mathrm{rev}}\big),$$

so that overall performance reflects accuracy in both reaction directions.

2.5 Data Splitting and Leakage Control

Reactions are grouped by the unordered donor–acceptor pair (14-character InChIKeys). Thus (donor = A, acceptor = B) and (donor = B, acceptor = A) share a group. We apply a GroupKFold cross-validation (K=3) over these groups so that no pair appears in more than one fold. All members of a pair-group remain together. Audits confirmed the absence of train–validation–test leakage.

Within each outer fold, model selection uses a pair-aware Kennard–Stone (KS) proce-

View Article Online
DOI: 10.1039/D6DD00113K

dure.[87] Group representatives are embedded as binary Morgan fingerprint vectors $\mathbf{x}, \mathbf{y} \in \{0,1\}^{2048}$ (radius 2) and compared using the generalized Tanimoto similarity,[88]

$$T(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - \mathbf{x} \cdot \mathbf{y}},$$

with distance $D = 1 - T$. The distance between two reactions is

$$D_{\text{pair}} = \min\big(D_{DD} + D_{AA},\ D_{DA} + D_{AD}\big),$$

i.e., the cheaper of aligned and swapped donor/acceptor matchings, ensuring role invariance. KS ordering yields an 85/15 train/validation split; we repeat this with different seeds to form replicates. For final retraining, a fresh 90/10 KS split is used only for early stopping.

2.6 Hyperparameter Optimization and Final Evaluation

Hyperparameters were optimized using the Optuna framework,[89] which implements Bayesian optimization via the Tree-structured Parzen Estimator procedure. The search was conducted on inner KS splits constructed within the training data, with pruning based on intermediate validation losses. The optimization objective was the mean validation $\text{MAE}_{\ln k,\text{avg}}$.

The search space spanned both architectural and training parameters, including message-passing depth and width, feedforward head width, dropout rate, learning-rate schedule, order-handling mode, and optional global features.

Following hyperparameter selection, the top-ranked configurations were evaluated using a fixed 10-fold, donor–acceptor pair–grouped cross-validation procedure. For each fold, the model was trained on nine folds and evaluated on the held-out fold, and out-of-fold (OOF) predictions were aggregated so that each reaction was evaluated exactly once as unseen data. We report mean ± standard deviation across test folds. Identical optimization protocols were applied to all encoder and feature variants.

View Article Online
DOI: 10.1039/D6DD00113K

2.7 Optimization and Reproducibility

Models were trained with Adam and a Noam-style learning-rate schedule comprising linear warm-up followed by exponential decay.[90] Target variables were transformed on training data only: A_{10} and n were standardized; E_a underwent Yeo-Johnson transformation[91] and standardization; and $\ln k(T)$ was standardized across $T \in [300, 3000]$ K. For RAD features, radii were expanded using radial basis functions and dihedrals were encoded as $(\sin, \cos)$. A single input scaler was fit on training data within each fold. Experiments used fixed random seeds and deterministic kernels, while split signatures and KS seeds were logged. Full hyperparameter search ranges, selected configurations, and training scripts are provided in the public repository (see Code Availability).

2.8 Practical Deployment Workflow

Application of the trained model requires only DFT-optimized ground-state reactant geometries and does not involve transition-state searches or high-level single-point refinements.

1. **Reactant optimization.** Optimize the hydrogen donor (R1H) and hydrogen acceptor (R2H) structures at the ωB97X-D/def2-TZVP level of theory, consistent with the training data.

2. **Electronic descriptor extraction.** Extract per-atom Mulliken charges and spin densities from the DFT output. Utility scripts provided in the repository parse Gaussian or ORCA log files and generate a tabular (CSV) file containing atom-indexed electronic descriptors.

3. **Structure export.** Export the optimized reactant geometries in Structure Data File (SDF) format, preserving atom indexing and connectivity.

4. **Model inference.** Provide the donor and acceptor SDF files together with the corresponding electronic-descriptor CSV files to the pretrained model. During featurization,

View Article Online
DOI: 10.1039/D6DD00113K

atomic descriptors are merged with graph-based and geometric features to construct the full node and edge representations. The model predicts the modified Arrhenius parameters (A_{10}, n, E_a) and deterministically computes temperature-dependent rate coefficients $\ln k(T)$.

Because the workflow requires only DFT geometry optimizations and descriptor extraction, the computational cost is substantially lower than a full transition-state-theory treatment involving TS optimization, vibrational frequency calculations, intrinsic reaction coordinate analysis, and high-level single-point refinements.

3 Results and Discussion

3.1 Chemistry represented in the dataset (donor vs. acceptor roles)

We characterize the chemical diversity of the dataset via RMG Atom Types[92,93] assigned to the reactive heavy atoms. Throughout this section, the terms "donor" and "acceptor" refer to the role labels stored in the SDF annotations for each reaction instance (corresponding to the R_1H and R_2H reactant structures), rather than to intrinsic chemical classes. Because the hydrogen-abstraction reaction family is formally reversible, this designation reflects the directional representation adopted in the dataset. Each reaction entry therefore carries contextual site annotations identifying the role-labeled reactive atom type and (when applicable) the associated abstracted hydrogen.

View Article Online
DOI: 10.1039/D6DD00113K

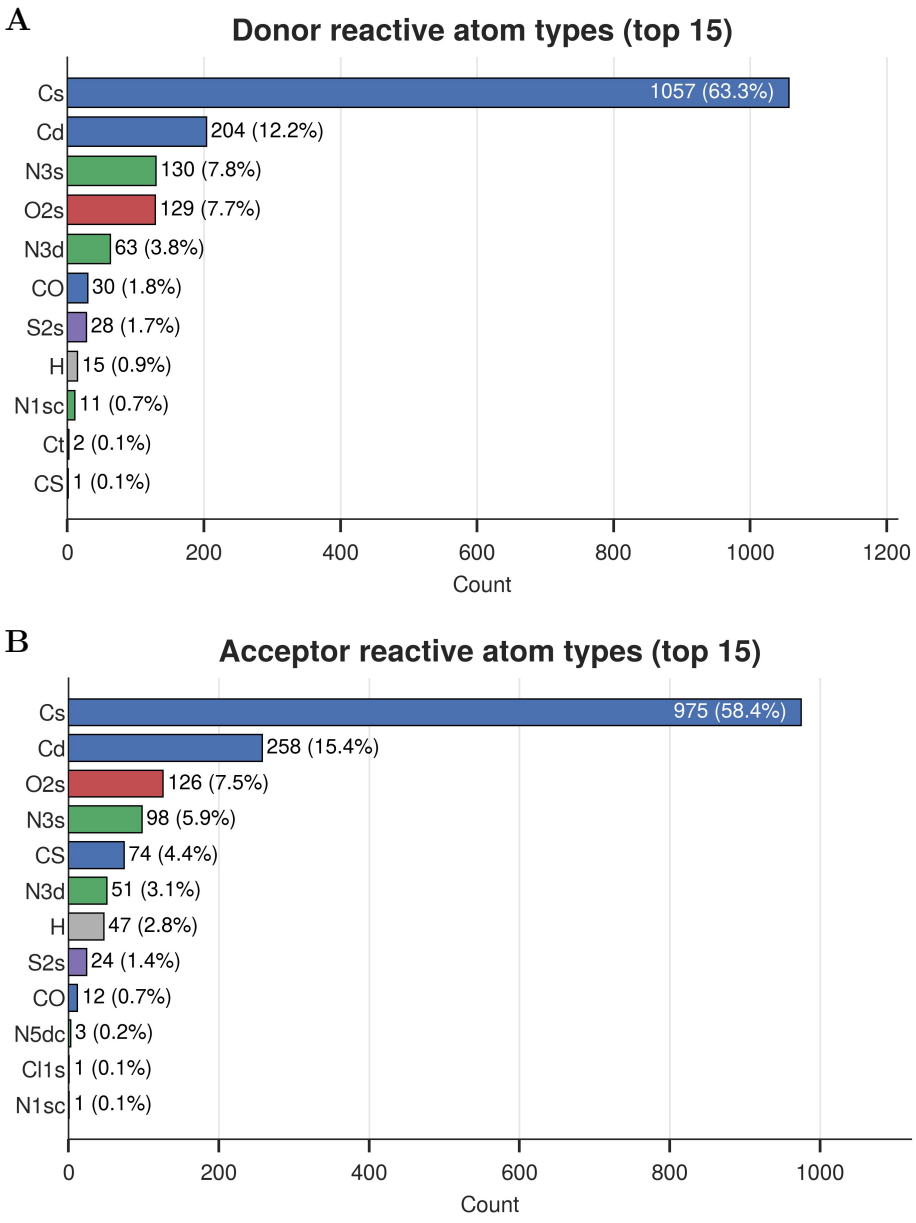

Figure 4: Reactive atom types in the dataset. **A:** Donor sites are dominated by `Cs` (1057) with `Cd` (204) as the main secondary class; `N3s` (130) and `O2s` (129) are also frequent, with smaller contributions from `N3d`, `CO`, and `S2s`. **B:** Acceptor sites show a similar ordering, led by `Cs` (975) and `Cd` (258), followed by `O2s` (126) and `N3s` (98); `CS` (74) is a notable additional class, while the remaining types occur at low frequency.

Donor-labeled sites are predominantly carbon-centered, with `Cs` as the most frequent atom type and `Cd` as the principal secondary class; `N3s` and `O2s` environments also contribute appreciably. Acceptor-labeled sites exhibit a broadly similar composition, likewise dominated

View Article Online
DOI: 10.1039/D6DD00113K

by carbon with oxygen and nitrogen contributions, but include a modest `CS` subset and a slightly longer low-frequency tail (e.g., `H`, `S2s`, and halogens). Figure 4 summarizes these empirical role-labeled distributions.

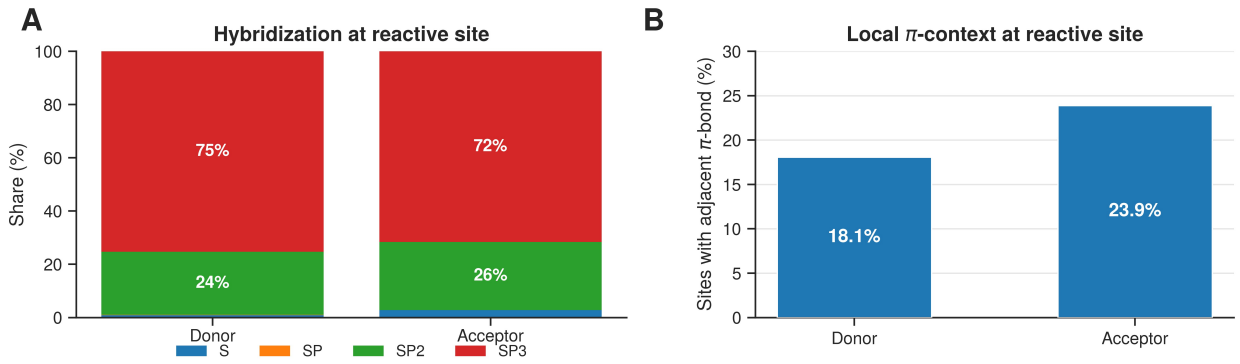

Figure 5: Local bonding at reactive atoms. **(A)** Hybridization at reactive sites: both donor and acceptor sites are dominated by sp^3 centers (75% and 72%, respectively), with comparable but slightly higher sp^2 character at acceptors (26% vs. 24%). **(B)** Adjacent π-bond frequency: acceptor sites are more often adjacent to π systems (23.9%) than donor sites (18.1%).

These site-level statistics indicate that within this dataset, hydrogen abstraction most frequently involves saturated (sp^3) centers on both sides of the labeled reaction role, with a substantial minority of cases involving sp^2 hybridized sites. Atoms labeled as acceptors are modestly more likely to be adjacent to π systems compared to donor-labeled sites. This difference reflects the statistical composition of the curated reactions rather than an intrinsic mechanistic asymmetry. Overall, the donor- and acceptor-labeled sites share largely similar hybridization profiles, with contextual differences confined primarily to π-adjacent environments.

3.2 Predictive Performance

3.2.1 Hyperparameter search and model selection

We evaluated model configurations that varied along three axes: (i) the feature set provided to the encoder (graph-only, atom-augmented, and geometry-enriched variants includ-

View Article Online
DOI: 10.1039/D6DD00113K

ing edge-level features and/or RAD descriptors), (ii) donor–acceptor order handling (`aware` vs. `invariant`), and (iii) core architectural and training hyperparameters (e.g., message-passing depth and width, aggregation type, dropout, and learning-rate schedule). To prevent leakage, the outer folds were grouped by *unordered* donor–acceptor InChIKey pairs so that both reaction orientations remained within the same fold. Candidate configurations were ranked by validation MAE($\ln k$), and the top-ranked settings were re-evaluated using repeated group-preserving cross-validation. Full cross-validation results across feature sets and training objectives (including runs with and without the Arrhenius-layer consistency term) are reported in the Table **??**.

Table 2 summarizes model-selection results across feature and architecture variants. We denote node-level reactive annotations as RA (reaction-anchored features), edge-level continuous geometry as Geom, and their combination as RA+Geom. OA and OI refer to order-aware and order-invariant donor–acceptor handling, respectively, while M-bin and M-cnt indicate binary and count-based Morgan fingerprint global features. Across all geometry- and annotation-enhanced configurations, performance differences are relatively modest. The lowest cross-validated error is achieved by a DMPNN using combined node-level and edge-level geometry (RA+Geom), order-aware pairing, and Arrhenius-layer supervision. Several alternative augmented representations—most notably those using only edge-level geometry (Geom) or atom-augmented descriptors without explicit geometry—exhibit statistically indistinguishable performance within one standard deviation. Thus, while RA+Geom yields the numerically best result, multiple geometry- and annotation-enhanced configurations provide comparably reliable predictive accuracy.

CMPNN achieves similar cross-validated error across feature modes. However, we adopt DMPNN as the primary architecture due to its bond-localized inductive bias, established benchmarking history, and consistent stability across data regimes. For hydrogen abstraction, an edge-focused message-passing scheme provides a chemically natural bias, as the reaction is governed by localized bond cleavage and radical-centered rearrangements. We

View Article Online
DOI: 10.1039/D6DD00113K

Table 2: 10-fold cross-validation (CV) on unscaled $\ln k(T)$ (300–3000 K), ranked by $\mathbf{MAE}_{\ln k,\mathrm{avg}}$.[a]

Mode[b]	Global[c]	Order[d]	Sup.[e]	MAE[f]	MSE	R^2
DMPNN						
RA+Geom	—	OA	On	**0.952 ± 0.072**	2.856 ± 0.682	0.952 ± 0.011
Geom	—	OA	On	0.979 ± 0.078	2.956 ± 0.765	0.950 ± 0.012
Atom	—	OA	On	0.999 ± 0.135	3.039 ± 1.004	0.948 ± 0.017
RA+Geom	M-bin	OI	On	1.005 ± 0.092	3.305 ± 0.822	0.944 ± 0.014
RA+Geom	M-bin	OA	Off	1.014 ± 0.113	3.294 ± 1.039	0.944 ± 0.017
RA	—	OA	On	1.015 ± 0.100	3.206 ± 0.817	0.945 ± 0.014
Baseline[g]	—	OI	On	1.364 ± 0.138	5.583 ± 1.465	0.909 ± 0.023
CMPNN						
RA+Geom	M-cnt	OI	On	**0.954 ± 0.028**	2.784 ± 0.121	0.931 ± 0.003
Path	M-cnt	OI	On	0.957 ± 0.017	2.969 ± 0.118	0.926 ± 0.003
Path+Geom	—	OI	On	0.974 ± 0.031	2.954 ± 0.157	0.927 ± 0.004
RA+Geom	—	OI	On	0.975 ± 0.033	2.821 ± 0.158	0.930 ± 0.004
RA+Geom	M-bin	OI	On	0.986 ± 0.029	2.881 ± 0.144	0.928 ± 0.004
RA	M-bin	OI	On	0.991 ± 0.015	2.938 ± 0.160	0.927 ± 0.004
Baseline[g]	M-cnt	OI	On	1.121 ± 0.013	3.524 ± 0.104	0.912 ± 0.003

a. Metrics are computed on physical $\ln k(T)$ and averaged over forward and reverse directions ($\mathrm{MAE}_{\ln k,\mathrm{avg}}$).

b. Mode denotes the geometry and annotation channels provided to the encoder. Atom = atom-augmented baseline without explicit geometry; RA = role-anchored annotations with full-graph node-level RAD descriptors but without edge-level geometry; Path = role-anchored annotations with path-restricted node-level RAD descriptors but without edge-level geometry; Geom = edge-level continuous geometric features (bond distances, angles, and dihedrals) without node-level reactive geometry (RAD); RA+Geom = full-graph node-level RAD combined with edge-level continuous geometry; RAD geometric values and existence flags are masked beyond 4 graph hops from the transferring hydrogen. Path+Geom is the analogous path-restricted variant.

c. Global features: M-bin/M-cnt = Morgan fingerprints (binary/count).

d. OA/OI denote order-aware and order-invariant donor–acceptor handling (see Figs. 2 and 3).

e. All models supervise ($\log_{10} A, n, E_a$); "Sup." indicates whether the additional Arrhenius-layer consistency loss on $\ln k(T)$ is applied during training.

f. Bold indicates the best-performing configuration within the DMPNN family.

g. Baseline rows are shown as a reference point and are not part of the MAE-ranked top six. The full Modes × MPNN result table including all evaluated configurations is provided as Table **??**.

View Article Online
DOI: 10.1039/D6DD00113K

therefore select the DMPNN configuration with RA+Geom features, order-aware handling, and Arrhenius-layer supervision (without global molecular descriptors) as the reference model for all subsequent analyses.

In contrast, geometry-free baseline models perform substantially worse, with $\mathrm{MAE}_{\ln k}$ values in the range 1.36–1.49, irrespective of donor–acceptor order handling or inclusion of global molecular fingerprints (see Table **??** in the Supporting Information). The systematic gap between baseline and augmented models demonstrates that enriching graph representations with reaction-aware geometric information yields a substantial and reproducible improvement in kinetic accuracy. While no single geometric encoding is uniquely decisive, the inclusion of local chemical context beyond topology alone is essential for accurate prediction of hydrogen-abstraction rates.

3.2.2 Directional Recovery of Kinetic Parameters

Following the model selection described in Sec. 3.2.1, we fix the top configuration and evaluate it using a fixed 10-fold, pair-grouped cross-validation partition constructed via GroupKFold. For each fold we train on nine folds and test on the held-out fold; OOF predictions are aggregated so that each reaction appears exactly once as unseen data. The network predicts *standardized* Arrhenius triplets $[\log_{10} A,\ n,\ E_a]$ separately for the forward and reverse directions, which are then de-standardized to their physical units before analysis. During training we also pass the predicted parameters through an Arrhenius layer to compute $\ln k(T)$ over 300–3000 K as an auxiliary supervision signal.

We first summarize predictive accuracy in $\ln k(T)$, where k is expressed in units of $\mathrm{cm^3\,mol^{-1}\,s^{-1}}$. Figure 6 presents a parity plot of predicted versus observed mean $\ln k(T)$, averaged over the temperature range 300–3000 K and both reaction directions, using the aggregated out-of-fold (OOF) predictions. Predictions are tightly clustered along the $y = x$ line across the full dynamic range, demonstrating accurate reproduction of absolute rate magnitudes with minimal systematic bias. The high density of points near parity indicates

View Article Online
DOI: 10.1039/D6DD00113K

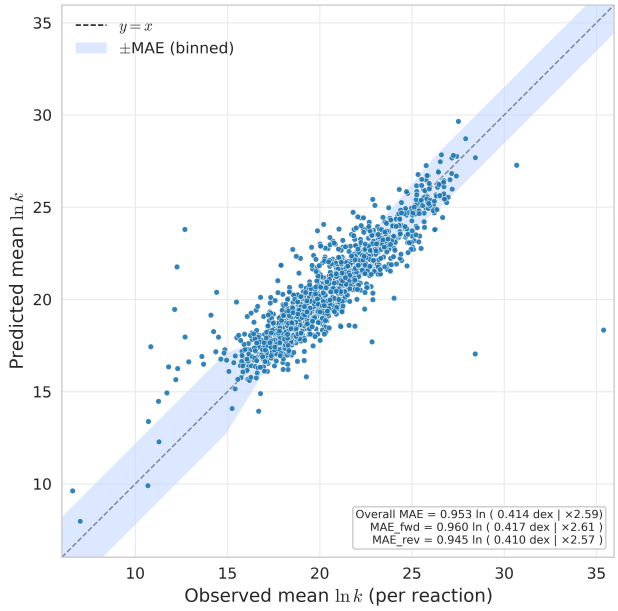

Figure 6: Parity plot of predicted versus observed temperature-averaged $\ln k(T)$, where k is expressed in units of $\mathrm{cm^3\,mol^{-1}\,s^{-1}}$. Values are averaged uniformly over the temperature range 300–3000 K and both reaction directions using aggregated out-of-fold (OOF) predictions. The shaded band denotes the binned ±MAE in $\ln k(T)$.

consistent performance across chemically diverse reactions rather than reliance on a narrow subset of the rate spectrum.

The shaded envelope represents the binned ±MAE in $\ln k$, computed per reaction and averaged within bins of true $\ln k$. This band remains narrow over most of the populated kinetic range, indicating stable predictive accuracy across multiple orders of magnitude in rate. A slight broadening is observed in the intermediate $\ln k$ regime, suggesting modest heteroscedasticity associated with chemically heterogeneous reactions in this region rather than a global degradation in model fidelity. Overall, the parity plot confirms that the selected model provides well-calibrated rate predictions suitable for downstream mechanistic and comparative analyses.

Analyzing prediction errors versus the local reaction environment (Table 3) shows that abstractions from sp^3-hybridized donor atoms are predicted relatively accurately (MAE below 0.9 in $\ln k$), while errors increase for sp^2 donors.

The largest errors are associated with reactions involving heteroatom-centered radicals

View Article Online
DOI: 10.1039/D6DD00113K

Table 3: Prediction error in $\ln k$ stratified by donor hybridization (RDKit assignment). Shown are the number of reactions (n), the mean absolute error (MAE), median absolute error, and 95% confidence interval (CI) of the mean.

Donor hybridization	n	MAE ($\ln k$)	Median ($\ln k$)	95% CI
sp^3	1227	0.880	0.633	±0.052
sp^2	380	1.137	0.780	±0.116

and strongly polarized donor–acceptor pairs, with MAEs in the range 1.5–1.9 in $\ln k$. In contrast, hydrocarbon-dominated abstractions are predicted relatively accurately. Alkyl–alkyl reactions, constituting the largest single dataset class, achieve an MAE of 0.75 in $\ln k$ ($n = 617$). These chemically homogeneous reaction classes dominate the central portion of the rate spectrum and contribute to the tight clustering observed near parity (Fig. 6).

On the aggregated OOF predictions from the 10-fold cross-validation, the model attains an overall error of MAE = 1.082 in $\ln k$, corresponding to 0.470 dex (decimal exponent) or a multiplicative uncertainty factor of approximately 2.95. Performance is well balanced between reaction directions, with MAEfwd = 1.048 and MAErev = 1.117 in $\ln k$, indicating no systematic directional bias. Table 4 reports mean absolute error (MAE), mean squared error (MSE), coefficient of determination (R^2), and bias (mean residual, predicted − true; positive indicates over-prediction) by reaction direction.

Table 4: Directional accuracy of Arrhenius parameters from OOF test predictions (10-fold CV). E_a in kJ mol^{-1}. Bias is mean residual (predicted − true).

Direction	Parameter	MAE	Bias	R^2
Forward	$\log_{10} A$	0.902	-0.0130	0.789
	n	0.268	-0.0012	0.662
	E_a	4.736	-0.9786	0.856
Reverse	$\log_{10} A$	0.885	-0.0033	0.760
	n	0.264	-0.0029	0.657
	E_a	5.393	-0.6005	0.909

For the forward direction, biases are small across all parameters, indicating negligible systematic over- or under-prediction. For the reverse direction, accuracy remains comparable for the prefactor and temperature exponent. Reverse-direction activation energies exhibit

View Article Online
DOI: 10.1039/D6DD00113K

a modest increase in MAE (5.39 kJ mol^{-1} versus 4.74 kJ mol^{-1} forward) while showing a slightly higher coefficient of determination (R^2 = 0.909 versus 0.856). Taken together, the comparable MAE, bias, and R^2 values across directions indicate balanced predictive performance without systematic directional asymmetry.

Parity plots (Fig. 7) show tight clustering around the $y{=}x$ line for $\log_{10} A$ and n in both directions. For E_a, both forward and reverse predictions align closely with the diagonal across the range of barriers, and no systematic directional deviation is apparent by visual inspection.

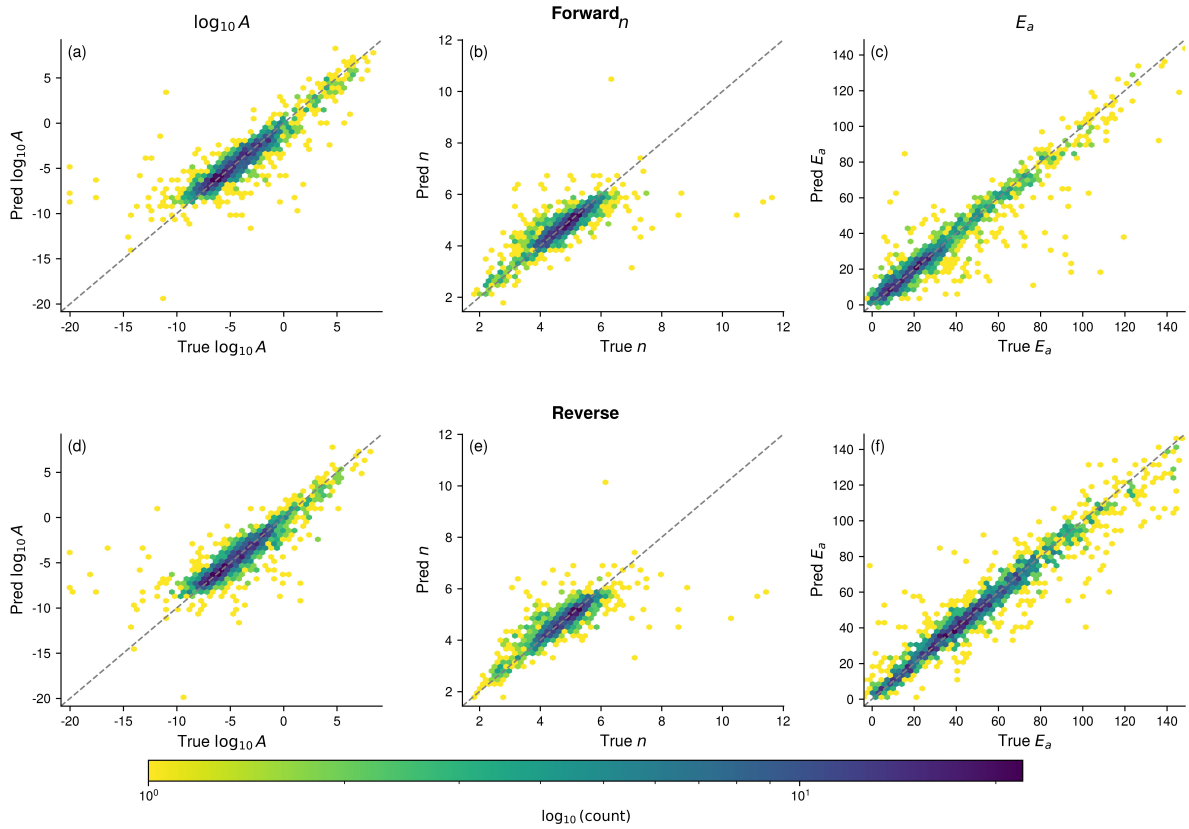

Figure 7: Parity plots for recovered Arrhenius parameters ($\log_{10} A, n, E_a$) from OOF test predictions (10-fold CV). All quantities are shown in physical (de-standardized) units; E_a is reported in kJ mol^{-1}. Top row: forward; bottom row: reverse. Panels: (a,d) $\log_{10} A$, (b,e) n, (c,f) E_a. Dashed line denotes $y{=}x$.

Biases are small in magnitude for all parameters and both directions, indicating limited systematic error. For the forward direction, biases are -0.013 for $\log_{10} A$, -0.001 for n, and

View Article Online
DOI: 10.1039/D6DD00113K

$-0.98\,\mathrm{kJ\,mol^{-1}}$ for E_a; for the reverse direction they are -0.003, -0.003, and $-0.60\,\mathrm{kJ\,mol^{-1}}$, respectively. Interpreted multiplicatively, the prefactor biases correspond to average scaling factors of $10^{-0.013} \approx 0.97$ (forward) and $10^{-0.003} \approx 0.99$ (reverse).

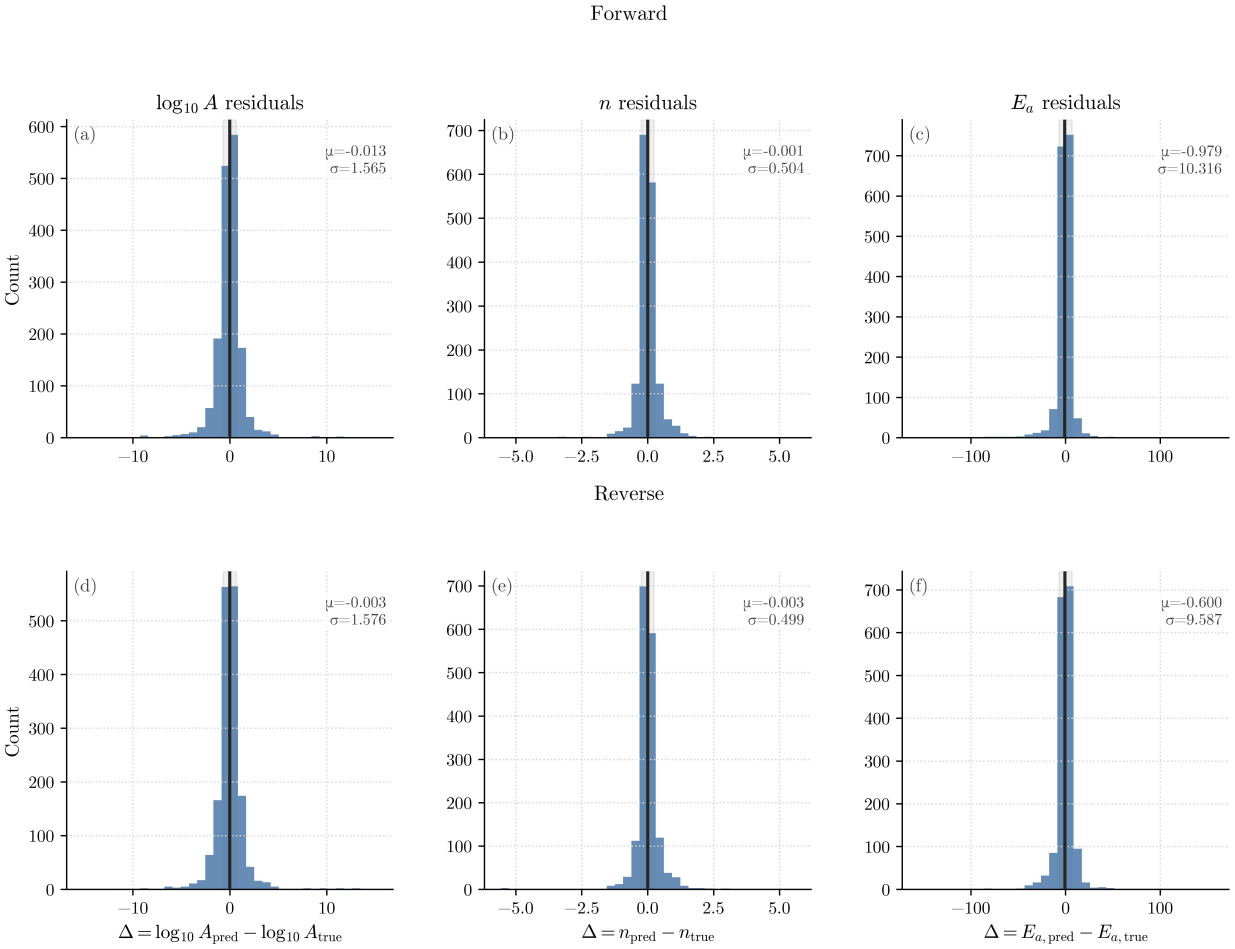

Figure 8: Residual distributions for recovered $(\log_{10} A, n, E_a)$ from out-of-fold test predictions (10-fold CV). Top row: forward; bottom row: reverse. Panels: (a,d) $\log_{10} A$ residuals, (b,e) n residuals, (c,f) E_a residuals ($\mathrm{kJ\,mol^{-1}}$). Histograms show $\Delta = \hat{y} - y$ for each parameter; per-panel insets report the mean (μ) and standard deviation (σ) of the residuals.

Residual distributions (Fig. 8) are centered near zero with approximately symmetric tails for all parameters. The broadest residual spread is observed for E_a, consistent with its wider dynamic range and greater sensitivity to small structural variations. Overall, these results indicate near-linear calibration for $\log_{10} A$ and n in both directions, and well-ranked but moderately noisier predictions for activation energies. Despite this asymmetry,

View Article Online
DOI: 10.1039/D6DD00113K

systematic biases remain small, and the resulting factor-of-few uncertainties are appropriate for downstream kinetic modeling when propagated explicitly. A quantitative mapping of parameter MAEs to multiplicative rate uncertainties is provided in Sec. **??**.

3.2.3 Predictive performance and representative Arrhenius behavior

We quantify per–reaction accuracy using the mean absolute deviation of $\log_{10} k(T)$ between the predicted and reference Arrhenius curves over $T \in [300, 3000]$ K. For reversible systems, forward and reverse directions are averaged to yield a single, balanced measure of model performance. Across the test set, the distribution of per–reaction MAE($\log_{10} k$) values over $T \in [300, 3000]$ K (28-point grid) is tightly concentrated (Table 5).

Table 5: Distribution summary of per–reaction MAE($\log_{10} k$) **over** $T \in [300, 3000]$ **K (28-point grid).**

Median (dex)	Q1	Q3	Robust σ	% within 1σ	% within 2σ
0.285	0.182	0.485	0.194	69%	86%

The median error is 0.285 dex, with an inter-quartile range of 0.182–0.485 dex. Using a robust normal-equivalent scale estimate ($\sigma \approx 1.48\,\mathrm{MAD}$, where MAD denotes the median absolute deviation), we obtain $\sigma \approx 0.194$ dex; 69% of reactions lie within $\pm 1\sigma$ of the median and 86% within $\pm 2\sigma$. In multiplicative terms, a median error of 0.285 dex corresponds to a factor of $10^{0.285} \approx 1.9$ in $k(T)$ across the full temperature range.

In the best–fit example, Fig. 9, the predicted and reference Arrhenius curves are nearly indistinguishable across 300–3000 K. Both the magnitude and curvature of $k(T)$ are reproduced, confirming that the learned (A, n, E_a) parameters capture the physical behavior with quantitative accuracy. The RMG estimates provide a reasonable physical baseline, though they deviate more notably than the ML predictions in this instance. This is expected, as the specific chemical environment in this reaction may not be explicitly represented in the current RMG rate-tree hierarchy, whereas the ML model generalizes across these structural nuances during training.

View Article Online
DOI: 10.1039/D6DD00113K

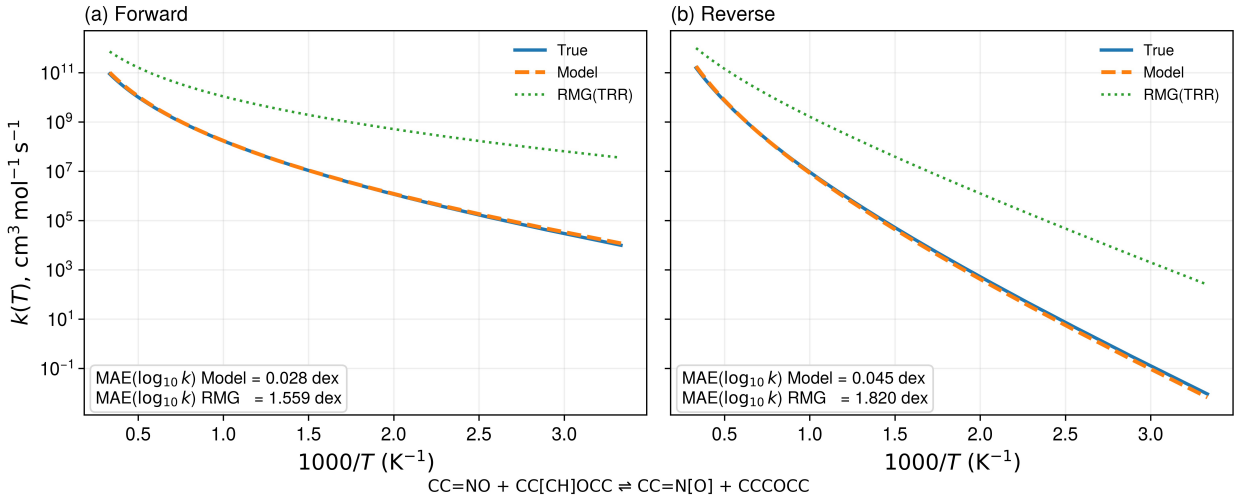

Figure 9: Best (lowest MAE) reaction. Predicted (orange, dashed), reference (blue, solid), and RMG estimated (green, dotted) Arrhenius curves over 300–3000 K in the forward (a) and reverse (b) directions, yielding MAE($\log_{10} k$) < 0.05 dex per direction. RMG(TRR) denotes RMG's template rate–rule estimate from its kinetics families, rather than a value retrieved from a matched training reaction.

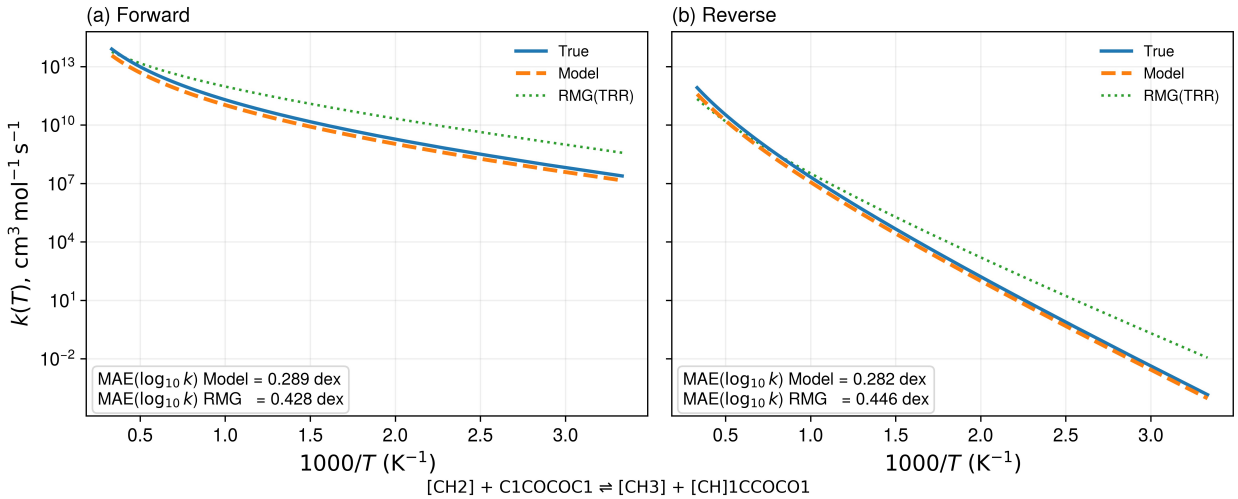

Figure 10: Median (typical) reaction. Predicted (orange, dashed), reference (blue, solid), and RMG estimated (green, dotted) Arrhenius curves across 300–3000 K in the forward (a) and reverse (b) directions. The resulting per–direction errors, MAE($\log_{10} k$) ≈ 0.29 dex, are representative of the median performance across the test set.

The median case, Fig. 10, illustrates the typical predictive behavior of the model. Across the full temperature range, deviations remain well below one order of magnitude, and the predicted Arrhenius curves closely track the reference slope in both directions. This indicates

View Article Online
DOI: 10.1039/D6DD00113K

that the model reliably captures the underlying activation energy trend while allowing for modest shifts in the overall rate prefactor. By comparison, the template–based RMG(TRR) estimates show larger systematic offsets and slight slope mismatches, highlighting the improved quantitative fidelity of the ML approach for representative reactions.

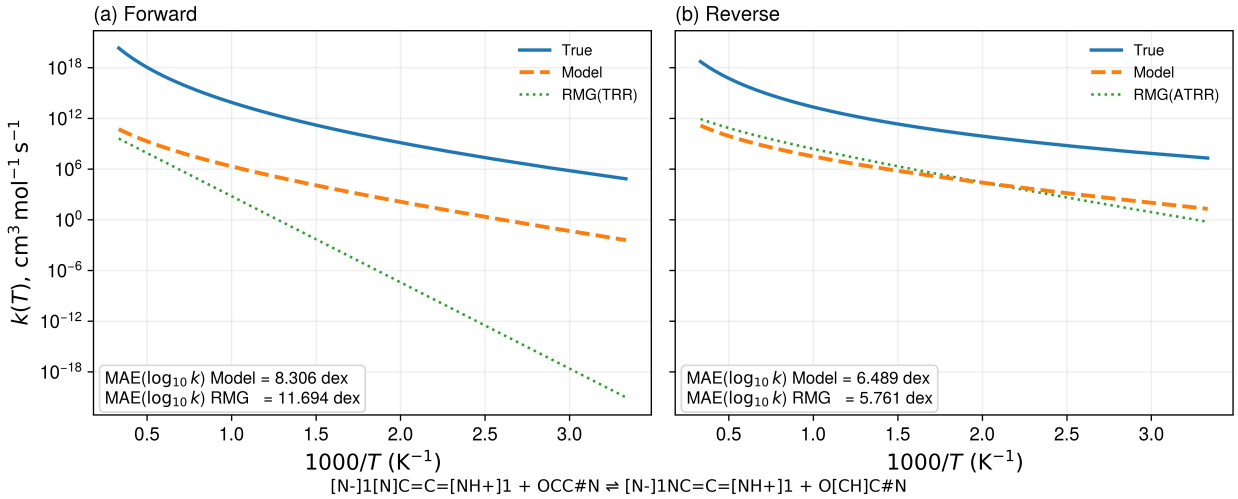

Figure 11: Worst (highest MAE) reaction. Predicted (orange, dashed), reference (blue, solid), and RMG estimated (green, dotted) Arrhenius curves exhibit large discrepancies in absolute magnitude in both the forward (a) and reverse (b) directions, leading to very high MAE($\log_{10} k$) values.

Even in the worst–performing example (Fig. 11), the model reproduces the correct qualitative Arrhenius behavior: in both directions, the predicted rates increase smoothly with increasing temperature, consistent with physically meaningful activation barriers. However, the predicted curves are systematically offset from the reference by several orders of magnitude, resulting in extremely large MAE($\log_{10} k$) values. The template–based RMG predictions exhibit even larger deviations in the forward direction, including noticeable slope mismatches, underscoring the limitations of rule–based extrapolation in this challenging regime.

Figure 12 illustrates a regime in which the ML model substantially outperforms RMG's template–based rate rule (TRR) estimate. The predicted Arrhenius curves closely track the reference in both magnitude and temperature dependence across the full temperature

View Article Online
DOI: 10.1039/D6DD00113K

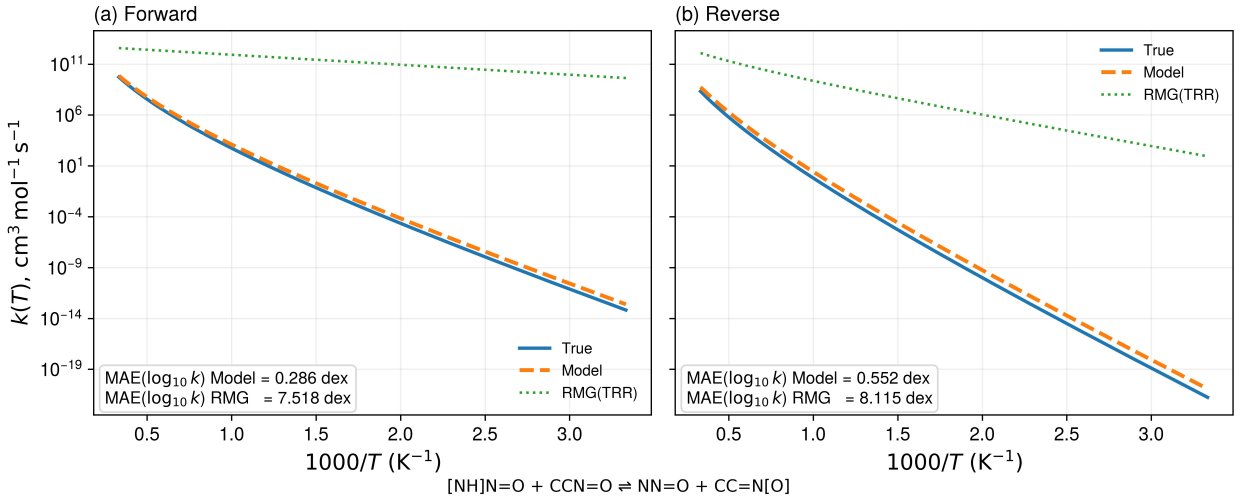

Figure 12: Reaction where the ML model outperforms the RMG(TRR) estimate. Predicted (orange, dashed) and reference (blue, solid) Arrhenius curves remain closely aligned across 300–3000 K in both the forward (a) and reverse (b) directions, yielding substantially lower MAE($\log_{10} k$) than the template–based RMG prediction (green, dotted).

range. In contrast, the RMG(TRR) estimate deviates by several orders of magnitude and exhibits a slightly shallower temperature dependence. Such behavior typically arises when the local chemical environment lies outside the effective coverage of existing RMG training data, forcing the hierarchy to interpolate from sparsely populated or weakly matched rules. In these cases, the learned model benefits from continuous feature representations and is able to recover accurate kinetics where discrete rule–based extrapolation fails.

Figure 13 illustrates a case in which the RMG(TRR) estimate achieves lower error than the ML model. In the forward direction, the ML prediction exhibits a large systematic offset from the reference and a noticeably weaker temperature dependence, leading to substantial error across the full temperature range. By contrast, the RMG prediction more closely follows both the magnitude and slope of the reference Arrhenius curve.

A similar pattern is observed in the reverse direction, where the ML model again deviates by several orders of magnitude and underestimates the temperature sensitivity. The RMG(TRR) estimate better captures the overall Arrhenius behavior in this regime, demonstrating that rule–based kinetics can remain competitive for specific reaction environments

View Article Online
DOI: 10.1039/D6DD00113K

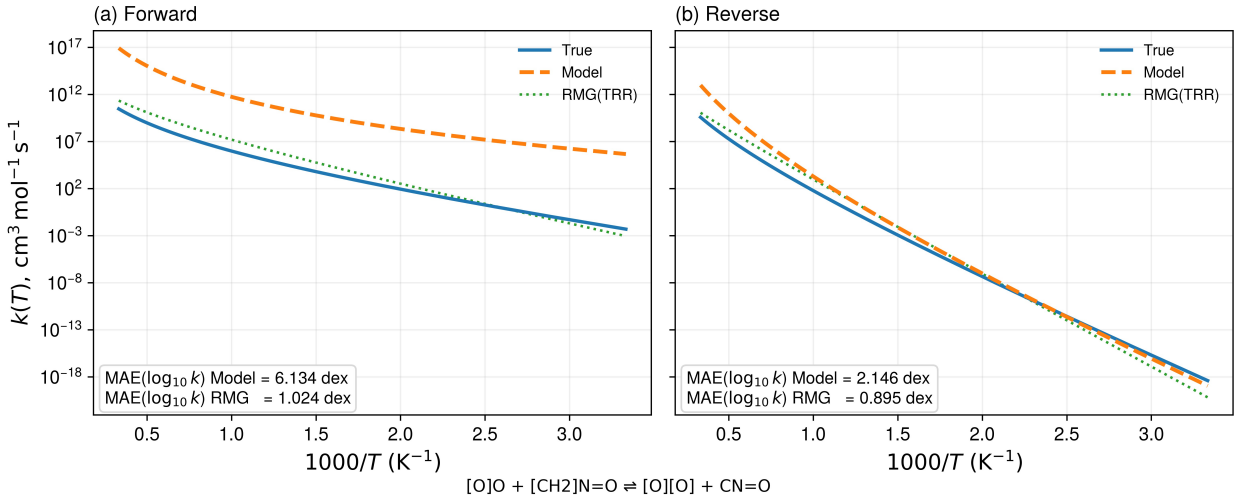

Figure 13: Reaction where the RMG(TRR) estimate outperforms the ML model. In the forward (a) and reverse (b) directions, the template–based RMG prediction (green, dotted) remains closer to the reference Arrhenius curves (blue, solid) than the ML prediction (orange, dashed), resulting in a lower MAE($\log_{10} k$) for RMG's estimate.

that are well represented within the existing template hierarchy.

Taken together, these representative reactions illustrate that the ML model typically achieves lower error than template–based estimates while maintaining smooth, physically consistent Arrhenius behavior across a wide temperature range. In the majority of cases, discrepancies between predicted and reference rates appear primarily as approximately uniform shifts in magnitude rather than distortions of the temperature dependence. This indicates that the model has learned continuous structure–kinetics relationships in the (A, n, E_a) parameter space.

A small number of reactions exhibit lower error for RMG than for the ML model. These cases are often associated with specific chemical environments that are well represented within existing RMG template hierarchies but sparsely sampled in the training data used here. Such failures therefore appear to arise from limited coverage of particular regions of chemical space, rather than from an intrinsic inability of the model to represent the underlying kinetics. This suggests that targeted data expansion or focused active–learning strategies in these regimes could further improve model robustness.

View Article Online
DOI: 10.1039/D6DD00113K

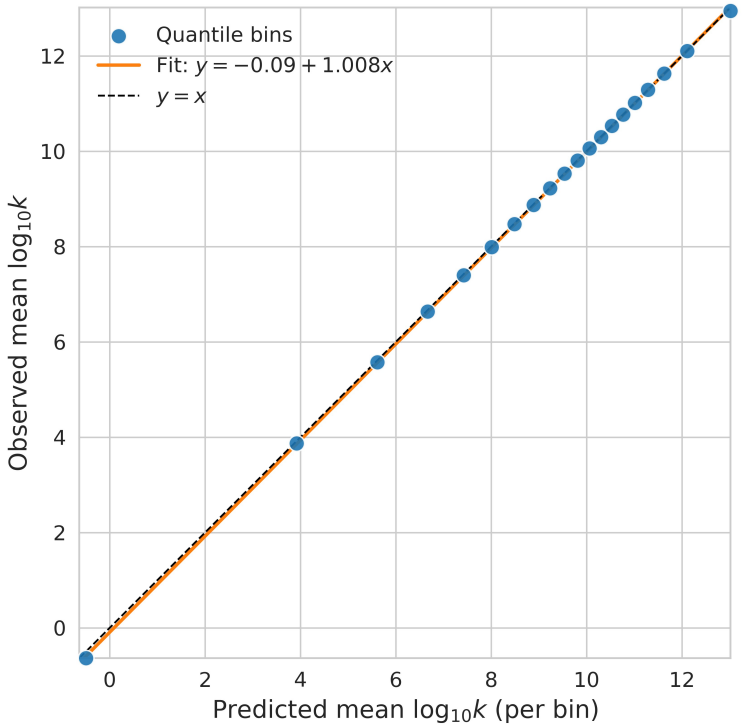

Figure 14: Calibration of predicted rate magnitudes. Mean observed versus predicted $\log_{10} k(T)$ values computed within quantile bins of the predictions. The fitted linear relationship (orange) closely follows the ideal $y = x$ line (dashed), indicating minimal global bias and near–unity scaling across the full dynamic range of predicted rate constants.

Figure 14 shows a reliability analysis of $\log_{10} k(T)$ obtained by pooling forward and reverse predictions over a 300–3000 K temperature grid and partitioning the predicted values into 20 equal-frequency bins. Bin-wise means of predicted and observed $\log_{10} k(T)$ are nearly collinear, with fitted relation $\overline{y}_{\text{obs}} = -0.09 + 1.008\,\overline{y}_{\text{pred}}$, indicating minimal global bias. The slope remains very close to unity and the intercept is small relative to the dynamic range of the data, demonstrating that the model preserves rate magnitudes without systematic compression or expansion across several orders of magnitude. The absence of structured deviation at low or high rate extremes indicates that residual errors are primarily reaction-specific rather than attributable to global miscalibration.

To examine how predictive accuracy varies across local chemical environments, we performed a site-resolved error analysis stratified by donor and acceptor RMG atom types and their donor–acceptor pairings; full results and figures are reported in Sec. **??**.

View Article Online
DOI: 10.1039/D6DD00113K

3.2.4 Comparison with experimental NIST kinetics

To assess external consistency, we compared Arkane-derived rates and model predictions against Arrhenius parameters reported in the NIST Chemical Kinetics Database[94] for reactions common to both datasets. Because curated experimental data are sparse, only a subset of reactions could be matched: 60 forward and 19 reverse reactions. Arrhenius curves were visualized over 300–3000 K, while deviations were evaluated within the NIST-reported temperature range for each fit and expressed in terms of $\log_{10} k(T)$ relative to the NIST reference.

For visualization, we report a representative (median) matched reaction in each direction, selected as the reaction whose MAE($\log_{10} k$) relative to the corresponding NIST fit is closest to the median across the matched subset, computed over the overlapping NIST temperature window. When multiple NIST entries were available for a reaction, we used the most recent recommended/evaluated fit with maximal overlap in the comparison range.

In the forward example (Fig. 15), Arkane exhibits a systematic magnitude offset from the critically evaluated Arrhenius fit,[95] yielding MAE($\log_{10} k$) = 0.794 dex, whereas the model deviation is 0.201 dex. This closer agreement arises from the model drifting relative to its Arkane training target rather than from explicit experimental calibration. Because training is performed exclusively on Arkane-derived kinetics, improved alignment with experiment in individual cases should be interpreted as incidental.

The reverse example (Fig. 16), derived from the same evaluated source,[95] shows the complementary situation: Arkane lies slightly closer to experiment (0.753 dex) than the model (0.851 dex). In both directions, discrepancies relative to experiment are dominated by magnitude shifts rather than curvature distortions.

Taken together, these representative reactions demonstrate that the model neither systematically improves nor degrades agreement with experiment relative to Arkane. Its objective is fidelity to the quantum-chemical reference, not correction of underlying electronic-structure errors.

View Article Online
DOI: 10.1039/D6DD00113K

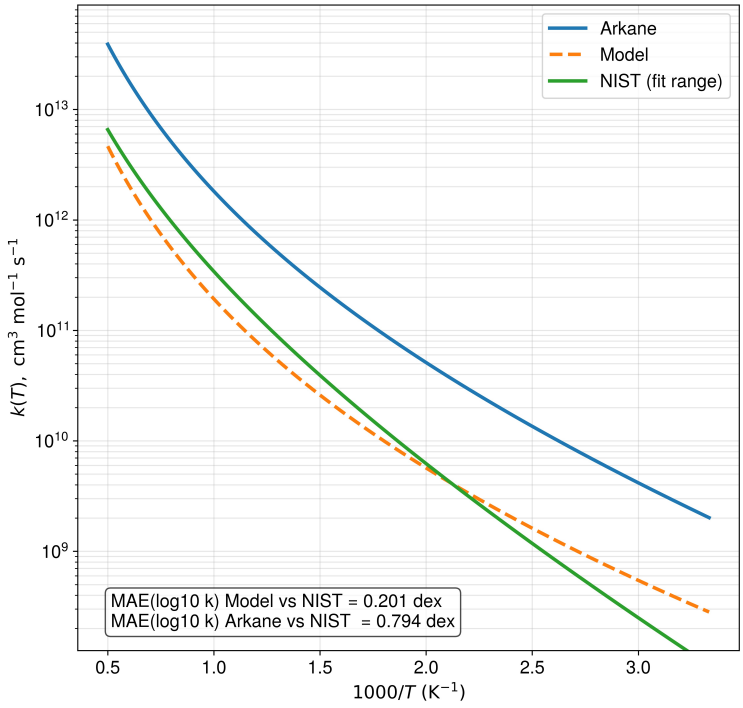

Figure 15: Forward-direction comparison with NIST kinetics. Arkane (blue, solid), model (orange, dashed), and NIST Arrhenius fit (green) over 300–3000 K.

View Article Online
DOI: 10.1039/D6DD00113K

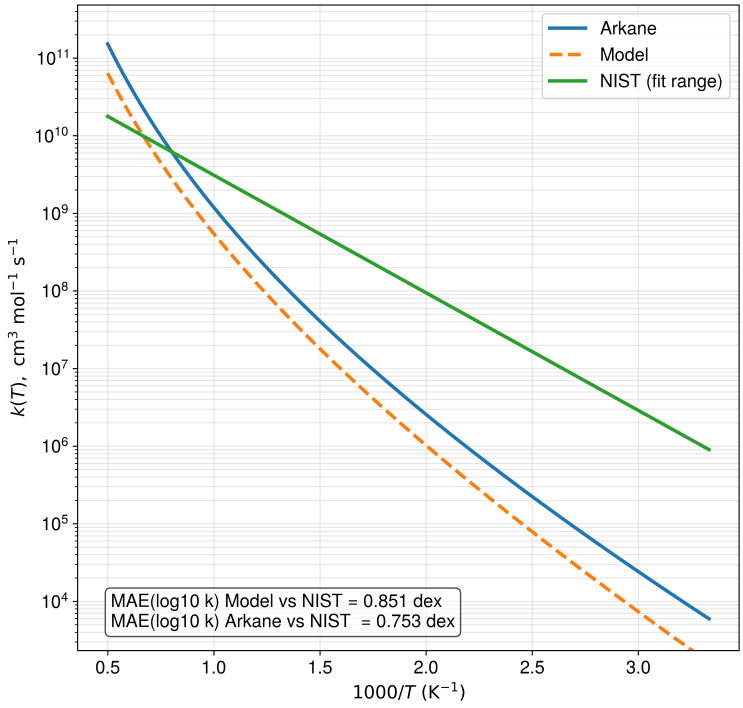

Figure 16: Reverse-direction comparison with NIST kinetics. Arkane (blue, solid), model (orange, dashed), and NIST Arrhenius fit (green) over 300–3000 K.

View Article Online
DOI: 10.1039/D6DD00113K

To quantify trends across the full NIST-matched subset, we computed the temperature-averaged absolute deviation

$$|\Delta| = \langle |\log_{10} k(T)_{\text{pred}} - \log_{10} k(T)_{\text{NIST}}| \rangle_{T \in [300,3000]\,\text{K}} .$$

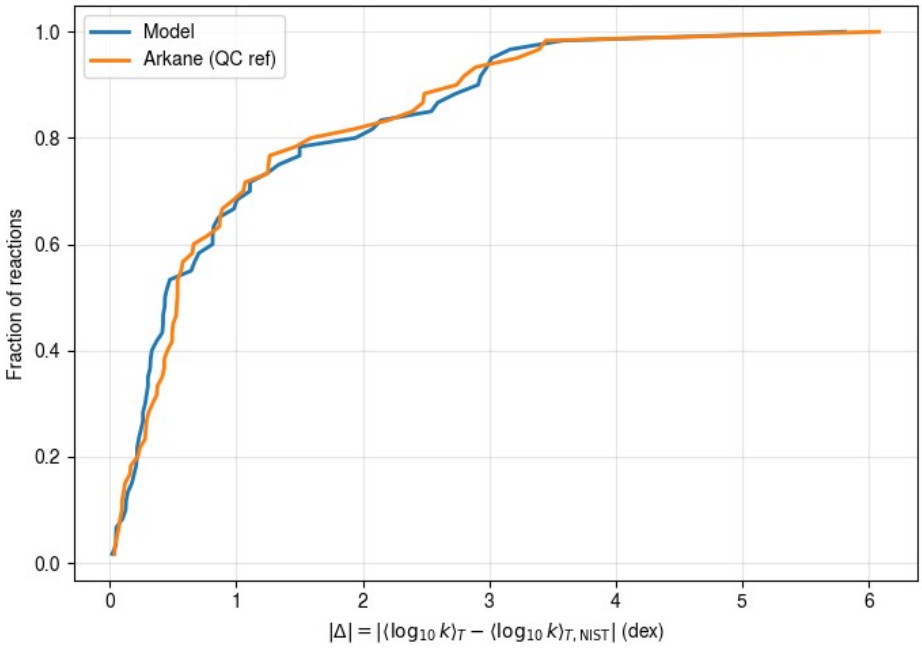

Figure 17: Forward-direction ECDF of deviation from NIST. Absolute $\log_{10} k$ error (dex) vs. cumulative fraction of reactions. Model (blue) and Arkane (orange) show comparable agreement.

For the forward direction (Fig. 17), approximately 70% of reactions lie within one order of magnitude of NIST for both Arkane and model predictions. The ECDFs are nearly overlapping, indicating no systematic dominance and only reaction-specific differences.

The reverse-direction ECDF (Fig. 18) shows similar behavior: roughly 70–75% of reactions fall within 1 dex of experiment, and the model and Arkane curves remain closely aligned across the full deviation range. Overall, despite the limited size of the NIST-matched subset, the model preserves the experimental kinetic scale encoded in the Arkane reference dataset without introducing systematic bias. Agreement with NIST is governed primarily by the underlying quantum-chemical reference quality rather than by the model approximation.

View Article Online
DOI: 10.1039/D6DD00113K

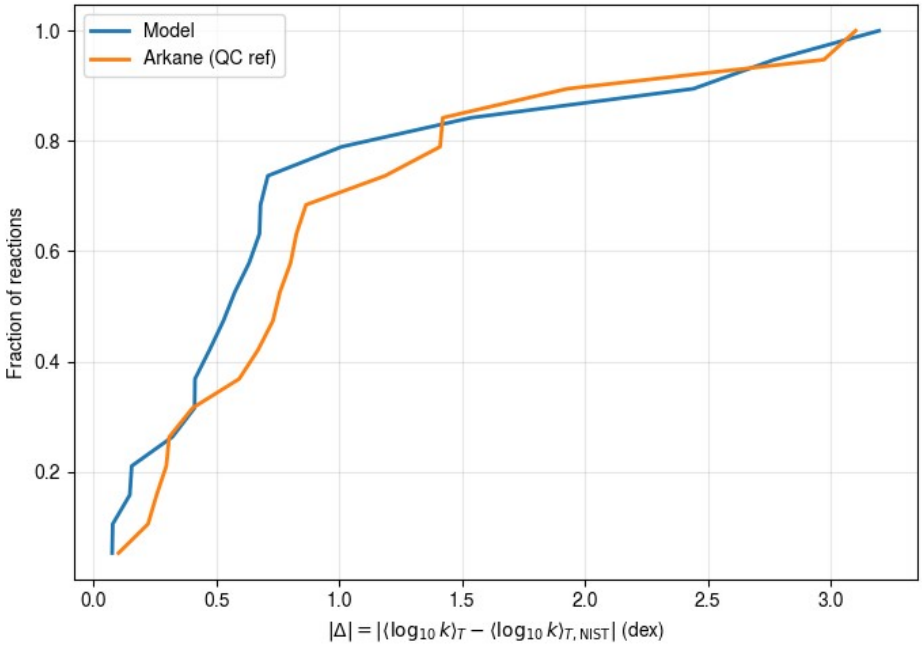

Figure 18: Reverse-direction ECDF of deviation from NIST. Absolute $\log_{10} k$ error (dex) vs. cumulative fraction of reactions. Model (blue) and Arkane (orange) show comparable agreement.

4 Conclusions

We demonstrate that hydrogen-abstraction kinetics can be predicted directly from reaction structure with factor-of-few accuracy by enriching message passing with reaction-anchored descriptors and local three-dimensional geometry. Across group-aware 10-fold cross-validation, the selected RA+Geom DMPNN configuration achieves MAE = 1.082 in $\ln k$ (0.470 dex; $\sim 3\times$), with balanced forward/reverse performance, minimal global bias, near-linear scaling across several orders of magnitude in rate, and a tightly concentrated per-reaction Arrhenius-curve error distribution (median 0.285 dex; $\sim 1.9\times$ over 300–3000 K). Geometry-free baselines already achieve strong predictive accuracy, indicating that topological and role-anchored information captures much of the kinetic signal. However, incorporating localized three-dimensional geometric context yields consistent and statistically robust improvements in error and variance across folds, demonstrating that explicit geometry provides complementary information beyond topology alone.

View Article Online
DOI: 10.1039/D6DD00113K

Errors are structured by chemical environment: sp^3 donor abstractions and hydrocarbon-dominated classes are predicted most reliably, whereas heteroatom-centered radicals and strongly polarized donor–acceptor pairs contribute the largest deviations, primarily as near-uniform magnitude shifts rather than distortions in temperature dependence. Relative to template-based RMG rate rules, the learned model typically achieves lower error while preserving smooth, physically consistent Arrhenius behavior. Comparison to NIST-matched reactions indicates that the model preserves the kinetic scale encoded in the Arkane quantum-chemical reference without introducing systematic bias, implying that residual disagreement with experiment is dominated by reference-level uncertainties rather than model artifacts. In addition, predicted forward/reverse equilibrium constants agree with the textbook DFT-derived reference to within half an order of magnitude despite no explicit detailed-balance constraint during training, indicating that thermodynamic consistency is largely recovered from the kinetic targets alone rather than being enforced by construction.

The present analysis also identifies clear routes for further improvement. Explicit one-dimensional hindered-rotor and large-amplitude-inversion treatment of the training labels, beyond the current rigid-rotor/harmonic-oscillator approximation, would improve fidelity for chemistry classes such as peroxide-mediated abstractions, multi-rotor ethers, and benzylic-tertiary radicals, where rotor corrections can exceed an order of magnitude in $k(T)$. Augmenting the loss with a detailed-balance-consistent term that draws on the Arkane-derived Gibbs free energies $\Delta G_{\mathrm{rxn}}(T)$, released alongside the dataset as NASA polynomials, would promote the implicit forward/reverse equilibrium-constant agreement observed here into an explicit training constraint. Transfer learning on a curated subset of experimental NIST rate coefficients would further recalibrate the present DLPNO-CCSD(T)-F12 anchored predictions to measured reality, complementing rather than replacing the systematically generated quantum-chemical training set.

Overall, our results demonstrate that this framework can significantly reduce quantum-chemical computation effort for reaction-mechanism development. The trained graph neural

View Article Online
DOI: 10.1039/D6DD00113K

network provides the factor-of-few accuracy required for large-scale mechanism generation without the iterative bottleneck of manual transition-state searching. By identifying the specific chemical environments that bound current accuracy, this framework establishes a clear and actionable path toward systematically improving data-driven kinetic prediction through targeted expansion of underrepresented reaction regimes.

5 Data and Software Availability

The datasets supporting this article are provided in the Supplementary Information and are archived on Zenodo at DOI: https://doi.org/10.5281/zenodo.20433305. The custom code associated with this work, including scripts for model training, evaluation, and analysis, is available from the GitHub repository https://github.com/calvinp0/chemprop_arrhenius. A persistent archived snapshot of the GitHub repository is available on Zenodo at DOI: https://doi.org/10.5281/zenodo.20485495.

Supporting Information Available

Supporting Information PDF: code and repository information; description of the ARC hydrogen-abstraction transition-state heuristic; dataset DOI and SDF schema; mathematical definition of the directed edge-geometric encoding; complete cross-validation and ablation results across D-MPNN and CMPNN model variants; interpretation of Arrhenius-parameter errors in terms of rate-constant uncertainty; quality analysis of the modified-Arrhenius fits; quantification of hindered-rotor and inversion corrections for a representative reaction subset; forward/reverse equilibrium-constant consistency analysis; and site-resolved error analyses by donor, acceptor, and donor–acceptor atom type.

The hydrogen abstraction reaction dataset, including optimized geometries, vibrational frequencies, high-level electronic energies, fitted Arrhenius parameters, SDF records, and auxiliary analysis data, is archived on Zenodo (DOI: 10.5281/zenodo.18597964).[?]

View Article Online
DOI: 10.1039/D6DD00113K

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgement

This work was supported in part by the Stephen and Nancy Grand Technion Energy Program (GTEP) and the Boeing-Technion SAF Innovation Center funded by the Boeing Company. We thank the Chemprop development team at MIT for releasing an open-source codebase that served as the foundation for this work.

References

(1) von Schneidemesser, E.; Monks, P. S.; Allan, J. D.; Bruhwiler, L.; Forster, P.; Fowler, D.; Lauer, A.; Morgan, W. T.; Paasonen, P.; Righi, M. et al. Chemistry and the Linkages between Air Quality and Climate Change. *Chemical Reviews* **2015**, *115*, 3856–3897.

(2) Miller, J. A.; Sivaramakrishnan, R.; Tao, Y.; Goldsmith, C. F.; Burke, M. P.; Jasper, A. W.; Hansen, N.; Labbe, N. J.; Glarborg, P.; Zádor, J. Combustion chemistry in the twenty-first century: Developing theory-informed chemical kinetics models. *Progress in Energy and Combustion Science* **2021**, *83*, 100886.

(3) Rotavera, B.; Taatjes, C. A. Influence of functional groups on low-temperature combustion chemistry of biofuels. *Progress in Energy and Combustion Science* **2021**, *86*, 100925.

(4) Grinberg Dana, A.; Van Geem, K. M.; Cavallotti, C.; Green, W. H. Predictive Chemical Kinetic Modeling: Where We Succeed, Where We Struggle, and What Comes Next. *ACS Engineering Au* **2026**, *6*, 1–19.

View Article Online
DOI: 10.1039/D6DD00113K

(5) Heald, C. L.; Kroll, J. H. The fuel of atmospheric chemistry: Toward a complete description of reactive organic carbon. *Science Advances* **2020**, *6*, eaay8967.

(6) Lakshmanan, S.; Bhati, M. Unravelling the atmospheric and climate implications of hydrogen leakage. *International Journal of Hydrogen Energy* **2024**, *53*, 807–815.

(7) Zhang, Y.; Jacob, D. J.; Maasakkers, J. D.; Sulprizio, M. P.; Sheng, J.-X.; Gautam, R.; Worden, J. Monitoring global tropospheric OH concentrations using satellite observations of atmospheric methane. *Atmospheric Chemistry and Physics* **2018**, *18*, 15959–15973.

(8) Yu, J.; Ruan, S.; Song, H.; Zhang, L.; Yang, M. Machine learning rate constants of hydrogen abstraction reactions between ester and H atom. *Combustion and Flame* **2023**, *255*, 112901.

(9) Skeie, R. B.; Sandstad, M.; Krishnan, S.; Myhre, G.; Sand, M. Sensitivity of climate effects of hydrogen to leakage size, location, and chemical background. *Atmospheric Chemistry and Physics* **2025**, *25*, 4929–4942.

(10) López-Comí, L.; Morgenstern, O.; Zeng, G.; Masters, S. L.; Querel, R. R.; Nedoluha, G. E. Assessing the sensitivity of the hydroxyl radical to model biases in composition and temperature using a single-column photochemical model for Lauder, New Zealand. *Atmospheric Chemistry and Physics* **2016**, *16*, 14599–14619.

(11) Jain, S.; Li, D.; Aggarwal, S. K. Effect of hydrogen and syngas addition on the ignition of iso-octane/air mixtures. *International Journal of Hydrogen Energy* **2013**, *38*, 4163–4176.

(12) Shaqiri, S.; Kaczmarek, D.; vom Lehn, F.; Beeckmann, J.; Pitsch, H.; Kasper, T. Experimental Investigation of the Pressure Dependence of Iso-Octane Combustion. *Frontiers in Energy Research* **2022**, *10*.

View Article Online
DOI: 10.1039/D6DD00113K

(13) Hartness, S. W.; Rotavera, B. Dependence of Biofuel Ignition Chemistry on OH-Initiated Branching Fractions. *Frontiers in Mechanical Engineering* **2021**, *7*.

(14) Zhang, Z.; Zhu, R.; Zhu, Y.; Weng, W.; He, Y.; Wang, Z. Experimental and Kinetic Study on Laminar Burning Velocities of High Ratio Hydrogen Addition to CH4+O2+N2 and NG+O2+N2 Flames. *Energies* **2023**, *16*, 5265, Number: 14.

(15) Osipova, K. N.; Sarathy, S. M.; Korobeinichev, O. P.; Shmakov, A. G. Laminar Burning Velocities of Formic Acid and Formic Acid/Hydrogen Flames: An Experimental and Modeling Study. *Energy & Fuels* **2021**, *35*, 1760–1767.

(16) Zhang, X.; Wang, J.; Chen, Y.; Li, C. Effect of CH4, Pressure, and Initial Temperature on the Laminar Flame Speed of an NH3–Air Mixture. *ACS Omega* **2021**, *6*, 11857–11868.

(17) Dong, W.; Hong, R.; Yao, J.; Wang, D.; Yan, L.; Qiu, B.; Chu, H. Soot formation and laminar combustion characteristics of anisole: ReaxFF MD simulation and kinetic analysis. *Carbon Neutrality* **2024**, *3*, 34.

(18) Baulch, D. L.; Bowman, C. T.; Cobos, C. J.; Cox, R. A.; Just, T.; Kerr, J. A.; Pilling, M. J.; Stocker, D.; Troe, J.; Tsang, W. et al. Evaluated Kinetic Data for Combustion Modeling: Supplement II. *Journal of Physical and Chemical Reference Data* **2005**, *34*, 757–1397.

(19) Srinivasan, N. K.; Su, M.-C.; Sutherland, J. W.; Michael, J. V. Reflected shock tube studies of high-temperature rate constants for OH + CH4 –> CH3 + H2O and CH3 + NO2 –> CH3O + NO. *The Journal of Physical Chemistry. A* **2005**, *109*, 1857–1863.

(20) Bystrov, N. S.; Emelianov, A. V.; Eremin, A. V.; Kurbatova, E. S.; Yatsenko, P. I. Joint Effect of Shock-Wave Heating and Laser Photolysis for the Generation of Active Atoms and Radicals in a Wide Temperature Range. *High Temperature* **2024**, *62*, 705–708.

View Article Online
DOI: 10.1039/D6DD00113K

(21) Blázquez, S.; González, D.; García-Sáez, A.; Antiñolo, M.; Bergeat, A.; Caralp, F.; Mereau, R.; Canosa, A.; Ballesteros, B.; Albaladejo, J. et al. Experimental and theoretical investigation on the OH + CH3C(O)CH3 reaction at interstellar temperatures (T=11.7-64.4 K). *ACS earth & space chemistry* **2019**, *3*, 1873–1883.

(22) Williams, P. J. H.; Boustead, G. A.; Heard, D. E.; Seakins, P. W.; Rickard, A. R.; Chechik, V. New Approach to the Detection of Short-Lived Radical Intermediates. *Journal of the American Chemical Society* **2022**, *144*, 15969–15976.

(23) Du, P.; Wang, J.; Sun, G.; Chen, L.; Liu, W. Hydrogen atom abstraction mechanism for organic compound oxidation by acetylperoxyl radical in Co(II)/peracetic acid activation system. *Water Research* **2022**, *212*, 118113.

(24) Lamberts, T.; Fedoseev, G.; Kästner, J.; Ioppolo, S.; Linnartz, H. Importance of tunneling in H-abstraction reactions by OH radicals - The case of CH4 + OH studied through isotope-substituted analogs. *Astronomy & Astrophysics* **2017**, *599*, A132.

(25) Zhang, L.; Ye, L.; Wang, F.; Gao, W.; Yu, J.; Zhang, L. Prediction of Hydrogen Abstraction Rate Constants at the Allylic Site between Alkenes and OH with Multiple Machine Learning Models. *The Journal of Physical Chemistry A* **2024**, *128*, 761–772.

(26) Fernández-Ramos, A.; Miller, J. A.; Klippenstein, S. J.; Truhlar, D. G. Modeling the Kinetics of Bimolecular Reactions. *Chemical Reviews* **2006**, *106*, 4518–4584.

(27) Lupi, J.; Puzzarini, C.; Cavallotti, C.; Barone, V. State-of-the-Art Quantum Chemistry Meets Variable Reaction Coordinate Transition State Theory to Solve the Puzzling Case of the H2S + Cl System. *Journal of Chemical Theory and Computation* **2020**, *16*, 5090–5104.

(28) Vereecken, L.; Glowacki, D. R.; Pilling, M. J. Theoretical Chemical Kinetics in Tropospheric Chemistry: Methodologies and Applications. *Chemical Reviews* **2015**, *115*, 4063–4114.

View Article Online
DOI: 10.1039/D6DD00113K

(29) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C. et al. CREST—A program for the exploration of low-energy molecular chemical space. *The Journal of Chemical Physics* **2024**, *160*, 114110.

(30) Cavallotti, C.; Pelucchi, M.; Georgievskii, Y.; Klippenstein, S. J. EStokTP: Electronic Structure to Temperature- and Pressure-Dependent Rate Constants—A Code for Automatically Predicting the Thermal Kinetics of Reactions. *Journal of Chemical Theory and Computation* **2019**, *15*, 1122–1145.

(31) Grinberg Dana, A.; Ranasinghe, D.; Wu, O. H.; Grambow, C.; Dong, X.; Johnson, M.; Goldman, M.; Liu, M.; Green, W. H. ReactionMechanismGenerator/ARC: ARC 1.1.0. 2019; https://zenodo.org/records/3356849.

(32) Marks, J.; Gomes, J. Efficient Transition State Searches by Freezing String Method with Graph Neural Network Potentials. 2025; http://arxiv.org/abs/2501.06159, arXiv:2501.06159 [physics].

(33) Michelbach, C. A.; Tomlin, A. S. Automatic mechanism generation for the combustion of advanced biofuels: A case study for diethyl ether. *International Journal of Chemical Kinetics* **2024**, *56*, 233–262.

(34) Cai, L.; Pitsch, H. Optimized chemical mechanism for combustion of gasoline surrogate fuels. *Combustion and Flame* **2015**, *162*, 1623–1637.

(35) Tomlin, A. S. The role of sensitivity and uncertainty analysis in combustion modelling. *Proceedings of the Combustion Institute* **2013**, *34*, 159–176.

(36) Wang, H.; Sheen, D. A. Combustion kinetic model uncertainty quantification, propagation and minimization. *Progress in Energy and Combustion Science* **2015**, *47*, 1–31.

View Article Online
DOI: 10.1039/D6DD00113K

(37) Carey, F. A.; Sundberg, R. J. *Advanced Organic Chemistry: Part A: Structure and Mechanisms*; Springer Science & Business Media, 2007; Google-Books-ID: g5dYyJMBhCoC.

(38) Dill, K. A.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*; Garland Science, 2011; Google-Books-ID: _DKAQgAACAAJ.

(39) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *The Journal of Chemical Physics* **1958**, *29*, 546–572.

(40) Benson, S. W. *Thermochemical Kinetics: Methods for the Estimation of Thermochemical Data and Rate Parameters*; Wiley, 1976; Google-Books-ID: qURRAAAAMAAJ.

(41) Saeys, M.; Reyniers, M.-F.; Van Speybroeck, V.; Waroquier, M.; Marin, G. B. Ab initio group contribution method for activation energies of hydrogen abstraction reactions. *Chemphyschem: A European Journal of Chemical Physics and Physical Chemistry* **2006**, *7*, 188–199.

(42) Sumathi, R.; Carstensen, H.-H.; Green, W. H. Reaction Rate Prediction via Group Additivity Part 1: H Abstraction from Alkanes by H and CH3. *The Journal of Physical Chemistry A* **2001**, *105*, 6910–6925.

(43) Liu, M.; Grinberg Dana, A.; Johnson, M. S.; Goldman, M. J.; Jocher, A.; Payne, A. M.; Grambow, C. A.; Han, K.; Yee, N. W.; Mazeau, E. J. et al. Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *Journal of Chemical Information and Modeling* **2021**, *61*, 2686–2696.

(44) Green, W. H. Perspective on automated predictive kinetics using estimates derived from large datasets. *International Journal of Chemical Kinetics* **2024**, *56*, 637–648.

View Article Online
DOI: 10.1039/D6DD00113K

(45) Johnson, M. S.; Dong, X.; Grinberg Dana, A.; Chung, Y.; Farina, D. J.; Gillis, R. J.; Liu, M.; Yee, N. W.; Blondal, K.; Mazeau, E. et al. RMG Database for Chemical Property Prediction. *Journal of Chemical Information and Modeling* **2022**, *62*, 4906–4915.

(46) Johnson, M. S.; Green, W. H. A machine learning based approach to reaction rate estimation. *Reaction Chemistry & Engineering* **2024**, *9*, 1364–1380.

(47) Grambow, C. A.; Pattanaik, L.; Green, W. H. Deep Learning of Activation Energies. *The Journal of Physical Chemistry Letters* **2020**, *11*, 2992–2997.

(48) Chang, H.-C.; Tsai, M.-H.; Li, Y.-P. Enhancing Activation Energy Predictions under Data Constraints Using Graph Neural Networks. 2024; https://chemrxiv.org/engage/chemrxiv/article-details/675a8aa6085116a1332391ed.

(49) Spiekermann, K. A.; Pattanaik, L.; Green, W. H. Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy. *The Journal of Physical Chemistry A* **2022**, *126*, 3976–3986.

(50) Yu, J.; Shan, D.; Song, H.; Yang, M. A novel hybrid machine learning model for predicting rate constants of the reactions between alkane and CH3 radical. *Fuel* **2022**, *322*, 124150.

(51) Xia, M.; Zhang, Y.; Song, H.; Jia, Y.; Yang, M. Predicting Rate Constants of Hydrogen Abstraction Reactions between OH/HO2 and Alkanes by Machine Learning Models. *The Journal of Physical Chemistry A* **2025**, *129*, 309–316.

(52) Houston, P. L.; Nandi, A.; Bowman, J. M. A Machine Learning Approach for Prediction of Rate Constants. *The Journal of Physical Chemistry Letters* **2019**, *10*, 5250–5258.

(53) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-selectivity prediction with a machine-learned reac-

View Article Online
DOI: 10.1039/D6DD00113K

tion representation and on-the-fly quantum mechanical descriptors. *Chemical Science* **2021**, *12*, 2198–2208.

(54) Stuyver, T.; Coley, C. W. Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. *The Journal of Chemical Physics* **2022**, *156*, 084104.

(55) Vargas, S.; Gee, W.; Alexandrova, A. High-throughput quantum theory of atoms in molecules (QTAIM) for geometric deep learning of molecular and reaction properties. *Digital Discovery* **2024**, *3*, 987–998.

(56) Karwounopoulos, J.; Landsheere, J. D.; Galustian, L.; Jechtl, T.; Heid, E. Graph-based prediction of reaction barrier heights with on-the-fly prediction of transition states. *Digital Discovery* **2025**, *4*, 3208–3216.

(57) Jian, Y.; Zhang, Y.; Wei, Y.; Fan, H.; Yang, Y. Reaction Graph: Towards Reaction-Level Modeling for Chemical Reactions with 3D Structures. 2025.

(58) van Gerwen, P.; Briling, K. R.; Bunne, C.; Somnath, V. R.; Laplaza, R.; Krause, A.; Corminboeuf, C. 3DReact: Geometric Deep Learning for Chemical Reactions. *Journal of Chemical Information and Modeling* **2024**, *64*, 5771–5785.

(59) Vijay, S.; Venetos, M. C.; Spotte-Smith, E. W. C.; Kaplan, A. D.; Wen, M.; Persson, K. A. CoeffNet: predicting activation barriers through a chemically-interpretable, equivariant and physically constrained graph neural network. *Chemical Science* **2024**, *15*, 2923–2936.

(60) Song, Y.; Zheng, S.; Niu, Z.; Fu, Z.-h.; Lu, Y.; Yang, Y. Communicative Representation Learning on Attributed Molecular Graphs. 2020; pp 2831–2838.

(61) Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; sriniker;

View Article Online
DOI: 10.1039/D6DD00113K

Gedeck, P.; Jones, G.; NadineSchneider et al. rdkit/rdkit: 2024_09_2 (Q3 2024) Release. 2024; https://doi.org/10.5281/zenodo.13990314.

(62) Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry* **1999**, *20*, 720–729.

(63) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. et al. Gaussian 09 (Revision A02). *Gaussian Inc. Wallingford CT* **2009**,

(64) Bhoorasingh, P. L.; Slakman, B. L.; Seyedzadeh Khanshan, F.; Cain, J. Y.; West, R. H. Automated Transition State Theory Calculations for High-Throughput Kinetics. *The Journal of Physical Chemistry A* **2017**, *121*, 6896–6904.

(65) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics* **2020**, *22*, 7169–7192.

(66) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *Journal of Chemical Theory and Computation* **2019**, *15*, 2847–2862.

(67) Pracht, P.; Grimme, S. Calculation of absolute molecular entropies and heat capacities made simple. *Chemical Science* **2021**, *12*, 6551–6568.

(68) Spicher, S.; Plett, C.; Pracht, P.; Hansen, A.; Grimme, S. Automated Molecular Cluster Growing for Explicit Solvation by Efficient Force Field and Tight Binding Methods. *Journal of Chemical Theory and Computation* **2022**, *18*, 3174–3189.

(69) Pracht, P.; Bannwarth, C. Fast Screening of Minimum Energy Crossing Points with Semiempirical Tight-Binding Methods. *Journal of Chemical Theory and Computation* **2022**, *18*, 6370–6385.

View Article Online
DOI: 10.1039/D6DD00113K

(70) Pavošević, F.; Peng, C.; Pinski, P.; Riplinger, C.; Neese, F.; Valeev, E. F. SparseMaps—A systematic infrastructure for reduced scaling electronic structure methods. V. Linear scaling explicitly correlated coupled-cluster method with pair natural orbitals. *The Journal of Chemical Physics* **2017**, *146*, 174108.

(71) Peterson, K. A.; Adler, T. B.; Werner, H.-J. Systematically convergent basis sets for explicitly correlated wavefunctions: The atoms H, He, B–Ne, and Al–Ar. *The Journal of Chemical Physics* **2008**, *128*, 084102.

(72) Neese, F. The ORCA program system. *WIRES Comput. Molec. Sci.* **2012**, *2*, 73–78, Type: journal Article.

(73) Neese, F. Software update: the ORCA program system, version 5.0. *WIRES Comput. Molec. Sci.* **2022**, *12*, e1606, Type: journal Article.

(74) Kossmann, S.; Neese, F. Efficient Structure Optimization with Second-Order Many-Body Perturbation Theory: The RIJCOSX-MP2 Method. *Journal of Chemical Theory and Computation* **2010**, *6*, 2325–2338.

(75) Weigend, F.; Köhn, A.; Hättig, C. Efficient use of the correlation consistent basis sets in resolution of the identity MP2 calculations. *The Journal of Chemical Physics* **2002**, *116*, 3175–3183.

(76) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Physical Chemistry Chemical Physics* **2006**, *8*, 1057–1065.

(77) Yousaf, K. E.; Peterson, K. A. Optimized auxiliary basis sets for explicitly correlated methods. *The Journal of Chemical Physics* **2008**, *129*, 184108.

(78) Grinberg Dana, A.; Johnson, M. S.; Allen, J. W.; Sharma, S.; Raman, S.; Liu, M.; Gao, C. W.; Grambow, C. A.; Goldman, M. J.; Ranasinghe, D. S. et al. Automated

View Article Online
DOI: 10.1039/D6DD00113K

reaction kinetics and network exploration (Arkane): A statistical mechanics, thermodynamics, transition state theory, and master equation software. *International Journal of Chemical Kinetics* **2023**, *55*, 300–323.

(79) Eckart, C. The Penetration of a Potential Barrier by Electrons. *Physical Review* **1930**, *35*, 1303–1309.

(80) Pieters, C. Hydrogen Abstraction Reaction Data. 2026; https://zenodo.org/records/18597964.

(81) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M. et al. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **2019**, *59*, 3370–3388.

(82) Spiekermann, K. A.; Dong, X.; Menon, A.; Green, W. H.; Pfeifle, M.; Sandfort, F.; Welz, O.; Bergeler, M. Accurately Predicting Barrier Heights for Radical Reactions in Solution Using Deep Graph Networks. *The Journal of Physical Chemistry A* **2024**, *128*, 8384–8403.

(83) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, *5*, 107–113.

(84) Kelley, B. DescriptaStorus. https://github.com/bp-kelley/descriptastorus.

(85) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *Journal of Chemical Information and Modeling* **2024**, *64*, 9–17.

(86) Huber, P. J. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* **1964**, *35*, 73–101.

View Article Online
DOI: 10.1039/D6DD00113K

(87) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137–148, _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1969.10490666.

(88) Tanimoto, T. T. *An Elementary Mathematical Theory of Classification and Prediction*; IBM Internal Report, 1957.

(89) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA, 2019; pp 2623–2631.

(90) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. Advances in Neural Information Processing Systems. 2017.

(91) Yeo, I.; Johnson, R. A. A new family of power transformations to improve normality or symmetry. *Biometrika* **2000**, *87*, 954–959.

(92) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Computer Physics Communications* **2016**, *203*, 212–225.

(93) Grinberg Dana, A.; Liu, M.; Green, W. H. Automated chemical resonance generation and structure filtration for kinetic modeling. *International Journal of Chemical Kinetics* **2019**, *51*, 760–776, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/kin.21307.

(94) Manion, J. A.; Huie, R. E.; Levin, R. D.; Burgess, D. R. J.; Orkin, V. L.; Tsang, W.; McGivern, W. S.; Hudgens, J. W.; Knyazev, V. D.; Atkinson, D. B. et al. *NIST Chemical Kinetics Database*; 2015.

View Article Online
DOI: 10.1039/D6DD00113K

(95) Tsang, W.; Hampson, R. F. Chemical kinetic data base for combustion chemistry. Part I. Methane and related compounds. *Journal of Physical and Chemical Reference Data* **1986**, *15*, 1087–1279.

View Article Online
DOI: 10.1039/D6DD00113K

Data availability

The datasets supporting this article are provided in the Supplementary Information and are archived on Zenodo at DOI: https://doi.org/10.5281/zenodo.20433305. The custom code associated with this work, including scripts for model training, evaluation, and analysis, is available from the GitHub repository https://github.com/calvinp0/chemprop_arrhenius. A persistent archived snapshot of the GitHub repository is available on Zenodo at DOI: https://doi.org/10.5281/zenodo.20485495.