Supporting Information

Every atom counts: Predicting sites of reaction based on chemistry within two bonds

Ching Ching Lam[1] and Jonathan M. Goodman[1]*

1. Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW
*jmg11@cam.ac.uk

Table of Contents

1. Dataset and Scripts

1.1 Data

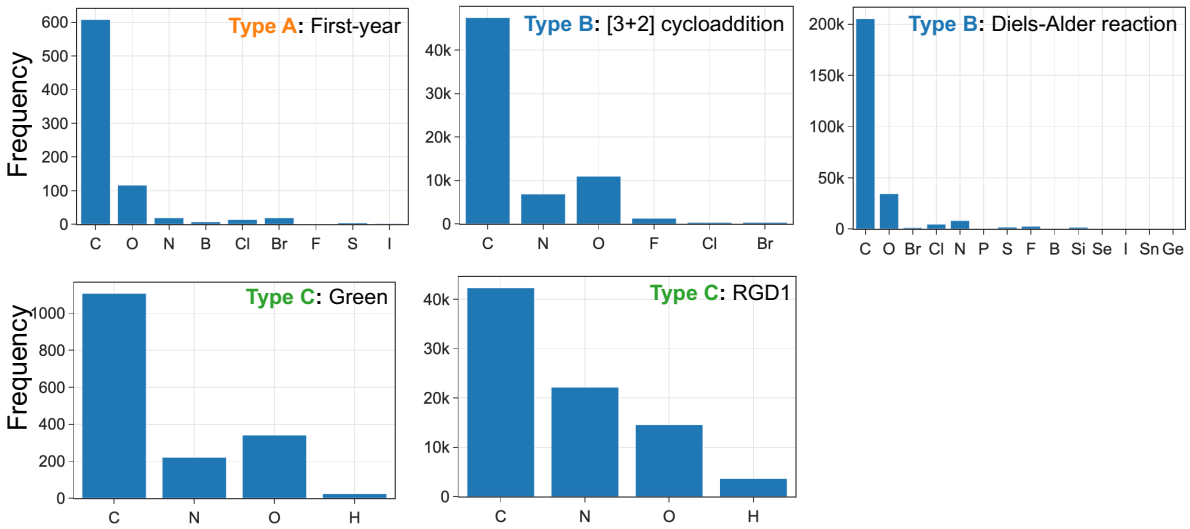


Figure S1. An overview of the reactants in each dataset. The bar charts summarize the nature of the non-hydrogen atoms in the molecules.

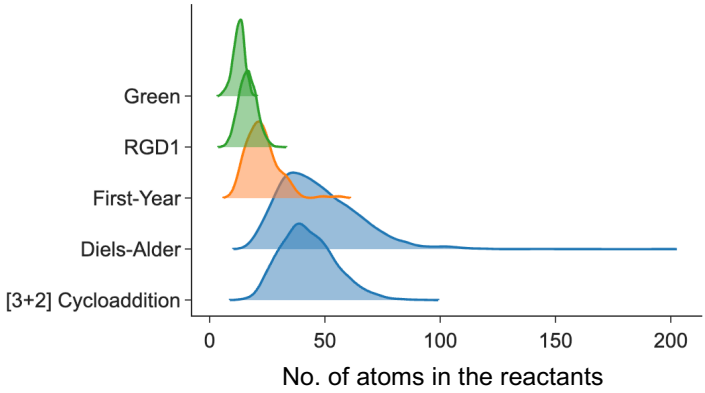


Figure S2. The distribution of the size of the chemical systems in each dataset.

Table S1. Numerical records on filtering reactions

Dataset	Number of filtered reactions	Number of reactions before filtering	Number of reactions after filtering
First-year	0	147	147
[3+2] cycloaddition	21	5974	5953
Diels-Alder	263	11274	11011
RGD1 (EA < 40 kcal mol^{-1})	125	11406	11281
Green (EA < 40 kcal mol^{-1})	11	332	321
RGD1 (EA < 60 kcal mol^{-1})	517	28038	27521
Green (EA < 60 kcal mol^{-1})	37	1049	1012

Source of errors:

1. **3D structure error**:

Unable to be processed by RDkit to generate 3D structure at the force field level with the default setting, *ie* return errors when encountering the following functions –

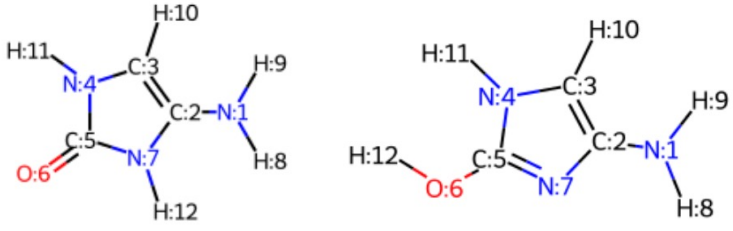
```
AllChem.EmbedMolecule(mol,randomSeed=0xf00d)
AllChem.MMFFOptimizeMolecule(mol)
mol.GetConformer()
```

2. **Re-indexing error**:

Atom-to-atom mapping needs to be carried out between the competitive pathways, reactions with the same reactants but different products. This ensures consistent atom indexing across the competitive pathways, which is important for generating labels in preparation for model training.

The atom-to-atom mapping across competitive pathways was done via mol1.GetSubstructMatch(mol2) function from RDkit. mol1 and mol2 are RDkit Chem.Mol objects generated from reactant SMILES strings that contain atom indexing. mol1.GetSubstructMatch(mol2) returns the indices of atoms in mol1 that match with mol2.

Reactions with the same reactants were grouped together via conversion to InChI string. Inevitably, reactions with tautomer structures with difference connectivity as reactants were grouped together, such as the example below:



InChI=1S/C3H5N3O/c4-2-1-5-3(7)6-2/h1H,4H2,(H2,5,6,7)

Figure S3. Tautomer structures with difference connectivity share the same InChI string.

The mol1.GetSubstructMatch(mol2) function was unable to match resonance structures with different connectivity. This led to the re-indexing error.

3. **Placeholder atom error:**

The reaction SMILES string contains placeholder atoms, '*'.

4. **Hypervalent error:**

The reaction involved hypervalent molecule.

A breakdown on the filtered reactions:

[3+2] cycloaddition: 21 reactions were filtered due to 3D structure errors.

Diels-Alder: 7 reactions contain hypervalent molecules. 8 reactions involve re-indexing errors during the atom-to-atom mapping for competitive pathways. 22 reactions contain placeholder atoms. 226 reactions were filtered due to 3D structure errors.

RGD1 (EA < 40 kcal mol^{-1}): 63 reactions were filtered due to 3D structure errors. 62 reactions encounter re-indexing errors during the atom-to-atom mapping for competitive pathways.
RGD1 (EA < 60 kcal mol^{-1}): 89 reactions were filtered due to 3D structure errors. 428 reactions encounter re-indexing errors during the atom-to-atom mapping for competitive pathways.

Green (EA < 40 kcal mol^{-1}): 11 reactions were filtered due to 3D structure errors.
Green (EA < 60 kcal mol^{-1}): 33 reactions were filtered due to 3D structure errors. 4 reactions encounter re-indexing errors during the atom-to-atom mapping for competitive pathways.

Combined dataset

The overlaps between the datasets with the 'chk_dup_v2.py' script via converting the reactant and product into InChI strings. The test shows that there are overlaps between the RGD1 and Green dataset (EA cut-off < 40 kcal mol^{-1}). Three reactions (idx in RGD1 dataset = 1308, 6573, 6574) in the RGD1 dataset also appear in the Green dataset. One reaction (idx in RGD1 dataset = 11190) from the RGD1 dataset is a competitive pathway, the same reactants but different products, to reactions in the Green dataset. These reactions were removed from the RGD1 dataset prior to selecting reactions of the combined dataset.

1.2 Scripts

GitHub repository TwoBondChem/
https://github.com/Goodman-lab/TwoBondChem

A. Directory tree of the TwoBondChem/

```
TwoBondChem/
├─ bond_strength_descriptor/
│    ├─ bond_pre_classifier_v2.py
│    └─ bondlength_fromsmi_v2.py
├─ data_processing/
│    ├─ competitive_pathway_atom_mapping.py
│    ├─ initial_treatment.py
│    └─ remove_emb_error.py
├─ dataset/
│    ├─ cyclo_data_v2_16072024.csv
│    ├─ da_08012024_vCorr.csv
│    ├─ diels_alder_data_v7_19052024.csv
│    ├─ exam_test.csv
│    ├─ exam_test_28122023.csv
│    ├─ fav_RGD1_13012024_v2.csv
│    ├─ first-year_data_all_27122023.csv
│    ├─ green_04012024.csv
│    ├─ green_ea60_c_23052024.csv
│    └─ rgd_ea60_c_23052024.csv
├─ model_evaluation/
│    ├─ chk_dup_v3.py
│    ├─ evaluation.py
│    ├─ random_sample_test.py
│    ├─ random_sample_test_all.py
│    ├─ random_sample_test_Bmix.py
│    ├─ random_sample_test_Cmix.py
│    ├─ save_model.py
│    ├─ save_model2.py
│    ├─ take_one_out_test.py
│    └─ train_size_test.py
├─ model_test_2.ipynb
├─ React.py
├─ README.md
├─ RF_model_yr1_28122023.sav
└─ scripts/
     ├─ atidx.py
     ├─ bond_classification_01112022.csv
     ├─ get_descriptors_v2.py
     ├─ get_label_v4.py
     ├─ model_training_v3.py
     └─ training_prepare_v5.py
├─ cyclo_2b+/
│    ├─ cyclo_0_13012024_label.csv
│    ├─ cyclo_1_13012024_label.csv
│    ├─ cyclo_2_13012024_label.csv
│    ├─ cyclo_3_13012024_label.csv
│    ├─ cyclo_4_13012024_label.csv
│    ├─ cyclo_5_13012024_label.csv
│    ├─ cyclo_6_13012024_label.csv
│    ├─ cyclo_7_13012024_label.csv
```

```
│   ├── cyclo_8_13012024_label.csv
│   └── cyclo_9_13012024_label.csv
├── random_sample_test_EXAMPLE.py
└── evaluate_EXAMPLE.ipynb
```

*Archive files have not been included

B. Environment

The scripts in this project are written in Python under the following environment:

- The Python (3.8.12) Standard Library
- Pandas (1.1.5)
- Numpy (1.19.2)
- Sklearn (0.24.2/1.0.1)
- RDkit (2021.9.4)
- Scipy (1.4.1)
- Rxnmapper (0.1.4)

C. Data

dataset/ folder:

All the datasets (Type A: the first-year reactions + three reactions from Part 1A exam at the University of Cambridge; Type B: [3+2] cycloaddition[1] and Diels-Alder reaction dataset;[2] Type C: the Reaction Graph Depth 1 (RGD1))[3] and the Green dataset).[4] are available in the dataset/ folder as csv files. The reaction data is processed and formatted in the same style for the purpose of this investigation. The atom-to-atom mapping numbering in the reaction SMILES is consistent for competitive pathways with the same reactants. There are three columns in each csv file:

- idx: index of the reaction
- code: reactions with the same code are competitive pathways (*ie* having the same reactants)
- reaction: mapped reaction SMILES

The RGD1 dataset csv file also has an 'Rind' column, which corresponds to the index in the 'reaction' column in the original datafile:
https://figshare.com/articles/dataset/model_reaction_database/21066901?file=40272727 (accessed Feb 5th, 2024).

'da_08012024_vCorr.csv' contains 100 Diels-Alder reactions, where the atom-to-atom mapping errors have been picked out and corrected manually.[5] Before the correction, reactions with index = 2, 30, 31, 49, 66, 70, 88 and 99 contain errors.

Three out of 147 reactions in the first-year reaction dataset ('first-year_data_all_27122023.csv') had mapping errors, which were subsequently corrected manually. Before the correction, reactions with index = 50, 74 and 124 contain errors. Reactions with a code greater than 77 belong to the testing data set.

We filtered and processed the RGD1 and the Green datasets to ensure that the reactions are thermodynamically favourable (*ie* $\Delta H_r < 0$ kcal mol^{-1}) with a low kinetic barrier (*ie* activation energy, EA < 40 kcal mol^{-1}). This gives the 'fav_RGD1_13012024_v2.csv' and 'green_04012024.csv'. To study the

effect of varying the EA cut-off, we also filtered and processed the RGD1 and the Green datasets with an EA cut-off of 60 kcal mol and ΔH_r < 0 kcal mol^{-1} – 'rgd_ea60_c_23052024.csv' and 'green_ea60_c_23052024.csv'.

data_processing/competitive_pathway_atom_mapping.py input and output example files:
'exam_test.csv' and 'exam_test_28122023.csv'

D. Scripts

React_v2.py: code for using the random forest model
model_test_3.ipynb: Jupyter notebook on how to use the 'React.py' script

Saved models:

Table S2. Saved models in the GitHub

File	Note
RF_model_yr1_28122023.sav	RF model trained from all reactions in the first-year dataset
RF_model_cyclo_11062024.sav	RF model trained from the first 100 reactions in the [3+2] cycloaddition dataset
RF_model_da_11062024.sav	RF model trained from the first 100 reactions in the Diels-Alder dataset
RF_model_green_11062024.sav	RF model trained from the first 100 reactions in the Green dataset
RF_model_rgd_11062024.sav	RF model trained from the first 300 reactions in RGD1 dataset
RF_model_Bmix_11062024.sav	RF model trained from 100 sets of randomly selected [3+2] cycloaddition reactions and 100 sets of randomly selected Diels-Alder reactions.
RF_model_Cmix_11062024.sav	RF model trained from 100 sets of randomly selected Green reactions and 100 sets of randomly selected RGD1 reactions.
RF_model_all_11062024.sav	RF model trained from 78 sets of first-year reactions, 100 sets of [3+2] cycloaddition reactions, 100 sets of Diels-Alder reactions, 100 sets of Green reactions and 100 sets of RGD1 reactions. All reactions were randomly selected from the corresponding full dataset.

scripts/ folder:

atidx.py: contains the functions for conducting atom-to-atom mapping and formatting the reaction SMILES strings for competitive pathways

get_descriptors_v2.py: contains functions for generating the atomistic descriptor components

- Associated file: **'bond_classification_01112022.csv'** – the parameters for generating the bond strength descriptors

get_label_v4.py: contains functions for generating the atomistic label
training_prepare_v5.py: compiles functions in 'get_descriptors_v2.py' and 'get_label_v4.py' to generate the descriptor arrays and labels for atoms in a set of reactants
model_training_v3.py: functions for training and evaluating the model

data_processing/ folder:

initial_treatment.py: group reactions with the same reactants together vis InChI strings; for the Diels-Alder dataset only, reactions with placeholder atoms are also removed. Reactions with hypervalent molecules lead to errors in the model training and testing. They are removed manually from the Diels-Alder dataset.

remove_emb_error.py: filter reactions with 3D structure errors

competitive_pathway_atom_mapping.py: This script executes the functions in atidx.py to conduct atom-to-atom mapping of individual reactions and between the competitive pathways. This process was conducted only on the first-year and Diels-Alder datasets.

- Associated files in the dataset/ folder: 'exam_test.csv' and 'exam_test_28122023.csv' – the input and output csv file from executing the script are provided for illustrations

bond_strength_descriptor/ folder:

bondlength_fromsmi_v2.py: gather bond length data from SMILES strings
bond_pre_classifier_v2.py: using the bond length data to generate bond strength descriptor classification criteria, *ie* the 'bond_classification_01112022.csv' file

model_evaluation/ folder:

All the scripts under the model_evaluation/ should be moved from TwoBondChem/model_evaluation/ to TwoBondChem/ before execution.

chk_dup_v3.py: check if there are overlaps between datasets

random_sample_test.py, random_sample_test_all.py, random_sample_test_Bmix.py, random_sample_test_Cmix.py: scripts for carrying out random sampling test

take_one_out_test.py: the script for carrying out the take-one-out cross-validation test

test_size_test.py: The script for investigating the effect of varying the testing or training dataset size

save_model.py: executing this script to train models using reactions from a single dataset and return the trained models in .sav files
save_model2.py: executing this script to train models using reactions from multiple datasets and return the trained models in .sav files

evaluation.py: This script takes the .csv output from the above script to calculate performance metrics by comparing the predictions with the actual labels.

hyperpara_test_v5.py and **para_df.csv**: the script and csv file for conducting hyperparameter tuning

The following files under TwoBondChem/ provide an example of the training and evaluation procedure:

random_sample_test_EXAMPLE.py: training and testing for 2-bond+ models trained from the [3+2] cycloaddition dataset. The code is taken from 'random_sample_test.py' under model_evaluation/ folder. This script can be executed as it is from the TwoBondChem/ folder
cyclo_2b+/ contains output files from executing the 'random_sample_test_EXAMPLE.py' script
evaluate_EXAMPLE.ipynb: evaluation of the predictions using the 'evaluation.py' under model_evaluation/ folder

1.3 Atom-to-Atom Mapping

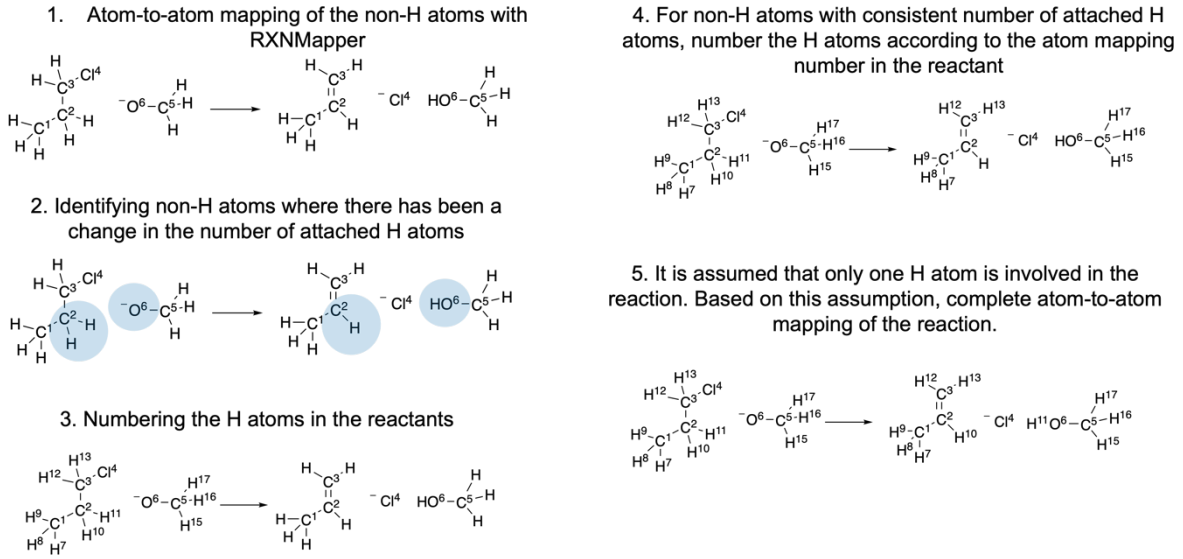


Figure S4. Atom-to-atom mapping procedures for the first-year and Diels-Alder reaction dataset

2. Bond strength descriptors

The bond strength component involves categorising the bonds to which the atom is connected into one of the 86 bond classes across 12 bond types (*ie* CC, CN, CH, CCl, CO, CS, CF, NO, HN, HO and OP). The complete list of bond classes is in SI Table S3. The classification of the 86 bond classes was established using the method employed in the MolE8 machine-learned potential energy surface model.[6] The classification of bonds is based on the bond length of the specific bond type. The parametrisation procedure was repeated and the bond length distributions of each bond type on MMFF-optimised structures were studied using a dataset of 100,000 molecules from ChEMBL-28.[7] Kernel density estimations (KDE) were applied to the histogram of the bond length distribution (SI Figure S5). The minima observed on the KDE curve were identified as the boundaries dividing each bond class.

The bond strength descriptors explicitly provide information on the chemical environment of the atom beyond the two-bond range. The bond strength descriptors can distinguish between the same bond (*ie* the same bond type and bond order) but located in different chemical environments due to subtle bond length differences. For example, the single C-C bond within a conjugated system and in a saturated chain have slightly different bond lengths and are categorised into two bond strength classes (SI Figure S6).

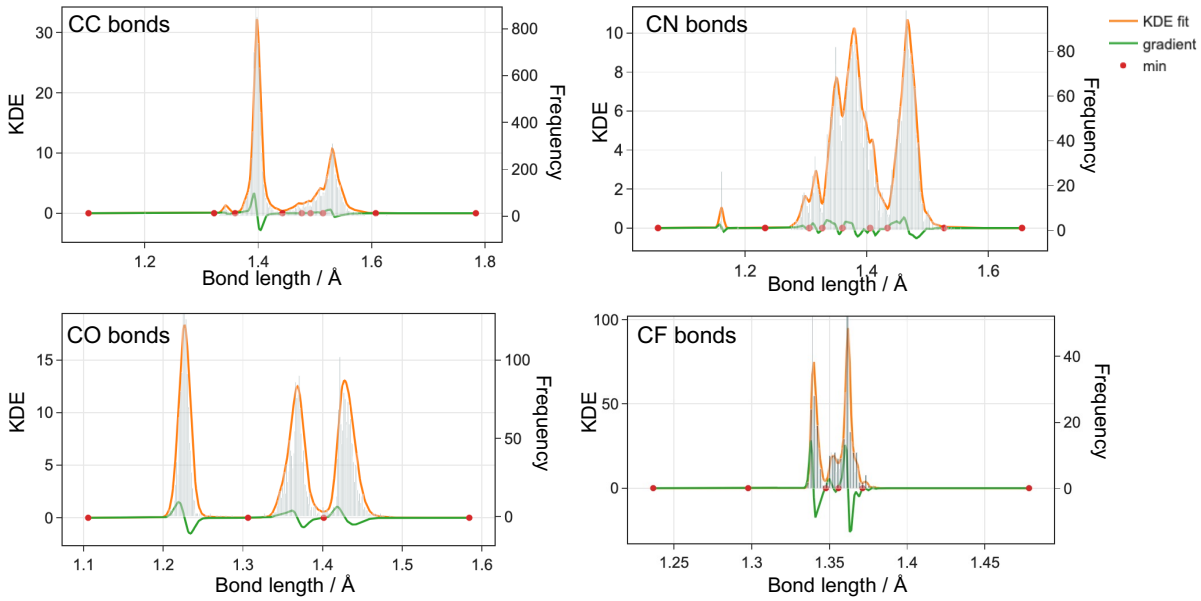


Figure S5. Derivation of the bond strength classes from bond length distribution histograms. The corresponding plot for CC, CN, CO and CF are given above as examples.[6,7]

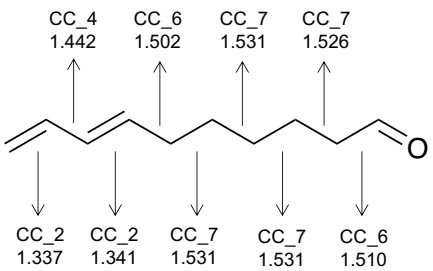


Figure S6. Bond strength class and bond length in Å for all the CC bonds in an example molecule.

Table S3. Classification of the bond strength class based on the bond length: bond classes and the associated parameters under each bond type.

Bond Type	**Bond strength classes** **{'bond class name':[minimum length in Å, maximum length in Å]}**
CC	{'CC_0': [0, 1.1007], 'CC_1': [1.1007, 1.3593], 'CC_2': [1.3593, 1.4769], 'CC_3': [1.4769, 1.3223], 'CC_4': [1.3223, 1.4429], 'CC_5': [1.4429, 1.4924], 'CC_6': [1.4924, 1.5144], 'CC_7': [1.5144, 1.6075], 'CC_8': [1.6075, 1.784], 'CC_9': [1.784, 3]}
CN	{'CN_0': [0, 1.0569], 'CN_1': [1.0569, 1.3056], 'CN_2': [1.3056, 1.3271], 'CN_3': [1.3271, 1.3602], 'CN_4': [1.3602, 1.233], 'CN_5': [1.233, 1.4057], 'CN_6': [1.4057, 1.4342], 'CN_7': [1.4342, 1.5273], 'CN_8': [1.5273, 1.6554], 'CN_9': [1.6554, 3]}
CH	{'CH_0': [0, 0.9363], 'CH_1': [0.9363, 1.0756], 'CH_2': [1.0756, 1.1072], 'CH_3': [1.1072, 1.2415], 'CH_4': [1.2415, 1.0912], 'CH_5': [1.0912, 3]}
CCl	{'CCl_0': [0, 1.5942], 'CCl_1': [1.5942, 1.7386], 'CCl_2': [1.7386, 1.7251], 'CCl_3': [1.7251, 1.713], 'CCl_4': [1.713, 1.8929], 'CCl_5': [1.8929, 3]}
CO	{'CO_0': [0, 1.1058], 'CO_1': [1.1058, 1.5842], 'CO_2': [1.5842, 1.3065], 'CO_3': [1.3065, 1.4016], 'CO_4': [1.4016, 3]}
HN	{'HN_0': [0, 0.8936], 'HN_1': [0.8936, 1.0002], 'HN_2': [1.0002, 1.0173], 'HN_3': [1.0173, 1.0113], 'HN_4': [1.0113, 1.0108], 'HN_5': [1.0108, 1.0243], 'HN_6': [1.0243, 1.0248], 'HN_7': [1.0248, 1.0293], 'HN_8': [1.0293, 1.1665], 'HN_9': [1.1665, 3]}
CS	{'CS_0': [0, 1.5101], 'CS_1': [1.5101, 1.7539], 'CS_2': [1.7539, 1.6793], 'CS_3': [1.6793, 1.7428], 'CS_4': [1.7428, 1.6302], 'CS_5': [1.6302, 1.8009], 'CS_6': [1.8009, 1.875], 'CS_7': [1.875, 1.9866], 'CS_8': [1.9866, 3]}
OS	{'OS_0': [0, 1.3371], 'OS_1': [1.3371, 1.7609], 'OS_2': [1.7609, 1.4747], 'OS_3': [1.4747, 3]}
CF	{'CF_0': [0, 1.3558], 'CF_1': [1.3558, 1.2366], 'CF_2': [1.2366, 1.3477], 'CF_3': [1.3477, 1.3713], 'CF_4': [1.3713, 1.2977], 'CF_5': [1.2977, 1.4784], 'CF_6': [1.4784, 3]}
NO	{'NO_0': [0, 1.1352], 'NO_1': [1.1352, 1.4542], 'NO_2': [1.4542, 1.5734], 'NO_3': [1.5734, 1.3306], 'NO_4': [1.3306, 3]}
HO	{'HO_0': [0, 0.8564], 'HO_1': [0.8564, 0.9611], 'HO_2': [0.9611, 0.9747], 'HO_3': [0.9747, 0.9752], 'HO_4': [0.9752, 0.9777], 'HO_5': [0.9777, 0.9792], 'HO_6': [0.9792, 0.9797], 'HO_7': [0.9797, 0.9872], 'HO_8': [0.9872, 1.1189], 'HO_9': [1.1189, 3]}
OP	{'OP_0': [0, 1.3884], 'OP_1': [1.3884, 1.7245], 'OP_2': [1.7245, 1.5467], 'OP_3': [1.5467, 3]}

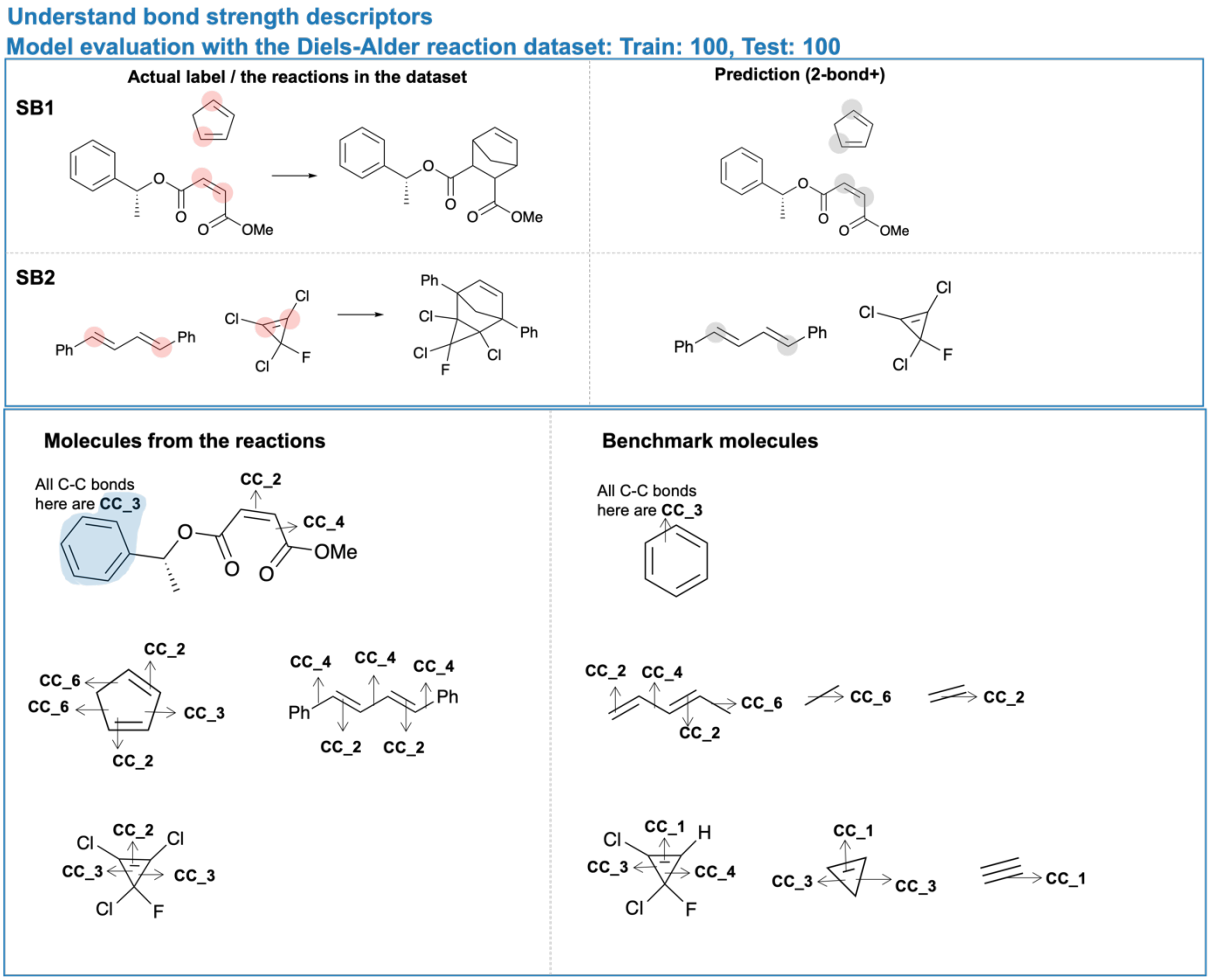


Figure S7. Additional case studies on model trained with Diels-Alder reaction data to understand bond strength descriptors.

The bond strength descriptor classification depends on the bond length measurements on the force field structure. In frequently appeared chemical motifs, the classifications should match our chemical intuition better. For example, single, aromatic, double and triple C-C bonds typically have the bond strength descriptor of CC_6, CC_3, CC_2 and CC_1, respectively. For the less common chemical motif, the classification tends to be harder to interpret and might be against our intuition. For example, the three C-C bonds within a cyclopropene are classified as CC_3, CC_3 and CC_1, respectively. Inevitably, this would impact the performance of the model. In the above study, the model trained on Diels-Alder reactions could not identify the cyclopropene motif as a potential dienophile.

3. Model Evaluation

Algorithm Benchmarking:

Table S4. Performance on benchmarking with different machine learning algorithms: random forest (RF), K-nearest neighbour (KNN), support vector (SVC), gaussian process (Gaussian) and multi-layer perceptron classifier (NN). The number of sets of reactions involved in training and testing are specified below in bracket. The '2-bond +' descriptor composition was used. The sets of reactions for training and testing are indicated in the brackets.

Model	**Accuracy**	**Precision**	**Recall**
	Train: year 1 A (78), Test: year 1 B (30)		
RF	92.0%	86.9%	70.5%
KNeigh	89.5%	80.4%	62.1%
SVC	86.5%	73.3%	50.0%
Gaussian	91.1%	86.9%	65.2%
NN	91.0%	84.6%	66.7%
	Train: [3+2] cycloaddition (100), Test: [3+2] cycloaddition (100)		
RF	100.0%	99.8%	100.0%
KNeigh	99.1%	99.8%	94.4%
SVC	97.7%	100.0%	85.2%
Gaussian	99.4%	99.8%	96.5%
NN	99.9%	99.8%	99.3%
	Train: RGD1 (300), Test: RGD1 (100)		
model	accuracy	precision	recall
RF	81.4%	76.4%	60.7%
KNeigh	80.6%	74.9%	59.2%
SVC	78.9%	74.3%	52.3%
Gaussian	81.3%	78.0%	58.0%
NN	81.1%	76.8%	58.8%

Take-one-out cross-validations procedures:

1. A set of reactions was taken out for testing and evaluation. The model training was carried out on the rest of the dataset. The '2-bond +' descriptor composition was used. The training is performed with random forest classifier.
2. Predictions were made and recorded on atoms in the set of reactions outside of the training set.
3. The above steps were repeated for all sets of reactions in the dataset.

The unit for training and testing data is 'set', *ie* a set of competitive reactions with the same reactants. The specific datasets and number of sets of reactions involved are specified in the tables below.

Table S5A: Results of the take-one-out cross-validation test. Predictions from the models were treated collectively to calculate the metrices below.

Dataset (no. of sets of reactions)	% by atoms			% sets of reactants with:		
	Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
First-year (107)	89.8%	80.6%	63.9%	24.3%	39.3%	41.1%
[3+2]-cycloaddition (100)	99.9%	99.5%	99.5%	96.0%	100.0%	98.0%
Diels-Alder (100)	97.3%	90.6%	82.4%	47.0%	76.0%	63.0%
Diels-Alder (100)*	98.5%	92.2%	91.6%	58.0%	85.0%	72.0%
RGD1 (100)	82.1%	74.3%	67.8%	7.0%	29.0%	28.0%
Green (100)	83.8%	83.9%	80.4%	27.0%	49.0%	50.0%

*After correcting the atom-to-atom mapping errors

Table S5B: Results of the take-one-out cross-validation test. The predictions for each set of reactions were treated individually to compute performance metrices, followed by calculating the mean and standard deviation for the entire dataset. As each set of results from testing only involve less than 50 atoms, the standard deviation is expected to be large.

Dataset (no. of sets of reactions)	% by atoms		
	Accuracy	Precision	Recall
First-year (107)	89.8±10.3%	72.4±27.2%	83.1±22.4%
[3+2]-cycloaddition (100)	99.8±0.7%	99.5±3.5%	99.7±2.4%
Diels-Alder (100)	97.1±4.9%	86.3±20.8%	92.7±13.9%
Diels-Alder (100)*	98.5±2.3%	91.6±15.3%	94.3±12.3%
RGD1 (100)	81.8±12.1%	70.5±25.7%	72.7±24.4%
Green (100)	83.6±15.1%	82.9±20.4%	86.3±17.8%

*After correcting the atom-to-atom mapping errors

Increasing the size of the training data:

Table S6: Model performance when varying the size of the training data. The RGD1 dataset was used for conducting this investigation. The '2-bond +' descriptor composition was used. The unit for training and testing data is 'set', *ie* a set of competitive reactions with the same reactants. The same set of data, 100 sets of RGD1 reactions (*ie* 100 reactions in total), was used for testing.

No. of sets of reactions in training	% by atoms			% sets of reactants with:		
	Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
100	76.3%	63.1%	63.0%	6.0%	14.0%	28.0%
300	81.2%	76.4%	59.9%	8.0%	30.0%	24.0%
600	81.7%	75.8%	62.8%	6.0%	27.0%	25.0%
900	82.6%	78.8%	62.4%	12.0%	35.0%	27.0%
3000	83.2%	79.6%	63.9%	16.0%	38.0%	25.0%
6000	84.9%	84.5%	64.7%	14.0%	47.0%	26.0%
10300	84.3%	83.5%	63.5%	15.0%	43.0%	28.0%

Combining the datasets:

Table S7: Performance of '2-bond +' models trained from the combined dataset. The specific datasets and number of sets of reactions involved are specified below in a consistent format. Let us take the first model presented below, 'Train: 50 (104) [3+2] cycloaddition + 50 (50) Diels-Alder', as an example. 50 sets of [3+2] cycloaddition reactions with the same reactants (*ie* equivalent to 104 reactions) and 50 sets of Diels-Alder reactions with the same reactants (*ie* equivalent to 50 reactions) are involved in training. In total, there are 154 reactions in the training dataset. Other than the overall results, the performance breakdown by the reactions from different datasets in the testing dataset is also given below.

	% by atoms			% sets of reactants with:		
Testing dataset	Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
Type B						
Train: 50 (104) [3+2] cycloaddition + 50 (50) Diels-Alder, Test: 50 (101) [3+2] cycloaddition + 50 (50) Diels-Alder						
Overall	97.6%	91.1%	89.7%	58.0%	85.0%	75.0%
[3+2] cycloaddition	99.2%	100.0%	95.3%	80.0%	100.0%	80.0%
Diels-Alder	96.5%	83.5%	84.6%	36.0%	70.0%	70.0%
Train: 100 (205) [3+2] cycloaddition + 100 (100) Diels-Alder, Test: 100 (219) [3+2] cycloaddition + 100 (100) Diels-Alder						
Overall	98.2%	94.1%	91.3%	68.0%	88.0%	80.5%
[3+2] cycloaddition	99.5%	99.8%	97.2%	88.0%	99.0%	89.0%
Diels-Alder	97.4%	88.6%	85.7%	48.0%	77.0%	72.0%
Type C						
Type C: Train: 50 (53) RGD1 + 50 (59) Green, Test: 50 (57) RGD1 + 50 (61) Green						
Overall	80.3%	76.6%	69.8%	11.0%	31.0%	36.0%
RGD1	79.3%	73.0%	62.3%	4.0%	18.0%	22.0%
Green	81.9%	80.1%	78.1%	18.0%	44.0%	50.0%
Train: 100 (107) RGD1 + 100 (126) Green, Test: 100 (110) RGD1 + 100 (120) Green						
Overall	79.8%	76.5%	67.3%	20.0%	35.0%	39.0%
RGD1	77.7%	66.0%	62.6%	7.0%	20.0%	31.0%
Green	82.8%	88.7%	72.0%	33.0%	50.0%	47.0%
Combining all datasets: the global model						
Train: 77(105) first-year + 100 (205) [3+2] cycloaddition + 100 (100) Diels-Alder + 100 (107) RGD1 + 100 (126) Green, Test: 30 (41) first-year + 100 (219) [3+2] cycloaddition + 100 (100) Diels-Alder + 100 (110) RGD1 + 100 (120) Green						
Overall	92.5%	81.1%	75.7%	35.3%	50.7%	54.2%
First-year	91.3%	83.0%	68.8%	33.3%	56.7%	53.3%
[3+2] cycloaddition	99.1%	96.8%	96.6%	80.0%	94.0%	87.0%
Diels-Alder	96.1%	78.8%	83.3%	38.0%	54.0%	65.0%
RGD1	77.6%	68.2%	59.0%	6.0%	19.0%	27.0%
Green	82.5%	81.1%	70.1%	18.0%	34.0%	38.0%

Hyperparameter tuning

Table S8. Performance on benchmarking with different hyperparameter setting for the RF model. The hyperparameter tuning test was conducted on three different datasets: first-year, [3+2] cycloaddition and RGD1. In the hyperparameter tuning using the first-year dataset, the split of the training and testing dataset is as presented in SI Figure S8. 78 sets of reactions were chosen for training and the rest for testing. For the hyperparameter tuning using the [3+2] cycloaddition and RGD1 dataset, the first 100 sets of reactions were used for training and the sequential 100 sets of reactions were used for testing. The results with the default hyperparameter setting from the scikit-learn package are highlighted in yellow.

Parameter max_features, criterion, n_estimators	Accuracy	Precision	Recall
Train: first-year (78), Test: first-year (30)			
entropy, log2, 100	92.5%	71.9%	86.8%
entropy, log2, 200	93.2%	73.4%	89.5%
entropy, log2, 20	92.8%	75.0%	85.7%
entropy, log2, 50	92.6%	71.9%	87.6%
entropy, log2, 75	93.8%	76.6%	89.9%
entropy, sqrt, 100	93.4%	75.0%	88.9%
entropy, sqrt, 200	92.9%	73.4%	87.9%
entropy, sqrt, 20	92.8%	70.3%	90.0%
entropy, sqrt, 50	93.1%	73.4%	88.7%
entropy, sqrt, 75	93.2%	75.0%	88.1%
gini, log2, 100	93.7%	75.0%	90.6%
gini, log2, 200	93.1%	72.7%	89.4%
gini, log2, 20	92.0%	72.7%	83.8%
gini, log2, 50	92.9%	75.0%	86.5%
gini, log2, 75	93.2%	75.8%	87.4%
gini, sqrt, 100	**93.4%**	**75.0%**	**88.9%**
gini, sqrt, 200	93.7%	74.2%	91.4%
gini, sqrt, 20	92.6%	71.9%	87.6%
gini, sqrt, 50	93.2%	73.4%	89.5%
gini, sqrt, 75	93.4%	75.8%	88.2%
Train: [3+2] cycloaddition (100), Test: [3+2] cycloaddition (100)			
entropy, log2, 100	99.9%	99.3%	99.8%
entropy, log2, 200	99.9%	99.8%	99.8%
entropy, log2, 20	99.7%	98.2%	99.8%
entropy, log2, 50	99.8%	99.1%	99.8%
entropy, log2, 75	99.9%	100.0%	99.1%
entropy, sqrt, 100	100.0%	100.0%	99.8%
entropy, sqrt, 200	99.9%	100.0%	99.1%
entropy, sqrt, 20	99.8%	99.3%	99.1%
entropy, sqrt, 50	100.0%	100.0%	99.8%
entropy, sqrt, 75	100.0%	100.0%	99.8%
gini, log2, 100	100.0%	100.0%	99.8%
gini, log2, 200	100.0%	100.0%	99.8%

gini, log2, 20	99.6%	97.7%	99.8%
gini, log2, 50	99.9%	100.0%	99.1%
gini, log2, 75	99.9%	99.5%	99.8%
gini, sqrt, 100	**100.0%**	**100.0%**	**99.8%**
gini, sqrt, 200	99.9%	99.8%	99.8%
gini, sqrt, 20	99.9%	99.3%	99.8%
gini, sqrt, 50	99.9%	99.8%	99.8%
gini, sqrt, 75	100.0%	100.0%	99.8%
	Train: RGD1 (100), Test: RGD1 (100)		
entropy, log2, 100	76.8%	65.7%	63.4%
entropy, log2, 200	77.0%	63.6%	64.3%
entropy, log2, 20	77.6%	63.6%	65.4%
entropy, log2, 50	76.3%	61.6%	63.3%
entropy, log2, 75	77.6%	65.3%	64.9%
entropy, sqrt, 100	76.9%	64.1%	64.0%
entropy, sqrt, 200	76.6%	63.6%	63.6%
entropy, sqrt, 20	77.2%	60.5%	65.6%
entropy, sqrt, 50	76.7%	63.4%	63.7%
entropy, sqrt, 75	77.0%	63.0%	64.5%
gini, log2, 100	76.1%	62.0%	62.9%
gini, log2, 200	76.2%	62.2%	63.1%
gini, log2, 20	77.2%	63.9%	64.6%
gini, log2, 50	76.5%	61.5%	63.9%
gini, log2, 75	76.0%	60.7%	63.0%
gini, sqrt, 100	**76.1%**	**61.1%**	**63.1%**
gini, sqrt, 200	76.6%	63.2%	63.7%
gini, sqrt, 20	76.3%	60.3%	63.7%
gini, sqrt, 50	76.8%	63.9%	63.7%
gini, sqrt, 75	77.6%	66.8%	64.5%

The random sampling test procedure for results in Table S9 and S10:

The following procedures were conducted on each dataset individually:

1. 100 sets of reactions with the same reactants were randomly selected for training and testing, respectively.
2. Model training was performed using the 'two-bond +' descriptor composition. The metrics from the evaluation were recorded after the testing.
3. The above steps were repeated five times. The mean and standard deviation of the performance metrics were calculated.

The effect of increasing the test data size:

Table S9: Investigating the effect of increasing the test data size: The mean and standard deviation of the performance metrics were calculated based on the results of random sampling tests.

	% by atoms			% sets of reactants with:		
Test data size	Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
[3+2] cycloaddition						
100	99.7±0.2%	97.2±1.5%	99.2±1.0%	87.4±6.3%	98.0±1.8%	89.2±5.0%
200	98.8±0.5%	92.1±4.3%	96.1±4.1%	64.9±11.7%	84.4±9.1%	74.0±11.2%
500	98.7±0.4%	93.0±2.8%	94.9±2.0%	64.2±6.8%	85.1±9.0%	73.2±9.4%
1000	98.1±0.2%	88.6±1.5%	93.2±2.1%	55.6±3.6%	72.5±3.2%	63.7±2.4%
1500	97.9±0.4%	89.6±2.5%	91.7±5.1%	53.4±7.4%	69.1±8.2%	67.0±4.3%
2000	97.7±0.4%	88.0±2.7%	90.6±3.3%	52.5±4.8%	64.6±6.8%	64.6±3.3%
Diels-Alder						
100	96.1±0.7%	72.2±4.9%	90.0±1.8%	42.8±3.5%	64.0±2.4%	54.4±1.9%
200	96.6±0.3%	75.9±2.8%	90.0±1.4%	41.4±4.6%	66.2±3.1%	56.5±8.6%
500	96.3±0.2%	74.1±1.3%	89.7±0.6%	43.7±1.9%	65.5±1.6%	60.8±5.4%
1000	96.3±0.3%	73.1±2.1%	90.8±1.1%	43.3±4.2%	65.1±3.6%	57.2±5.4%
1500	96.3±0.2%	73.8±0.7%	89.9±1.4%	43.2±2.5%	65.8±2.0%	58.2±1.2%
2000	96.4±0.1%	74.6±1.1%	89.4±0.2%	43.6±2.4%	65.6±0.3%	60.4±2.4%
RGD1						
100	79.2±2.0%	58.1±5.7%	68.0±1.7%	6.6±1.2%	21.6±4.3%	28.8±5.9%
200	78.0±0.5%	55.8±2.2%	66.8±2.3%	4.6±1.4%	18.6±1.4%	23.7±3.3%
500	78.0±0.6%	57.6±2.4%	65.8±1.6%	5.5±0.4%	17.0±1.5%	23.4±1.3%
1000	78.0±0.9%	60.3±3.8%	64.8±1.5%	4.7±0.3%	17.9±1.5%	27.1±2.4%
1500	78.2±0.6%	60.3±2.5%	65.1±2.0%	4.7±1.0%	17.9±1.6%	28.3±1.7%
2000	77.7±0.6%	58.4±1.7%	64.8±1.3%	4.5±0.7%	17.2±1.4%	27.0±3.6%

The effect of changing the activation energy (EA) cut-off:

Table S10: Investigating the effect of changing the activation energy (EA) cut-off: The mean and standard deviation of the performance metrics were calculated based on the results of random sampling tests.

		% by atoms		% sets of reactants with:		
EA cut-off (kcal mol^{-1})	Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
			Green			
37	85.1±0.2%	78.0±0.6%	87.9±0.6%	27.6±1.5%	49.6±1.0%	41.4±2.2%
40	84.3±0.6%	76.8±1.4%	87.3±0.9%	27.4±3.5%	47.2±2.6%	43.4±3.3%
45	84.0±0.6%	76.4±0.8%	86.6±0.8%	22.4±2.5%	46.4±0.5%	39.2±2.6%
50	83.0±1.0%	77.5±2.2%	84.2±2.5%	19.6±2.5%	43.0±3.3%	44.0±4.2%
55	81.9±0.5%	77.2±1.0%	82.4±1.6%	19.6±1.9%	38.4±1.7%	44.4±2.7%
60	78.6±1.6%	70.7±3.3%	82.5±0.7%	15.4±3.2%	31.8±2.4%	33.6±7.4%
			RGD1			
30	79.5±1.8%	50.7±2.3%	70.8±5.0%	7.4±3.3%	24.8±6.8%	23.2±5.7%
35	79.1±0.7%	57.8±3.2%	66.7±1.8%	6.8±2.6%	20.0±1.4%	25.8±4.9%
40	78.5±1.7%	57.0±5.8%	66.3±2.7%	5.6±1.7%	18.6±2.6%	26.2±9.2%
45	77.6±1.0%	55.4±1.9%	67.2±2.3%	4.0±1.4%	15.8±2.7%	22.2±2.3%
50	76.6±2.4%	55.1±6.6%	65.3±4.2%	4.2±1.3%	16.0±6.8%	24.2±6.6%
55	76.8±1.6%	56.2±5.5%	65.7±2.0%	2.4±1.2%	14.4±2.9%	18.2±3.4%

Using the Eyring equation (eq S1), the rate constant (k) and the half-life ($t_{1/2}$, eq S2) at the room temperature, 298 K, were calculated (Table S11). It is assumed that the reaction is simple and monomolecular, A→B. k_B, h, R and T are Boltzmann's constant, Planck's constant, ideal gas constant and temperature respectively.

Table S11. The rate constant (k) and the half-life ($t_{1/2}$) based on the Eyring equation at 298 K

$\Delta G^{\ddagger}$ (kcal mol^{-1})	k (s^{-1})	$t_{1/2}$ (s)
20	1.33E-02	52.01
30	6.17E-10	1.12E+09
40	2.86E-17	2.42E+16
50	1.32E-24	5.23E+23

$$k = \frac{k_B T}{h} e^{\frac{-\Delta G^{\ddagger}}{RT}} \quad \text{(S1)}$$

$$t_{1/2} = \ln(2)\,/k \quad \text{(S2)}$$

The above calculation shows that the reaction will take days when the barrier is above 30 kcal mol^{-1}. Therefore, our chosen cut-off is sensible for leaving out the less kinetically favourable reactions at room temperature.

Table S12. Performance of the models with first-year reactions. 'Train: 78 (105)' implies that 78 sets of reactions with the same reactants, equivalent to 105 reactions, were involved in training. 'Test: 30 (42)' means 30 sets of reactions with the same reactants, equivalent to 42 reactions, were involved in testing. In total, 147 reactions were involved in the training and testing of the model. This corresponds to 100% of the first-year reaction dataset. For each set of data, 'one-bond', 'two-bond' and 'two-bond +' descriptor compositions were considered.

Entry	Descriptors	% by atoms			% sets of reactants with:		
		Accuracy	Precision	Recall	No fault predictions	No more than one fault prediction	All reactive atoms predicted correctly
				Type A			
	First-year reactions: Train: 78 (105), Test: 30 (42) → total 147 reactions, *ie* 100% of the dataset						
1	One-bond	88.59%	73.64%	63.28%	10.0%	56.7%	43.3%
2	Two-bond	91.59%	83.96%	69.53%	26.7%	56.7%	56.7%
3	Two-bonds +	93.09%	87.27%	75.00%	40.0%	70.0%	63.3%

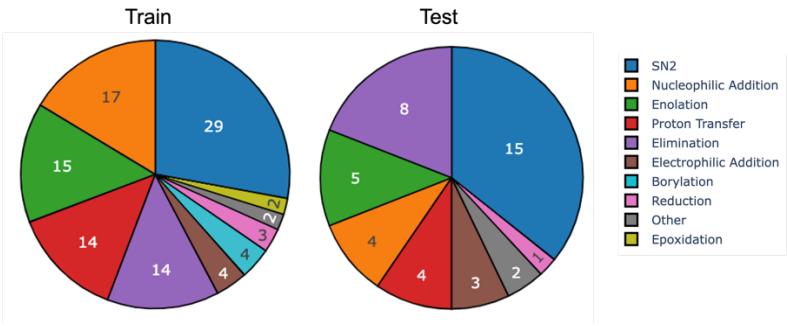


Figure S8. Composition of the training and testing first-year dataset in model evaluation for producing results in Table S12

4. Performance on non-elementary reaction examples

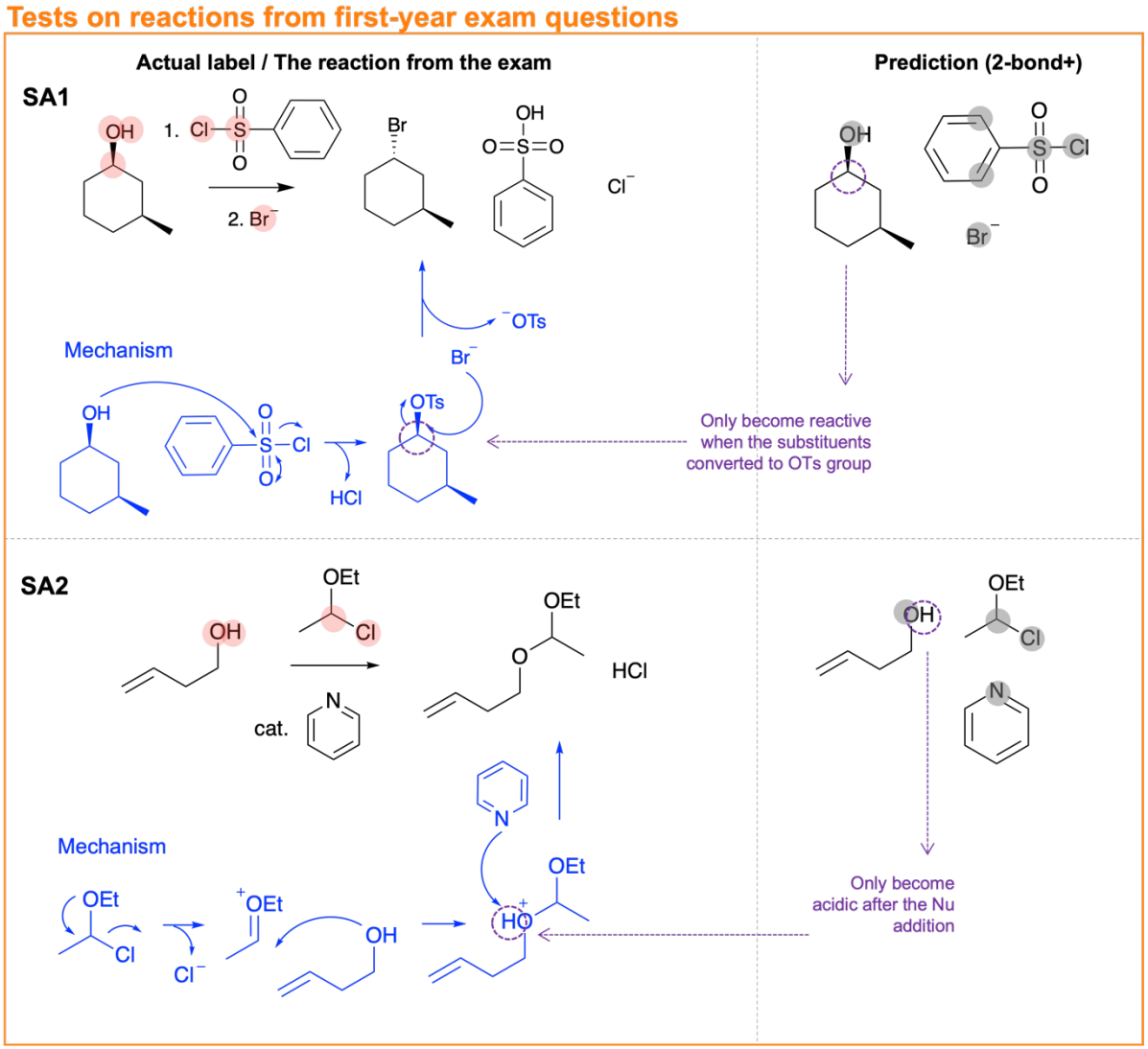


Figure S9. Additional examples: Predictions from the 'two-bond +' model trained on all first-year reactions using reactions from first-year exams at the University of Cambridge. These reactions are non-elementary.

It is assumed that reactions within the first-year reaction data set are elementary. This categorisation is based on chemical intuition rather than computations. Thus, we expect a few non-elementary reactions within the dataset and a better performance of the models on the elementary reactions compared to non-elementary reactions. In the above cases, the model trained on the first-year reaction dataset cannot pick out atoms that become more reactive in the intermediate than in the reactant.

5. Reference

1 T. Stuyver, K. Jorner and C. W. Coley, *Sci. Data*, 2023, **10**, 66.
2 S. Li, X. Wang, Y. Wu, H. Duan and L. Tang, *RSC Adv.*, 2022, **12**, 33801–33807.
3 Q. Zhao, S. M. Vaddadi, M. Woulfe, L. A. Ogunfowora, S. S. Garimella, O. Isayev and B. M. Savoie, *Sci. Data*, 2023, **10**, 145.
4 C. A. Grambow, L. Pattanaik and W. H. Green, *Sci. Data*, 2020, **7**, 137.
5 P. Schwaller, B. Hoover, J.-L. L. Reymond, H. Strobelt and T. Laino, *Sci. Adv.*, 2023, **7**, eabe4166.
6 S. Lee, K. Ermanis and J. M. Goodman, *Chem. Sci.*, 2022, **13**, 7204–7214.
7 F. Berenger and K. Tsuda, *J. Cheminform.*, 2021, **13**, 88.