Supporting Information

Multitask Prediction of Site Selectivity in Aromatic C-H Functionalization Reactions

Thomas J. Struble,[a‡] Connor W. Coley,[a‡] and Klavs F. Jensen*[a]
E-mail: kfjensen@mit.edu

[a] *Department of Chemical Engineering*
[‡] *Equal contribution*

Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02139.

S1 (Additional) Materials and methods

S1.1 Details of data set

A breakdown of the number of examples for each task in the whole dataset is provided below. The tasks are loosely categorized based on the "type" of reaction but this does not imply that the reactions with the same description share an underlying mechanism. A representative example is in the borylation category there are reactions using an organolithium reagent and examples using palladium coupling.

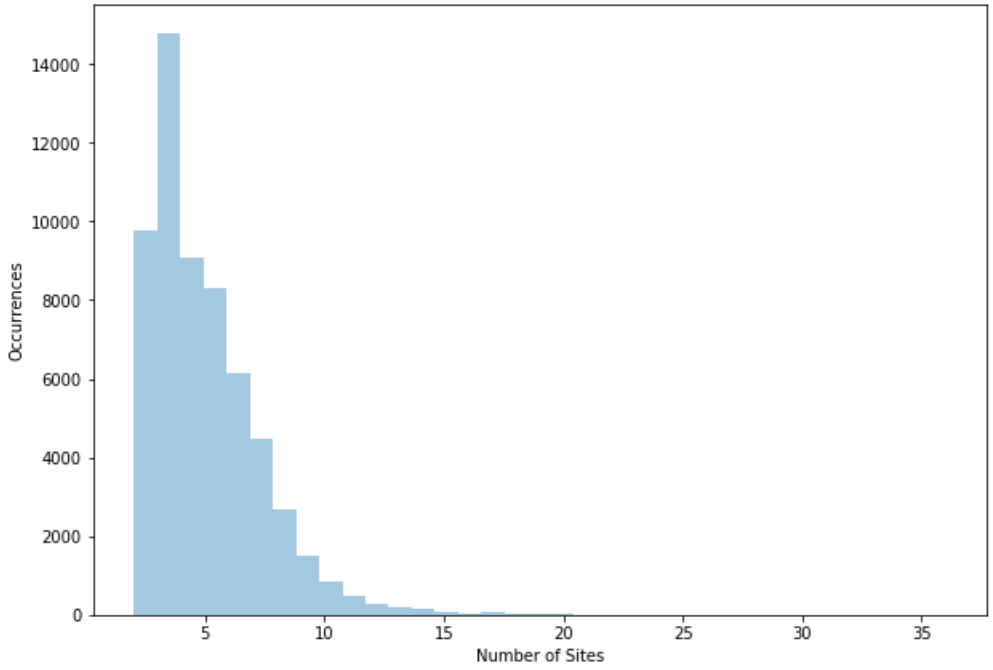
Table S1: Description of all C-H activation tasks used in this study. The number of examples reflects the total number, divided in an 80:10:10 split for training:validation:testing.

Num. Examples	Description	SMILES of other reactant
16152	bromation	BrBr
5791	nitration	[N+](=O)(O)[O-]
4569	formylation	CN(C)C=O
2895	chlorination	ClCl
1311	Friedel-Crafts acylation	CC(=O)Cl
983	carbonylation	C=O
833	carboxylation	O=C=O
768	Friedel-Crafts acylation	CC(=O)OC(C)=O
743	olefination	C=CC(=O)OCC
743	arylation	Ic1ccccc1
690	borylation	CC1(C)OB(B2OC(C)(C)C(C)(C)O2)OC1(C)C
688	Friedel-Crafts acylation	O=C(Cl)c1ccccc1
659	olefination	C=CC(=O)OCCCC
595	1,2 addition	O=Cc1ccccc1
573	1,4 addition	C=CC(=O)OC
558	methylation	CI
547	silylation	C[Si](C)(C)Cl
513	formylation	COC(Cl)Cl
487	1,2 addition	C=Cc1ccccc1
470	1,4 addition	O=[N+]([O-])C=Cc1ccccc1
457	acylation	CC(=O)O
454	sulfonyl azide addition	Cc1ccc(S(=O)(=O)N=[N+]=[N-])cc1
415	arylation	COc1ccc(I)cc1
400	arylation	Brc1ccccc1
394	arylation	OB(O)c1ccccc1
360	arylation	Cc1ccc(I)cc1
345	1,4 addition	C=CC(C)=O
333	formylation	C1N2CN3CN1CN(C2)C3

326	trifluoroacetylation	O=C(OC(=O)C(F)(F)F)C(F)(F)F
315	methylation	CO
285	acetylation	O=C1CCC(=O)O1
284	acetylation	O=C(Cl)CCl
274	sulfonylation	Cc1ccc(S(=O)(=O)Cl)cc1
268	thiolation	CSSC
262	arylation	N#Cc1ccc(Br)cc1
258	acetylation	O=C(O)C(=O)c1ccccc1
256	arylation	COc1ccc(Br)cc1
255	arylation	Cc1ccc(Br)cc1
240	borylation	CC1(C)OBOC1(C)C
236	alkyne coupling	CC(C)[Si](C#CBr)(C(C)C)C(C)C
232	arylation	Nc1ccc([N+](=O)[O-])cc1
227	stannylation	CCCC[Sn](Cl)(CCCC)CCCC
223	acylation	O=C(Cl)C(=O)Cl
220	borylation	CC(C)OB1OC(C)(C)C(C)(C)O1
216	1,2 addition	C#Cc1ccccc1
210	1,2 addition	CCOC(=O)C(=O)C(F)(F)F
209	acylation	O=C(O)c1ccccc1
207	olefination	C=CC(=O)OC(C)(C)C
204	acylation	CCOC(=O)C(=O)Cl
201	diazotization	Nc1ccccc1
199	1,2 addition	O=C(C=Cc1ccccc1)c1ccccc1
196	arylation	Brc1cccnc1
191	amination	C1COCCN1
191	arylation	Clc1ccc(I)cc1
186	1,2 addition	C(#Cc1ccccc1)c1ccccc1
182	phosphine synthesis	ClP(c1ccccc1)c1ccccc1
178	1,2 addition	CCOC(=O)C=O
178	arylation	O=Cc1ccc(Br)cc1
172	arylation	FC(F)(F)c1ccc(Br)cc1
170	acylation	CCC(=O)Cl
167	1,2 addition	COc1ccc(C=O)cc1
167	trichloroacylation	O=C(Cl)C(Cl)(Cl)Cl
164	alkylation	OCc1ccccc1
164	alkylation	CCOC(=O)C(F)(F)Br
163	arylation	O=[N+]([O-])c1ccc(Br)cc1
159	alkylation	C=CCBr
159	1,2 addition	CC=O
158	alkylation	CN(C)CN(C)C
156	arylation	FC(F)(F)c1ccc(I)cc1
155	methylation	CS(C)=O
155	amidation	O=c1onc(-c2ccccc2)o1
155	arylation	CC(=O)c1ccc(Br)cc1
151	acylation	O=Cc1ccc(Cl)cc1
150	methylation	Sc1ccccc1
149	arylation	CCOC(=O)c1ccc(I)cc1
147	phosphonate synthesis	CCO[PH](=O)OCC
147	allylation	C=CCOC(C)=O
138	alkylation	O=S(=O)(CCl)c1ccccc1
138	arylation	N#Cc1ccccc1Br
135	acylation	O=C1OC(=O)c2ccccc21
134	acylation (oxidative)	Cc1ccccc1
133	arylation	Cc1ccc(Cl)cc1
132	alkylation (oxidative)	C1COCCO1
131	arylation	N#Cc1ccc(I)cc1
128	acylation	O=C(Cl)CCCl

126	alkylation	OC(C=Cc1ccccc1)c1ccccc1
126	silylation	CC[SiH](CC)CC
125	arylation	Fc1ccc(Br)cc1
125	acylation	COc1ccc(C(=O)Cl)cc1
124	borylation	COB(OC)OC
121	1,2 addition	CC(=O)c1ccccc1
120	1,2 addition	O=C(C(F)(F)F)C(F)(F)F
120	arylation	Cc1ccc(B(O)O)cc1
118	alkylation	OC(c1ccccc1)c1ccccc1
117	arylation	Clc1ccc(Br)cc1
116	acylation	O=C(Cl)c1ccc(Cl)cc1
115	alkylation	COC(=O)C(=[N+]=[N-])C(=O)OC
114	sulfonylation	O=S(=O)(Cl)c1ccccc1
113	alkylation	OC(C#Cc1ccccc1)c1ccccc1
112	amination	O=S(=O)(c1ccccc1)N(F)S(=O)(=O)c1ccccc1
111	1,2 addition	O=C(c1ccccc1)c1ccccc1
111	phosphine oxide synthesis	O=[PH](c1ccccc1)c1ccccc1
110	amination	Cc1ccc(S(=O)(=O)NN)cc1
110	1,2 addition	COC(=O)C(=O)C(F)(F)F
109	amination	CCOC(=O)N=NC(=O)OCC
109	acylation	Cc1ccc(C(=O)Cl)cc1
108	alkylation	c1ccc(C2CO2)cc1
108	borylation	CC(C)OB(OC(C)C)OC(C)C
107	acylation	CC(C)(C)C(=O)Cl
107	stannylation	C[Sn](C)(C)Cl
107	1,2 addition	CCCC#CCCC
106	1,4 addition	O=C1C=CCCC1
105	1,4 addition	O=C1C=CCC1
105	isocyanate addition	O=C=Nc1ccccc1
104	arylation	O=[N+]([O-])c1ccc(I)cc1
103	1,2 addition	O=C1CCCCC1
103	alkylation	COC(=O)C(=[N+]=[N-])c1ccccc1
103	phosphonate synthesis	CC(C)O[PH](=O)OC(C)C
103	arylation	Clc1ccccc1
101	acylation	Cc1ccc(C=O)cc1
101	silylation	C[SiH](O[Si](C)(C)C)O[Si](C)(C)C
100	alkylation	OC12CC3CC(CC(C3)C1)C2
100	acylation	O=C(Cl)Cc1ccccc1

Figure S1: Distribution of dataset (train/valid/test) by number of unique sites in the reactant



S1.2 Details of model architecture

The description of the WLN model is presented here with minimal modification from Coley *et al.*[1].

S1.2.1 Notation

Symbol	Meaning
u, v	atoms
$N(v)$	Set of atoms adjacent to v
$\tau(\cdot)$	ReLU activation function
$\sigma(\cdot)$	Sigmoid function
U, V, W, M, P, Q	learned matrices in WLN

S1.2.2 Weisfeiler-Lehman Network (WLN)

Weisfeiler-Lehman Network[2] is a type of graph convolutional network derived from Weisfeiler-Lehman (WL) graph kernel[3]. The architecture is designed to embed the computations inherent in WL graph kernel to learn isomorphism invariant representation of atoms. The atom representation is computed by iteratively augmenting the representation of adjacent atoms. Specifically, each atom v is initialized with a feature vector f_v indicating its atomic number, formal charge, degree of connectivity, explicit and implicit valence, and aromaticity. Each bond (u, v) is associated with a feature vector f_{uv} indicating its bond order and ring status. In each iteration, we updated atom representations as follows:

$$f_v^l = \tau\left(U_1 f_v^{l-1} + U_2 \sum_{u \in N(v)} \tau(V_1 f_u^{l-1} + V_2 f_{uv})\right) \qquad (1 \le l \le L)$$

where f_v^l is the atom representation at the lth iteration, initialized with $f_v^0 = f_v$ atom features. U_1, U_2, V_1, V_2 are model parameters to be learned, shared across all L iterations. The final local atom representations are computed as

$$c_v = \sum_{u \in N(v)} W_1 f_u^L \odot W_2 f_{uv} \odot W_3 f_v^L$$

We refer the reader to 2 for more details about the mathematical intuition and justification of the WLN.

S1.2.3 Attention Mechanism

The atom embedding c_v only record local chemical environment, namely atoms and bonds accessible within L steps from atom v. Even if L were very large, c_v could not encode any information about other reactant molecules, as information cannot be propagated between two reactant molecules that are disconnected. We argue that it is important to enable information to flow between distant or disconnected atoms. For example, the reaction center may be influenced by certain reagents that are disconnected from reactant molecules. In this case, it is necessary for atom representation c_v to encode such distal chemical effects. Therefore, we propose to enhance the model in previous section with an attention mechanism. [4]

Specifically, let α_{vz} be the attention score of atom v upon atom z. The "global" atom representation $\tilde{c}_v$ of atom v is calculated as the weighted sum of all reactant atoms where the weight comes from the attention module:

$$\alpha_{vz} = \sigma(u^T \tau(P_a c_v + P_a c_z + P_b b_{vz}))$$

$$\tilde{c}_v = \sum_v \alpha_{vz} c_z$$

The attention score is computed based on "local" atom representations c_v from WLN. σ is the sigmoid activation function.

S1.2.4 Reaction Site Prediction

The WLN is trained to predict the likelihood that a specific atom will be the favored site in a specific C-H activation reaction. We denote this likelihood as $p_{t,v}$, where t is the prediction task and v is the atom. The likelihoods are not normalized within a molecule to sum to one, but instead are computed using an elementwise sigmoid action σ to produce a vector p_v across prediction tasks.

$$p_v = \sigma\big(Q\ \tau(M_a \tilde{c}_v + P_a c_v)\big)$$

The above neural network is jointly optimized with the WLN to minimize the sigmoid cross entropy loss for each reaction example

$$-\sum_t \sum_v y_{t,v} \log p_{t,v} + (1 - y_{t,v}) \log(1 - p_{t,v})$$

where $y_{t,v} = 1$ iff v is the atom undergoing C-H activation for task t.

S1.3 Inclusion of reagents

Including the reagents as part of the input was tested to see if the accuracy of the model could be improved. The data was further filtered by removing any atom mapping from reagents, and confirming all of the recorded reagents can be parsed by RDKit. The benefit would be that better accuracy could be achieved but with the trade-off of the end user having to provide reagents at prediction time. The model performed marginally better with the reagents included but still do not capture drastic changes in selectivity based on very specific conditions. However, care should be taken to compare these results directly to the model that does not include reagents. The data set that includes reagents has multiple reactions that have the same outcome but use different reagents and thus is slightly different than the data set used in the multitask model without reagents.

Table S2: Results for inclusion of reagents in training

Model	Validation Set[a] (%)	Test Set[a] (%)
With Reagents	89/94	87/92

[a] Reported as top 1 accuracy / mean reciprocal rank

S2 RegioSQM comparison

RegioSQM predictions include all sites that are within a threshold of the lowest energy carbocation conformer (in this case 1 kcal/mol) which allows for multiple predictions in each molecule. The WLN methodology accuracy is based on the top 1 atom score which cannot be directly compared. An analysis is performed where the accuracy is based on how

many sites that that RegioSQM predicts. For example if RegioSQM predicts 3 sites that are all within 1 kcal/mol of the lowest energy conformer, then the accuracy for the WLN is relaxed to if the top 3 predictions include the correct site. The results are grouped into two categories 1) direct comparison of the top 1 predictions, filtered to include only examples where RegioSQM predicts one site, and 2) comparison of the top 2 or 3 sites for both methodologies when RegioSQM predicts multiple sites. Also included in Table S3 in the column 2 or 3 sites, is the top 1 accuracy of that subset for the WLN. Interstingly, the top 1 accuracy is not much lower than when RegioSQM has 2 or 3 sites it had chosen.

Table S3: Comparison to RegioSQM[5] on a random subset of 494 bromination reactions from our test set. Performance is divided into two columns according to the number of sites RegioSQM believes to be equally likely.

	1 site[a] (%)	2 or 3 sites[a] (%)	Time (12 CPU's)
RegioSQM[5]	86.7	74.2	>10 days
WLN	87.9	71.0[b]/84.7[c]	6.3s

[a]Number of sites predicted by RegioSQM, [b]Reported as top 1 accuracy, [c]Reported as top 2 or 3 accuracy

S3 (Additional) Results

Initial hyperparameter search is shown in **Table S4**. Intermediate values between entries 1 and 2 (hidden size of 300, learning rate of 0.003, and depth of 5) were chosen for further comparison between different model architectures and is outlined in **Table S5**. Performance is also broken down by number of available symmetric reaction sites in the molecule (the data distribution broken down by number of sites is shown in **Figures S2** and **S3**. Batch size did not impact accuracy, so for the hyperparameter search, the batch size was set to 20 and the data was randomly shuffled at the beginning of each epoch.

Table S4: Additional hyperparameter optmization

entry	WLN depth	hidden	learning rate	multitask	validation accuracy (%)
1	5	512	0.00100	True	86.6
2	5	256	0.00050	True	86.2
3	4	512	0.00050	True	85.9
4	3	512	0.00050	True	85.7
5	3	512	0.00100	True	85.6
6	4	256	0.00050	True	85.5
7	5	256	0.00100	True	85.5
8	4	512	0.00100	True	85.4
9	3	256	0.00100	True	85.3
10	5	512	0.00009	True	85.3
11	4	256	0.00100	True	85.3
12	5	128	0.00100	True	85.1
13	4	128	0.00050	True	85.1
14	5	512	0.00050	True	84.9
15	3	256	0.00050	True	84.9
16	5	128	0.00050	True	84.8
17	4	64	0.00100	True	83.8
18	4	128	0.00100	True	83.8
19	4	512	0.00009	True	83.6
20	3	128	0.00100	True	83.4
21	5	64	0.00100	True	83.1
22	3	128	0.00050	True	82.9
23	5	64	0.00050	True	81.6
24	3	512	0.00009	True	81.5
25	5	256	0.00009	True	81.3
26	4	64	0.00050	True	81.2
27	3	64	0.00100	True	81.2
28	4	256	0.00009	True	80.2
29	4	256	0.01000	True	80.1
30	3	512	0.01000	True	79.9
31	4	512	0.01000	True	79.9
32	5	64	0.01000	True	79.7
33	5	512	0.01000	True	79.3
34	5	256	0.01000	True	78.8
35	3	64	0.00050	True	78.8
36	3	256	0.01000	True	78.1
37	5	128	0.00009	True	78.0
38	3	64	0.01000	True	77.3
39	5	128	0.01000	True	77.1
40	4	64	0.01000	True	75.9
41	4	128	0.00009	True	75.9
42	4	128	0.01000	True	75.4
43	3	256	0.00009	True	75.2
44	3	128	0.01000	True	72.6
45	5	64	0.00009	True	71.8
46	3	128	0.00009	True	71.3
47	4	64	0.00009	True	70.0
48	3	64	0.00009	True	66.3

Table S5: Additional results

Model[b]	Baseline	Hidden size	Added Features	Validation accuracy[b] (%)
Single-task	yes	300	Yes	46.7
Single-task	no	100	no	84.4
Single-task	no	300	no	86.4
Single-task	no	100	yes	83.3
Single-task	no	300	yes	84.6
Multitask	yes	300	no	21.3
Multitask	yes	300	yes	49.0
Multitask	**no**	**300**	**no**	**87.0**
Multitask	no	300	yes	87.6

[a]A depth of 5 was used for the WLN with a lr of 0.003. [b]Reported as top 1 accuracy.

Figure S2: Performance of the validation set by number of sites

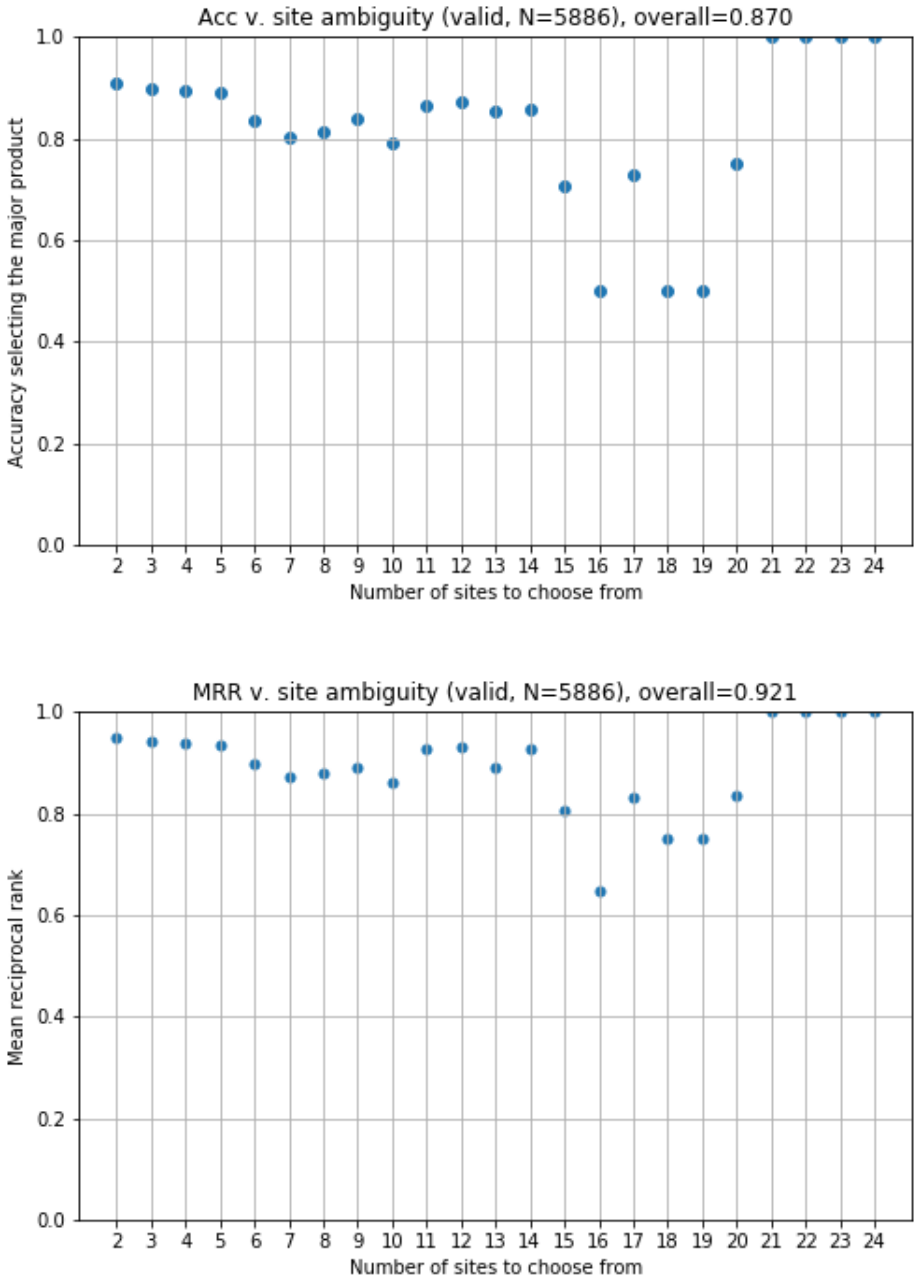
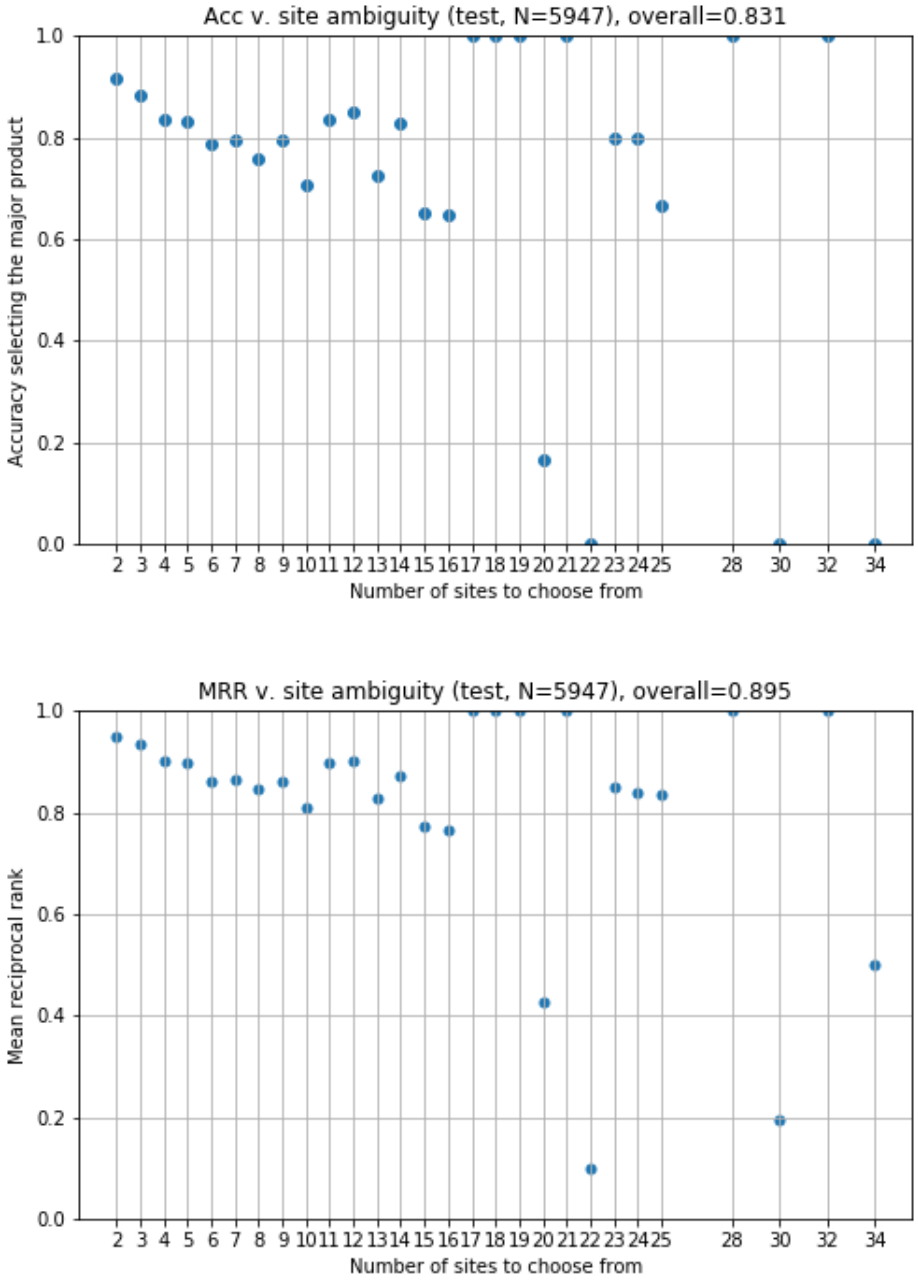


Figure S3: Performance of the test set by number of sites



S3.1 Examples of predictions

One possible application of the multitask model is in late stage functionalization of aromatics. The model would allow a chemist to view some reactions that give a high probability to be site selective. However, these scores are not indicative of reaction yield, only the probability for that site to be functionalized. The synthesis workflow would still require chemists to decide whether protections/deprotections would be needed to avoid functional group interactions with catalysts or reagents. The first example in **Figure S4** shows that it **S1** would be possible to access two different sites with various reactions. Also shown are examples that would not give selectivity or have low probaliity of accessing any site. If a chemist wants another site on **S1** to be functionalized then they could go one step back in the synthetic sequence and run selectivity predictions. **Figure S5** shows one retrosynthetic suggestion that breaks the molecule into **S2** and another site on the molecule could be selectively functionalized. The final example in **Figure S6** demonstrates again that on **S3** there are some reactions that give high probability for the highlighted site and there are often many that give a low probability which would likely not be routes to be executed.

Figure S4: Example 1

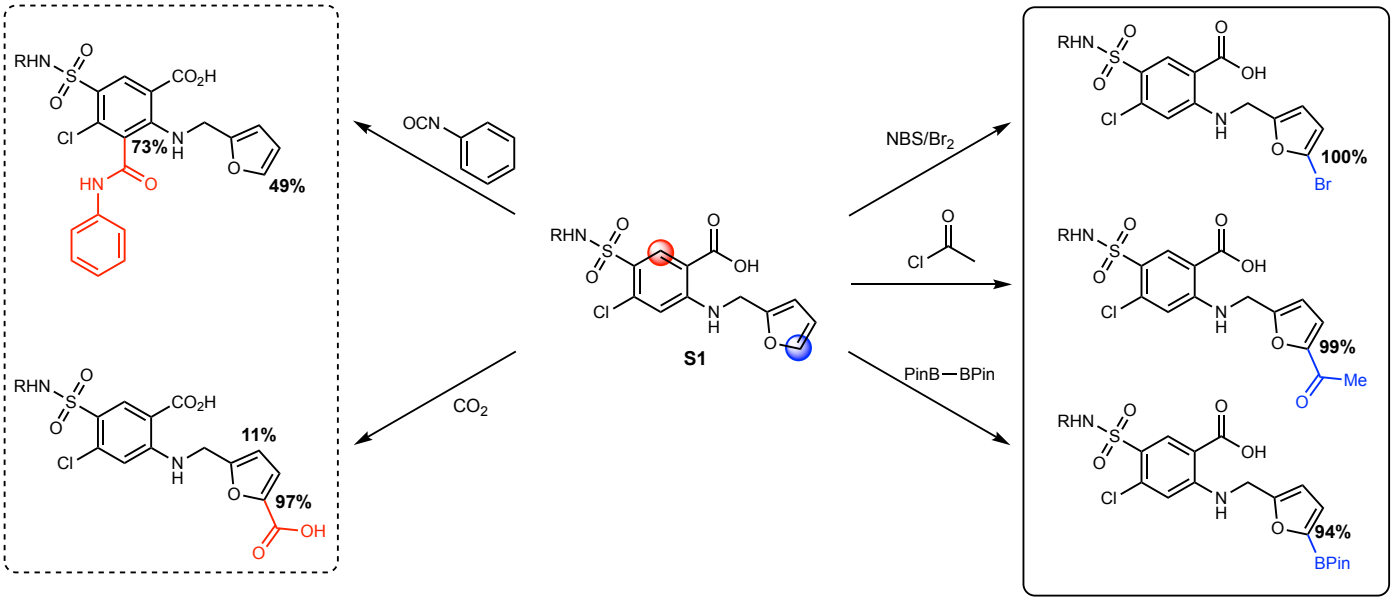


Figure S5: Example after one step retrosynthesis

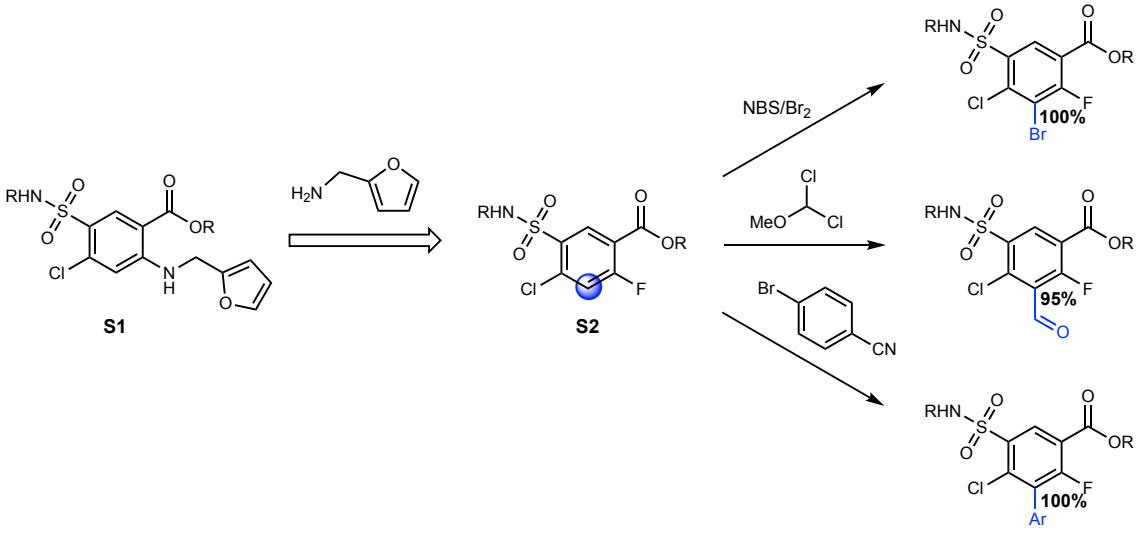
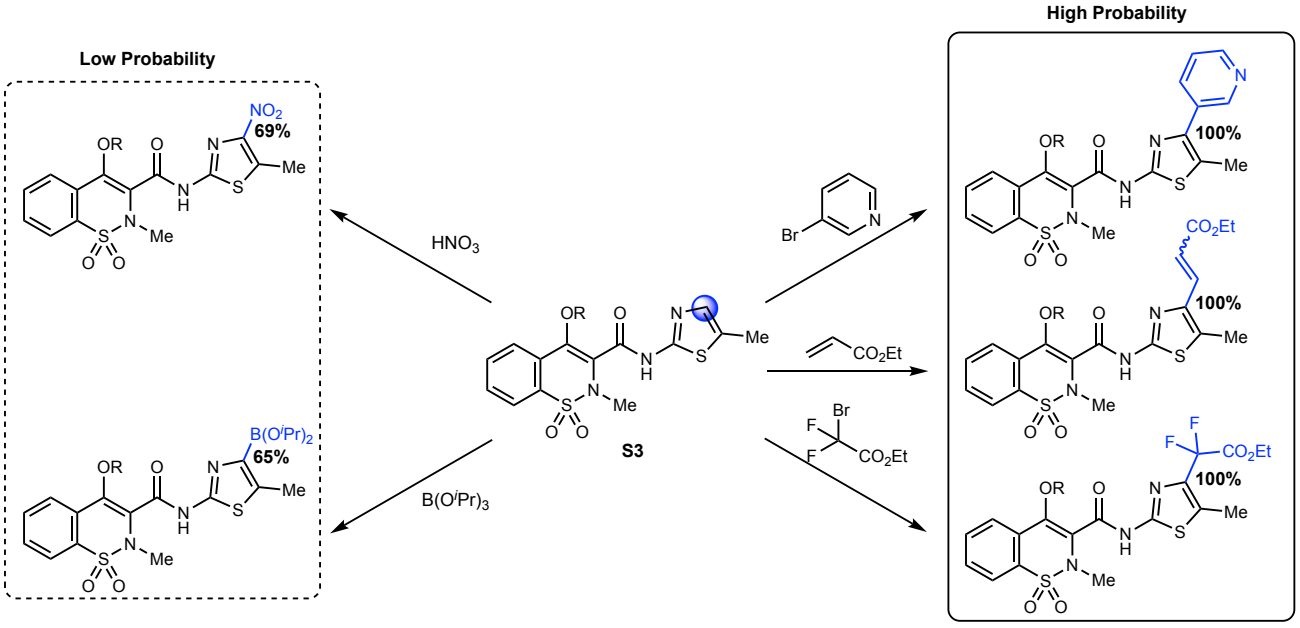


Figure S6: Example 2



S3.2 Examples of failed predictions

Below are examples of failed predictions. There are 9 tasks that have accuracy below 50%. **Table S6** shows the tasks that have poor test accuracy. These low accuracies are generally attributed to the time-split validation we use for the dataset. For example, for the task CCOC(=O)C(F)(F)Br which has 0% accuracy on the test set, all 17 test examples are from a substrate scope where a new catalyst/ligand system was developed to alter selectivity. Examples of failed predictions are grouped by their task and drawn below.

Table S6: Tasks with low accuracy

Task	N_{test}	top-1 accuracy (%)
CCOC(=O)C(F)(F)Br	17	0.0
O=C=Nc1ccccc1	11	9.1
CC(=O)c1ccccc1	13	15.3
C#Cc1ccccc1	22	36.3
OB(O)c1ccccc1	40	37.5
CC(=O)O	46	43.4
C=Cc1ccccc1	49	44.9
C=CC(C)=O	35	45.7
Cc1ccc(B(O)O)cc1	12	50.0

Figure S7: Failed predictions for task CCOC(=O)C(F)(F)Br. Reaxys ID's A) 44846829 B) 44846838 C) 44846844 D) 44846850 E) 44846862

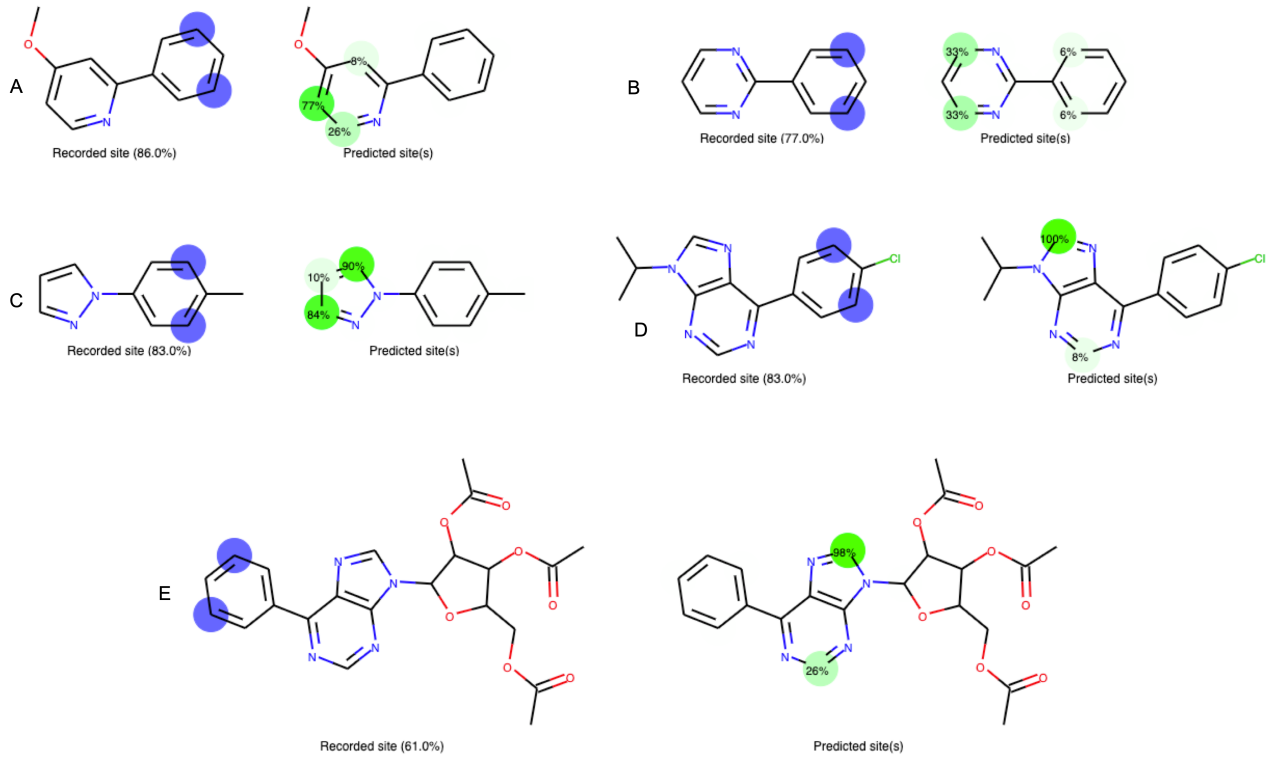


Figure S8: Failed predictions for task O=C=Nc1ccccc1. Reaxys ID's A) 44164716 B) 44164728 C) 44164741

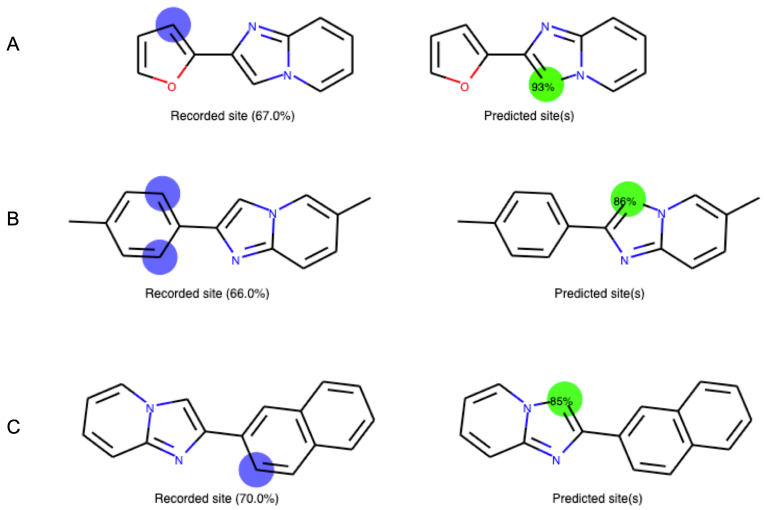


Figure S9: Failed predictions for task CC(=O)c1ccccc1. Reaxys ID's A) 42571625 B) 43419010 C) 43419015 D) 43419013

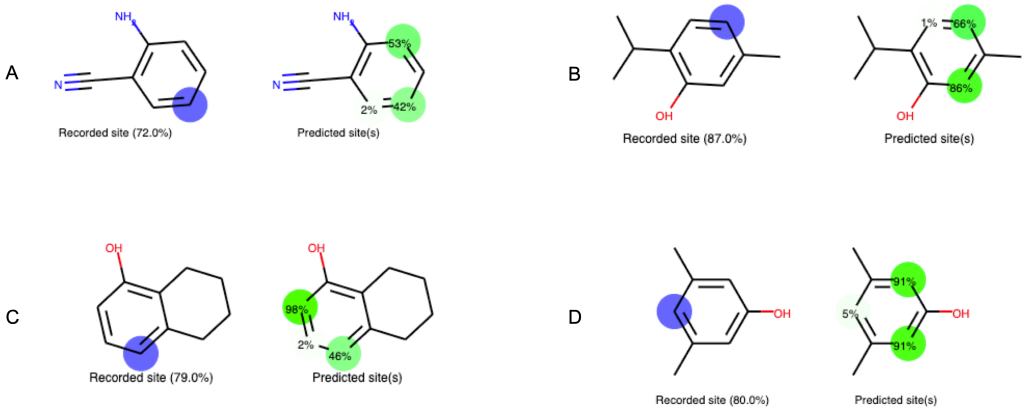


Figure S10: Failed predictions for task C#Cc1ccccc1. Reaxys ID's A) 44761859 B) 43805454 C) 43420046 D) 42092979

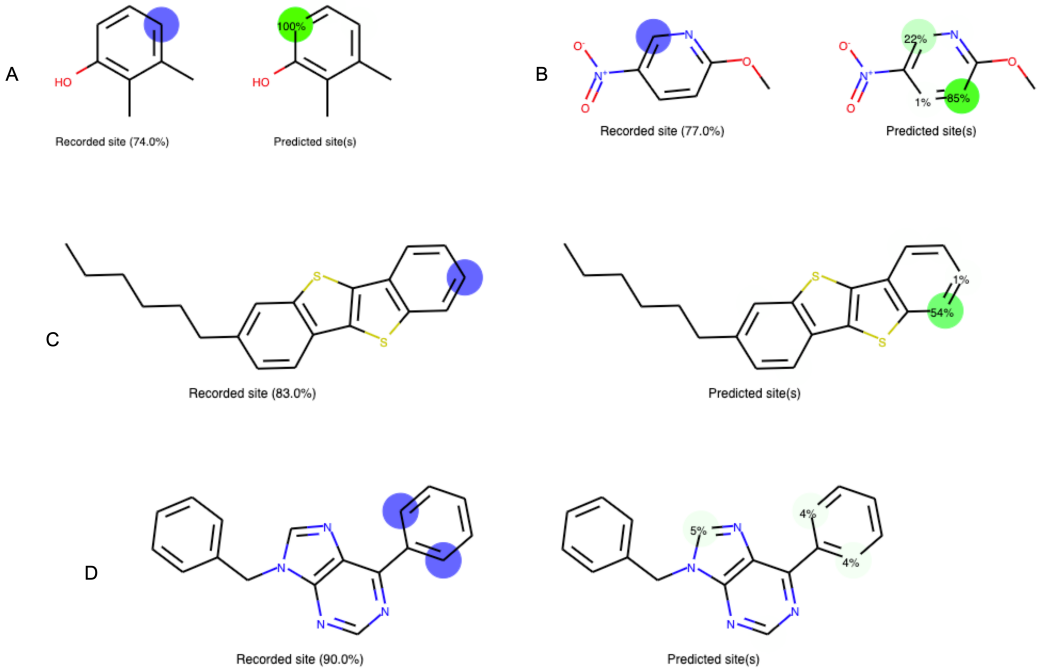


Figure S11: Failed predictions for task OB(O)c1ccccc1. Reaxys ID's A) 43473813 B) 43905584 C) 44346055

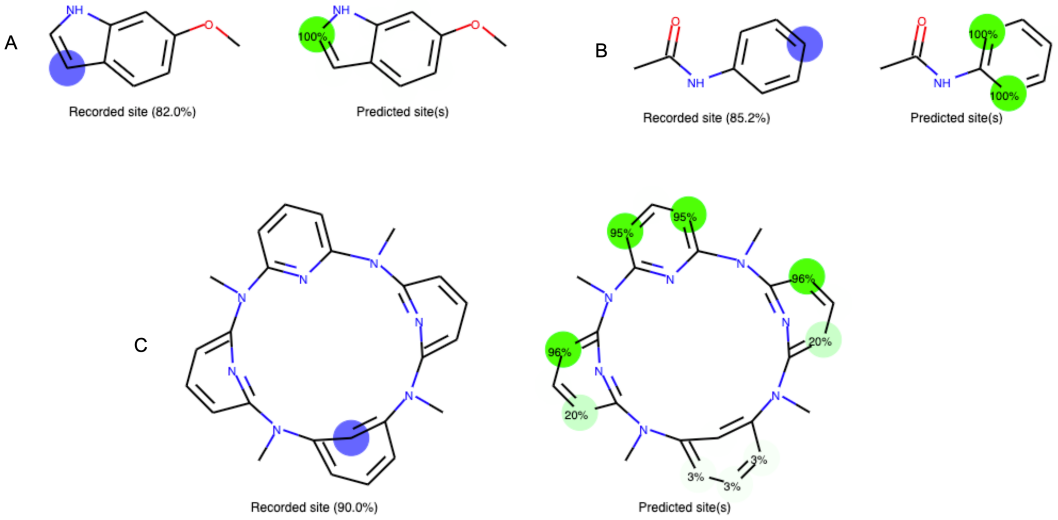


Figure S12: Failed predictions for task CC(=O)O. Reaxys ID's A) 44326647 B) 44447392 C) 44461717

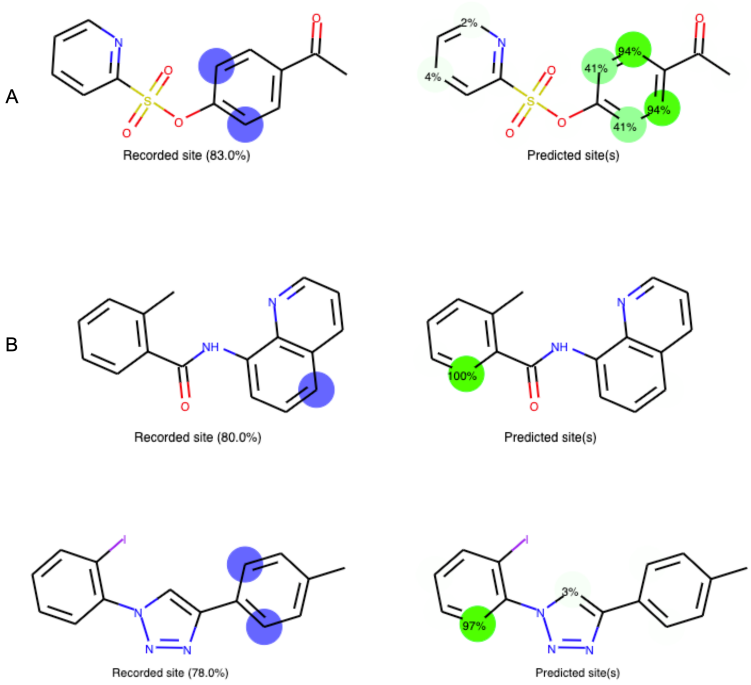


Figure S13: Failed predictions for task C=Cc1ccccc1. Reaxys ID's A) 42799511 B) 43106440 C) 43644703 D) 44151514

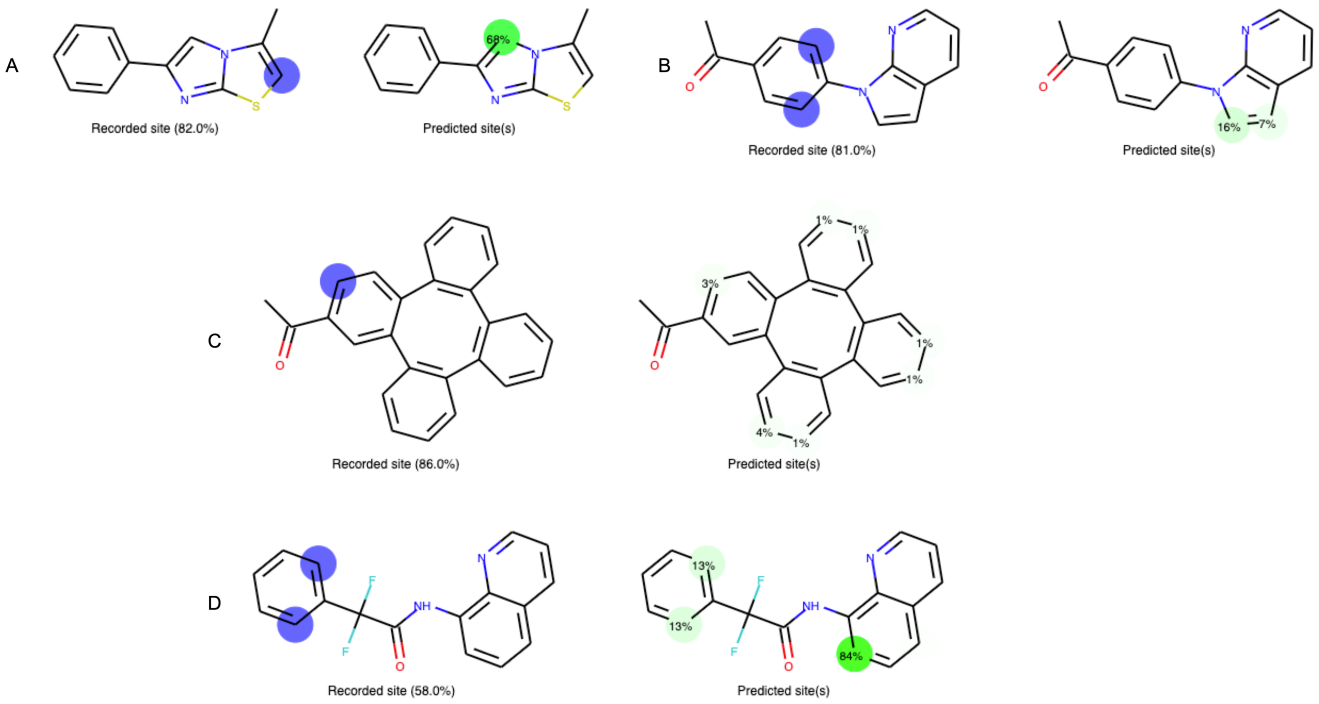


Figure S14: Failed predictions for task C=CC(C)=O. Reaxys ID's A) 35555246 B) 37544015 C) 40982653

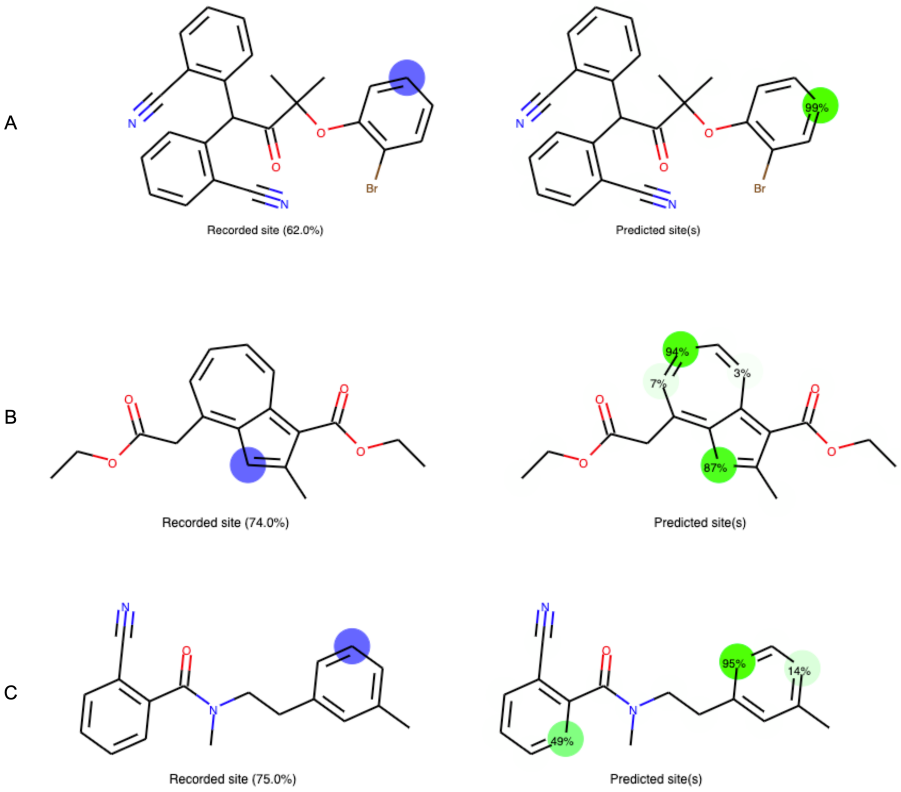
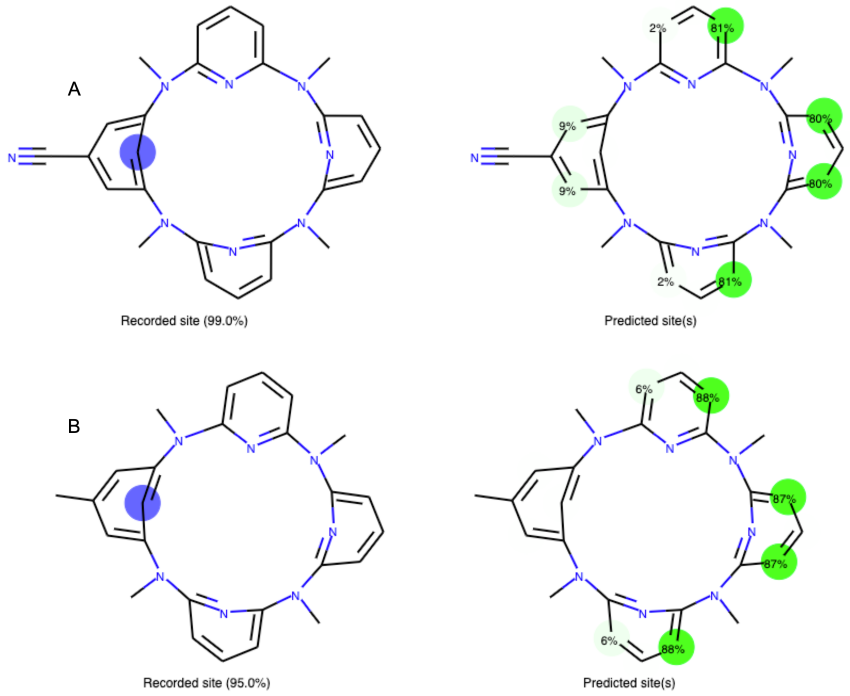


Figure S15: Failed predictions for task Cc1ccc(B(O)O)cc1. Reaxys ID's A) 44346052 B) 44346091



References

[1] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, *Chemical Science*, 2019, **10**, 370–37.

[2] T. Lei, W. Jin, R. Barzilay and T. S. Jaakkola, ICML, 2017.

[3] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn and K. M. Borgwardt, *Journal of Machine Learning Research*, 2011, **12**, year.

[4] D. Bahdanau, K. Cho and Y. Bengio, *CoRR*, 2014, **abs/1409.0473**, year.

[5] J. C. Kromann, J. H. Jensen, M. Kruszyk, M. Jessing and M. Jørgensen, *Chemical Science*, 2017, **9**, 660–665.