doi.org/10.26434/chemrxiv.9735599.v1

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

Thomas J. Struble, Connor W. Coley, Klavs F. Jensen

Submitted date: 27/08/2019 • Posted date: 28/08/2019

Citation information: Struble, Thomas J.; Coley, Connor W.; Jensen, Klavs F. (2019): Multitask Prediction of Site Selectivity in Aromatic C-H Functionalization Reactions. ChemRxiv. Preprint.

Aromatic C-H functionalization reactions are an important part of the synthetic chemistry toolbox. Accurate prediction of site selectivity can be crucial for prioritizing target compounds and synthetic routes in both drug discovery and process chemistry. However, selectivity may be highly dependent on subtle electronic and steric features of the substrate. We report a generalizable approach to prediction of site selectivity that is accomplished using a graph-convolutional neural network for the multitask prediction of 123 C-H functionalization tasks. In an 80/10/10 training/validation/testing pseudo-time split of about 58,000 aromatic C-H functionalization reactions from the Reaxys database, the model achieves a mean reciprocal rank of 92%. Once trained, inference requires approximately 200 ms per compound to provide quantitative likelihood scores for each task. This approach and model allow a chemist to quickly determine which C-H functionalization reactions-if any-might proceed with high selectivity.

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

Thomas J. Struble,[†,‡] Connor W. Coley,[†,‡] and Klavs F. Jensen[*,†]

†*Department of Chemical Engineering, Massachusetts Institute of Technology*

‡*Equal contribution*

E-mail: kfjensen@mit.edu

Abstract

Aromatic C-H functionalization reactions are an important part of the synthetic chemistry toolbox. Accurate prediction of site selectivity can be crucial for prioritizing target compounds and synthetic routes in both drug discovery and process chemistry. However, selectivity may be highly dependent on subtle electronic and steric features of the substrate. We report a generalizable approach to prediction of site selectivity that is accomplished using a graph-convolutional neural network for the multitask prediction of 123 C-H functionalization tasks. In an 80/10/10 training/validation/testing pseudo-time split of about 58,000 aromatic C-H functionalization reactions from the Reaxys database, the model achieves a mean reciprocal rank of 92%. Once trained, inference requires approximately 200 ms per compound to provide quantitative likelihood scores for each task. This approach and model allow a chemist to quickly determine which C-H functionalization reactions–if any–might proceed with high selectivity.

Introduction

Aromatic and heterocyclic ring systems are ubiquitous in approved drugs and natural products. Because a substitution at any position of the aromatic ring can greatly alter the biological activity profile of a compound, it is common to investigate many substitution patterns during drug development to build a structure-activity relationship. The compounds that are easiest to include in such an analogue library are those that can be accessed from a common late-stage intermediate; often, an analogue of interest might require carrying that substitution through the whole synthetic process. In an ideal scenario, we would be able to selectively functionalize a late-stage intermediate (or other member of the compound library) at a specific position of interest.

Highly site selective reactions are more broadly useful for planning, prioritizing, and executing efficient routes in synthetic chemistry. Achieving high selectivity requires the use of conditions or substrates that can differentiate multiple similar reactive sites within the same molecule. Furthermore, molecules or routes that include steps with unclear site selectivity may be discarded by chemists due to the lack of a robust method to access the target. The concept of site selectivity has been highlighted recently with the development and application of methods for late stage C-H functionalization of drug targets[1–5] which has spurred new methods research[6–9] and HTS campaigns.[10,11] Here, we focus on the subset of functionalization reactions that target aromatic C-H motifs.

Historically, prediction of site selectivity for aromatic C-H bonds has focused on electrophilic aromatic substitution (EAS) reactions. EAS is acatalytic and proceeds through a relatively simple mechanism whereby selectivity is determined primarily by the stability of the intermediate cation, allowing for relatively straightforward predictivity.[12] Early methods include using Hammett and Taft Parameters to approximate the nucleophilicity of aromatic rings.[13] Later, models using estimated ^{1}H and ^{13}C NMR shifts (using ChemDraw's linear additivity rules), motivated by the same principle of estimating nucleophilicity, achieved 80% accuracy on a collection of 130 EAS reactions limited to electron rich aromatics and

heterocycles.[14] When supplemented with density functional theory (DFT) calculations,[15] the accuracy of predictions was reported to be >95% on the same dataset and formed the basis for a follow-up semi-empirical quantum mechanical (SQM) method. The SQM model predicts selectivity based on estimated energies of carbocations generated from protonation at each site, meant to represent potential intermediates.[16] Any carbocation that is within thresholds of 1 or 3 kcal/mol of the lowest energy structure is marked as a possible site for reactivity. The RegioSQM method reaches 90% accuracy within 1 kcal/mol and 96% accuracy within 3 kcal/mol on their test set of 525 reactions, although this definition of accuracy does not penalize the prediction of *multiple* reactive sites. More recently, a machine learning approach was applied to prediction of EAS site selectivity prediction using calculated descriptors, including RegioSQM carbocation calculations, and a random forest model to show 90% accuracy on an external validation set.[17] These methods achieve good accuracy for EAS reactions but require several minutes per prediction, primarily because multi-conformer DFT or PM6 calculations are necessary as inputs.

Computational prediction of reaction outcomes has has a plethora of approaches through many years. Early methods used hand coded heuristics to determine reaction outcomes and include the programs CAMEO,[18,19] EROS,[20] IGOR,[21] SOPHIA,[22] and Robia.[23] Later, machine learning methods have addressed the issue of reaction prediction by using synthetically-generated mechanistic data,[24–26] by scoring predictions based on heuristically extracted templates from synthetic[27] or experimental data,[28,29] or by making direct predictions of product species also using experimental data.[30–32] These approaches can generalize across a large range of different reaction types and reaction outcomes, but more subtle aspects of site, stereo, and regio-selectivity are lost.

A chemist's intuition for site selectivity is based on *structure* and rarely reliant on precise calculations. There have been many approaches to directly learn molecular function from structure for various prediction tasks, without relying on fixed descriptors, fingerprints, or other feature engineering. One such example is the Weisfeiler-Lehman network, a type of

graph convolutional neural network model, used by Coley et al.[32] and Jin et al..[31] This model, based on the Weisfeiler-Lehman graph kernel, operates directly on molecular graphs containing atom- and bond-level features to learn a meaningful numerical representation of structure.

We sought to address prediction of site selectivity using two basic hypotheses. 1) Selectivity can be learned from two-dimensional structure without calculated atom features or an explicit 3D conformer, thus reducing the computational overhead. 2) *Any* reaction class can be learned simultaneously using multitask learning, expanding the scope of predictions beyond EAS reactions.

Methods

Data preparation

We use Reaxys as our source of reaction data. We extract two disjoint subsets of C-H functionalization reactions before pooling them: the first is a focused set of four EAS reactions (bromination, chlorination, nitration, and sulfonylation); the second is a broader set of many different reaction types.

The EAS dataset was extracted by identifying all reactions where the sole structural difference between reactants and products is the replacement of an aromatic C-H with a bromo, chloro, nitro, or sulfonyl group. Matching reactions were further filtered to only include ones likely to proceed through EAS by only allowing certain reagents. Brominations were restricted to reactions using Br_2 or *N*-bromosuccinimide; chlorinations using *N*-chlorosuccinimide, sulfuryl chloride, phosphorus oxychloride, Cl_2, or thionyl chloride; Nitrations using nitric acid; sulfonylations using chlorosulfonic acid or sulfuric acid.

The more diverse dataset was extracted by identifying all bimolecular reactions where the difference between *one* reactant and the product is the replacement of an aromatic C-H with a single-bonded heavy atom (preserving the aromaticity of the ring). Matching reactions were categorized into distinct "tasks" based on the identity of the *other* reactant species.

Note that tasks are not defined based on the fragment contributed by the other reactant species, but by its full identity; this is to separate different means of introducing the same functional group, as there may be different selection criteria. Only tasks with at least 100 precedent examples were kept.

Both datasets were further filtered to exclude any reactions with <50% yield, due to our inability to know with certainty that the reported product was the major product (i.e., the site of C-H activation was the most favorable site). The data was further filtered to include only reactions that have more than one aromatic C-H site in the reactant which left 58k examples. It is important to note that the training data is not uniform in its distribution across tasks. For example, of the 127 tasks there are 16k bromination examples with the next highest task having 5.7k examples and many only have 100 examples; a detailed description of the different tasks can be found in the SI (Table S1). Reactant symmetry was taken into account when preparing ground truth labels of the most favorable site. We use an 80:10:10 split for training, validation, and testing within each task using a pseudo-time split[33] validation based on the date each reaction appeared in Reaxys; this is intended to simulate a prospective prediction of site selectivity based on our current body of knowledge. Performance on the validation set was used for early analysis of hyperparameter settings as well as early stopping during training.

Model architecture

The overall architecture uses the WLN encoder coupled to a feed forward neural network as the site predictor.[32] All atom descriptors are calculated using RDKit.[34] The local atom environment is calculated by initializing each atom v with a feature vector f_v in one of two ways: 1) including only structural features representing atomic number, formal charge, explicit and implicit valence, and aromaticity, or 2) also including features representing the total number of hydrogens, aromaticity of neighbors, Crippen LogP and molar refractivity,[35] total polar surface area, accessable surface area, electrotopological state,[36] and Gasteiger

partial charges.[37] For each bond, its bond order and ring status are included in a feature vector f_{uv}.

The atom-centered feature vectors are iteratively updated L times by sum-pooling a learned embedding of neighboring atoms at each iteration. The final representation of *local* atomic structure is calculated using another learned network. An attention mechanism is included to capture the influence of atoms further than L bonds away, including atoms on disconnected molecules; it is used to calculate a weighted sum of all reactant atoms is to give a global representation of each atom. Together, the local and global atom features are used to predict atom reactivity scores for each task, scaled to between zero and one by a sigmoid activation function. The sigmoid activation function is used so that multiple sites can be predicted as in the case with symmetry or non-selective reactions and does not force the model to predict a site if it is not likely as would be the case with a softmax activation. The WLN and the multitask predictor are optimized together by minimizing the cross entropy loss of

$$-\sum_{v} y_v \log p_{t,v} + (1 - y_v) \log(1 - p_{t,v})$$

where $y_v = 1$ if and only if v is the most favorable site of C-H activation for task t and $p_{t,v}$ is the score assigned to atom v for that task t. The full mathematical details of the model can be found in the Supplemental Information.

During inference, a prediction for every atom per task is made and the overall architecture is shown in Figure 1.

As a baseline model, we include a multilayer feed forward neural network that operates only on atom feature vectors f_v to make reactivity predictions in isolation. This model does not contain any pooling of information from neighboring atoms.

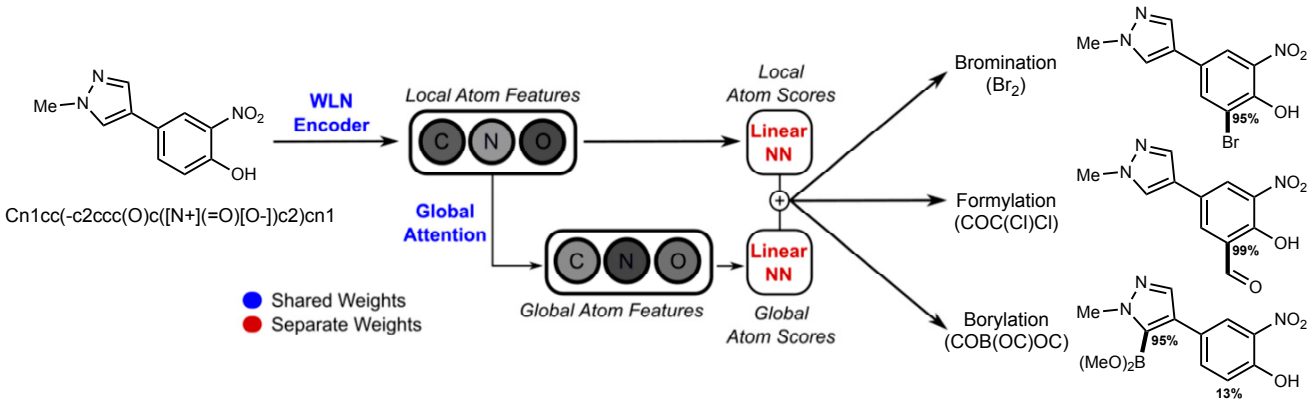


Figure 1: Overall model workflow. A Weisfeiler-Lehmann Network (WLN) learns to encode a molecular structure into atom-level feature vectors, which are used as the basis for a multitask prediction of site-selectivity for many reaction types simultaneously.

Results and discussion

Single-task performance

Previous studies employing the Weisfeiler-Lehman graph convolutional neural network[28,31,32,38] formed the basis for using structure, in the form of a molecular graph, as the input to site predictions. The Weisfeiler-Lehman neural network (WLN) is perfectly suited to this task; a local representation for each atom is calculated and a global context based using an attention mechanism can be employed. As proof of concept that the WLN encoder would allow for learning site selectivity, a single task network was constructed to predict a score for each atom in a molecule for four EAS reactions: bromination, chlorination, nitration, and sulfonylation.

Table 1: Top-1 accuracy (%) for single-task models

Model	Validation Set	Test Set[a]
Br_2 ($N = 13028$)	89.5 ($N = 1629$)	83.6
Cl_2 ($N = 2264$)	75.6 ($N = 283$)	82.0
NO_2 ($N = 4660$)	88.7 ($N = 583$)	87.3
SO_3 ($N = 57$)	71.4 ($N = 8$)	50.0

[a]Number of test set examples are the same as validation set examples

We first evaluate the performance of our model when training on each selectivity task

individually. Table 1 shows the accuracy of the 4 different EAS reactions on the validation and test sets. Overall, average to good performance is seen for each reaction. The highest accuracy was seen with bromination and nitration reactions which correlates to both of these reactions having more training data than chlorination or sulfonylation. These results demonstrate that using the SMILES as an input to construct the molecular graph allows the WLN network to learn from the molecular structure without the need for computationally expensive atom features.

Cross-task performance

Since the four previous EAS reactions operate under a similar mechanism, a model with better performance should transfer to the other similar tasks. Sulfonylation was removed from this study since the performance was only 50% for its own test set. A separate network was trained for each task (reaction type) and then tested on the validation set for all the other tasks. Results in 2 demonstrate that the accuracy varies when a model is transferred to another tasks's validation set. Overall, bromination has the most data and seems to generalize the best across tasks but nitration with the second most amount of data does not seem to follow this trend. Additionally, aromatic scaffolds in each task might not be very diverse since they are a subset of C-H functionalization reactions and each task would not generalize well to others.

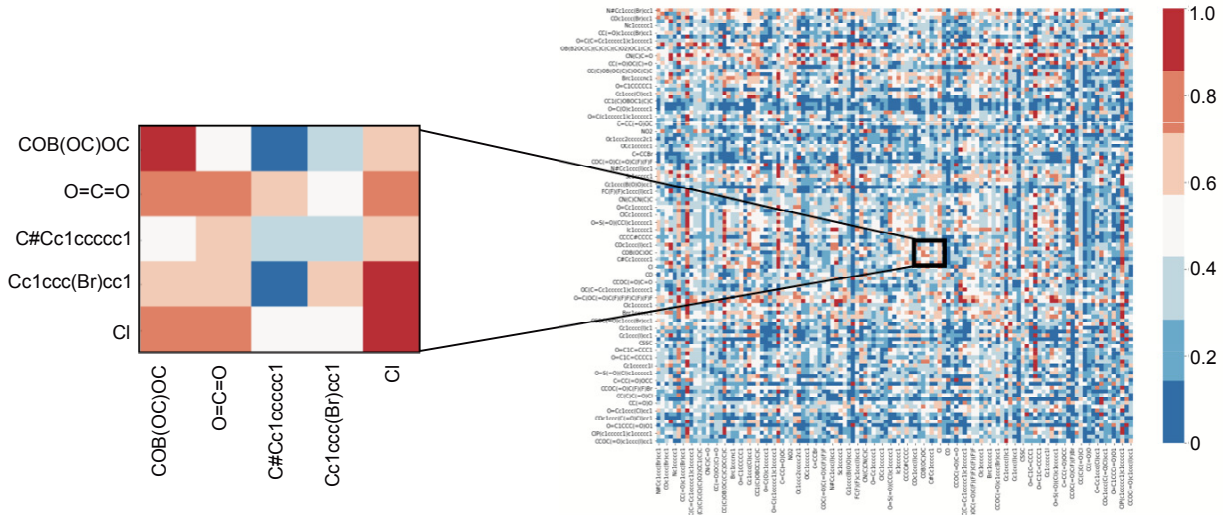
Table 2: Top-1 accuracy (%) when training on a single EAS task and testing on another.

		Trained on		
		Br ($N = 13028$)	Cl ($N = 2264$)	NO_2 ($N = 4660$)
Tested on	Br ($N = 1629$)	89.5	75.8	70.3
	Cl ($N = 283$)	79.9	75.6	66.1
	NO_2 ($N = 583$)	81.3	64.3	88.7

To further demonstrate that single task network training/testing was not a viable option for a larger corpus of reactions, all 127 tasks were trained/tested on one another. A heat-map shown in Figure 2 demonstrates that direct transfer across tasks is not achievable. A small

cross-section is blown up to show that there are reactions that might not operate under the same mechanism in the dataset. This is due to two main factors, 1) not all the tasks are similar in mechanism and 2) some tasks only have around 100 examples which might not be sufficient for the structure encoding to be learned and generalized.

Figure 2: Top-1 accuracy when training on a single C-H functionalization task and testing on another. Columns correspond to the training task; rows correspond to the testing task.



Multi-task performance

To address the issue of low data availability for some tasks, a multitask network was constructed so that the encoder weights are shared. A multitask network imitates the chemists intuition that within the broad scope of all C-H functionalization reactions, there are shared mechanisms driving selectivity (e.g., nucleophilicity, electrophilicity, steric hindrance, catalyst-directing groups). After calculating learned atom representation, the representations are fed through a fully-connected linear layer, summed, and separated into the individual tasks. A baseline model was also constructed that performs the atom embedding without the graph convolution step. Without iterative updating of neighboring atoms, the baseline model cannot directly learn structure and results are poor (Table 3). To demonstrate a baseline model benefits from atom descriptors, simple features calculated by RDKit[39] were

included in the input representation and indeed a significant boost in accuracy is observed (Table 3). The hypothesis that the atom embedding calculated by the WLN encoder are a good representation of the local area around the atom is supported by the similar performance regardless of whether additional atom features are included in the input of the WLN encoder (i.e., pre-calculating these atom properties provides little benefit in performance).

Table 3 shows that the overall accuracy for single task prediction is close to or the same as the multitask predictor. Generally, within each task, the operating mechanism that defines selectivity is inherent to the task and there is not a wide diversity of conditions that would lead to differing selectivity. For example, in the task of coupling bromobenzene (task labeled Brc1ccccc1), some form of palladium is listed in the reagents for over 70% of the reactions. Similarly, reaction conditions (reagents, catalysts, solvents, etc.) are relatively consistent within a task, so adding the reagents to the input does not greatly improve the model performance for *most* tasks (see SI for complete details). Including reagents can help improve accuracy and is desirable for cases such a virtual screens or if the chemist already knows the conditions, but for a chemist's idea generation it is beneficial to also have a model that would not require the additional input of reagents.

Table 3: Comparison of results on the test set of different model architectures

Model	Without atom features		With atom features	
	Top-1 Acc. (%)	MRR (%)	Top-1 Acc. (%)	MRR (%)
Baseline (No WLN)	21.3	45.9	47.6	64.9
Single Task	81.8	88.6	81.0	88.3
Multitask	83.1	89.5	84.0	90.1

Comparison to other methodology

Previous methodology, namely RegioSQM,[16] achieves high site prediction accuracy based on enumeration and calculation of protonated carbocation intermediates in the EAS pathway.[40] Comparison of the WLN methodology to RegioSQM was limited to bromination reactions for a fair comparison to the intended application of RegioSQM. 500 reactions were selected

at random from our test set for comparison. 6 of the 500 had either had a structure that failed to converge or had a proton transfer during optimization, and were thrown out leaving 494 total reactions for comparison. The RegioSQM methodology outputs the lowest energy carbocation and any other of enumerated carbocations that are within a set threshold of 1 kcal/mol as the top predictions. This means that there can be multiple predictions by RegioSQM. In contrast, analysis of the WLN uses only the top 1 pick to calculate accuracy. To overcome this disparity in evaluation between the two methodologies in how accuracy is defined, analysis was performed by ranking the predictions. With RegioSQM, the predictions were ranked by lowest to highest energy conformer and in the WLN ranking was based on the final atom scores for brominaion (a comparison based on how the authors of RegioSQM performed analysis is detailed in the SI). The accuracy of the top 1, 2, and 3 choices and the mean reciprocal rank are calculated and reported.

In Table 4 are the results of this comparison which shows that the WLN predictions for bromination are more accurate than those of RegioSQM methodology. Finally and most important is that 300 predictions by RegioSQM takes over 10 days to complete while the neural network model makes the same predictions in 6.3 seconds both using 12 CPU cores. Even when training is factored in for the WLN, the total time is less than 4 hours for *all* 130 tasks and 58k total examples (train/valid/test) on a single GTX 1080 GPU. Although not surprising that a neural net will make predictions faster than semi-emperical methods, our approach now makes application to thousands of molecules tractable.

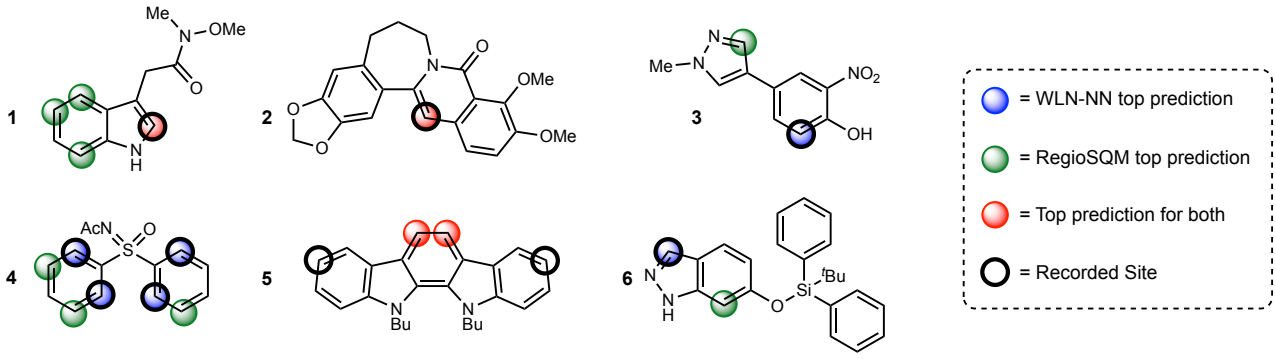
Table 4: Comparison to RegioSQM[16] on a random subset of 494 bromination reactions from our test set.

	Top 1 (%)	Top 2 (%)	Top 3 (%)	MRR (%)	Time (12 CPU's)
RegioSQM[16]	79.7	89.2	93.3	87.1	>10 days
WLN	85.0	92.3	95.7	90.5	6.3 s

Some examples of predictions for both methodologies are shown in Figure 3. Only the top 1 prediction is drawn for the WLN and only sites that are within 1 kcal/mol of the lowest energy conformer in the RegioSQM method are drawn. The first two examples show that the

WLN is effective at finding the correct site and often that correlates with the prediction from RegioSQM. An advantage of the WLN is demonstrated by the sites chosen for compound **1** where the WLN chooses one site and RegioSQM gives 4 predictions one of which is correct. Predictions for compounds **3** and **4** show that when the sites with the lowest calculated energies are not the true reactive site, the WLN can recognize the structure and furnish the correct position. An interesting pattern is seen when both methodologies predict the incorrect site where the WLN agrees with the RegioSQM prediction demonstrating that the structure is indeed learned using the multitask network. The final examples is one that was shown in the original publication of RegioSQM methodology where presumably sterics on the phenolic oxygen disfavor the lowest energy site from reacting and the WLN correctly identifies the reactive site.

Figure 3: Comparative examples between the WLN and RegioSQM predictions. All RegioSQM predictions within 1 kcal/mol are highlighted.



Conclusion

In conclusion, a multitask network was developed to predict site selectivity of aromatic C-H functionalization reactions using a Weisfeiler-Lehman encoder for learning predictions based on structure. Top 1 accuracy of the model on the test set is 84% with a mean reciprocal rank of 90%. The WLN achieves similar or better accuracy when compared to semi-empirical methods and drastically reduces the time for predictions from minutes to less than a second

per molecule. In addition, the tool can be used to prioritize compounds or intermediates in a synthetic route that could be accessed selectively, leading to a more diversified collection of compounds. Finally, a large corpus of general reactions was previously used with the WLN architecture for reaction prediction and this study extends the WLN's capability to learn very specific site selectivity and reactivity.

Acknowledgement

We thank the DARPA Make-It program under contract ARO W911NF-16-2-0023 and the Machine Learning for Pharmaceutical Discovery and Synthesis consortium and for funding this research. We also thank Elsevier for the invaluable curation of the Reaxys data set.

Notes and References

(1) Joanna, W.; Glorius, F. *Nature Chemistry* **2013**, *5*, 369.

(2) Cernak, T.; Dykstra, K. D.; Tyagarajan, S.; Vachal, P.; Krska, S. W. *Chemical Society Reviews* **2015**, *45*, 546–576.

(3) Boström, J.; Brown, D. G.; Young, R. J.; Keserü, G. M. *Nature Reviews Drug Discovery* **2018**, *17*, 709.

(4) Yao, H.; Liu, Y.; Tyagarajan, S.; Streckfuss, E.; Reibarkh, M.; Chen, K.; Zamora, I.; Fontaine, F.; Goracci, L.; Helmy, R.; et al., *European Journal of Organic Chemistry* **2017**, 7122–7126.

(5) Durak, L. J.; Payne, J. T.; Lewis, J. C. *ACS Catalysis* **2016**, *6*, 1451–1454.

(6) Lyons, T. W.; Sanford, M. S. *Chemical Reviews* **2010**, *110*, 1147–1169, PMID: 20078038.

(7) Davies, H. M. L.; Morton, D. *ACS Central Science* **2017**, *3*, 936–943, PMID: 28979934.

(8) Neufeldt, S. R.; Sanford, M. S. *Accounts of Chemical Research* **2012**, *45*, 936–946.

(9) Toste, F. D.; Sigman, M. S.; Miller, S. J. *Accounts of Chemical Research* **2017**, *50*, 609–615, PMID: 28945415.

(10) Boga, S. B.; Christensen, M.; Perrotto, N.; Krska, S. W.; Dreher, S.; Tudge, M. T.; Ashley, E. R.; Poirier, M.; Reibarkh, M.; Liu, Y.; et al., *Reaction Chemistry & Engineering* **2017**, *2*, 446–450.

(11) Krska, S. W.; DiRocco, D. A.; Dreher, S. D.; Shevlin, M. *Accounts of Chemical Research* **2017**, *50*, 2976–2985, PMID: 29172435.

(12) Olah, G. A. *Accounts of Chemical Research* **1971**, *4*, 240–248.

(13) Hansch, C.; Leo, A.; Taft, R. W. *Chemical Reviews* **1991**, *91*, 165–195.

(14) Kruszyk, M.; Jessing, M.; Kristensen, J. L.; Jørgensen, M. *The Journal of Organic Chemistry* **2016**, *81*, 5128–5134, PMID: 27176051.

(15) Liljenberg, M.; Brinck, T.; Herschend, B.; Rein, T.; Rockwell, G.; Svensson, M. *The Journal of Organic Chemistry* **2010**, *75*, 4696–4705.

(16) Kromann, J. C.; Jensen, J. H.; Kruszyk, M.; Jessing, M.; Jørgensen, M. *Chemical Science* **2017**, *9*, 660–665.

(17) Tomberg, A.; Johansson, M. J.; Norrby, P. *The Journal of Organic Chemistry* **2019**, *84*, 4695–4703.

(18) Salatin, T. D.; Jorgensen, W. L. *Journal of Organic Chemistry* **1980**, *45*, 2043–2051.

(19) Bures, M. G.; Roos-Kozel, B. L.; Jorgensen, W. L. *The Journal of Organic Chemistry* **1985**, *50*, 4490–4498.

(20) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K. A new treatment of chemical reactivity: Development of EROS, an expert system for reaction prediction and synthesis design. Organic Synthesis, Reactions and Mechanisms. Berlin, Heidelberg, 1987; pp 19–73.

(21) Ugi, I.; Bauer, J.; Bley, K.; Dengler, A.; Dietz, A.; Fontain, E.; Gruber, B.; Herges, R.; Knauer, M.; Reitsam, K.; Stein, N. *Angewandte Chemie International Edition in English* **1993**, *32*, 201–227.

(22) Satoh, H.; Funatsu, K. *Journal of Chemical Information and Computer Sciences* **1995**, *35*, 34–44.

(23) Socorro, I. M.; Taylor, K.; Goodman, J. M. *Organic Letters* **2005**, *7*, 3541–3544, PMID: 16048337.

(24) Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. *Journal of Chemical Information and Modeling* **2011**, *51*, 2209–2222, PMID: 21819139.

(25) Kayala, M. A.; Baldi, P. *Journal of Chemical Information and Modeling* **2012**, *52*, 2526–2540, PMID: 22978639.

(26) Molga, K.; Gajewska, E. P.; Szymkuć, S.; Grzybowski, B. A. *React. Chem. Eng.* **2019**, –.

(27) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. *ACS Central Science* **2016**, *2*, 725–732, PMID: 27800555.

(28) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. *ACS Central Science* **2017**, *3*, 434–443.

(29) Segler, M. H. S.; Waller, M. P. *Chemistry – A European Journal* **2017**, *23*, 5966–5971.

(30) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Bekas, C.; Lee, A. A. *arXiv e-prints* **2018**, arXiv:1811.02633.

(31) Jin, W.; Coley, C. W.; Barzilay, R.; Jaakkola, T. S. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. NIPS. 2017.

(32) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. *Chemical Science* **2019**, *10*, 370–37.

(33) Sheridan, R. P. *Journal of Chemical Information and Modeling* **2013**, *53*, 783–790, PMID: 23521722.

(34) Landrum, G. RDKit: Open-source cheminformatics. `http://www.rdkit.org`.

(35) Wildman, S. A.; Crippen, G. M. *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 868–873.

(36) Hall, L. H.; Kier, L. B. *Journal of Chemical Information and Computer Sciences* **1995**, *35*, 1039–1045.

(37) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219 – 3228.

(38) Lei, T.; Jin, W.; Barzilay, R.; Jaakkola, T. S. Deriving Neural Architectures from Sequence and Graph Kernels. ICML. 2017.

(39) Added atom features to the initial Weisfeiler-Lehman implementation include Atomic number, Crippen logP, Crippen molar refractivity, total polar surface area, accessible surface area, Gasteiger Charge, atom electronegativity.

(40) Other methodology by Kruszyk et. al. using neural networks was not implemented for comparison due to the restriction of using inaccessible software such as Gaussian for DFT inputs. RegioSQM uses free software that is easily obtained and deployed and furthermore is an input in Kruszyk et. al. neural network implementation.

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.

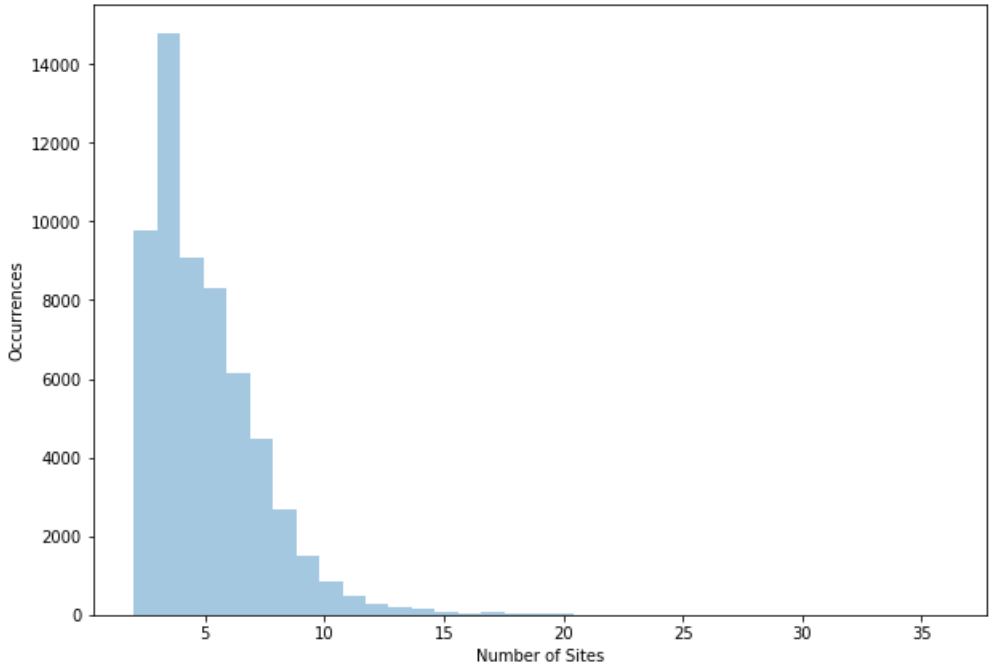
Table S1: Description of all C-H activation tasks used in this study. The number of examples reflects the total number, divided in an 80:10:10 split for training:validation:testing.

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**.

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

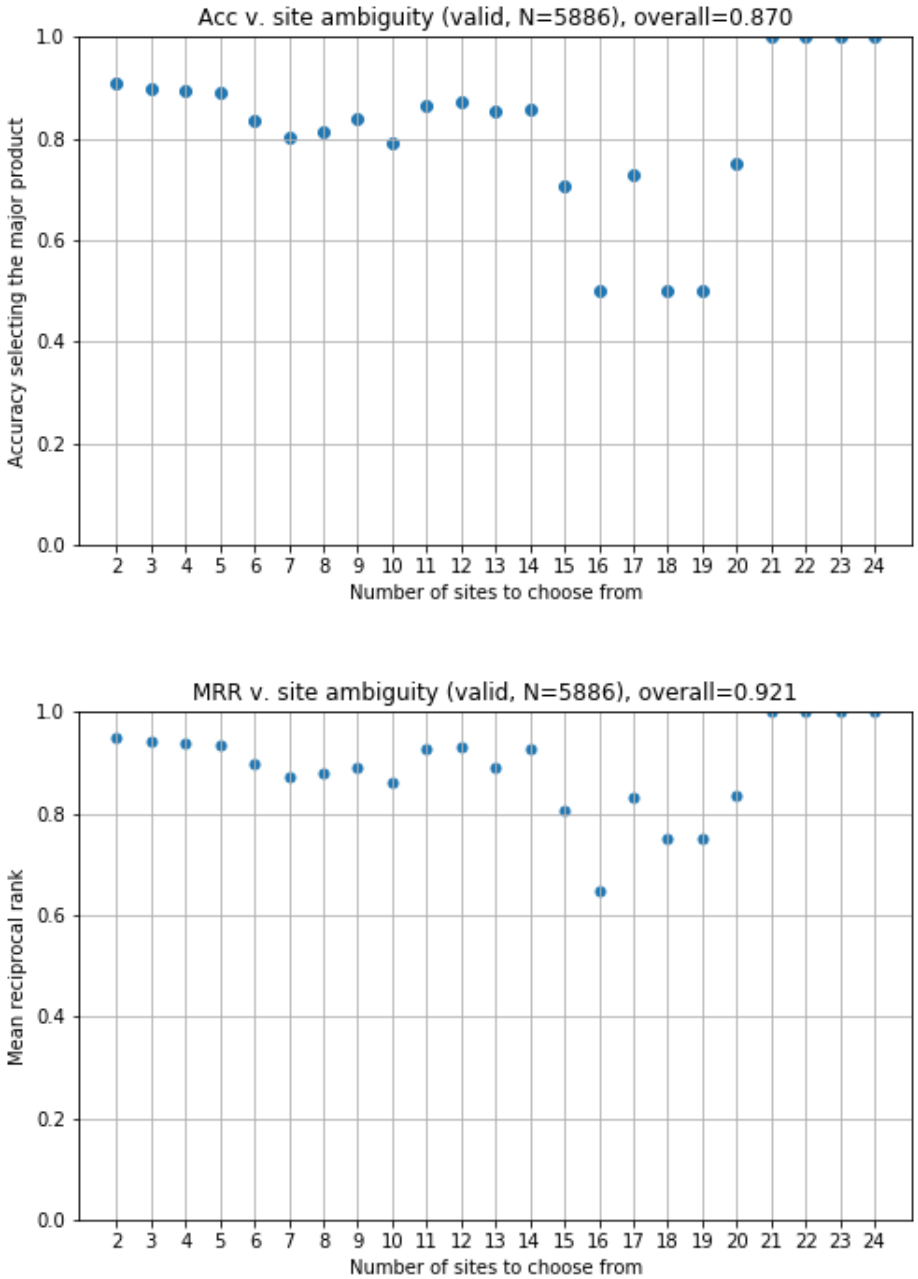
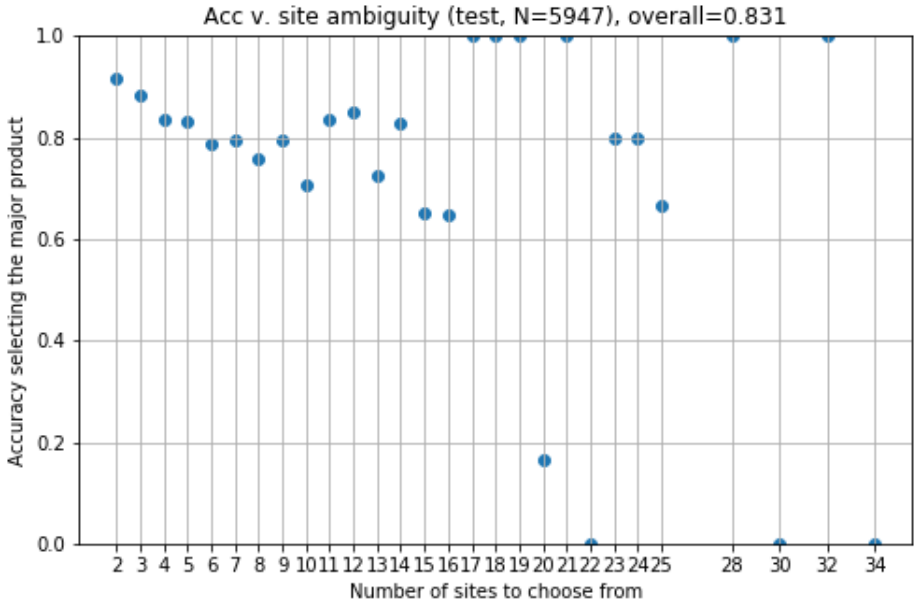
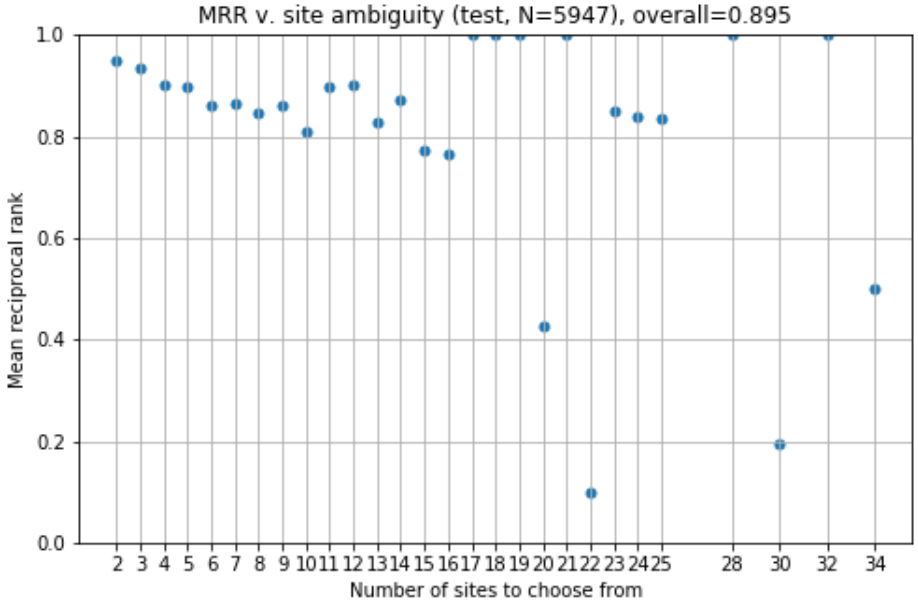


Figure S3: Performance of the test set by number of sites





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

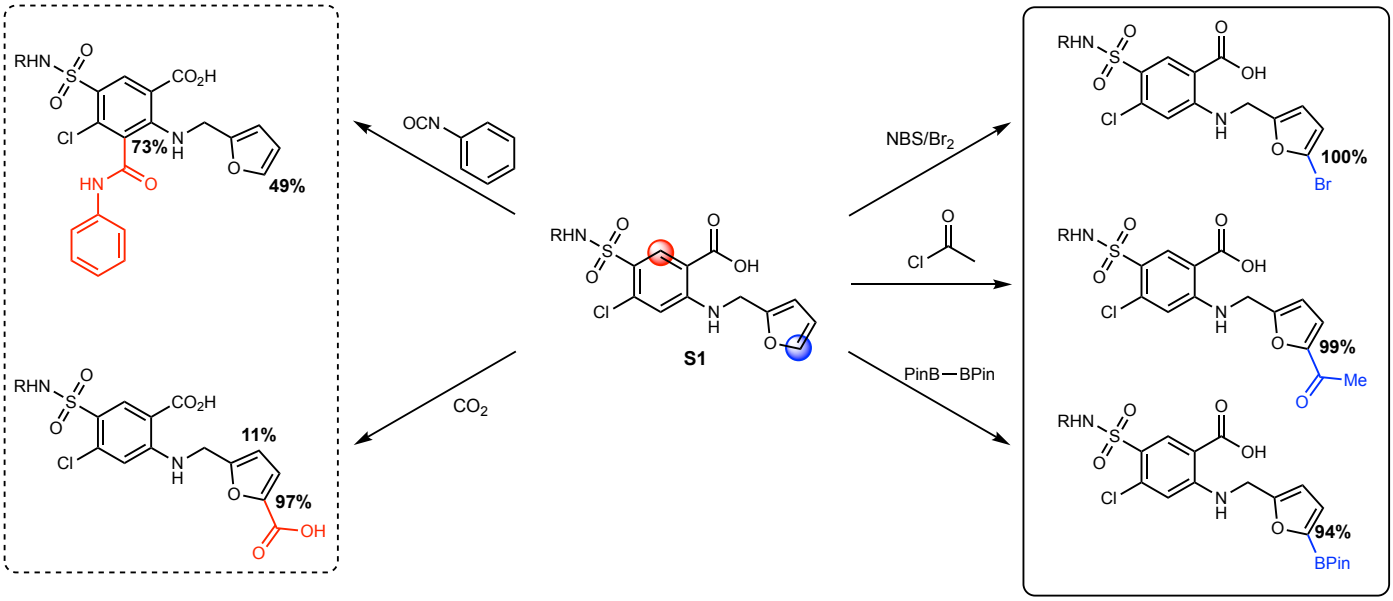


Figure S5: Example after one step retrosynthesis

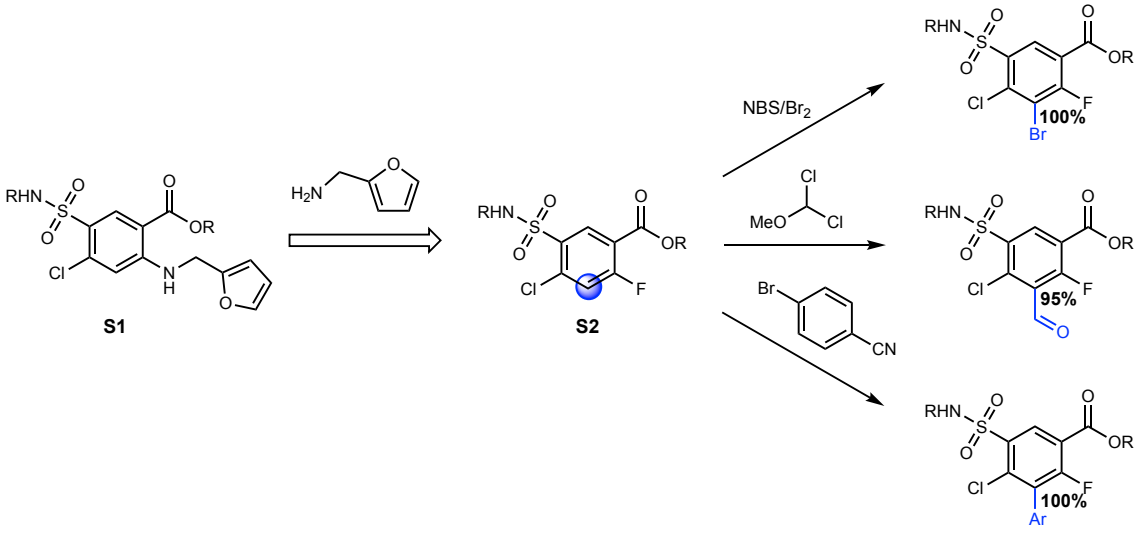
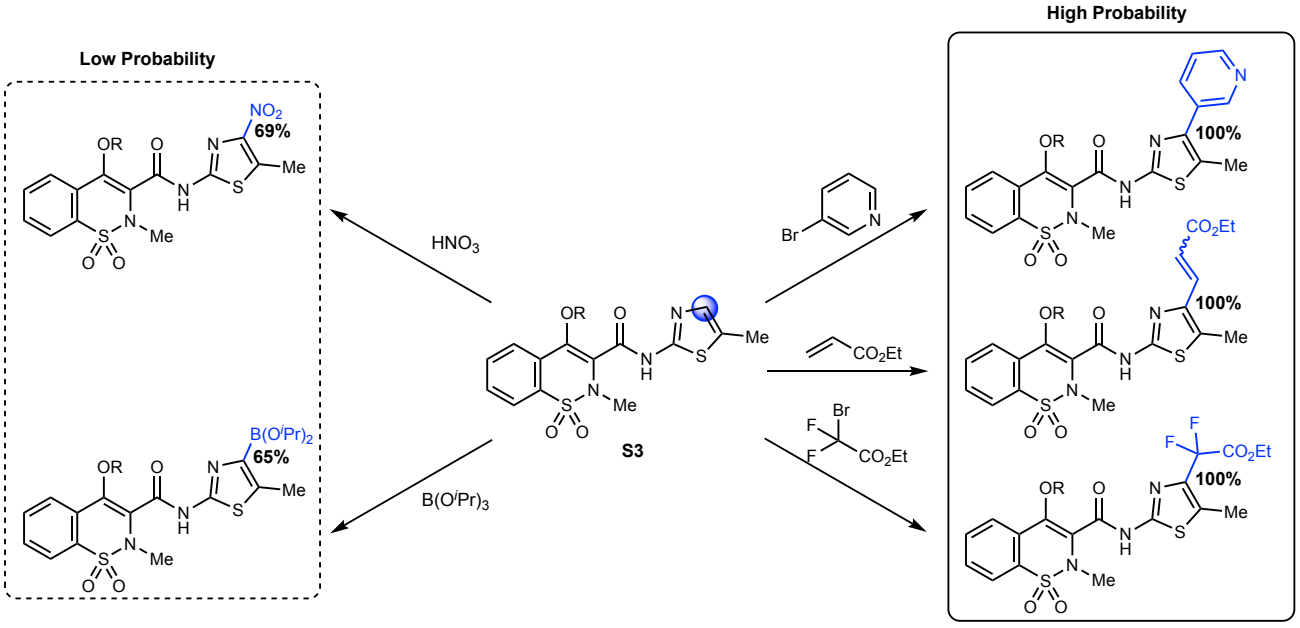


Figure S6: Example 2



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.