Electronic Supplementary Material (ESI) for Digital Discovery.

Repurposing Quantum Chemical Descriptor Datasets for on the Fly Generation of Informative Reaction Representations: Application to Hydrogen Atom Transfer Reactions

Javier E. Alfonso-Ramos,[†] Rebecca M. Neeser,[‡,¶] and Thijs Stuyver[*,†]

†*Ecole Nationale Supérieure de Chimie de Paris, Université PSL, CNRS, Institute of Chemistry for Life and Health Sciences, 75 005 Paris, France*

‡*Massachusetts Institute of Technology, Department of Chemical Engineering, 02139 Cambridge (MA), United States.*

¶*ETH Zürich, Institute of Pharmaceutical Sciences, 8093 Zürich, Switzerland.*

E-mail: thijs.stuyver@chimieparistech.psl.eu

Contents

S1 HAT dataset construction workflow

Any pair of C-H bond dissociation reactions[1] can be combined into a HAT reaction as follows:

$$\textbf{BDE reaction 1:} \quad X - H \rightarrow X\bullet + H\bullet \quad \Delta G_r = \mathrm{BDFE}_1 \tag{1}$$

$$\textbf{BDE reaction 2:} \quad Y - H \rightarrow Y\bullet + H\bullet \quad \Delta G_r = \mathrm{BDFE}_2 \tag{2}$$

$$\textbf{resulting HAT reaction:} \quad X-H + Y\bullet \rightarrow X\bullet + Y-H \quad \Delta G_r = \mathrm{BDFE}_1 \text{ - } \mathrm{BDFE}_2 \tag{3}$$

where BDFE stands for the bond dissociation free energy.

Consequently, the BDE-db dataset by St. John et al.[1] enables in principle the combinatorial construction of almost 40 billion HAT reactions – an intractable number, even for regular enumeration and reaction SMILES parsing. To make the data set size manageable, we generated two distinct samples of C-H bond dissociation reactions, one million reactions each. These samples were then combined on an entry-by-entry basis, resulting in a random sample of 1 million HAT reactions across the full 40B reaction space.

This final chemical reaction space has a $\Delta \mathrm{G}_{rxn,mean}$ = - 0.014 kcal/mol and standard deviation of 14.50 kcal/mol (Fig 1). As the initial dataset contained only the C, H, N and O elements and focused on organic compounds (which tend to be dominated by carbon chains), almost 70 percent of the dataset consists of HAT reactions involving two carbon atoms (Table 1)

From this full reaction dataset, a computationally tractable subset of 2000 data points was extracted for explicit reaction profile computation. During reaction sampling, we aimed to cover as much structural diversity present in the chemical space as possible, to maximize the generalizability of the eventual machine learning model trained on the computed data

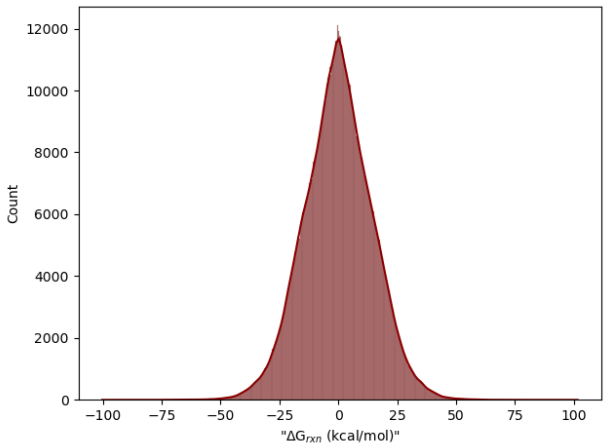

Figure 1: Distribution of ΔG across the hypothetical chemical space.

Table 1: Statistics of the HAT reactions across the hypothetical chemical space.

Bond Broken	Bond Formed	Count	Percent
C-H	C-H	717901	71.79
C-H	H-N	78873	7.87
C-H	H-O	50403	5.04
H-N	C-H	79235	7.92
H-N	H-N	8815	0.88
H-N	H-O	5548	0.55
H-O	C-H	50093	5.01
H-O	H-N	5508	0.55
H-O	H-O	3616	0.36

across the full chemical space. To this end, each reaction was first encoded into a binary reaction fingerprint, more specifically the differential reaction fingerprint.[2] Next, an initial sample of 2000 reactions was selected, and the distance between each reaction pair was determined. For each pair of reactions for which the cosine distance amounted to less than 0.85, the latter reaction was discarded from the sample. Subsequently, new reactions were added to the sample until the sample size reached 2000 again, and the same procedure was replicated. This iterative process was repeated until all sampled reactions were sufficiently different so that their mutual reaction fingerprint distances amounted to the threshold set, i.e., 0.85.

To construct the dataset of reaction profiles for HAT reactions in a fully automated

manner, the Python package autodE[3] in combination with Gaussian16[4] was used. The first step taken by autodE is the generation of the 3D structures for reactants and products based on the provided reaction SMILES string.[5] More specifically, conformers are generated with the ETKDGv3 algorithm[6] as implemented in RDKit,[7] after which an optimization at GFN2-xTB level of theory is performed.[8] Next, the conformers are ranked based on their relative energy. If the *ade.Config.hmethod_conformers* keyword is set to False, then the energy values obtained during the xTB optimization are used for the ranking. Otherwise, a refined value obtained through a single-point low-level DFT (M06-2X/def2-SVP) calculation is used as the ranking criterion (we selected the latter, *vide infra*).[9,10] For the transition state location, from the molecular graphs of reactants and product species, the set of bond rearrangements that leads from the reactants graph to the products graph is identified. A guess transition state (TS) is obtained through a series of constrained optimizations of a truncated system. The guess TS is then refined through an optimization, followed by an analysis of the imaginary vibration mode, to ensure that the correct bonds are being broken/formed throughout the reaction. TS conformers are generated with the help of the randomize-and-relax algorithm, the conformer with the lowest energy is optimized and its imaginary vibration mode is checked once again. Next, geometry optimization and frequency/thermal corrections for reactants, products, and transition state structures are performed at the M06-2X/def2-SVP level of theory at standard conditions (298.15 K and 1 atm). Single-point energy refinements were finally computed at M06-2X/def2-TZVP level of theory.[11] This final DFT level of theory was selected based on a previous benchmarking study of St. John et al.[12] All calculations were performed in the gas phase.

By default, autodE generates 300 conformers and uses a root mean squared displacement (RMSD) threshold of 0.3Å to exclude identical conformers. Several tests to ensure reproducibility and accuracy were performed on a small set of 30 reactions(cf. Fig. 2). The 30 reactions were first computed twice with respectively [300, 600, 1000, 1500, 2000] conformers generated, an RMSD threshold of 0.1Å and with both options for ranking the conformers

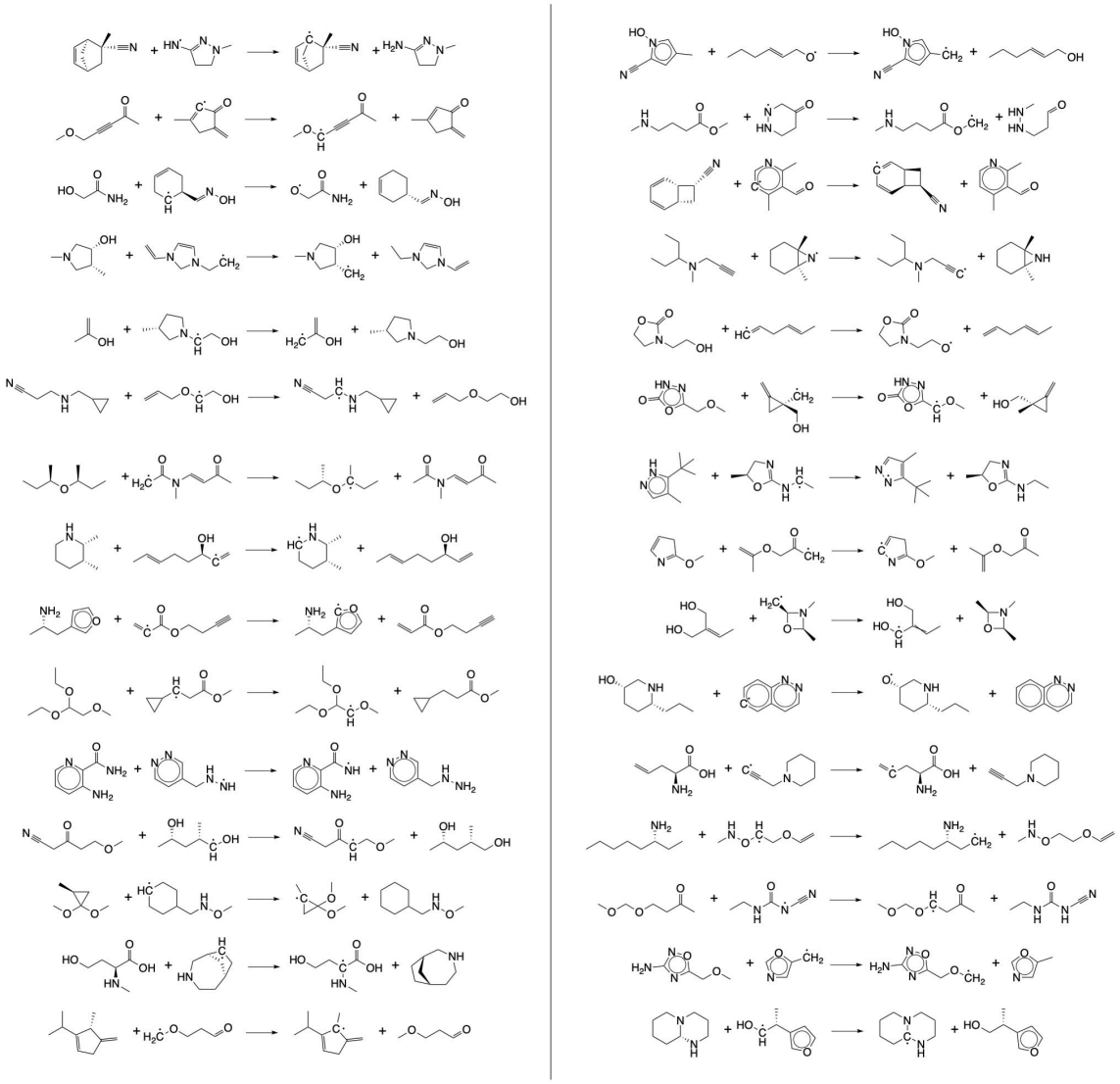

Figure 2: The 30 reactions used for determining the autodE parameters.

(i.e., single-point low-level DFT or xTB level of theory). The best results were obtained when 1000 conformers were generated, and hence we selected these settings throughout our analyses below.

Next, we aimed to assess the effect of the different ranking criteria. In first instance, the reaction profiles for the 30 reactions were computed twice with conformer ranking based on single-point DFT energies. From Fig. 3a-b, it is clear that both activation and reaction energies for our test set are reproduced well (MAE $\sim$ 1.0 kcal/mol and RMSE $\sim$ 1.4 kcal/mol

for the activation energies and MAE $\sim$ 0.6 kcal/mol and RMSE $\sim$ 0.9 kcal/mol for the reaction energies). Subsequently, two consecutive runs, one using conformers ranked based on GFN2-xTB energies and a second using (single-point) DFT energies, were performed. As can be seen from Fig. 3c-d, the change in the level of theory for the conformers ranking diminishes significantly the reproducibility (MAE $\sim$ 1.9 kcal/mol and RMSE $\sim$ 3.7 kcal/mol for the activation energies and MAE $\sim$ 0.7 kcal/mol and RMSE $\sim$ 1.0 kcal/mol for the reaction energies). As a final test, we aimed to check the reproducibility of the workflow with conformer selection consistently at the GFN2-xTB level of theory. As can be observed from Fig 3e-f, a significant difference can be found between both activation energies (MAE $\sim$ 2.1 kcal/mol and RMSE $\sim$ 4.2 kcal/mol) and reaction energies (MAE $\sim$ 0.5 kcal/mol and RMSE $\sim$ 1.0 kcal/mol), indicating that conformer ranking based on GFN2-xTB energies is not accurate nor reproducible. As such, based on the results above, we decided to generate 1000 conformers for each species and select the lowest energy conformer for reactant and product from single-point DFT computations.

We also checked whether reactant and product complexes should be considered in the computed reaction profiles. For the same subset of 30 reactions, we computed reaction profiles both with and without complexes. For 19 out of the 22 successful profiles completed, the complexes ended up higher in energy than the isolated reactants when thermal corrections at room temperature were included, indicating that complexation is irrelevant in these cases (cf. Table 2). For the three cases for which a stable complex could be located, the stabilization energy relative to the separated reactants amounted to a mere 1-2 kcal/mol. Considering the computational cost of systematically computing the reaction complex, in combination with the fact that attempts to identify complexes for our radical reactions inherently have a particularly high risk of resulting in an unintentional change in bonding due to a barrierless addition of the radical to unsaturated bonds, we decided not to compute complexes altogether, and consistently use the isolated reactants and products as the reference for activation and reaction energy computation.

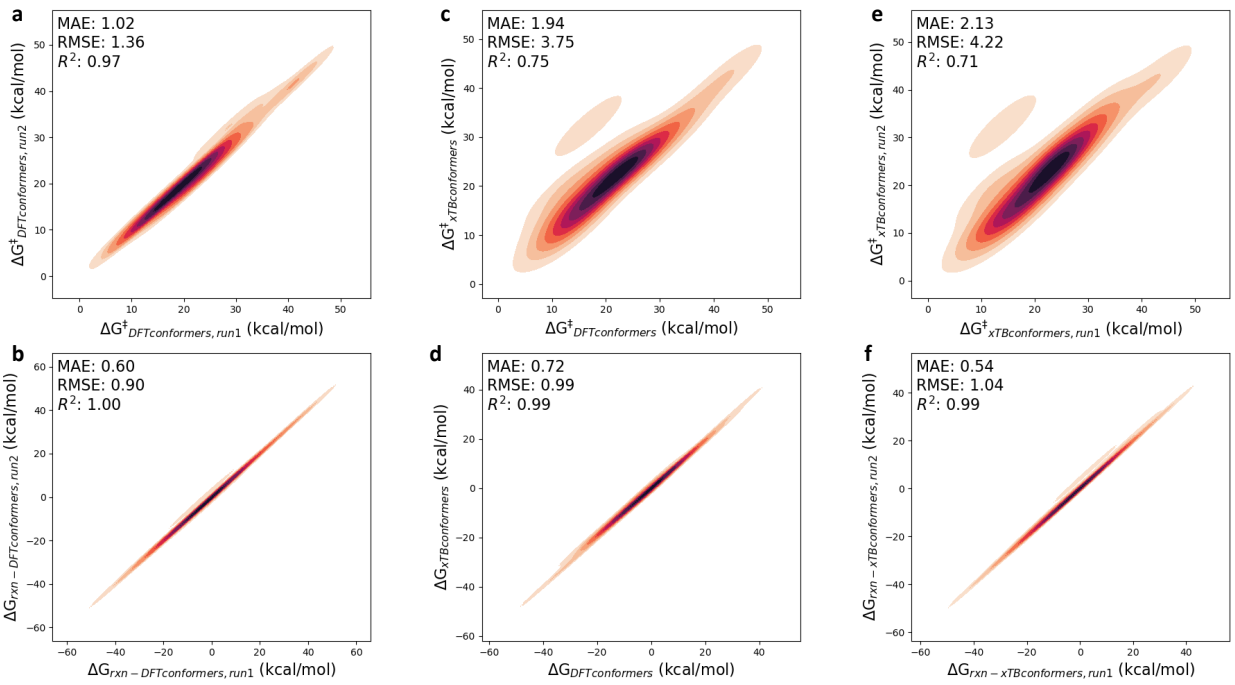

Figure 3: Correlation between (a) the activation energies (ΔG_{act}) and (a) the reaction energies (ΔG_{rxn}) for two consecutive autodE runs with conformer selection at DFT level of theory. Correlation between (c) ΔG_{act} and (d) ΔG_{rxn} for an autodE run with conformer selection at DFT level of theory and a consecutive run with conformer selection at GFN2-xTB level of theory. Correlation between (e) ΔG_{act} and (f) ΔG_{rxn} for two consecutive autodE runs with conformer selection at GFN2-xTB level of theory.

By default, autodE[3] does not exchange stereochemical information between reactants and products in a reaction SMILES; it simply searches for the conformation that has the lowest energy globally for each species independently. The transition state (TS) on the other hand inherits the stereochemistry from one side; by default this is the reactant side. In the case of our dataset of HAT reactions, in the TS, the hydrogen atom will be abstracted by an atom with sp^2 hybridization. Consequently, stereochemical compatibility between products and TSs is not inherently guaranteed.

A conflicting stereochemistry can emerge in two manners in our data. In the first case, only the product side contains a stereocenter, i.e.,

$$\text{Case 1} \quad X - H \; + \; Y\bullet \rightarrow X\bullet \; + Y^{*} - H \tag{4}$$

Table 2: Complexation energies for the subset of 30 reactions, where $\Delta E_{complex}$ stands for the electronic energy, and $\Delta G_{complex}$ stands for the Gibbs Free energy (at standard conditions, i.e., 298.15 K and 1 atm).

Reaction Index	$\Delta E_{complex}$ (kcal/mol)	$\Delta G_{complex}$ (kcal/mol)
456222	-6.43	5.36
138021	-7.94	5.35
518224	-10.02	3.46
618981	-9.20	2.75
569328	-8.48	3.79
909735	-13.51	-1.41
694905	-12.43	0.40
573275	-9.49	3.14
87255	-8.72	3.82
966587	-6.27	5.39
859423	-14.66	-2.79
936714	-9.72	2.77
1454896	-8.42	3.95
387409	-10.34	1.62
281679	-12.18	1.20
184059	-10.97	2.38
98148	-11.00	0.67
1249926	-14.68	-1.98
247298	-6.39	5.73
831637	-8.36	5.07
720962	-10.59	2.43
379829	-13.99	0.34
909565	-4.66	7.74
957024	-10.65	2.30

where stereocenters are indicated with a * sign. For this simple case, we decided to simply invert the input reaction SMILES. Since the reactant now carries the stereocenter, the TS will inherit the correct stereochemistry.

In the second case, both sides of the reaction contain a stereocenter involved in the reaction.

$$\text{Case 2} \quad X^* - H \; + \; Y\bullet \rightarrow X\bullet \; + Y^* - H \tag{5}$$

Given the reactant and TS geometries, the compatible product conformer can be deter-

mined for this case as well in principle, but this can only be done after an initial version of the profile has already been generated.

As such, to deal with this second case, a script to correct potential stereochemical incompatibilities for the products was executed once the full reaction profile was obtained. In a first step of this script, the coordinates of the subset of atoms in the TS geometry corresponding to the product molecule are extracted. The extracted geometry is then optimized with GFN2-xTB and converted to a SMILES string. If the stereochemistry is not the same for the original product and the SMILES string obtained from the TS geometry, then the latter geometry is fully optimized according to the regular autodE workflow, and the reaction profile is updated. In this step, one error was found and removed.

Despite the TS quality checks present in autodE, some erroneous reaction profiles were not recognized as such and consequently, had to be filtered out manually. In first instance, reactions with negative activation energies were removed (2 in total). Since all of these reactions involved low imaginary frequencies, we subsequently inspected visually the normal modes of all the TSs with imaginary frequencies ($|\nu_{imag}| < 500$ cm^{-1}), resulting in the removal of another 13 reactions.

Once autodE generates the conformers of the TS, the connectivity is not revisited and can lead to erroneous structures. Using the same autodE functions, the graphs for both reactants and TS were generated and the connectivity between every atom was checked. In 24 cases, erroneous structures were found and removed.

Overall, out of the 2000 reactions considered, 1511 profile reactions were computed successfully (75.6%) according to the workflow described above. For 44 cases, no TS could be found. The main alternative source of failures was related to errors in the optimizations in several instances of the autodE workflow. For 46 cases, time limit was reached.

In order to take into account the tunneling effect, the semi-classical rate constant is defined as

$$k^{SC}(T) \;=\; \kappa(T) k^{TST}(T) \tag{6}$$

where κ is the tunneling transmission coefficient and k^{TST} is the conventional Transition State Theory (TST) rate constant.

The k^{TST} is computed through the Eyring-Polanyi equation as

$$k^{TST} \;=\; \frac{k_B T}{h} e^{(-\Delta G^{\ddagger}/RT)} \tag{7}$$

where k_B is the Boltzmann constant, T is the absolute temperature, h is the Planck constant and R is the gas constant.

The tunneling transmission coefficient (κ) is defined as the ratio of the thermally averaged quantum tunneling probability and the quasiclassical transmission probability. Under the assumption that the shape of the potential energy along the reaction coordinate s can be approximated by an Eckart potential:

$$V(s) \;=\; \frac{y\Delta V}{1+y} \;+\; \frac{By}{(1+y)^2} \tag{8}$$

where ΔV is the reaction energy, y is a parameter that depends on the force constant of the normal mode of the transition and the value of the potential at its maximun and B is

$$B \;=\; [V_{max}^{1/2} + (V_{max} - \Delta V)^{1/2}]^2 \tag{9}$$

where V_{max} is the barrier height, the tunneling probability $P^T(E)$ can be analytically evaluated as

$$P^T(E) \;=\; \frac{cosh(a+b) - cosh(a-b)}{cosh(a+b) + cosh(d)} \tag{10}$$

where

$$a \;=\; 2\pi(2\mu E)^{1/2}/(\hbar\alpha) \tag{11}$$

$$b \;=\; 2\pi[2\mu(E-\Delta V)]^{1/2}/(\hbar\alpha) \tag{12}$$

$$d \;=\; 2\pi[2\mu B-\hbar^2\alpha^2/4]^{1/2}/(\hbar\alpha) \tag{13}$$

The expression of the tunneling transmission coefficient is

$$\kappa(T) \;=\; 1 \;+\; \frac{2}{k_BT}\int_{E_0}^{V^{max}} sinh[(V_{max}-E)/k_BT]P^T(E)dE \tag{14}$$

For the numerical integration of $\kappa(T)$, 10-point Gauss-Legendre quadrature with a change of integration interval was used. In those cases where the barrier height is negative, the term B is undefined and the Wigner tunneling approximation was used. $\kappa(T)$ is expressed as

$$\kappa(T) \;=\; 1 \;+\; \frac{1}{24}(\frac{h|\nu^{\ddagger}|}{k_BT})^2 \tag{15}$$

where $\nu^{\ddagger}$ is the frequency of the normal mode of the transition.

The final activation energy with tunneling correction included is calculated using the Eyring-Polanyi equation with the semi-classical rate constant k^{SC}.

S1.1 Data Records

All data files produced as part of this study are accessible through (`https://figshare.com/projects/Hydrogen_atom_transfer_reactions/188007`). Reaction IDs and SMILES, activation energies ($\Delta G^{\ddagger}$; in kcal/mol), activation energies with tunneling corrections ($\Delta G^{\ddagger}_{corr}$; in kcal/mol), and reaction energies (ΔG_{rxn}; in kcal/mol) for each computed reaction profile are provided in CSV format. Gaussian log-files for both the final frequency and single-point calculation for each reactant (both the original and stereo-constrained versions), TS and product species, XYZ-files of final geometries as well as a CSV file containing computed electronic energies and thermal corrections are available in a compressed archive file, full_dataset_profiles.tar.gz.

The files have been organized per reaction profile, identified through the reaction ID.

Within each directory, reactant XYZ-files are of the form r_#######.xyz, product XYZ-files are of the form p_#######.xyz, and transition state XYZ-files are of the form TS_#######.xyz. If the product had to be corrected to enforce stereochemical compatibility, the latter XYZ-files are included under to form of alt_p_#######.xyz. The frequency log-files can be found in a frequency_logs directory, and the single-point log-files can be found in a single_point_logs directory. The energies for all of these species are summarized per directory in energies.csv.

Additionally, the benchmarking data are made available in the benchmarking_data.tar.gz directory (*vide supra*).

S2 In-depth technical description of the surrogate model for QM descriptor prediction

First, the SMILES representation of the molecule is transformed into a graph-based representation. In this graph-based representation, each atom is represented as a node and each bond as an edge. The neighboring atoms $\mathcal{N}(v)$ of atom v are those nodes connected to v by edges. The initial featurization of each atom includes atom type, degree, explicit and implicit valence, formal charge, information on aromaticity and number of radical electrons on the atomic level and on the bond-level the bond type (single, double, triple or aromatic) and information about whether the bond is conjugated or in a ring. Atom features x_v and bond features e_{vw} are fed to a D-MPNN. The D-MPNN operates in two phases: a message-passing phase, which transmits information across the molecule to build a neural representation of the molecule, and a readout phase, which uses the final representation of the molecule to make predictions about the properties of interest.

The message-passing phase starts with the initialization of the edge hidden state according to

$$h_{vw}^{0} \;=\; \tau(W_i \; cat(x_v, e_{vw})) \tag{16}$$

where τ is the ReLU activation function, $W_i \in \mathbb{R}^{hXh_i}$ is a learned matrix and $\text{cat}(x_v, e_{vw} \in \mathbb{R}^{h_i})$ is the concatenation of the atom features x_v for atom v and the bond features e_{vw} for bond vw.

Next, the hidden states h_{vw}^{t} and the messages m_{vw}^{t} associated with each bond e_{vw} are updated at each step t until maximal depth T is reached using a message function M_t. The corresponding message-passing equations are

$$m_{vw}^{t+1} \;=\; \sum_{k \in \{N(v) \backslash w\}} M_t(x_v, x_w, h_{kv}^{t}) \tag{17}$$

$$h_{vw}^{t+1} \;=\; U_t(h_{vw}^{t}, m_{vw}^{t+1}) \tag{18}$$

for $t \in \{1, ..., T\}$. The current model defines $M_t(x_v, x_w, h_{vw}^{t}) = h_{vw}^{t}$ and implements U_t with shared parameters across layers t

$$U_t(h_{vw}^{t}, m_{vw}^{t+1}) \;=\; U(h_{vw}^{t}, m_{vw}^{t+1}) \;=\; \tau(h_{vw}^{0} + W_m m_{vw}^{t+1}) \tag{19}$$

where $W_m \in \mathbb{R}^{hXh_i}$ is a learned matrix with hidden size h.

In the last stage of the message-passing phase, the atom representation of the molecule is recovered by summing the incoming bond features according to

$$m_v \;=\; \sum_{w \in N(v)} h_{vw}^{T} \tag{20}$$

$$h_v \;=\; \tau(W_a \; cat(x_v, m_v)) \tag{21}$$

and similarly for the bond representation:

$$h_{vw} \;=\; \tau(W_b\; cat(e_{vw}, h_{vw}^{T})) \tag{22}$$

where $W_{a,b} \in \mathbb{R}^{hXh_i}$ is a learned matrix.

The learned atomic/bond representations are then converted into the corresponding descriptors through a multi-task readout layer.

With respect to the atom-level descriptors, a value for every atom is predicted independently by the FFNN, which means the sum of all values does not necessarily equal the global value of the property, e.g. in a doublet radical species, the sum of the predicted spin density of all the atoms may not be equal to 1 for a radical species. To fix this issue, an attention-based mechanism was implemented. This mechanism determines a weight factor for each atom, indicating how much its predicted value needs to be corrected. The final predicted value is generated from the initial atomic value and this weight.

As such, for the atom-level descriptors, e.g. atomic charges, the uncorrected descriptors are calculated first as:

$$q_v \;=\; \mathrm{FFNN}(h) \tag{23}$$

where h is the corresponding atomic/bond feature vector. The final corrected descriptor subject to the constraint can then be calculated as:

$$\hat{a}_v \;=\; \mathrm{FFNN}(h) \tag{24}$$

where $\hat{a}_v$ is an atom-level vector of the same dimensions as h, followed by:

$$w_v \;=\; \frac{\exp(u\hat{a}_v)}{\sum_v \exp(u\hat{a}_v)} \tag{25}$$

where u is a learnable atom-level vector that can be seen as a high level representation of a fixed query "which atom needs more correction?", and get a normalized weight w_i through

a softmax function, and then:

$$q_v^{final} \;=\; q_v \;+\; \frac{w_v(Q - \sum_v q_v)}{\sum_v w_v} \tag{26}$$

where Q is the constraint applied on the descriptor such that:

$$\sum_v q_v^{final} \;=\; Q \tag{27}$$

To predict global properties, the atom-centered feature vectors are either sum-pooled to predict extensive molecular properties such as total enthalpy, total free energy, total entropy and formation energy or mean-pooled to use as input to predict intensive properties such as HOMO and LUMO energy and the HOMO-LUMO gap. Since an extensive molecular property is dependent on the number of atoms the pooling function should also be dependent on the number of aggregated nodes requiring summation and vice versa for extensive properties

$$c_{ext} \;=\; \sum_v h_v \tag{28}$$

$$c_{int} \;=\; \frac{\sum_v h_v}{V} \tag{29}$$

where V is the number of atoms in the molecule.

S2.1 Hyperparameter optimization

To find a reasonable set of hyperparameters for the surrogate model, a grid search was performed in which the depth of the D-MPNN $t \in \{5, 6, 7, 8\}$, the number of FFNN layers $L \in \{2, 3, 4\}$, the D-MPNN hidden size $n_{h,D-MPNN} \in \{300, 600, 900, 1200\}$, and the FFNN hidden size $n_{h,FFNN} \in \{300, 600, 900\}$ was varied. The models were trained for 100 epochs with a batch size of 50 in a fixed, random 80/10/10-split. A SINEXP learning rate scheduler was used. The best combination of hyper-parameters was $t = 5$, $L = 4$, $n_{h,D-MPNN} =$

1200 and $n_{h,FFNN} = 900$. These were chosen based on the performance of the model on the validation set.

Once the hyper-parameters were selected, a new data split was generated, in which all the radical species and molecules present in the reactivity dataset (*vide supra*) were taken as the test set (around 1% of the data), and the remaining data was split randomly into 80% training and 20% validation. Because of the distinct splits used during hyperparameter optimization and actual training, overfitting of the hyperparameters to the test set is avoided.

S3 Reactivity model architectures and performances

As indicated in the main text, several architectures were tested for the downstream reactivity model. Whenever applicable, hyper-parameters were selected through a Bayesian Optimization-based (BO) search.[13] The accuracy on the $\Delta G^{\ddagger}$ prediction task was determined in 10-fold cross-validation, on a fixed data split (cf. `https://github.com/chimie-paristech-CTM/bde_hat/tree/master/scripts/baseline_models/splits`). During BO, a test set of 20% of the data was held out, after which 4-fold cross-validation on the remaining data was performed. The average root mean square error across the different folds was selected as the metric to optimize.

S3.1 Baseline models

Three model architectures were tested using the differential reaction fingerprints (DRFP; radius = 3; nbits = 2048) as input.[2] The DRFP algorithm takes a reaction SMILES as an input and creates a binary fingerprint based on the symmetric difference of two sets containing the circular molecular n-grams generated from the molecules listed left and right from the reaction arrow, respectively, without the need for distinguishing between reactants and reagents.

The first fingerprint-based model tested was k-nearest neighbors. The number of neigh-

bors (k) was set as a hyperparameter. Across 32 iterations of BO, the best accuracy was obtained for $k = 11$.

The next model architecture tested was a random forest. The hyperparameters for this model architecture are the maximal fraction of features ($max_{features}$), the minimal number of samples per leaf ($min_{samples-leaf}$), and the number of estimators ($n_{estimators}$). Optimal performance during BO optimization (64 iterations) was achieved for $max_{features} = 0.1$, $min_{samples-leaf} = 20$ and $n_{estimators} = 30$.

An XGBoost model was set up as well. The hyperparameters for this model architecture are γ, learning rate (lr), maximal depth of a tree (max_{depth}), the minimum sum of instance weights needed in a child ($min_{child-weight}$) and the number of estimators ($n_{estimators}$). After 128 iterations of Bayesian optimization, optimal performance was achieved for $\gamma = 4$, $lr = 0.05$, $max_{depth} = 5$, $min_{child-weight} = 4$ and $n_{estimators} = 300$.

Next to fingerprint-based models, a couple of graph neural network (GNN) models were considered as well. The first model tested was a Weisfeiler Lehman GNN (WL-GNN) one, adapted from our recent work on cycloaddition reactions.[14] A 7-dimensional hyperparameter search space was defined and 256 iterations of Bayesian optimization were performed. Since the training process of this model is time-consuming, the same parameters were used for the QM-augmented WL-GNN (*vide infra*).

We also tested a second GNN model architecture, namely Chemprop. Here, we used the condensed graph of reaction (CGR) representation, and performed 64 iterations of Bayesian optimization to determine the hyperparameters. The best results were obtained for *depth* = 6, *dropout* = 0.2, *FFN-hidden-size* = 500, *FFN-num-layers* = 2, *hidden-size* = 500.

Finally, a kernel ridge regression (KRR) model architecture were tested as well. Here, we focused on the recently introduced Bond-Based Reaction Representation.[15] For this model, the hyperparameters are the Gaussian kernel width σ and the regularisation parameter λ. After a grid search of 81 combinations, the best results were obtained for $\sigma = 100000$ and $\lambda = 10^{-9}$.

In Table 3, the performance of the various baseline models is presented.

Table 3: Performance on $\Delta G^{\ddagger}$, in terms of mean absolute error (MAE) and root mean square error (RMSE), of the baseline model architectures tested.

model	MAE (kcal/mol)	RMSE (kcal/mol)
k-nearest neighbors	5.33	7.13
random forest	5.25	6.96
XGBoost	5.32	7.03
KRR	3.55	4.71
WL-GNN	3.48	4.92
Chemprop	2.98	4.14

S3.2 Models based on the predicted valence bond-inspired representation

For the representation based on descriptors, the valence bond-inspired representation was used. For the atom-level descriptors, the spin densities on the radical centers on the reactant and product side were selected. Additionally, the partial charges on the radical and abstraction sites, as well as on the abstracted hydrogens, on both sides of the reaction were included. Furthermore, 4 reaction-level descriptors (relaxed $\mathrm{BDFE}_{forward/reverse}$ and frozen $\mathrm{BDE}_{forward/reverse}$) were also selected. Finally, the buried volumes, V_{bur} of both radicals were included, yielding 14 input features in total.

The simplest model tested was multivariate linear regression with descriptors. Subsequently, a k-nearest neighbors model with fingerprints was set up. Across 32 iterations of BO, the best value obtained for k was 5.

The next model architecture tested was a random forest. Optimal performance across 64 iterations of the BO optimization campaign was achieved for $max_{features} = 0.8$, $min_{samples-leaf} = 1$ and $n_{estimators} = 600$.

An XGBoost model was set up as well. After 128 iterations of Bayesian optimization, optimal hyperparameters for the descriptor-based featurization were determined to be $\gamma = 2$, $lr = 0.2$, $max_{depth} = 2$, $min_{child-weight} = 10$ and $n_{estimators} = 700$.

For the QM-augmented WL-GNN, optimal hyperparameters were *depth WLN* = 6, *weight factor atom vectors* = 0.4, *weight factor reaction vector* = 0.7, *initial learning rate* = 0.00219, *learning rate ratio* = 0.93, *depth FFNN* = 1 and *hidden size multiplier* = 10.

Finally, a feedforward neural network (FFNN) was constructed. A 4-dimensional search space was set up for this architecture, and 256 iterations of BO were performed (see Table 4 for a summary).

Table 4: Definition of the search space and the optimal parameter values emerging from the Bayesian optimization for the descriptor-based feed-forward neural network.

hyperparameter	min	max	distribution	optimal
layers	0	3	quniform	0
hidden size	10	300	quniform	230
initial learning rate	ln(0.01)	ln(0.08)	log uniform	0.0277
learning rate ratio	0.90	0.99	quniform	0.95

In Table 5, the performance of the various model architectures is presented. Comparison between Tables 3 and 5 leads to the conclusion that descriptor-based models outperform the fingerprint-based models by several kcal/mol.

Table 5: Performance on $\Delta G^{\ddagger}$, in terms of mean absolute error (MAE) and root mean square error (RMSE), of the descriptor-based model architectures tested.

model	MAE (kcal/mol)	RMSE (kcal/mol)
linear regression	2.28	3.14
k-nearest neighbors	2.60	3.67
random forest	2.10	3.01
XGBoost	2.24	3.14
WL-GNN	2.46	3.36
FFNN	1.98	2.78

S3.3 Δ-ML model

Because of the fairly good correlation between reaction and activation energies (cf. Figure 4c in the main text), a Δ-ML model to predict deviations from the thermodynamic-kinetic trend line was designed as well.

To this end, estimated reaction energies (ΔG^{est}_{rxn}) are first determined based on the BDFE values outputted by the surrogate model (cf. Eq. 3). Subsequently, a linear model between $\Delta G^{\ddagger}$ and the ΔG^{est}_{rxn} is fitted according to:

$$\Delta G^{\ddagger} = a\Delta G^{est}_{rxn} + b \tag{30}$$

An ML model is subsequently trained to predict the deviations in the actual activation energy relative to this trendline, i.e., the model target becomes,

$$\Delta\Delta G^{\ddagger} = \Delta G^{\ddagger} - \Delta G^{\ddagger} \tag{31}$$

We considered two architectures for this model: a fingerprint representation in combination with a RF, and an FFNN based on the predicted valence bond-inspired representation. The same hyperparameters were used for these models as in the corresponding non-delta learning models (*vide supra*). The former model reaches an MAE of 2.84 and RMSE of 3.73 on our in-house dataset, the latter reaches an MAE of 1.97 and RMSE of 2.76. As such, we unequivocally recover the benefit of using our surrogate-predicted representation, and the Δ-ML with access to the entire surrogate representation reaches an equivalent accuracy as the regular FFNN model.

S4 Analysis of descriptor importance in our in-house dataset of HAT reactions

In Table 6, an overview of the performance of linear regression models trained with only subsets of the full informative representation is provided. For the other model descriptor-based architectures, similar trends were obtained.

Table 6: Performance on $\Delta G^{\ddagger}$, in terms of mean absolute error (MAE) and root mean square error (RMSE), for multivariate linear models based on a subset of the descriptors. Trained on the in-house HAT dataset.

descriptor set	MAE (kcal/mol)	RMSE (kcal/mol)
BDFE	2.56	3.52
spin densities	4.33	5.78
charges	5.19	6.94
V_{bur}	5.35	7.08
frozen BDE	4.26	5.50
BDFE, charges	2.35	3.24
BDFE, spin densities	2.55	3.52
BDFE, V_{bur}	2.56	3.52
BDFE, frozen BDE	2.37	3.32
charge, frozen BDE	3.77	5.02
charge, V_{bur}	5.17	6.89
charge, spin densities	3.94	5.40
spin densities, V_{bur}	4.30	5.74
spin densities, frozen BDE	4.04	5.17
V_{bur}, frozen BDE	4.27	5.50
BDFE, charges, spin densities	2.36	3.22
frozen BDE, charges, spin densities	3.43	4.56
BDFE, charges, frozen BDE	2.27	3.16
charges, spin densities, V_{bur}	3.87	5.29
BDFE, charges, spin densities, V_{bur}	2.35	3.21
frozen BDE, charges, spin densities, V_{bur}	3.41	4.55
BDFE, charges, spin densities, frozen BDE	2.27	3.14

S5 Predictive models trained on the 238 alkoxy HAT dataset

S5.1 Data split

In Figure 4, the substrates for alkoxy radicals abstracting hydrogens from hydrocarbons and heterosubstituted compounds in an acetonitrile solution and the 44 experimental substrates compiled by Bietti *et al.*[16]

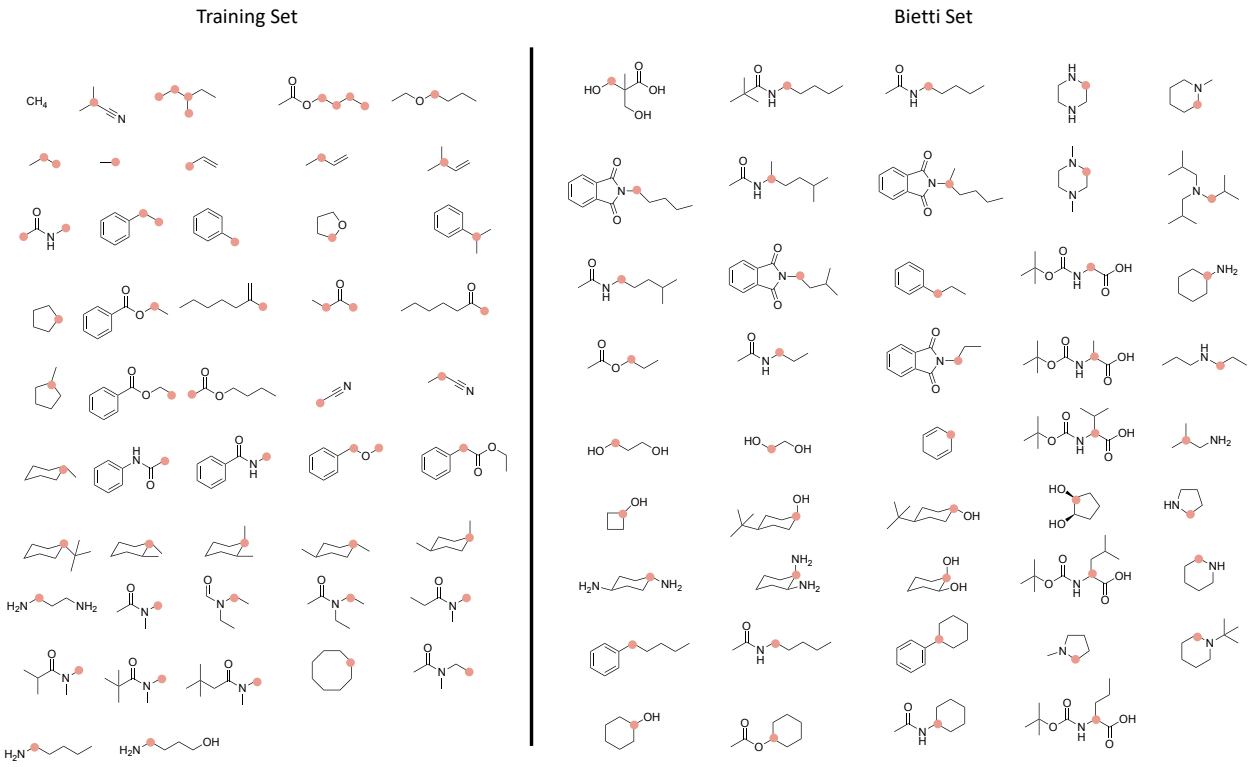

Figure 4: Training and Test set for the HAT reactions compiled by Bietti *et al.*[16] Abstraction sites are highlighted.

S5.2 Analysis of descriptor importance for the predictive models trained on the synthesis dataset

In Table 7, a detailed overview of the performance of linear regression models trained with only subsets of the full informative representation is provided.

S5.3 Re-scaling procedure

The experimental dataset collection 45 HAT reactions mediated by the cumyloxyl radical collected by Bietti *et al* reported the rate constants, the activation energy is calculated using the Eyring-Polanyi equation (7). For the calculations of the MAE and RMSE, the experimental activation energies were re-scaled. To this end, we predicted all the activation energies using the FFNN with 4 ensembles, trained in the alkoxy HAT dataset, and fit a

Table 7: Performance on $\Delta G^{\ddagger}$, in terms of mean absolute error (MAE) and root mean square error (RMSE), for multivariate linear models based on a subset of the descriptors. Trained on the alkoxy radical dataset.

descriptor set	MAE (kcal/mol)	RMSE (kcal/mol)
BDFE	2.25	2.57
spin densities	2.73	3.10
charges	1.83	2.28
V_{bur}	2.56	2.94
frozen BDE	2.28	2.62
BDFE, spin densities	1.95	2.27
BDFE, V_{bur}	2.13	2.39
BDFE, charges	1.44	1.80
BDFE, frozen BDE	2.06	2.40
charges, spin densities	1.74	2.19
charges, V_{bur}	1.75	2.15
charges, frozen BDE	1.86	2.25
spin densities, frozen BDE	2.21	2.55
spin densities, V_{bur}	2.23	2.52
V_{bur}, frozen BDE	2.19	2.49
BDFE, charges, V_{bur}	1.16	1.49
BDFE, charges, spin densities	1.27	1.61
frozen BDE, charges, spin densities	1.81	2.22
BDFE, charges, spin densities, V_{bur}	1.18	1.51
frozen BDE, charges, spin densities, V_{bur}	1.45	1.85
BDFE, charges, spin densities, frozen BDE	1.37	1.71

linear model between the predicted activation energies ($\Delta G^{\ddagger}_{pred}$) and the experimental ones ($\Delta G^{\ddagger}_{exp}$) as:

$$\Delta G^{\ddagger}_{pred} \;=\; a\Delta G^{\ddagger}_{exp} + b \tag{32}$$

for this model, b = 5.99 and a = 0.93. Once the model was fitted, we predicted the values of the experimental activation energies, and the MAE and RMSE were calculated with respect to this re-scaled activation energy.

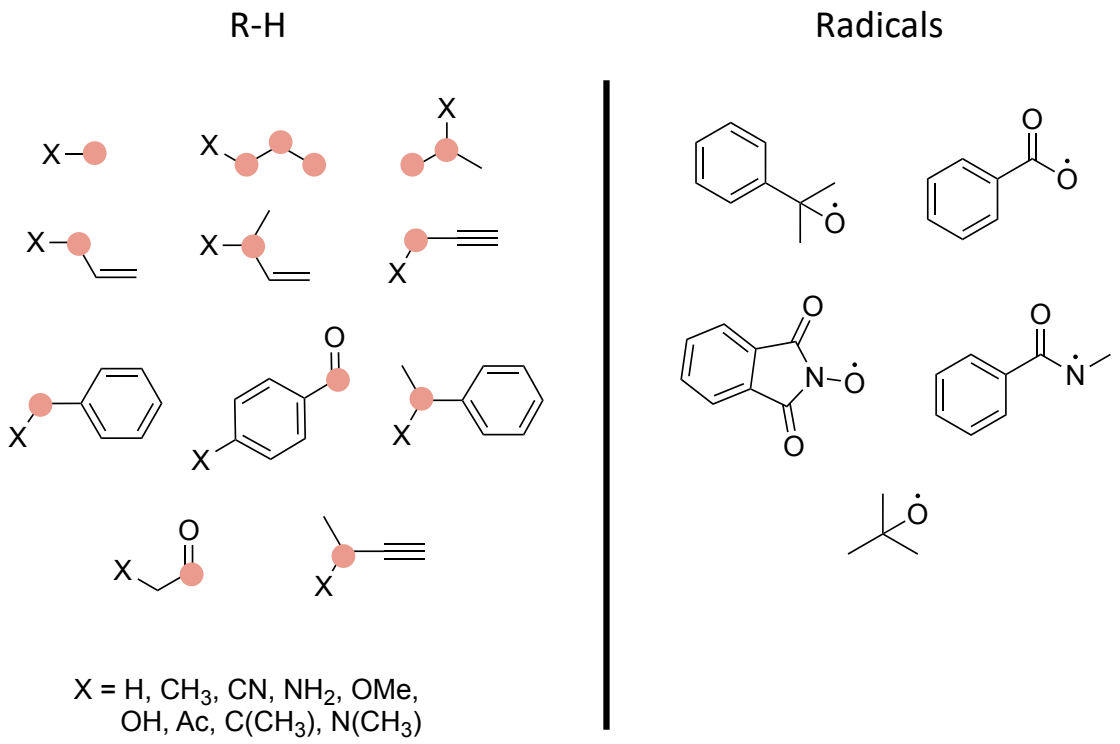

Figure 5: Chemical space for the photoredox HAT dataset. Abstraction sites are highlighted.

S6 Predictive models trained on the 564 photoredox HAT dataset

In Figure 5, the chemical space contained in this dataset is defined. In the original work,[17] a total of 17 radicals were selected, based on their synthetic importance in photoredox HAT reactivity. However, most of these contain halogen, sulfur or charged atoms – which cannot be treated by our surrogate model – and consequently, they were filtered out. A similar filter was applied to the substrates. 630 reactions from the original work of Yang *et. al.* are retained in this manner, for 35 of those, transition states were not located during the computation of the dataset. From the remaining 595 reactions, 31 reactions could not be atom mapped. The mapping step is necessary for the predictions of the surrogate model, so these reactions were also discarded.

In Figure 6, the performance of several models on the final dataset is presented: 'Ad-

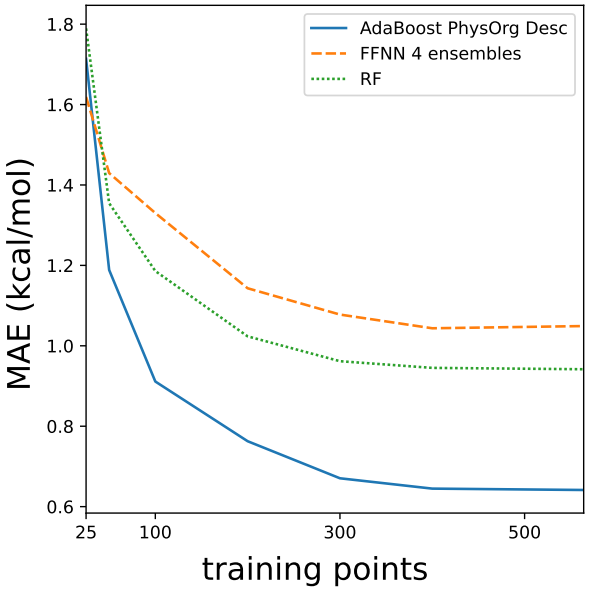

Figure 6: Learning curves for the $\Delta G^{\ddagger}$ for several models based in an intermediate VB learned representation and in the 50 physical organic descriptors.

aBoost PhysOrg Desc' corresponds to the original model with explicitly computed descriptors by Yang et al.,[17] 'FFNN 4 ensembles' corresponds to the ensembled, pre-trained feedforward neural network based on the predicted VB representation, and 'RF' corresponds to a random forest trained based on the latter. It should be clear that while the AdaBoost model outperforms our models on the full dataset, this advantage vanishes in the low data regime (at $N_{train} = 25$, our FFNN model even achieves the best accuracy, with an MAE of 1.62 kcal/mol).

S7 Feature selection for the linear model trained on the P450 metabolism dataset

As mentioned before, our VB-inspired representation of the single reaction constitutes 14 descriptors, meaning that the number of descriptors is almost equal to the number of samples (18 points, cf. Fig. 7). This would lead to overfitting even in the most simple architecture,

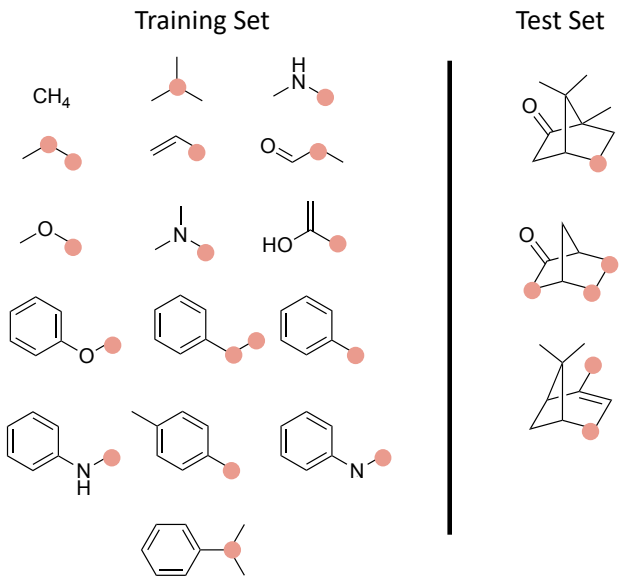

Figure 7: Training and Test set. Abstraction sites are highlighted.

i.e., the multivariate linear model.

In Table 8, an overview of the performance of linear regression models trained with only subsets of the full informative representation is provided.

S8 Atmospheric reactions extracted from RMechDB

The RMechDB dataset encompass of over 5,300 elementary radical reactions and 591 are HAT reactions.[19] We filtered out all reactions involving halogens or Sulfur, because the surrogate model was not trained on compounds with these elements. At the end, reaction profiles were computed for 268 reactions. 76 finished successfully. The main source of failure was the optimization of the identified transition state. 5 reactions did not finish in time and for 20 cases a barrierless reaction was found. 2 reactions presented a negative $\Delta G^{\ddagger}$ and for one case, the displacement of the normal mode of the transition did not correspond to the formation of the bond. As such, 73 reaction profiles were collected. In Figure 8 the distribution of the computed activation and reaction energies, as well as the correlation between both quantities, is presented. The relevant files are provided with the same description of

Table 8: Performance on $\Delta G^{\ddagger}$, in terms of mean absolute error (MAE) and root mean square error (RMSE), for multivariate linear models based on a subset of the descriptors. Trained on the data set by Tantillo and co-workers of 18 hydrogen atom transfer by the cytochrome P450 enzyme.[18]

descriptor set	MAE (kcal/mol)	RMSE (kcal/mol)
BDFE	1.42	1.79
spin densities	1.28	1.18
charges	1.44	1.56
V_{bur}	1.94	2.21
frozen BDE	1.81	2.12
spin densities, BDFE	1.69	1.99
spin densities, V_{bur}	1.12	1.32
spin densities, charges	1.25	1.35
spin densities, frozen BDE	1.60	2.31
BDFE, charges	1.98	2.40
BDFE, V_{bur}	2.09	2.43
BDFE, frozen BDE	1.60	2.25
charges, frozen BDE	1.75	2.31
charges, V_{bur}	3.61	4.10
V_{bur}, frozen BDE	1.33	1.59
spin densities, V_{bur}, charges	2.63	2.93
spin densities, V_{bur}, BDFE	2.17	2.41
spin densities, V_{bur}, frozen BDE	2.12	2.32

our in-house dataset, in a compressed archive file, RMechDB_profiles.tar.gz.

In Table 9, the performance of the various model architectures is presented.

Table 9: Performance on $\Delta G^{\ddagger}$, in terms of mean absolute error (MAE) and root mean square error (RMSE), of the descriptor-based model architectures tested for the RMechDB dataset.

model	MAE (kcal/mol)	RMSE (kcal/mol)
linear regression	1.20	1.57
random forest	1.38	1.83
FFNN	1.76	2.06
FFNN (4 ensembles)	1.29	1.51

References

(1) St. John, P. C.; Guan, Y.; Kim, Y.; Etz, B. D.; Kim, S.; Paton, R. S. Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell

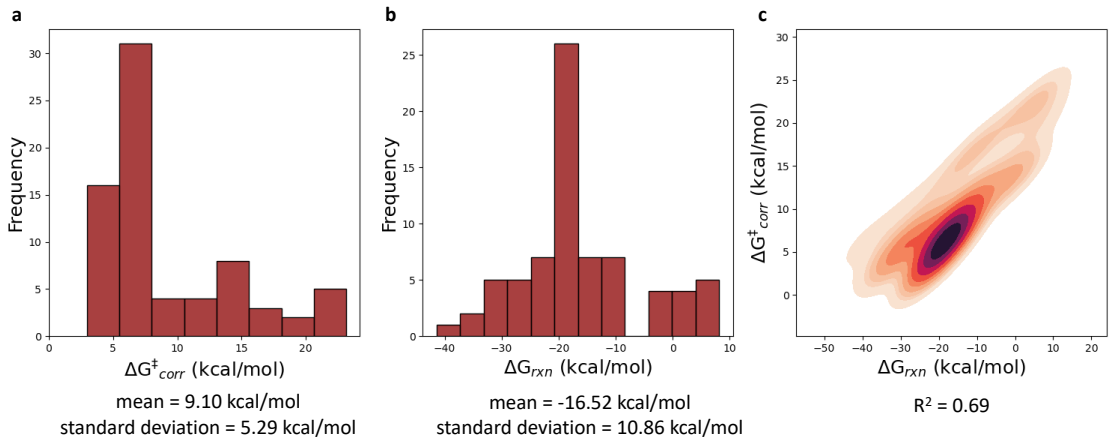

Figure 8: (a) Histogram presenting the distribution of the computed activation energies with tunneling corrections included ($\Delta G^{\ddagger}_{corr}$). (b) Histogram representing the distribution of the computed reaction energies (ΔG_{rxn}). (c) Correlation plot between $\Delta G^{\ddagger}_{corr}$ and ΔG_{rxn}.

molecules. *Sci. Data* **2020**, *7*, 244.

(2) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digit. Discov.* **2022**, *1*, 91–97.

(3) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. autodE: automated calculation of reaction energy profiles—application to organic and organometallic reactions. *Angew. Chem., Int. Ed.* **2021**, *133*, 4312–4320.

(4) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H.; others Gaussian 16. 2016.

(5) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(6) Riniker, S.; Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

(7) Landrum, G.; others RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**,

(8) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theor. Comput.* **2019**, *15*, 1652–1671.

(9) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(10) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *J. Chem. Phys.* **1992**, *97*, 2571–2577.

(11) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

(12) St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat. Commun.* **2020**, *11*, 2328.

(13) Bergstra, J.; Komer, B.; Eliasmith, C.; Yamins, D.; Cox, D. D. Hyperopt: a python library for model selection and hyperparameter optimization. *Comput. Sci. Discov.* **2015**, *8*, 014008.

(14) Stuyver, T.; Coley, C. W. Machine Learning-Guided Computational Screening of New Candidate Reactions with High Bioorthogonal Click Potential. *Chem. Eur. J.* **2023**, e202300387.

(15) van Gerwen, P.; Fabrizio, A.; Wodrich, M. D.; Corminboeuf, C. Physics-based representations for machine learning properties of chemical reactions. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045005.

(16) Salamone, M.; Galeotti, M.; Romero-Montalvo, E.; van Santen, J. A.; Groff, B. D.; Mayer, J. M.; DiLabio, G. A.; Bietti, M. Bimodal Evans–Polanyi relationships in hydrogen atom transfer from C (sp3)–H bonds to the cumyloxyl radical. A combined time-resolved kinetic and computational study. *J. Am. Chem. Soc.* **2021**, *143*, 11759–11776.

(17) Yang, L.-C.; Li, X.; Zhang, S.-Q.; Hong, X. Machine learning prediction of hydrogen atom transfer reactivity in photoredox-mediated C–H functionalization. *Org. Chem. Front.* **2021**, *8*, 6187–6195.

(18) Gingrich, P. W.; Siegel, J. B.; Tantillo, D. J. Regression Modeling for the Prediction of Hydrogen Atom Transfer Barriers in Cytochrome P450 from Semi-empirically Derived Descriptors. *Chemistry-Methods* **2022**, *2*, e202100108.

(19) Tavakoli, M.; Chiu, Y. T. T.; Baldi, P.; Carlton, A. M.; Van Vranken, D. RMechDB: A Public Database of Elementary Radical Reaction Steps. *J. Chem. Inf. Model.* **2023**, *63*, 1114–1123.