Electronic Supplementary Material (ESI) for Organic Chemistry Frontiers.

Supporting Information

Machine Learning Prediction of Hydrogen Atom Transfer Reactivity in Photoredox-Mediated C–H Functionalization

Li-Cheng Yang, Xin Li, Shuo-Qing Zhang, and Xin Hong*

Center of Chemistry for Frontier Technologies, Department of Chemistry, State Key Laboratory of Clean Energy Utilization, Zhejiang University

38 Zheda Road, Hangzhou, 310027 (China)

Email: hxchem@zju.edu.cn

Abstract: Photoredox-mediated hydrogen atom transfer (HAT) catalysis has reshaped the synthetic strategy of C–H bond functionalization. The rationalization and prediction of HAT reactivity are critical for the reaction design of photoredox-mediated C–H functionalization. In this work, we report the development of a machine learning model that can predict the HAT barrier of photoredox-mediated HAT catalysis using the physical organic descriptors of the ground state substrate and radical. Based on 2926 DFT-computed HAT barriers of the designed chemical space, the trained AdaBoost model is able to predict the HAT barrier with a mean absolute error of 0.60 kcal/mol in the out-of-sample test set. The applicability of the machine learning model is further vali-dated by comparing the prediction against the DFT-computed reactivities on scaffolds and substituents that are not present in the designed chemical space, as well as experimental kinetics data of HAT reaction with cumyloxyl radical. This work provides a ma-chine learning approach for reactivity prediction from physical organic descriptors and DFT-computed statistics, offering a useful tool that can be directly applied in the experimental designs of photoredox-mediated HAT catalysis.

DOI:

Table of Contents

Section S1. DFT computational details

All DFT calculations were performed with Gaussian 09 software package.[1] Geometry optimization of all minima and transition states (TS) were carried out at the B3LYP[2,3] level of theory with the 6-31+G(d,p) basis set. Vibrational frequencies were computed at the same level to evaluate its zero-point vibrational energy (ZPVE) and thermal corrections at 298 K and to check whether each optimized structure is an energy minimum or a transition state. The single-point energies were computed at the M06-2X[4] level of theory with the def2-TZVPP[5,6] basis set, using the gas-phase optimized structures. Solvation energies corrections were evaluated by a self-consistent reaction field (SCRF) using the SMD model[7] with M06-2X functional and 6-31G(d) basis set, based on the gas-phase optimized structures. Conformational searches for the intermediates and transition states have been conducted to ensure that the lowest energy conformers were located (Section S1.2). The 3D diagrams of molecules were generated using CYLview.[8]

S1.1. Solvation energy calculation

The acetonitrile solvent, DMSO solvent, and acetone solvent were chosen for the calculation of solvation energy due to its wide applications in the photoredox-mediated HAT catalysis. The computed barriers in acetonitrile were compared with those in acetone and DMSO for a representative set of HAT reactions. The details of the selected 56 reactions are elaborated in Figure S1a. Satisfying linear correlations were identified between the results in the three solvents (Figure S1b). Therefore, we believe the DFT-computed barriers and the derived machine learning model are also applicable for DMSO, acetone and other polar solvents. The computations of HAT barriers in solution were performed with the gas-phase optimized geometries and subsequent single-point energy calculations in solution.

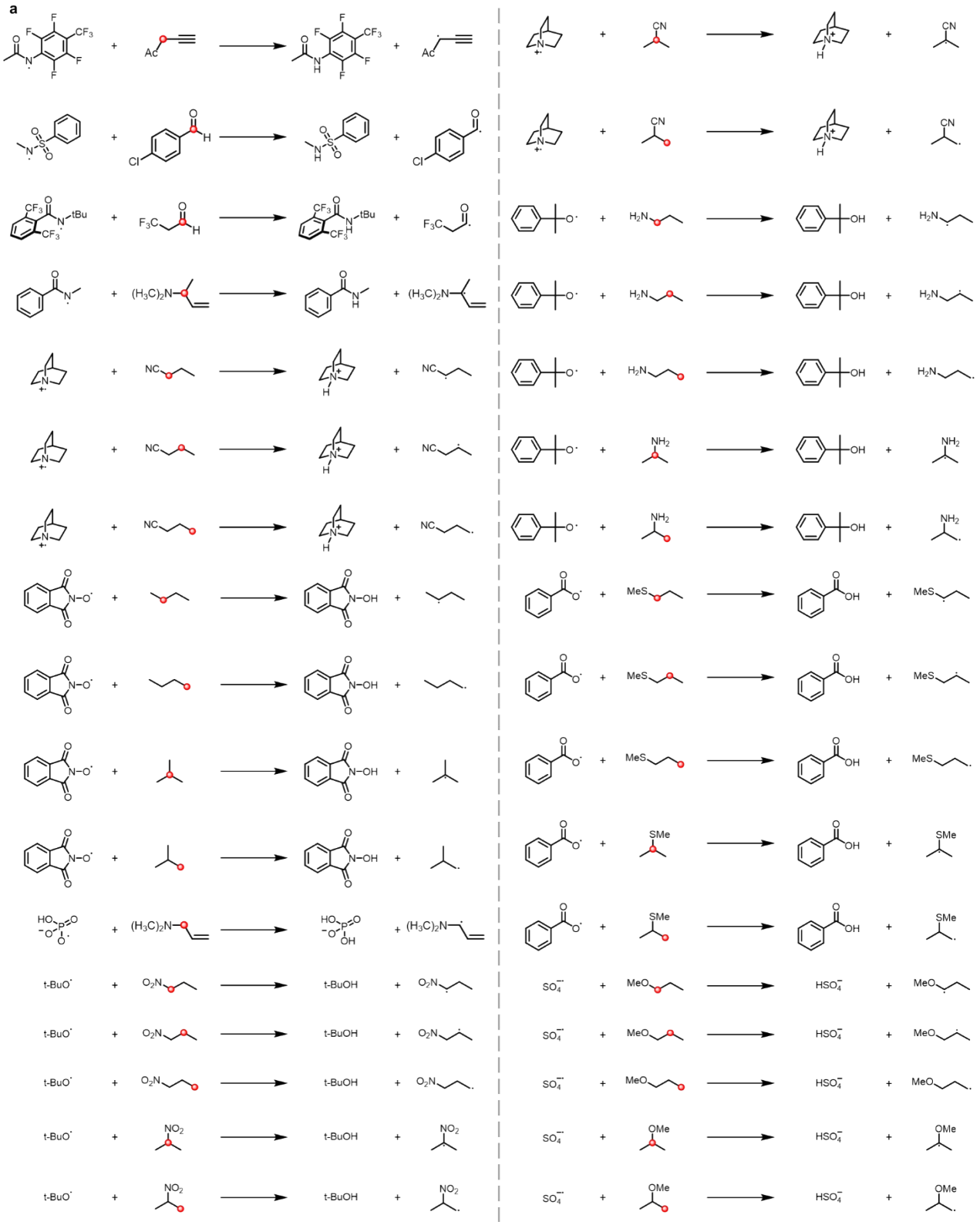


Figure S1. Comparisons of the DFT-computed HAT barriers in acetonitrile with those in DMSO and acetone for a selected set of representative reactions. (a) Testing reactions. The reacting C–H positions are labelled in red. (b) Correlation between DFT-computed HAT barriers in different solvents.

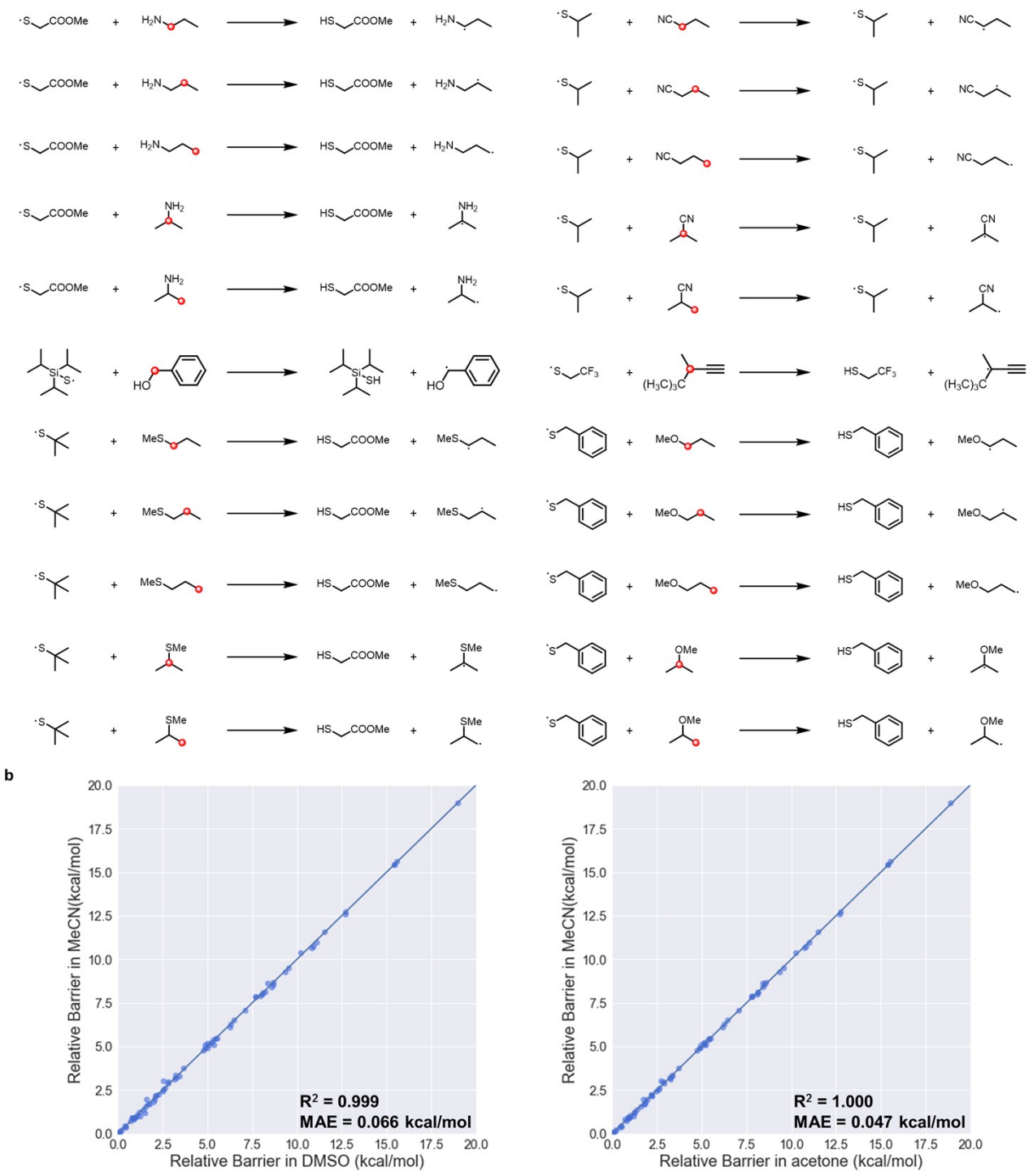


Figure S1. (continued) Comparisons of the DFT-computed HAT barriers in acetonitrile with those in DMSO and acetone for a selected set of representative reactions. (a) Testing reactions. The reacting C–H positions are labelled in red. (b) Correlation between DFT-computed HAT barriers in different solvents.

S1.2. Computational data generation

The flow chart of the DFT-computed data generation is elaborated in Figure S2. The substrates and radicals were optimized individually first, which led to the optimized structures and DFT-computed energies of the reactants. Through conformational search of the individual substrates and radicals, the most stable conformer was identified and used in the construction of initial guess in the transition state calculations. For each target transition state, combining the corresponding fragments of substrate and radical generated the initial guess. In addition to the generated geometry of the initial guess, several rotamers around the C–H–radical axial were also considered. These generated guess geometries for the transition state were subjected to the optimization of transition state. Each output file of the TS optimization was checked manually to ensure that correct HAT transition state was

located. For the error output files of the TS optimization, manual adjustment of the TS guess was performed to see if further optimization is successful. Subsequent single-point energy calculation was performed on all the optimized transition state structures. The most favorable conformer of the transition state was accounted for the DFT-computed HAT barrier.

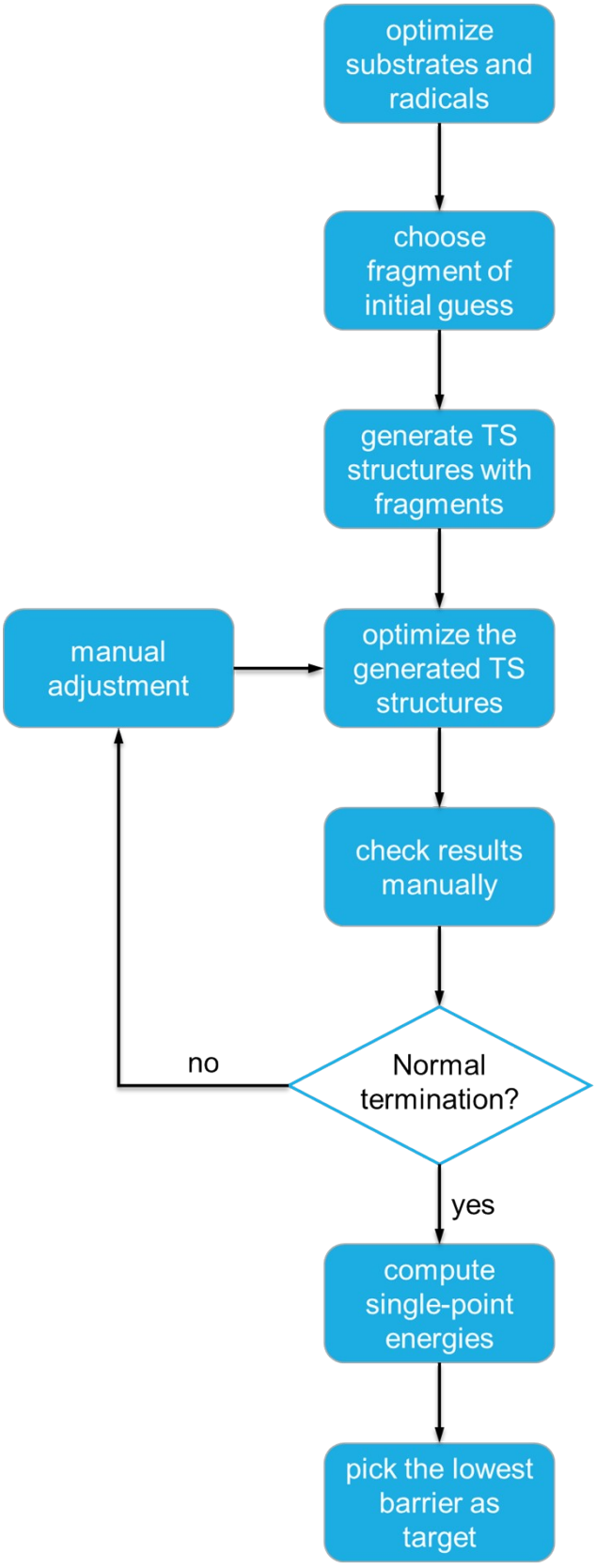


Figure S2. Flow chart of the DFT-computed data generation.

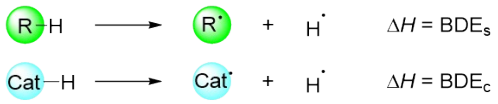
Section S2. Details and calculations of PhysOrg descriptors

For PhysOrg descriptors, we selected 56 descriptors in five categories of bond dissociation energy (BDE), frontier orbital energy, atomic charge, bond order and steric effect. These features describe the site-specific local properties (BDE, atomic charge, bond order and steric effect) as well as the global properties (frontier orbital energy). The details and the generation procedure of the descriptors are elaborated in this section.

All the descriptors, except for steric effect, were obtained through DFT calculations under the M06-2X/def2-TZVPP level of theory. Based on the gas phase-optimized geometry, nature bond orbital (NBO)[9] calculations were performed to obtain the atomic charge and Wiberg bond order[10-12] information.

S2.1. Bond dissociation energy (BDE)

The difference between of the cleaving/forming X–H bonds of the substrates and radicals, ΔBDE, were used in model training. The BDE difference, ΔBDE, was calculated as:



$$\Delta BDE = BDE_s - BDE_c$$

S2.2. Frontier orbital energy

Table S1 includes the definition and calculation details of the FMO energy-based descriptors.[13] Among HAT reactants and products, the HOMO and LUMO energies for singlet molecules and the SOMO energy for doublet molecules were obtained from single-point energy calculation. The HOMO and LUMO energies lead to a series of derived descriptors (chemical potential,[13,16] chemical hardness,[13,17-19] chemical softness[17,20,21] and electrophilicity[22]), and the calculation formulas are included in Table S1. The ionization energy (IE) was computed by comparing the energy of the neutral radical and that of the corresponding cation with the same geometry.[23, 24] The electron affinity (EA) was computed by comparing the energy of the neutral radical and that of the corresponding anion with the same geometry. IE and EA lead to the electronegativity and chemical softness of radical (Table S1).

Table S1. Symbol, definition and calculation formula of FMO energy-based descriptors.

Molecule	Feature	Definition	Calculation Formula
substrate	E01	HOMO energy of substrate of reactant	E_srHOMO
substrate	E02	LUMO energy of substrate of reactant	E_srLUMO
substrate	E03	Chemical potential (negative of electronegativity) of substrate of reactant	$\mu\_sr = \frac{E\_srHOMO + E\_srLUMO}{2}$
substrate	E04	Chemical hardness of substrate of reactant	$\eta\_sr = \frac{E\_srLUMO - E\_srHOMO}{2}$
substrate	E05	Chemical softness of substrate of reactant	$\sigma\_sr = \frac{1}{\eta\_sr}$
substrate	E06	Electrophilicity of substrate of reactant	$\omega\_sr = \frac{\mu\_sr^2}{2\eta\_sr}$
substrate	E07	SOMO energy of substrate of product	E_spSOMO

substrate	E08	Ionization energy of substrate of product	$IE\_sp = E\_spR^{+} - E\_spR$
substrate	E09	Electron affinity of substrate of product	$EA\_sp = E\_spR - E\_spR^{-}$
substrate	E10	Electronegativity of substrate of product	$\chi\_sp = -\mu\_sp = \frac{IE\_sp + EA\_sp}{2}$
substrate	E11	Chemical softness of substrate of product	$\sigma\_sp = \frac{1}{IE\_sp - EA\_sp}$
catalyst	E12	SOMO energy of catalyst of reactant	E_crSOMO
catalyst	E13	Ionization energy of catalyst of reactant	$IE\_cr = E\_crR^{+} - E\_crR$
catalyst	E14	Electron affinity of catalyst of reactant	$EA\_cr = E\_crR - E\_crR^{-}$
catalyst	E15	Electronegativity of catalyst of reactant	$\chi\_cr = -\mu\_cr = \frac{IE\_cr + EA\_cr}{2}$
catalyst	E16	Chemical softness of catalyst of reactant	$\sigma\_cr = \frac{1}{IE\_cr - EA\_cr}$
catalyst	E17	HOMO energy of catalyst of product	E_cpHOMO
catalyst	E18	LUMO energy of catalyst of product	E_cpLUMO
catalyst	E19	Chemical potential (negative of electronegativity) of catalyst of product	$\mu\_cp = \frac{E\_cpHOMO + E\_cpLUMO}{2}$
catalyst	E20	Chemical hardness of catalyst of product	$\eta\_cp = \frac{E\_cpLUMO - E\_cpHOMO}{2}$
catalyst	E21	Chemical softness of catalyst of product	$\sigma\_cp = \frac{1}{\eta\_cp}$
catalyst	E22	Electrophilicity of catalyst of product	$\omega\_cp = \frac{\mu\_cp^2}{2\eta\_cp}$

S2.3. Atomic charge

Table S2 includes the definition and calculation details of the atomic charge-based descriptors. In addition to the atomic charges computed with NBO calculation, the condensed-to-atom Fukui functions of reacting carbon atom (f_C^+, f_C^0 and f_C^-)[14,25,26] were also calculated. Once the condensed-to-atom Fukui function was evaluated, the condensed dual descriptor (Δf_C).[10,27]

When Q_C is the electronic population of reacting carbon atom in the molecule under consideration and N is the number of electrons, the f_C^+ is calculated by comparing the atomic charge in the anionic state and neutral state, using the same geometry optimized in the neutral state by definition.[15, 26]

$f_C^+ = Q_C^{N+1} - Q_C^N$

The f_C^0 is calculated by comparing the atomic charge in the anionic state and cationic state, using the same geometry optimized in the neutral state by definition.[15, 26]

$f_C^0 = (Q_C^{N+1} - Q_C^{N-1})/2$

The f_C^- is calculated by comparing the atomic charge in the neutral state and cationic state, using the same geometry optimized in the neutral state by definition.[15, 26]

$f_C^- = Q_C^N - Q_C^{N-1}$

The use of a dual descriptor defined in terms of the variation of hardness with respect to the external potential. The condensed dual descriptor of reacting carbon atom (Δf_C) is written as the difference between f_C^+ and f_C^-, can also be used as an alternative to rationalize the site reactivity[15,27-28]:

$$\Delta f_C = f_C^+ - f_C^-$$

Table S2. Symbol, definition and calculation formula of atomic charge-based descriptors.

Molecule	Feature	Definition	Calculation Formula
substrate	Q01	Atomic charge of the reacting C atom of substrate of reactant	$Q\_sr_C^N$
substrate	Q02	Fukui functions value f^+ of the reacting C atom of substrate of reactant	$Q\_srf_C^+ = Q\_sr^{N+1}_C - Q\_sr_C^N$
substrate	Q03	Fukui functions value f^- of the reacting C atom of substrate of reactant	$Q\_srf_C^- = Q\_sr_C^N - Q\_sr^{N-1}_C$
substrate	Q04	Fukui functions value Df of the reacting C atom of substrate of reactant	$\Delta f\_sr = Q\_srf_C^+ - Q\_srf_C^-$
substrate	Q05	Atomic charge of the reacting C atom of substrate of product	$Q\_sp_C^N$
substrate	Q06	Fukui functions value f^0 of the reacting C atom of substrate of product	$Q\_spf_C^0 = Q\_sp^{N+1}_C - Q\_sp^{N-1}_C$
catalyst	Q07	Atomic charge of the reacting atom of catalyst of reactant	$Q\_cr_C^N$
catalyst	Q08	Fukui functions value f^0 of the reacting atom of catalyst of reactant	$Q\_crf_C^0 = Q\_cr^{N+1}_C - Q\_cr^{N-1}_C$
catalyst	Q09	Atomic charge of the reacting atom of catalyst of product	$Q\_cp_C^N$
catalyst	Q10	Fukui functions value f^+ of the reacting atom of catalyst of product	$Q\_cpf_C^+ = Q\_cp^{N+1}_C - Q\_cp_C^N$
catalyst	Q11	Fukui functions value f^- of the reacting atom of catalyst of product	$Q\_cpf_C^- = Q\_cp_C^N - Q\_cp^{N-1}_C$
catalyst	Q12	Fukui functions value Df of the reacting atom of catalyst of product	$\Delta f\_cp = Q\_cpf_C^+ - Q\_cpf_C^-$

S2.4. Bond order

Raw data of bond order descriptors are obtained from Wiberg bond indexes.[10-12] Since the reaction is hydrogen atom transfer reaction, it is necessary to take C-H bond orders into consideration. For each substrate of reactant, the smallest C-H bond order of the reacting C atom is considered, while the average number of the bond order between H and the reacting atom, average number of the bond order without H are included for the reacting atom of catalyst of product. Besides, for both reactant and product of substrates, the biggest C-X bond order of the reacting C atom are picked. In addition, the average number of the bond order of the reacting C atom of each kind of molecules is regarded as important descriptor. (Table S3)

Table S3. Symbol, definition and calculation formula of bond order-based descriptors.

Molecule	Feature	Definition	Calculation Formula
substrate	B01	The biggest C-X bond order of the reacting C atom of substrate of reactant (X can not be H)	$B^{max}_{C-X}\_sr$
substrate	B02	The smallest C-H bond order of the reacting C atom of substrate of reactant	$B^{min}_{C-H}\_sr$

substrate	B03	The average number of the bond order of the reacting C atom of substrate of reactant	$B^{ave}_{C}\_sr$
substrate	B04	The average number of the bond order of the reacting C atom of substrate of product	$B^{ave}_{C}\_sp$
substrate	B05	The biggest C-X bond order of the reacting C atom of substrate of product (X can not be H)	$B^{max}_{C-X}\_sp$
catalyst	B06	The average number of the bond order of the reacting atom of catalyst of reactant	B^{ave}_cr
catalyst	B07	The average number of the bond order of the reacting atom of catalyst of product	B^{ave}_cp
catalyst	B08	The average number of the bond order between H and the reacting atom of catalyst of product	$B^{ave}_{C-H}\_cp$
catalyst	B09	The average number of the bond order without H of the reacting atom of catalyst of product	$B^{ave}_{C-X}\_cp$

S2.5. Buried volume

Buried volume was calculated using a written Matlab script based on its original definition.[29-32] Our Matlab script is available at GitHub: https://github.com/HFLSpopcorn/HAT-ReactivityPredictor. The sphere radius of 3Å, 4Å, and 5Å were used to measure the steric effects of various distances. (Table S4)

Table S4. Symbol, definition and calculation formula of buried volume-based descriptors.

Molecule	Feature	Definition	Calculation Formula
substrate	V01	Buried volume of the reacting C atom of substrate of reactant, sphere radius = 3Å	$\%V^{3Å}_{Bur}\_sr$
substrate	V02	Buried volume of the reacting C atom of substrate of reactant, sphere radius = 4Å	$\%V^{4Å}_{Bur}\_sr$
substrate	V03	Buried volume of the reacting C atom of substrate of reactant, sphere radius = 5Å	$\%V^{5Å}_{Bur}\_sr$
substrate	V04	Buried volume of the reacting C atom of substrate of product, sphere radius = 3Å	$\%V^{3Å}_{Bur}\_sp$
substrate	V05	Buried volume of the reacting C atom of substrate of product, sphere radius = 4Å	$\%V^{4Å}_{Bur}\_sp$
substrate	V06	Buried volume of the reacting C atom of substrate of product, sphere radius = 5Å	$\%V^{5Å}_{Bur}\_sp$
catalyst	V07	Buried volume of the reacting atom of catalyst of reactant, sphere radius = 3Å	$\%V^{3Å}_{Bur}\_cr$
catalyst	V08	Buried volume of the reacting atom of catalyst of reactant, sphere radius = 4Å	$\%V^{4Å}_{Bur}\_cr$
catalyst	V09	Buried volume of the reacting atom of catalyst of reactant, sphere radius = 5Å	$\%V^{5Å}_{Bur}\_cr$
catalyst	V10	Buried volume of the reacting atom of catalyst of product, sphere radius = 3Å	$\%V^{3Å}_{Bur}\_cp$
catalyst	V11	Buried volume of the reacting atom of catalyst of product, sphere radius = 4Å	$\%V^{4Å}_{Bur}\_cp$
catalyst	V12	Buried volume of the reacting atom of catalyst of product, sphere radius = 5Å	$\%V^{5Å}_{Bur}\_cp$

Section S3. Details of tested machine learning (ML) models

Model selection plays a crucial role in training machine learning models. We used 5-fold cross-validation to test the performance of candidate machine learning algorithms, and MAE, MSE and R^2 to examine the regression performance of these models. In *k*-fold cross validation, the dataset was randomly split to *k* parts. For each model training and evaluation, *k-1* parts were used for model training, and the rest one part was used for validation. In *k* times of error evaluations, the mean score of the *k* validation results is the final validation result for *k*-fold cross validation. The candidate algorithms include 17 regression models commonly used in Scikit-learn package,[33] which include Logistic Regression (LR), Naïve Bayes, Decision Tree, k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), Extra Trees, extreme Gradient Boosting (XGB), AdaBoost and Neural Network (NN). Based on the default hyperparameter settings, with certain modifications of hyperparameters, the details are shown as Table S5~10.

Table S5. 5-fold cross validation performance and hyperparameters of tested machine learning (ML) models (MSE is in $kcal^2/mol^2$ and MAE is in kcal/mol).

Learner name	Slearn regressor names	R^2	MSE	MAE	Main hyperparameters
AdaBoost	AdaBoostRegressor	0.971	0.73	0.61	base_estimator=ExtraTreesRegressor(n_estimators=30, criterion='mse', min_samples_split=5), n_estimators=50
NN	MLPRegressor	0.971	0.73	0.61	hidden_layer_sizes=(100, 100,), activation='relu'
XGB	XGBRegressor	0.971	0.73	0.61	max_depth=5, n_estimators=150
ExtraTrees	ExtraTreesRegressor	0.970	0.76	0.62	n_estimators=150, criterion='mse', min_samples_split=5
RF	RandomForestRegressor	0.961	0.98	0.70	n_estimators=150, criterion='mse', min_samples_split=5
GB	GradientBoostingRegressor	0.959	1.03	0.74	loss='ls', n_estimators=100
SVR	SVR	0.933	1.68	0.89	kernel='rbf', degree=3
DTree	DecisionTreeRegressor	0.921	1.98	0.99	criterion='mse', min_samples_split=2
K-Neighbors	KNeighborsRegressor	0.901	2.48	1.18	n_neighbors=5, weights='uniform', leaf_size=30
LR	LinearRegression	0.878	3.07	1.33	--
Ridge	Ridge	0.878	3.07	1.34	alpha=0.5
BR	BayesianRidge	0.877	3.10	1.34	--
LSVR	LinearSVR	0.874	3.18	1.33	--
SGD	SGDRegressor	0.871	3.22	1.38	loss='squared_loss'

Lasso	Lasso	0.645	8.92	2.35	alpha=1.0
GP	GaussianProcessRegressor	0.000	151.94	10.73	optimizer='fmin_l_bfgs_b',
Kernel Ridge	KernelRidge	0.000	307.17	17.41	alpha=1, kernel='linear'

Table S6. 5-fold cross validation performance of XGBoost model.

estimators	50	100	150 *(best)*	200
MAE	0.630	0.646	0.610	0.615
MSE	0.797	0.798	0.728	0.764
R^2	0.968	0.968	0.971	0.970

Table S7. 5-fold cross validation performance of Neural Network model.

layer	100,100 *(best)*	50,100,50	64,64	128,128
MAE	0.614	0.660	0.631	0.623
MSE	0.726	0.821	0.762	0.747
R^2	0.971	0.967	0.970	0.970

Table S8. 5-fold cross validation performance of AdaBoost model with ExtraTrees as base model.

estimators	10*200	20*30	20*50	20*70	30*30	30*50 *(best)*	30*70	200*10
MAE	0.614	0.618	0.619	0.616	0.613	0.612	0.618	0.617
MSE	0.751	0.749	0.770	0.758	0.757	0.728	0.749	0.770
R^2	0.970	0.970	0.969	0.970	0.970	0.971	0.970	0.969

Table S9. 5-fold cross validation performance of AdaBoost model with Decision Tree as base model.

estimators	50	70	100	200 *(best)*
MAE	0.712	0.717	0.710	0.700
MSE	0.977	1.014	0.982	0.978
R^2	0.961	0.960	0.961	0.961

Table S10. 5-fold cross validation performance of AdaBoost model with Random Forest as base model.

estimators	30*30	30*50	30*70	30*100 *(best)*
MAE	0.661	0.664	0.659	0.654
MSE	0.852	0.865	0.866	0.835
R^2	0.966	0.965	0.966	0.967

Section S4. Details of machine learning (ML)

The 2926 HAT barriers and corresponding descriptors were used to train the model. In order to establish a predictive model with the best performance, the performance of AdaBoost models with different base estimators were tested. The AdaBoost model performs best when ExtraTrees is used as the base estimator, and was further improved by feature selection. For hyperparameters, MSE is used as the loss function, and the numbers of estimators of ExtraTrees and AdaBoost are 30 and 50, respectively.

Table S11. Training and validation performances using AdaBoost model with ExtraTrees as base model.

Performance	R^2 of Training Set	MAE of Training Set	MSE of Training Set	R^2 of Validation Set	MAE of Validation Set	MSE of Validation Set
Scores	0.999	0.101	0.013	0.971	0.612	0.728

S4.1. Regression performance metrics

Mean absolute error (MAE) represents the difference between the original and predicted values, extracted by averaging the absolute differences over the data set.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y - \hat{y}|$$

Mean squared error (MSE) represents the difference between the original and predicted values, extracted by averaging the squared differences over the data set.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y - \hat{y})^2$$

Coefficient of determination (R^2) represents how well the predicted values fit comparing with the true values. The value ranges from 0 to 1. The higher the value is, the better performance the model has.

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

The adjusted R^2 compares the descriptive ability of regression models (two or more variables) that include a diverse number of independent variables, which corrects the influence of variable number.

The adjusted R^2 is defined as:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

Where n is the number of samples, p is the number of variables in the model.

As the definition of adjusted R^2 allows it to be negative, which means the model doesn't fit the data, we assigned the negative R^2 value as zero following literature precedent.[34]

In this work, we always used adjusted R^2 as the measure of R^2 in the main text and SI.

S4.2. Data preprocessing and model generation

All machine learning methods were implemented with Python3 scripts using Scikit-learn package. Prior to training, all descriptors were scaled by standard score method, which is calculated as

$$z = \frac{x - \mu}{\sigma}$$

where μ is the mean of the population and σ is the standard deviation of the population. The descriptors and the target property (reaction barrier) were subjected to candidate machine learning algorithms to evaluate the performance and select the best machine learning algorithm for subsequent feature selection. The detailed performances using five-fold cross validation are elaborated in Table S5.

S4.3. 5-fold cross validation performances with varying model complexities

In order to confirm that our model does not have overfitting issue, we tested the performance of the base ExtraTrees model under different hyperparameter "min_samples_split", and the results are shown in Figure S3. The smaller this parameter is, the higher the model complexity will be. Through the decreasing 'min_samples_split', the cross-validation score starts to decrease. Beyond the regime of min_samples_split < 4, this is the overfitting regime. The hyperparameter min_samples_split=5 is used in our final model, which does not belong to the overfitting regime.

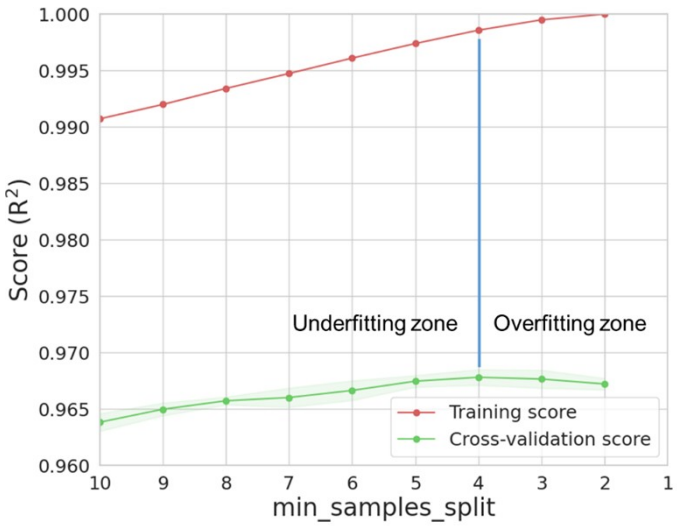


Figure S3. Training and validation scores with decreasing 'min_samples_split' for the base ExtraTrees model.

S4.4. Learning curve

The learning curve of the AdaBoost model is shown in Figure S4. The learning curve is built by random extraction of subsets from the total dataset for cross validation. In each subset, we performed 100 times five-fold cross validations to obtain the average mean value as well as the error (Table S12). It appears that smaller subset has poor prediction ability, which is probably due to the incapability of machine learning model to extract the meaningful pattern from the limited data in small subset. However, R^2 score exceeds 0.95 after an amount of 1000 data are used to train the model. With increasing number of the training examples, the following limited improvement of R^2 suggested that a sample space of near 3000 data is sufficient to get the model close to convergence.

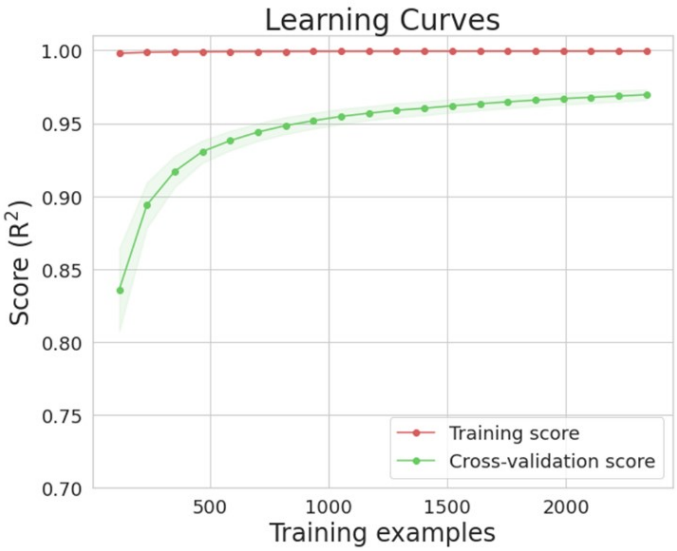


Figure S4. Learning curves of the final PhysOrg-AdaBoost model.

Table S12. Average and SD value of the performance of 100 times 5-fold cross validation.

Percentage of used data	Training scores		Validation scores	
	average	SD	average	SD
5%	0.998039	0.000408906	0.834551	0.027683
10%	0.998751	0.000182712	0.89606	0.012361
15%	0.998953	0.000118332	0.918265	0.009305
20%	0.999073	0.0000921	0.930829	0.007785
25%	0.999151	0.0000775	0.938682	0.006091
30%	0.999206	0.0000601	0.943898	0.006234
35%	0.999245	0.0000563	0.948288	0.006119
40%	0.999281	0.0000456	0.951642	0.005767
45%	0.999304	0.0000478	0.954531	0.005764
50%	0.999326	0.0000438	0.956678	0.005603
55%	0.999343	0.0000415	0.958662	0.005415
60%	0.999359	0.0000366	0.96069	0.005216
65%	0.999375	0.0000324	0.962209	0.005045
70%	0.999388	0.0000305	0.963591	0.004744
75%	0.999398	0.0000275	0.964708	0.004534
80%	0.999412	0.0000236	0.9659	0.00431
85%	0.999419	0.000021	0.966953	0.004201
90%	0.99943	0.0000195	0.967952	0.004054
95%	0.999436	0.0000196	0.968716	0.003948
100%	0.999442	0.0000171	0.96948	0.003898

S4.5. Features of established machine learning model

The importance score of all features is shown in Table S13. The feature E04, E05, E11, Q10, Q11, B02 were eliminated during feature selection.

Table S13. Details of feature importance score.

No	Feature	Score	No	Feature	Score	No	Feature	Score	No	Feature	Score	No	Feature	Score
1	BDE	0.2447	11	B04	0.0205	21	B07	0.0125	31	E03	0.0089	41	V02	0.0057
2	E10	0.0909	12	E16	0.0196	22	E18	0.0117	32	E13	0.0089	42	Q02	0.0057
3	E07	0.0743	13	B01	0.0163	23	E15	0.0111	33	E01	0.0086	43	E17	0.0056
4	E09	0.0482	14	V07	0.0151	24	Q05	0.0108	34	V09	0.0077	44	Q04	0.0053
5	Q01	0.0380	15	B09	0.0142	25	Q07	0.0107	35	E06	0.0067	45	Q08	0.0051
6	E08	0.0331	16	V08	0.0135	26	B05	0.0105	36	V12	0.0066	46	E02	0.0051
7	E21	0.0283	17	Q09	0.0131	27	E19	0.0103	37	V03	0.0065	47	Q12	0.0049
8	B03	0.0219	18	B08	0.0127	28	Q06	0.0095	38	V06	0.0064	48	Q03	0.0048
9	B06	0.0208	19	V11	0.0126	29	E22	0.0093	39	E12	0.0063	49	V04	0.0046
10	E20	0.0207	20	V10	0.0126	30	E14	0.0090	40	V05	0.0059	50	V01	0.0041

The correlation between the DFT-computed HAT barriers and the corresponding reaction free energies is shown in Figure S5. It can be clearly seen that the kinetics-thermodynamics correlation of HAT reaction is not a simple analytic function.

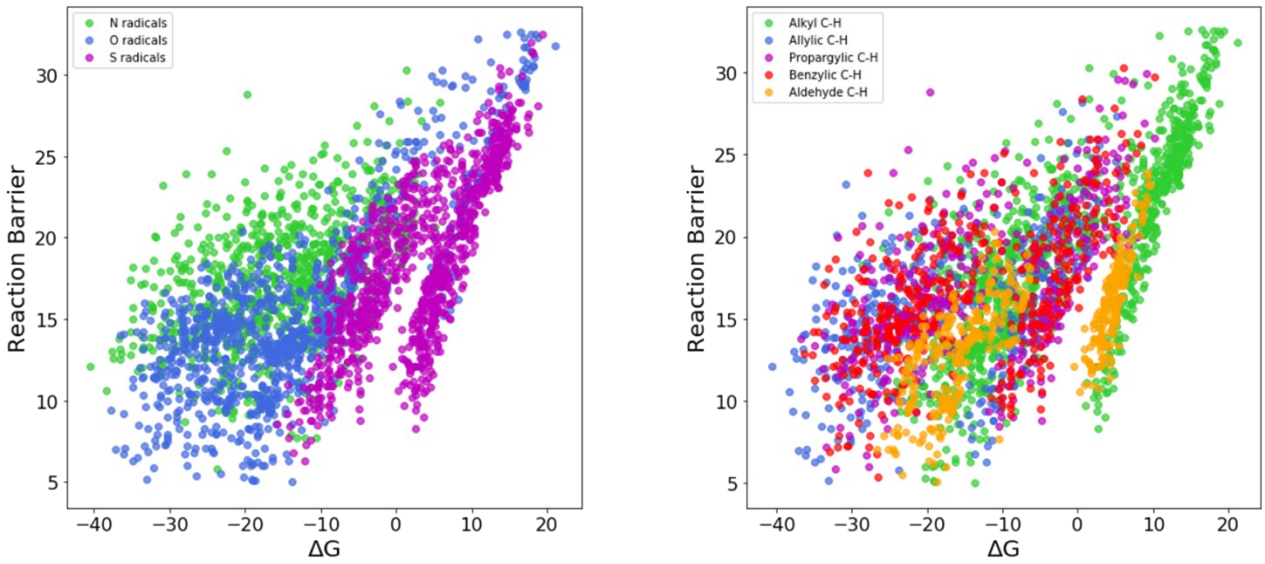


Figure S5. Correlation between the DFT-computed HAT barriers and the corresponding reaction free energies (ΔG). Different colors refer to different types of radicals or substrate scaffolds.

S4.6. Labels in training set

Each radical or fragment corresponds to a specific label, which can be seen in Figure S6. The specific reaction barriers can be found in the provided file according to the combined labels.

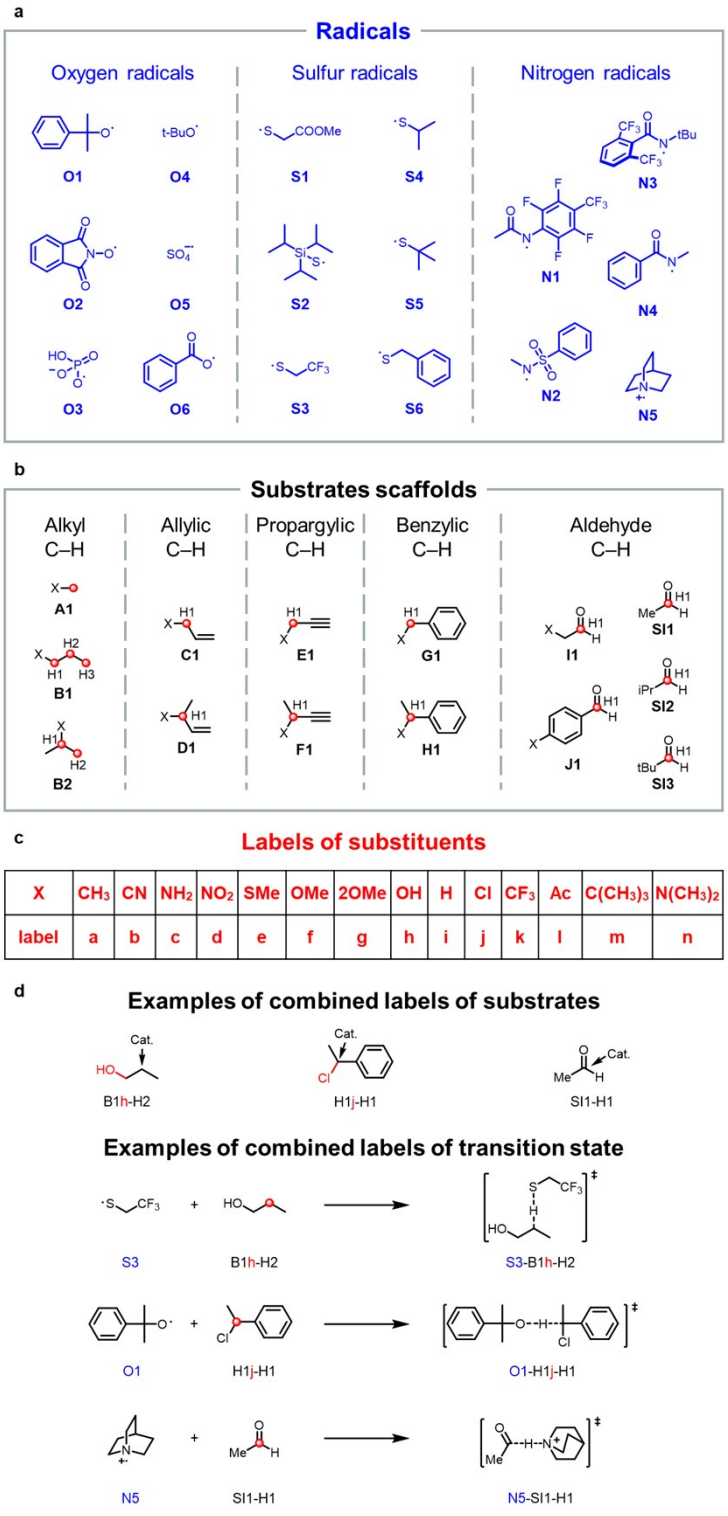


Figure S6. Labels of radical, substrate and transition state. (a) Labels of radicals. (b) Labels of substrate scaffolds. (c) Labels of substituent groups of substrates. (d) Examples of combined labels of substrates and transition states, the latter were named in the order of firstly the radical, secondly the substrate, last the reaction site.

Section S5. Detailed information in testing the conformational dependence of the PhysOrg-AdaBoost model

The structurally flexible substrates and their corresponding radicals were picked from published articles[35-38] on C–H functionalization

via HAT reaction. We used the program molclus[39] in the conformational search for each flexible substrate. The most favorable and tested high-energy conformers were both used to obtain the corresponding PhysOrg descriptors, in order to have two ML-predicted barriers. The geometries of the most favorable and tested high-energy conformers are shown in Figure S7. Structure RMSD of the conformers were calculated using VMD.[40] The vectorial angles were calculated from cosine acquired by scipy[41] package.

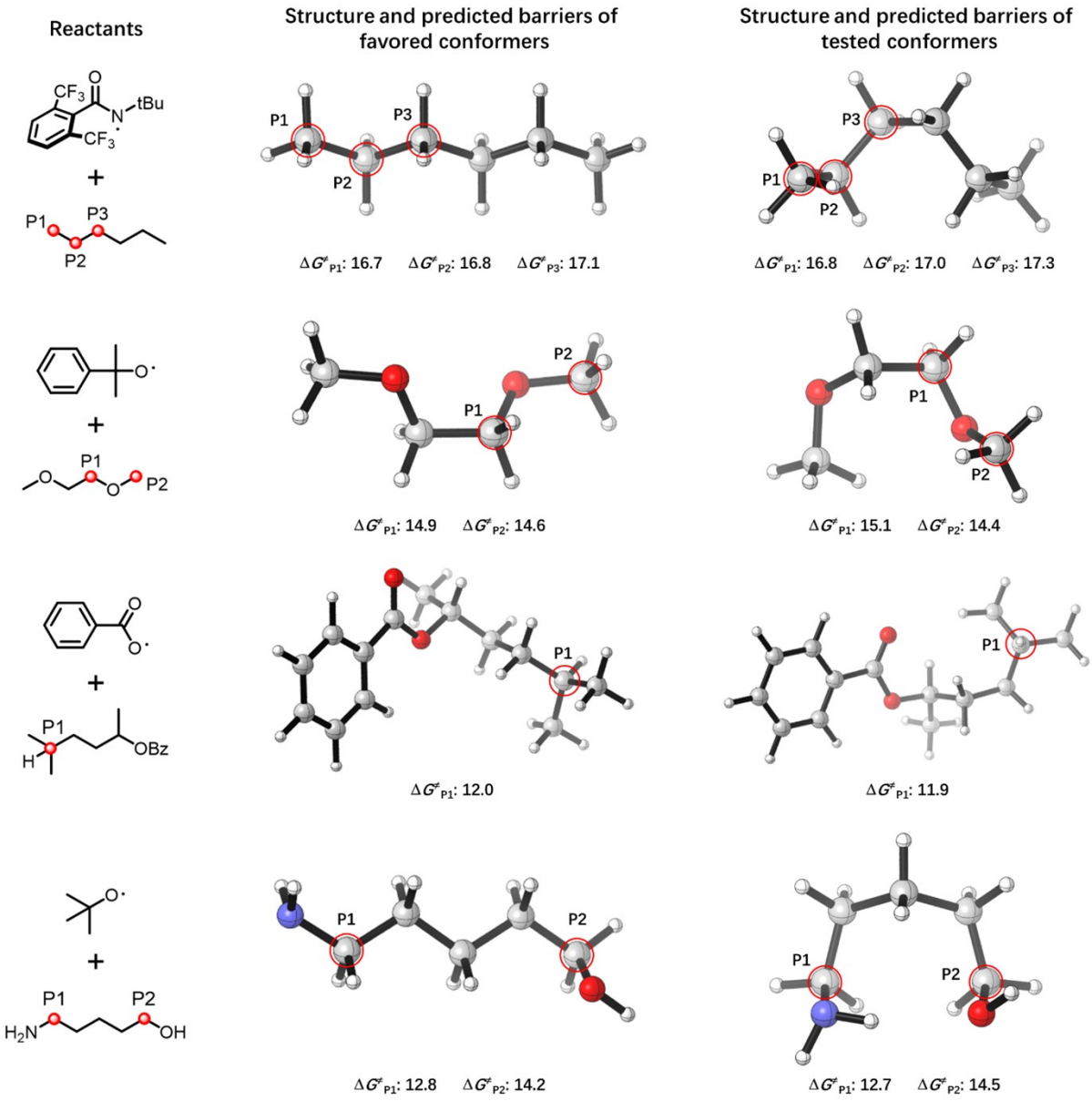


Figure S7. The DFT-optimized structures and predicted reaction barriers of the tested conformationally flexible substrates.

Section S6. Detailed performances of the PhysOrg-AdaBoost model in the test set of C–H functionalization substrates that were not present in the designed chemical space

Details of DFT-computed and ML-predicted barriers are shown in Table S14 and Figure S8. Labels of radicals are shown in Figure S6a. The substrates and their corresponding radicals were picked from published articles[35-38,42-52] on the C–H functionalization via HAT reaction.

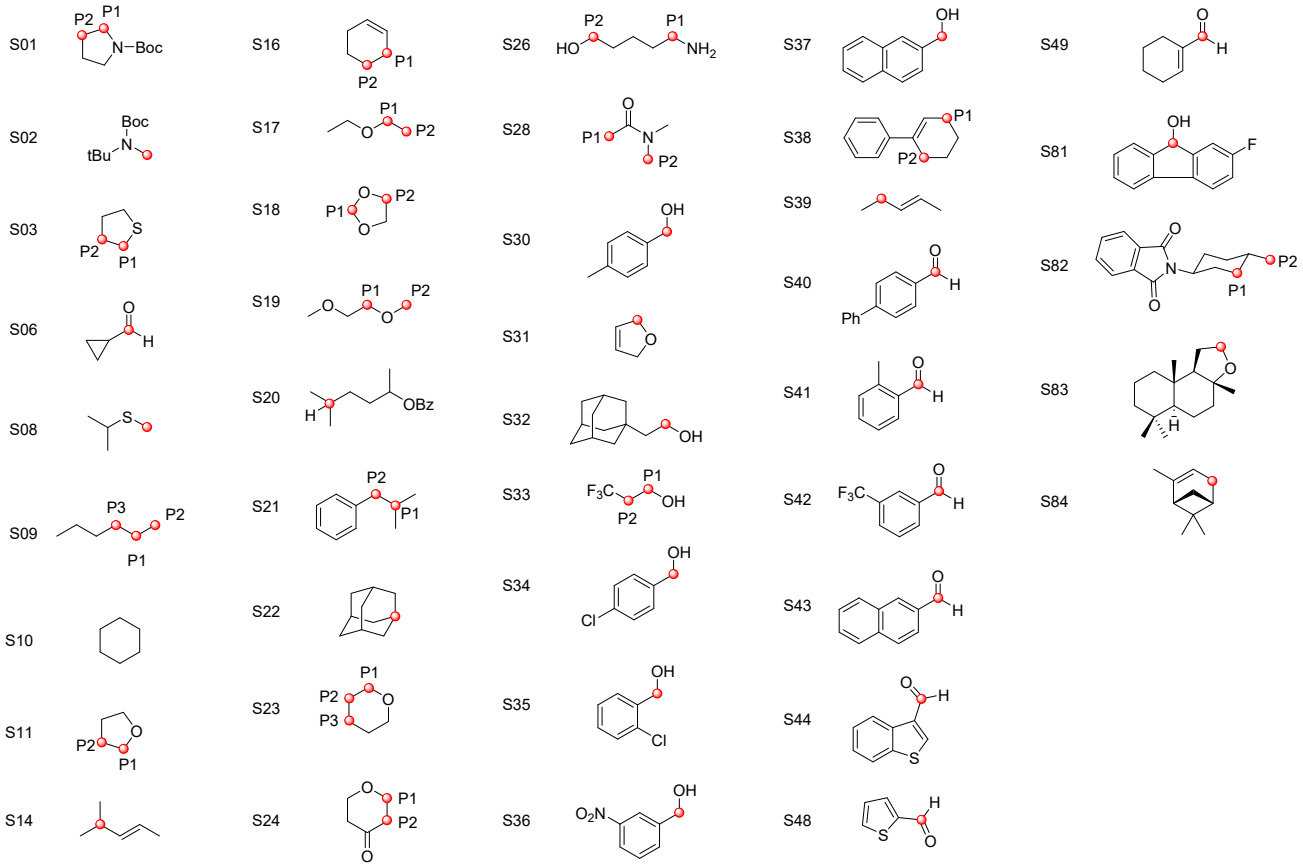


Figure S8. The labels of substrates in the test set. There are 41 substrates and 59 reaction sites studied in total.

Table S14. Details of DFT-computed and ML-predicted barriers in the test set of C–H functionalization substrates (barriers are in kcal/mol).

No	Substrate	Position	Radical	DFT-Barrier	ML-Barrier	No	Substrate	Position	Radical	DFT-Barrier	ML-Barrier
1	S01	P1	N5	12.5	11.7	35	S23	P3	O5	10.6	12.1
2	S01	P2	N5	17.9	18.8	36	S24	P1	O5	12.9	11.1
3	S02	P1	N5	13.4	13.5	37	S24	P2	O5	14.7	14.9
4	S03	P1	N5	9.4	11.9	38	S26	P1	O1	12.9	12.8
5	S03	P2	N5	18.8	18.4	39	S26	P2	O1	12.8	14.2
6	S06	P1	N5	14.1	13.9	40	S28	P1	O1	18.9	18.7
7	S08	P1	N5	12.6	13.6	41	S28	P2	O1	14.6	14.7
8	S09	P1	N3	16.2	16.7	42	S30	P1	S1	11.7	12.0
9	S09	P2	N3	16.8	16.8	43	S31	P1	S1	11.8	11.7
10	S09	P3	N3	17.2	17.1	44	S32	P1	S1	15.2	16.0
11	S10	P1	O3	10.9	11.1	45	S33	P1	S1	16.0	18.9
12	S11	P1	O4	13.2	13.5	46	S33	P1	S3	19.0	18.3
13	S11	P1	O5	8.4	7.4	47	S33	P2	S1	26.7	27.1
14	S11	P2	O4	16.1	16.3	48	S33	P2	S3	27.4	26.4
15	S11	P2	O5	12.6	12.4	49	S34	P1	O4	14.7	14.3
16	S14	P1	S2	12.7	13.1	50	S35	P1	O4	15.2	14.7
17	S16	P1	S1	13.8	14.8	51	S36	P1	O4	15.6	16.5
18	S16	P1	S2	12.5	14.6	52	S37	P1	O4	14.7	14.4
19	S16	P1	S3	12.6	14.1	53	S38	P1	S2	12.9	14.7

20	S16	P2	S1	20.2	20.5	54	S38	P2	S2	16.4	16.4
21	S16	P2	S2	19.7	20.3	55	S39	P1	S2	13.7	14.1
22	S16	P2	S3	19.3	19.8	56	S40	P1	O6	9.4	10.0
23	S17	P1	O4	13.5	13.6	57	S41	P1	O6	8.9	9.4
24	S17	P2	O4	18.9	17.8	58	S42	P1	O6	10.2	10.2
25	S18	P1	O4	12.8	13.9	59	S43	P1	O6	10.3	9.9
26	S18	P2	O4	14.6	14.8	60	S44	P1	O6	10.0	9.6
27	S19	P1	O4	14.5	14.9	61	S48	P1	O6	10.7	9.7
28	S19	P2	O4	15.2	14.6	62	S49	P1	O6	8.7	8.9
29	S20	P1	O6	10.1	12.0	63	S81	P1	O4	13.5	16.2
30	S21	P1	O3	9.7	10.5	64	S82	P1	N4	16.5	18.7
31	S21	P2	O3	11.1	12.2	65	S82	P2	N4	18.9	19.7
32	S22	P1	O3	8.0	10.7	66	S83	P1	O6	9.2	10.8
33	S23	P1	O5	9.5	8.9	67	S84	P1	S2	13.4	14.1
34	S23	P2	O5	12.2	13.1	□	□	□	□	□	□

For the above HAT reactions, there are 20 reactions with substrate that contains multiple C–H sites. $\Delta\Delta G$ were obtained from the DFT-computed and ML-predicted barriers. The results are shown in Figure S9. 19 out of the 20 reactions were accurately predicted.

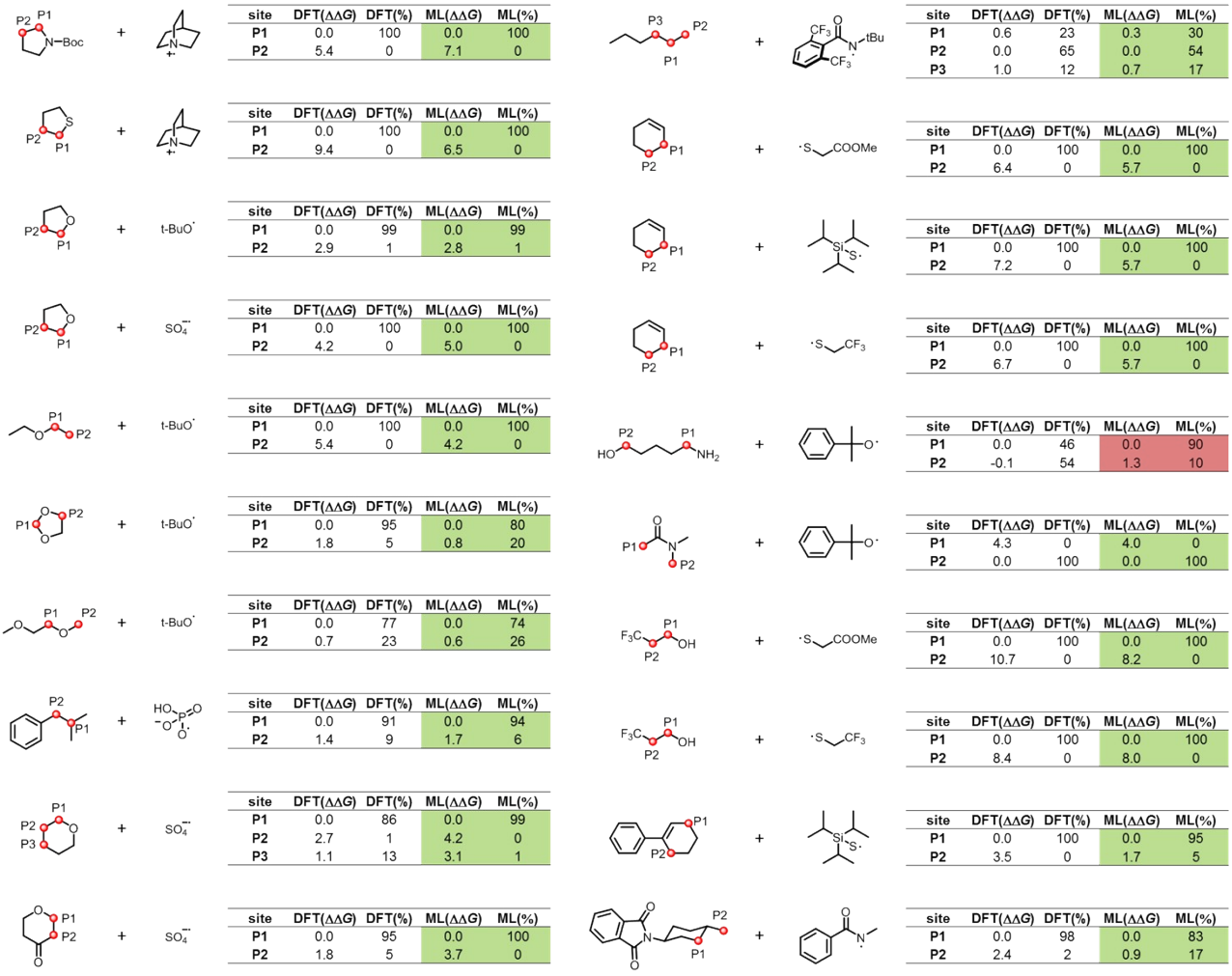


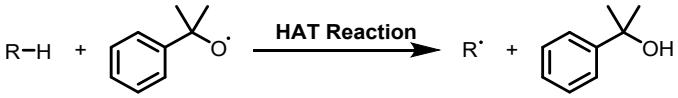
Figure S9. DFT-computed and ML-predicted HAT selectivities for selected transformations involving substrate containing multiple C–H sites.

Section S7. Detailed performances of the PhysOrg-AdaBoost model in test set of HAT reactions that have experimental kinetics data

Details of experimental and ML-predicted relative barriers are shown in Table S15 and Figure S10. The experimental barriers $\Delta G^{\neq}$ (EXP-Barrier) were calculated as

$$\Delta G^{\neq} = -RTln\frac{kh}{k_B T}$$

where k is experimentally determined reaction rate, h is Planck constant, k_B is Boltzmann constant, R is gas constant and T is temperature in K. The substrates and the cumyloxyl radical were picked from published articles[53-67] which provided experimentally determined HAT reaction rates. The reaction between substrates and cumyloxyl radical is listed below.



For 117 substrates with HAT reaction rate in this part, every potential active site of each substrates was picked out and used to get ML-predicted HAT barrier. C–H in tertiary C or α to electron withdrawing groups were regarded as reactive, and every C–H was taken into consideration for those substrates which have no obviously active site. The lowest ML-predicted barrier was considered as the overall reaction barrier of the substrate, which was compared with the experimental results. The experimentally determined HAT reaction rates were transformed to rate-transformed barriers. In consideration of the systematic errors, we used relative barrier to demonstrate the model performance in test set of HAT reaction rate. The substrate 1-(tert-butyl)piperidine(K64) was chose as reference substrate, and relative barriers(the barrier of each substrate minus the barrier of 1-(tert-butyl)piperidine) were calculated.

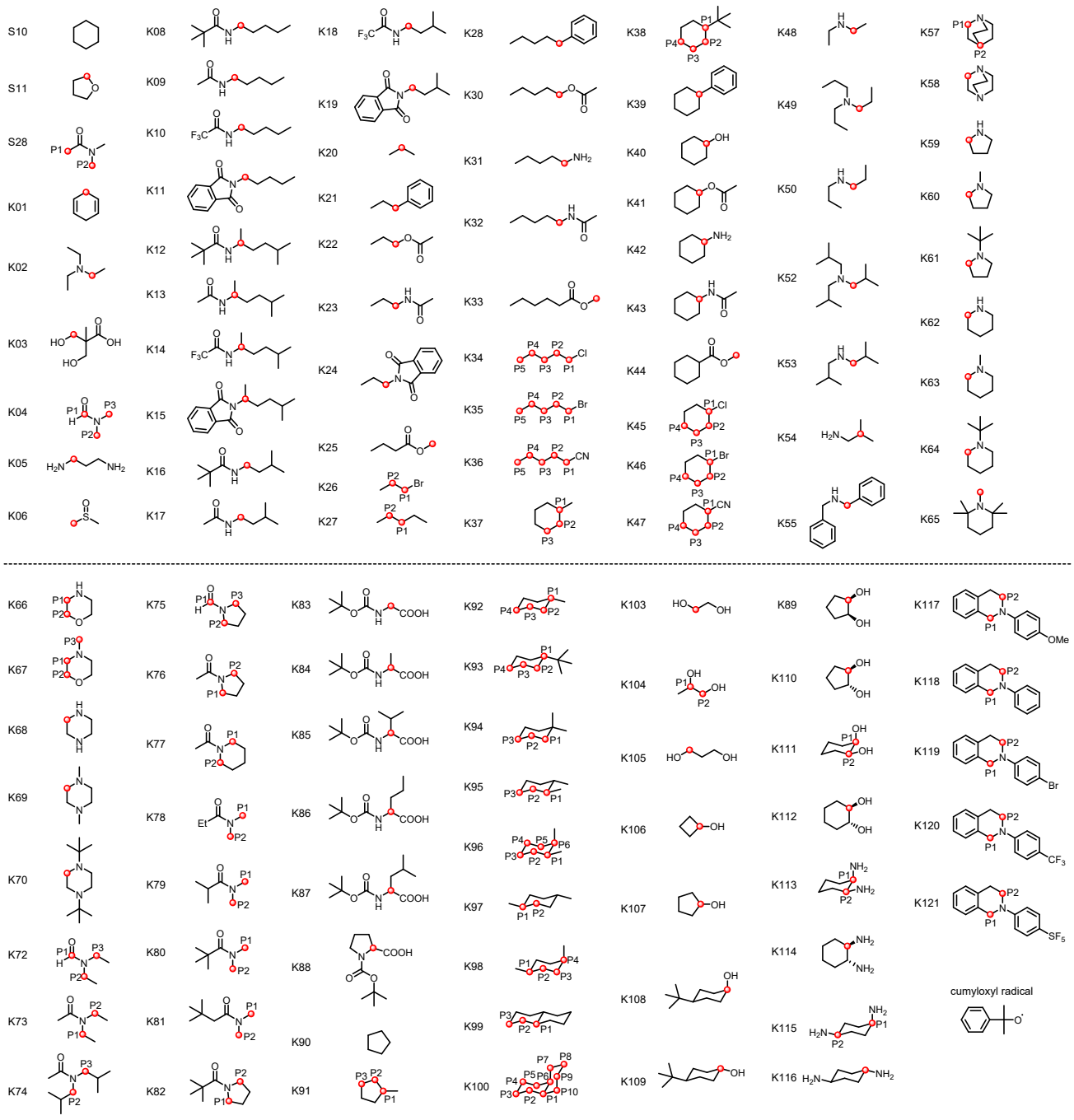


Figure S10. The radical (bottom right) and the labels of substrates in test set of HAT reaction rate. There are 117 substrates and 208 reaction sites concerned in total.

Table S15. Details of ML-predicted and Exp-determined reaction activities in test set of HAT reaction rate (barriers are in kcal/mol, the radical is cumyloxyl radical).

Substrate	Position	ML-Barrier	ML-Min	**ML-Relative Barrier (K64)**	EXP-k ($10^6 \cdot s^{-1}M^{-1}$)	EXP-Barrier	**EXP-Relative Barrier (K64)**
S10	P1	16.0	16.0	**3.7**	1.1[51]	9.2	**2.8**
S11	P1	13.4	13.4	**1.1**	5.8[51]	8.2	**1.8**
S28	P1	18.7	14.7	**2.4**	1.2[52]	9.2	**2.8**
S28	P2	14.7					
K01	P1	14.0	14.0	**1.6**	65.6[53]	6.8	**0.4**
K02	P1	12.5	12.5	**0.2**	219[53]	6.1	**-0.3**

K03	P1	16.2	16.2	**3.8**	26.9[54]	7.3	**0.9**
K04	P1	13.8	13.8	**1.5**	1.24[55]	9.1	**2.7**
K04	P2	15.1					
K04	P3	15.0					
K05	P1	12.9	12.9	**0.6**	35.5[56]	7.1	**0.7**
K06	P1	18.8	18.8	**6.4**	0.018[57]	11.6	**5.2**
K08	P1	14.2	14.2	**1.9**	0.76[58]	9.4	**3.0**
K09	P1	14.2	14.2	**1.8**	0.958[58]	9.3	**2.9**
K10	P1	16.3	16.3	**4.0**	0.295[58]	10.0	**3.6**
K11	P1	16.5	16.5	**4.1**	0.23[58]	10.1	**3.7**
K12	P1	13.8	13.8	**1.5**	0.34[58]	9.9	**3.5**
K13	P1	13.7	13.7	**1.4**	0.41[58]	9.8	**3.4**
K14	P1	15.5	15.5	**3.1**	0.39[58]	9.8	**3.4**
K15	P1	16.5	16.5	**4.1**	0.26[58]	10.1	**3.7**
K16	P1	14.6	14.6	**2.2**	0.86[58]	9.3	**3.0**
K17	P1	14.4	14.4	**2.1**	1.08[58]	9.2	**2.8**
K18	P1	16.4	16.4	**4.1**	0.41[58]	9.8	**3.4**
K19	P1	16.6	16.6	**4.3**	0.25[58]	10.1	**3.7**
K20	P1	15.4	15.4	**3.0**	0.1[58]	10.6	**4.2**
K21	P1	15.2	15.2	**2.9**	0.616[58]	9.5	**3.1**
K22	P1	16.1	16.1	**3.8**	0.03[58]	11.3	**4.9**
K23	P1	14.1	14.1	**1.7**	0.57[58]	9.6	**3.2**
K24	P1	16.5	16.5	**4.2**	0.03[58]	11.3	**4.9**
K25	P1	17.3	17.3	**5.0**	0.03[58]	11.3	**4.9**
K26	P1	16.9	16.8	**4.5**	0.04[58]	11.2	**4.8**
K26	P2	16.8					
K27	P1	16.0	15.8	**3.5**	0.31[58]	10.0	**3.6**
K27	P2	15.8					
K28	P1	15.3	15.3	**2.9**	0.79[58]	9.4	**3.0**
K30	P1	16.2	16.2	**3.9**	0.206[58]	10.2	**3.8**
K31	P1	12.8	12.8	**0.5**	15.5[58]	7.6	**1.2**
K32	P1	14.2	14.2	**1.8**	0.958[58]	9.3	**2.9**
K33	P1	17.3	17.3	**4.9**	0.197[58]	10.2	**3.8**
K34	P1	17.0	16.3	**4.0**	0.184[58]	10.3	**3.9**
K34	P2	17.1					
K34	P3	17.0					
K34	P4	16.3					
K34	P5	17.3					
K35	P1	16.9	16.3	**4.0**	0.176[58]	10.3	**3.9**
K35	P2	17.3					
K35	P3	17.3					
K35	P4	16.3					
K35	P5	17.4					

K36	P1	17.9	16.7	**4.4**	0.182[58]	10.3	**3.9**
K36	P2	17.6					
K36	P3	17.2					
K36	P4	16.7					
K36	P5	17.5					
K37	P1	15.0	15.0	**2.7**	1.01[58]	9.3	**2.9**
K37	P2	16.6					
K37	P3	16.2					
K37	P4	16.3					
K38	P1	15.8	15.8	**3.4**	0.82[58]	9.4	**3.0**
K38	P2	16.8					
K38	P3	16.5					
K38	P4	16.5					
K39	P1	15.1	15.1	**2.8**	0.91[58]	9.3	**2.9**
K40	P1	14.1	14.1	**1.8**	2.66[58]	8.7	**2.3**
K41	P1	16.4	16.4	**4.1**	0.42[58]	9.8	**3.4**
K42	P1	12.6	12.6	**0.3**	21[58]	7.5	**1.1**
K43	P1	14.2	14.2	**1.9**	0.69[58]	9.5	**3.1**
K44	P1	17.3	17.3	**4.9**	0.614[58]	9.5	**3.2**
K45	P1	16.2	16.2	**3.9**	0.462[58]	9.7	**3.3**
K45	P2	17.7					
K45	P3	17.3					
K45	P4	17.1					
K46	P1	16.4	16.4	**4.0**	0.33[57]	9.9	**3.5**
K46	P2	17.7					
K46	P3	17.3					
K46	P4	17.0					
K47	P1	16.8	16.8	**4.5**	0.39[58]	9.8	**3.4**
K47	P2	17.7					
K47	P3	17.3					
K47	P4	17.3					
K48	P1	12.6	12.6	**0.3**	110[59]	6.5	**0.1**
K49	P1	12.4	12.4	**0.0**	230[59]	6.0	**-0.4**
K50	P1	12.6	12.6	**0.3**	101[59]	6.5	**0.1**
K52	P1	12.7	12.7	**0.4**	127[59]	6.4	**0.0**
K53	P1	12.8	12.8	**0.5**	91[59]	6.6	**0.2**
K54	P1	12.9	12.9	**0.6**	9.6[59]	7.9	**1.5**
K55	P1	13.3	13.3	**0.9**	37.5[59]	7.1	**0.7**
K57	P1	15.4	15.2	**2.9**	3.7[59]	8.5	**2.1**
K57	P2	15.2					
K58	P1	15.4	15.4	**3.1**	9.6[59]	7.9	**1.5**
K59	P1	12.4	12.4	**0.1**	124[60]	6.4	**0.0**
K60	P1	12.2	12.2	**-0.1**	191[60]	6.1	**-0.2**

K61	P1	12.1	12.1	**-0.2**	300[60]	5.9	**-0.5**
K62	P1	12.5	12.5	**0.2**	107[60]	6.5	**0.1**
K63	P1	12.7	12.7	**0.4**	122[60]	6.4	**0.0**
K64(ref)	P1	12.3	12.3	**0.0**	126[60]	6.4	**0.0**
K65	P1	12.1	12.1	**-0.3**	171[60]	6.2	**-0.2**
K66	P1	14.2	13.0	**0.6**	50[60]	6.9	**0.5**
K66	P2	13.0					
K67	P1	14.4	12.2	**-0.1**	43.2[60]	7.0	**0.6**
K67	P2	12.6					
K67	P3	12.2					
K68	P1	12.4	12.4	**0.1**	226[60]	6.1	**-0.3**
K69	P1	12.5	12.5	**0.2**	116[60]	6.4	**0.0**
K70	P1	12.6	12.6	**0.3**	132[60]	6.4	**0.0**
K72	P1	14.3	14.1	**1.7**	1.25[61]	9.1	**2.7**
K72	P2	14.3					
K72	P3	14.1					
K73	P1	14.1	13.9	**1.5**	0.664[61]	9.5	**3.1**
K73	P2	13.9					
K74	P1	15.0	14.8	**2.5**	0.314[61]	9.9	**3.5**
K74	P2	14.8					
K75	P1	13.9	13.8	**1.5**	4.93[61]	8.3	**1.9**
K75	P2	13.9					
K75	P3	13.8					
K76	P1	14.1	13.6	**1.3**	9[61]	8.0	**1.6**
K76	P2	13.6					
K77	P1	14.2	14.2	**1.9**	3.17[61]	8.6	**2.2**
K77	P2	14.5					
K78	P1	14.0	14.0	**1.7**	1.55[61]	9.0	**2.6**
K78	P2	14.6					
K79	P1	14.2	14.2	**1.8**	1.69[61]	8.9	**2.6**
K79	P2	14.5					
K80	P1	14.3	14.3	**1.9**	1.41[61]	9.1	**2.7**
K80	P2	14.4					
K81	P1	14.1	14.1	**1.8**	1.6[61]	9.0	**2.6**
K81	P2	14.4					
K82	P1	14.4	13.6	**1.3**	5.17[61]	8.3	**1.9**
K82	P2	13.6					
K83	P1	17.2	17.2	**4.8**	0.396[62]	9.8	**3.4**
K84	P1	16.3	16.3	**4.0**	0.276[62]	10.0	**3.6**
K85	P1	16.6	16.6	**4.2**	0.199[62]	10.2	**3.8**
K86	P1	16.4	16.4	**4.1**	0.33[62]	9.9	**3.5**
K87	P1	16.6	16.6	**4.3**	0.59[62]	9.6	**3.2**
K88	P1	14.4	14.4	**2.1**	2.51[62]	8.7	**2.3**

K89	P1	14.5	14.5	**2.2**	2.49[63]	8.7	**2.3**
K90	P1	15.2	15.2	**2.9**	0.954[64]	9.3	**2.9**
K91	P1	14.3	14.3	**2.0**	1.31[64]	9.1	**2.7**
K91	P2	15.8					
K91	P3	15.4					
K92	P1	15.0	15.0	**2.7**	1.01[64]	9.3	**2.9**
K92	P2	16.6					
K92	P3	16.2					
K92	P4	16.3					
K93	P1	15.8	15.8	**3.4**	0.82[64]	9.4	**3.0**
K93	P2	16.8					
K93	P3	16.5					
K93	P4	16.5					
K94	P1	17.3	16.4	**4.1**	0.77[64]	9.4	**3.0**
K94	P2	16.4					
K94	P3	16.4					
K95	P1	15.4	15.4	**3.1**	1.03[64]	9.2	**2.8**
K95	P2	16.8					
K95	P3	16.4					
K96	P1	15.3	15.0	**2.7**	2.34[64]	8.8	**2.4**
K96	P2	16.6					
K96	P3	16.3					
K96	P4	16.3					
K96	P5	16.9					
K96	P6	15.0					
K97	P1	15.1	15.1	**2.8**	1.1[64]	9.2	**2.8**
K97	P2	16.7					
K98	P1	15.2	15.2	**2.8**	2.05[64]	8.8	**2.4**
K98	P2	16.6					
K98	P3	16.6					
K98	P4	15.2					
K99	P1	15.7	15.7	**3.4**	1.58[64]	9.0	**2.6**
K99	P2	16.6					
K99	P3	16.2					
K100	P1	15.5	15.5	**3.1**	2.85[64]	8.6	**2.2**
K100	P10	16.5					
K100	P2	16.5					
K100	P3	16.1					
K100	P4	16.1					
K100	P5	16.5					
K100	P6	15.5					
K100	P7	16.5					
K100	P8	16.1					

K100	P9	16.1					
K103	P1	15.4	15.4	**3.1**	0.84[63]	9.4	**3.0**
K104	P1	14.0	14.0	**1.7**	1.55[63]	9.0	**2.6**
K104	P2	15.1					
K105	P1	14.2	14.2	**1.8**	1.95[63]	8.9	**2.5**
K106	P1	13.8	13.8	**1.5**	2.08[63]	8.8	**2.4**
K107	P1	13.5	13.5	**1.1**	2.5[63]	8.7	**2.3**
K108	P1	14.1	14.1	**1.8**	5.06[63]	8.3	**1.9**
K109	P1	14.3	14.3	**2.0**	2.37[63]	8.7	**2.4**
K110	P1	14.2	14.1	**1.8**	0.93[63]	9.3	**2.9**
K110	P2	14.1					
K111	P1	14.6	14.6	**2.3**	2.7[63]	8.7	**2.3**
K111	P2	15.1					
K112	P1	15.2	15.2	**2.9**	1.43[63]	9.0	**2.7**
K113	P1	12.5	12.5	**0.1**	65.8[63]	6.8	**0.4**
K113	P2	13.2					
K114	P1	12.7	12.7	**0.4**	34.4[63]	7.2	**0.8**
K115	P1	12.6	12.4	**0.1**	55.4[63]	6.9	**0.5**
K115	P2	12.4					
K116	P1	12.7	12.7	**0.4**	33.1[63]	7.2	**0.8**
K117	P1	13.5	12.7	**0.4**	480[65]	5.6	**-0.8**
K117	P2	12.7					
K118	P1	13.7	13.0	**0.7**	280[65]	5.9	**-0.5**
K118	P2	13.0					
K119	P1	14.2	13.7	**1.3**	217[65]	6.1	**-0.3**
K119	P2	13.7					
K120	P1	14.8	14.5	**2.2**	201[65]	6.1	**-0.3**
K120	P2	14.5					
K121	P1	15.3	15.1	**2.7**	165[65]	6.2	**-0.2**
K121	P2	15.1					

Section S8. Detailed performances of the PhysOrg-AdaBoost model on selectivity prediction

Details of DFT-computed and ML-predicted barriers are shown in Table S16 and Table S17. Labels of radicals are shown in Figure S6a.

Table S16. Details of ML-predicted and DFT-computed barriers of menthol in test set of selectivity prediction (barriers are in kcal/mol)

Substrate	Position	Radical	DFT-Barrier	**DFT-Percentage**	ML-Barrier	**ML-Percentage**
Menthol	P1	O5	10.6	**0.0%**	10.8	**2.9%**
Menthol	P2	O5	8.1	**1.0%**	11.2	**1.4%**
Menthol	P3	O5	7.8	**1.7%**	9.2	**44.2%**
Menthol	P4	O5	5.4	**97.3%**	9.1	**51.5%**

Menthol	P1	N5	15.2	**0.3%**	16.3	**2.5%**
Menthol	P2	N5	15.3	**0.2%**	16.6	**1.5%**
Menthol	P3	N5	11.8	**89.5%**	14.5	**51.8%**
Menthol	P4	N5	13.1	**10.0%**	14.6	**44.2%**
Menthol	P1	S5	18.8	**1.2%**	19.9	**0.5%**
Menthol	P2	S5	19.5	**0.4%**	19.7	**0.6%**
Menthol	P3	S5	17.1	**20.3%**	17.4	**28.3%**
Menthol	P4	S5	16.3	**78.2%**	16.9	**70.6%**

Table S17. Details of ML-predicted and DFT-computed barriers of (+)-Camptothecin in test set of selectivity prediction (barriers are in kcal/mol)

Substrate	Position	Radical	DFT-Barrier	**DFT-Percentage**	ML-Barrier	**ML-Percentage**
(+)-Camptothecin	P1	O5	10.0	**44.6%**	14.1	**62.2%**
(+)-Camptothecin	P2	O5	9.9	**52.8%**	15.1	**12.3%**
(+)-Camptothecin	P3	O5	11.7	**2.5%**	14.9	**15.7%**
(+)-Camptothecin	P4	O5	14.0	**0.1%**	15.2	**9.8%**
(+)-Camptothecin	P1	N5	18.2	**89.6%**	18.8	**92.1%**
(+)-Camptothecin	P2	N5	19.5	**10.0%**	20.4	**6.3%**
(+)-Camptothecin	P3	N5	21.8	**0.2%**	21.6	**0.8%**
(+)-Camptothecin	P4	N5	21.8	**0.2%**	21.6	**0.8%**
(+)-Camptothecin	P1	S5	18.3	**79.4%**	19.1	**87.8%**
(+)-Camptothecin	P2	S5	19.1	**20.6%**	20.3	**12.2%**
(+)-Camptothecin	P3	S5	23.3	**0.0%**	24.4	**0.0%**
(+)-Camptothecin	P4	S5	25.8	**0.0%**	25.5	**0.0%**

Section S9. Dataset and cartesian coordinates of structures

The csv files of training set and test sets along with the details of cartesian coordinates of DFT-computed structures are available at the zip file of **supplementary information** and https://github.com/HFLSpopcorn/HAT-ReactivityPredictor.

Section S10. Performances of the PhysOrg-AdaBoost model using dataset splitting

The generalization ability of AdaBoost model was further examined by splitting the dataset following different substrate scaffolds, substituents, and radicals. The dataset with 2926 reaction barriers and corresponding physorg features was split into training set and test set, and an AdaBoost model was trained on training set (missing one scaffold, substituent type or radical) to predict the reaction barriers in the test set with specific type of substrate scaffolds, substituents, or radicals that were not present in the training set. Details of the size of training set, test set, and prediction abilities are shown in Table S18, Table S19 and Table S20. Labels of scaffolds, substituents and radicals are shown in Figure S6.

Table S18. Details of splitting test on substrate scaffolds (MAE and RMSE are in kcal/mol, MSE is in $kcal^2/mol^2$).

Scaffold	train	test	R^2	MAE	MSE	RMSE
A	2715	211	0.91	1.50	3.60	1.90
B	1922	1004	0.92	1.19	2.26	1.50
C	2709	217	0.94	0.70	0.91	0.96

D	2729	197	0.92	0.89	1.37	1.17
E	2705	221	0.97	0.63	0.64	0.80
F	2731	195	0.86	1.25	2.83	1.68
G	2707	219	0.91	1.07	1.91	1.38
H	2731	195	0.95	0.69	0.78	0.89
I	2676	250	0.27	2.61	8.23	2.87
J	2709	217	0.80	1.23	2.14	1.46

Table S19. Details of splitting test on substrate substituent (MAE and RMSE are in kcal/mol, MSE is in $kcal^2/mol^2$).

Substituent	train	test	R^2	MAE	MSE	RMSE
a	2705	221	0.96	0.66	0.69	0.83
b	2688	238	0.97	0.61	0.64	0.80
c	2705	221	0.95	0.81	1.04	1.02
d	2689	237	0.77	1.69	4.66	2.16
e	2691	235	0.93	0.81	1.06	1.03
f	2691	235	0.95	0.69	0.91	0.95
g	2878	48	0.77	1.59	4.01	2.00
h	2696	230	0.91	0.90	1.81	1.35
i	2841	85	0.96	0.68	0.85	0.92
j	2702	224	0.92	0.95	1.41	1.19
k	2703	223	0.86	1.46	3.30	1.82
l	2702	224	0.84	1.36	2.79	1.67
m	2688	238	0.81	1.17	3.72	1.93
n	2709	217	0.80	1.65	4.99	2.23

Table S20. Details of splitting test on radicals (MAE and RMSE are in kcal/mol, MSE is in $kcal^2/mol^2$).

Radical	train	test	R^2	MAE	MSE	RMSE
O1	2751	175	0.50	1.33	2.28	1.51
O2	2754	172	-1.68	6.98	49.83	7.06
O3	2759	167	0.85	0.89	1.48	1.22
O4	2744	182	0.70	0.85	1.44	1.20
O5	2748	178	0.87	0.89	1.43	1.20
O6	2764	162	-5.09	6.26	41.32	6.43
S1	2786	140	0.98	0.61	0.66	0.81
S2	2744	182	0.89	1.36	2.49	1.58
S3	2751	175	0.97	0.72	0.78	0.88
S4	2751	175	0.99	0.44	0.28	0.53
S5	2751	175	0.95	0.94	1.08	1.04
S6	2751	175	0.98	0.50	0.38	0.61
N1	2751	175	0.70	2.02	5.05	2.25
N2	2751	175	0.58	2.07	4.91	2.21
N3	2751	175	0.57	1.79	5.14	2.27
N4	2744	182	0.91	0.68	0.74	0.86
N5	2765	161	0.81	1.17	2.62	1.62

Section S11. Availability of the developed ML model for HAT reactivity prediction

The features in ML training, the reactivity calculation data and the code for developing machine Learning model is freely available at https://github.com/HFLSpopcorn/HAT-ReactivityPredictor.

References

1 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian 09, Revision D.01, Gaussian, Inc., Wallingford, CT, 2013.
2 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.
3 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
4 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
5 F. Weigend, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1057–1065.
6 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
7 A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B* 2009, **113**, 6378–6396.
8 C. Y. Legault, *CYLView*, 1.0b, Université de Sherbrooke, 2009 (http://www.cylview.org).
9 J. P. Foster and F. Weinhold, *J. Am. Chem. Soc.*, 1980, **102**, 7211–7218.
10 K. B. Wiberg, *Tetrahedron*, 1968, **24**, 1083–1096.
11 I. Mayer, *Chem. Phys. Lett.*, 1983, **97**, 270–274.
12 I. Mayer, *J. Comput. Chem.*, 2007, **28**, 204–221.
13 R. G. Parr and W. Yang, *Density Functional Theory of Atoms and Molecules Vol. 16*, Oxford Univ. Press, New York, 1989.
14 W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory*, Wiley-VCH, Weinheim, 2001.
15 P. K. Chattaraj, *Chemical reactivity theory : a density functional view*, CRC Press, Taylor & Francis, 2009.
16 R. G. Parr, R. A. Donnelly, M. Levy and W. E. Palke, *J. Chem. Phys.*, 1978, **68**, 3801–3807.
17 R. G. Parr and R. G. Pearson, *J. Am. Chem. Soc.*, 1983, **105**, 7512–7516.
18 R. G. Pearson, *Proc. Nati. Acad. Sci. U.S.A.* 1986, **83**, 8440.
19 M. Torrent-Sucarrat, M. Duran and M. Solà, *J. Phys. Chem. A* 2002, **106**, 4632–4638.
20 W. Yang and R. G. Parr, *Proc. Nati. Acad. Sci. U.S.A.* 1985, **82**, 6723.
21 M. Berkowitz and R. G. Parr, *J. Chem. Phys.*, 1988, **88**, 2554–2557.
22 R. G. Parr, L. v. Szentpály and S. Liu, *J. Am. Chem. Soc.*, 1999, **121**, 1922–1924.
23 J. P. Perdew, R. G. Parr, M. Levy and J. L. Balduz, *Phys. Rev. Lett.*, 1982, **49**, 1691–1694.
24 A. K. Chandra and M. T. Nguyen, *Faraday Discuss.*, 2007, **135**, 191–201.
25 R. G. Parr and W. Yang, *J. Am. Chem. Soc.*, 1984, **106**, 4049–4050.
26 W. Yang and W. J. Mortier, *J. Am. Chem. Soc.*, 1986, **108**, 5708–5711.
27 C. Morell, A. Grand and A. Toro-Labbé, *J. Phys. Chem. A* 2005, **109**, 205–212.
28 C. Morell, A. Grand and A. Toro-Labbé, *Chem. Phys. Lett.*, 2006, **425**, 342–346.
29 A. Poater, F. Ragone, S. Giudice, C. Costabile, R. Dorta, S. P. Nolan and L. Cavallo, *Organometallics*, 2008, **27**, 2679–2681.
30 A. Poater, B. Cosenza, A. Correa, S. Giudice, F. Ragone, V. Scarano and L. Cavallo, *Eur. J. Inorg. Chem.*, 2009, **2009**, 1759–1766.
31 H. Clavier and S. P. Nolan, *Chem. Commun.*, 2010, **46**, 841–861.
32 A. Poater, F. Ragone, R. Mariz, R. Dorta and L. Cavallo, *Chem. Eur. J.* 2010, **16**, 14348–14353.
33 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay and G. Louppe, *J. Mach. Learn. Res.* 2012, **12**.
34 J. Karch, *Improving on Adjusted R-Squared*, 2019.
35 M. Salamone, G. Carboni and M. Bietti, *J. Org. Chem.*, 2016, **81**, 9269–9278.
36 S. Feng, X. Xie, W. Zhang, L. Liu, Z. Zhong, D. Xu and X. She, *Org. Lett.*, 2016, **18**, 3846–3849.
37 A. M. Carestia, D. Ravelli and E. J. Alexanian, *Chem. Sci.*, 2018, **9**, 5360–5365.
38 S. Mukherjee, B. Maji, A. Tlahuext-Aca and F. Glorius, *J. Am. Chem. Soc.*, 2016, **138**, 16200–16203.
39 T. Lu, *Molclus program,* http://www.keinsci.com/research/molclus.html.
40 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics* 1996, 14, 33. VMD Official website. http://www.ks.uiuc.edu/Research/vmd/.
41 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R.

Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, Y. Vázquez-Baeza and C. SciPy, *Nat. Methods*, 2020, **17**, 261–272.
42 C. Le, Y. Liang, R. W. Evans, X. Li and D. W. C. MacMillan, *Nature*, 2017, **547**, 79–83.
43 X. Zhang and D. W. C. MacMillan, *J. Am. Chem. Soc.*, 2017, **139**, 11353–11356.
44 K. A. Margrey, W. L. Czaplyski, D. A. Nicewicz and E. J. Alexanian, *J. Am. Chem. Soc.*, 2018, **140**, 4213–4217.
45 J. Jin and D. W. C. MacMillan, *Angew. Chem. Int. Ed.*, 2015, **54**, 1565–1569.
46 J. D. Cuthbertson and D. W. C. MacMillan, *Nature*, 2015, **519**, 74–77.
47 M. Salamone and M. Bietti, *Acc. Chem. Res.*, 2015, **48**, 2895–2903.
48 K. Qvortrup, D. A. Rankic and D. W. C. MacMillan, *J. Am. Chem. Soc.*, 2014, **136**, 626–629.
49 J. Jin and D. W. C. MacMillan, *Nature*, 2015, **525**, 87–90.
50 S. Devari, M. A. Rizvi and B. A. Shah, *Tetrahedron Lett.*, 2016, **57**, 3294–3297.
51 S. Mukherjee, T. Patra and F. Glorius, *ACS Catal.*, 2018, **8**, 5842–5846.
52 S. Mukherjee, R. A. Garza-Sanchez, A. Tlahuext-Aca and F. Glorius, *Angew. Chem. Int. Ed.*, 2017, **56**, 14723–14726.
53 M. Bietti, R. Martella and M. Salamone, *Org. Lett.*, 2011, **13**, 6110–6113.
54 M. Salamone, M. Milan, G. A. DiLabio and M. Bietti, *J. Org. Chem.*, 2013, **78**, 5909–5917.
55 M. Bietti and M. Salamone, *Org. Lett.*, 2010, **12**, 3654–3657.
56 M. Salamone, I. Giammarioli and M. Bietti, *J. Org. Chem.*, 2011, **76**, 4645–4651.
57 M. Salamone, L. Mangiacapra and M. Bietti, *J. Org. Chem.*, 2015, **80**, 1149–1154.
58 M. Milan, M. Salamone and M. Bietti, *J. Org. Chem.*, 2014, **79**, 5710–5716.
59 M. Salamone, G. A. DiLabio and M. Bietti, *J. Org. Chem.*, 2012, **77**, 10479–10487.
60 M. Milan, M. Salamone, M. Costas and M. Bietti, *Acc. Chem. Res.*, 2018, **51**, 1984–1995.
61 M. Salamone, G. A. DiLabio and M. Bietti, *J. Am. Chem. Soc.*, 2011, **133**, 16625–16634.
62 M. Salamone, R. Martella and M. Bietti, *J. Org. Chem.*, 2012, **77**, 8556–8561.
63 M. Salamone, M. Milan, G. A. DiLabio and M. Bietti, *J. Org. Chem.*, 2014, **79**, 7179–7184.
64 M. Salamone, F. Basili and M. Bietti, *J. Org. Chem.*, 2015, **80**, 3643–3650.
65 M. Salamone, V. B. Ortega, T. Martin and M. Bietti, *J. Org. Chem.*, 2018, **83**, 5539–5545.
66 M. Salamone, V. B. Ortega and M. Bietti, *J. Org. Chem.*, 2015, **80**, 4710–4715.
67 E. Boess, L. M. Wolf, S. Malakar, M. Salamone, M. Bietti, W. Thiel and M. Klussmann, *ACS Catal.*, 2016, **6**, 3253–3261.