An Investigation Into the Potential of Machine Learning for the Prediction of Regioselectivity of C(sp^3)–H Bond Functionalization

Amitesh Pandey[1], **Gina Lee**[1], **Jules Schleinitz**[2], **Alba Carretero**[2], **Sarah Reisman**[2]
[1]Computing + Mathematical Sciences, Caltech, [2]Chemistry and Chemical Engineering, Caltech

Introduction

C–H functionalization refers to the class of reactions wherein a C–H bond is cleaved and replaced with a bond between the carbon atom and a desirable functional group or molecule. Organic reactants have many C–H bonds, often with very similar physical properties, as well as other sites prone to oxidation. In these cases, the prediction of the site most prone to functionalization is critical to the practical application of these reactions [1]. Thus, we present a digital machine learning-based tool for the prediction of reaction outcomes for this class of reactions.

Figure 1: Example of C–H Functionalization (Oxidation)

Methodology

Data Collection and Processing:

- Data mining from published literature and the Reaxys database were the primary sources for C–H functionalization reactions for the model [2].
- Quantum mechanical descriptors for each **carbon** atom in these reactions were retrieved and binary value was generated for each carbon to indicate if that carbon had been oxidized. [3]

Training Data
Vector Size: 8 (mapped to binary)
Num. Pairs: ~5000

	Carbon Atom Descriptors	Selectivity Value	
X	[1.256, 1.546, 1.955, . . .]	0	Y
	[2.439, 1.456, 1.958, . . .]	1	
	[1.455, 2.955, 0.365, . . .]	0	
	⋮	⋮	

Figure 2: Generated Training Data for the Model

Development of the Model:
The target function follows the expression:

$$f : \mathbb{R} \rightarrow \{0,\ 1\}, \text{such that, } f\,(\,\boldsymbol{x}) \ = y, \text{ for all } \boldsymbol{x} \in X$$

- Random forest, Linear and RidgeCV regression models were tested on the data and then the results of these models were compared to finalize on a learning model.
- A **Random forest** model with highly tailored architecture was deployed.
- A K-Fold Cross Validation based hyperparameter search for the model was carried out.

Architecture of the Model:

- The model's prediction is in the format of an array holding the predicted probabilities of oxidation for each carbon atom.
- The predicted most likely carbon is retrieved from the array and the reactant is manipulated to indicate the predicted oxidation.

Libraries and software Used:

- RDKit [4]
- Sci-kit Learn [5]
- xTB
- Pandas
- RXNMapper
- Morfeus [6]

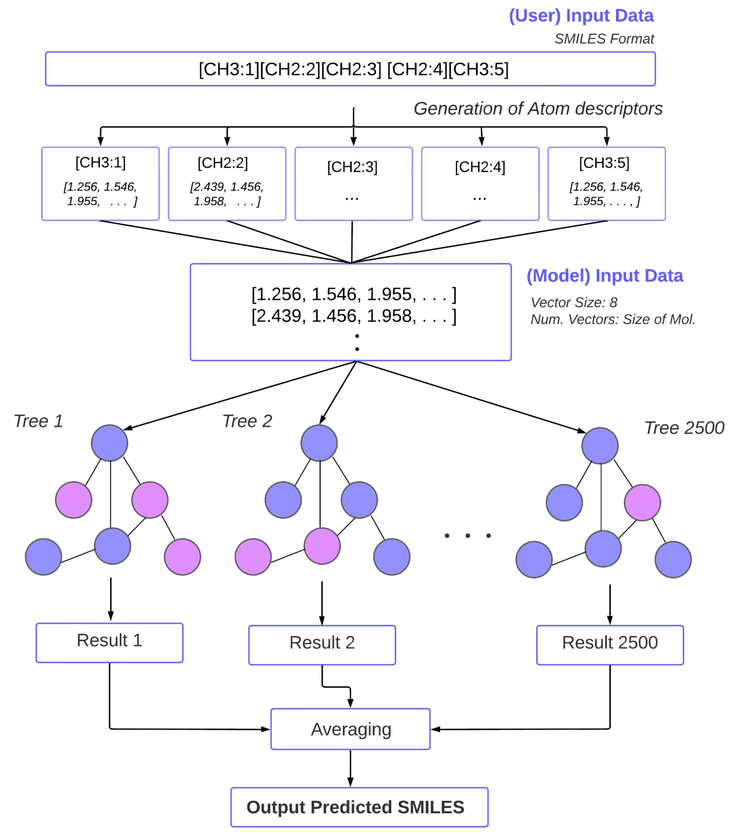


Figure 3: Pipeline of the Model

Alcohol *vs* Ketone Issue: The situation in which subsequent oxidations of the reactant resulted in the production of an aldehyde or ketone instead of an alcohol was a significant obstacle.

Solution: First, the predicted regioselectivity for the alcohol is retrieved. Then, this alcohol is passed through the model and the predicted regioselectivity for the production of the ketone is retrieved. Finally, these two are compared and whichever product has a higher likelihood of occurring is finalised.

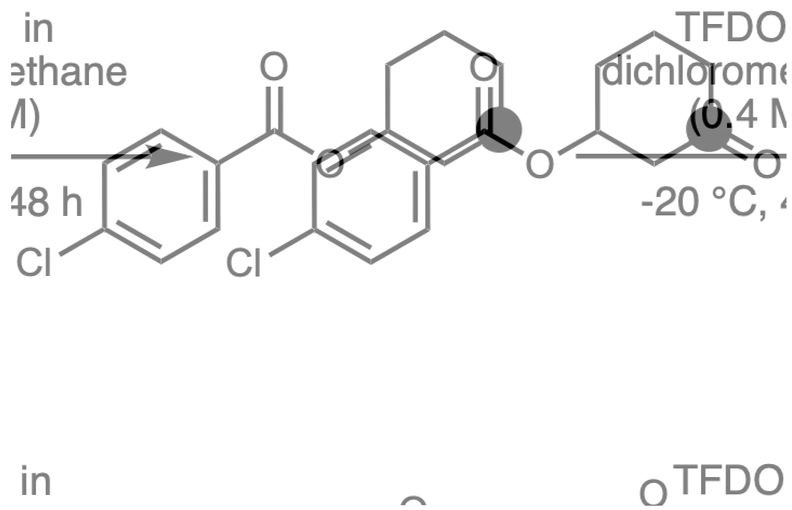
Results

- Site prediction accuracy was **70%**
- Product (SMILES) prediction accuracy was **40%**
- The Root Mean Square Percentage Error (RMSPE) was **33%**

Random Forest	SITE ACCURACY: 70% SMILES ACCURACY: 40% *Trained On: xTB Descriptors*	Linear	SITE ACCURACY: 22% SMILES ACCURACY: 16% *Trained On: xTB Descriptors*	
RidgeCV	SITE ACCURACY: 30% SMILES ACCURACY: 18% *Trained On: xTB Descriptors*	Random Forest	SITE ACCURACY: 54% SMILES ACCURACY: 1.5% *Trained On: Gasteiger Charges*	Baseline Metric

Table 1: Site and SMILES Accuracy using different models and training datasets

Examples of reactions where the regioselectivity was accurately predicted by the model:



Conclusion and Discussion

- Despite using computationally inexpensive descriptors and a considerably simple machine learning model, our system beat the set baseline significantly.
- By making minor changes to the prediction mechanism and the generation of descriptors, the model can be expanded to the prediction of the addition of other functional groups to carbon atoms.

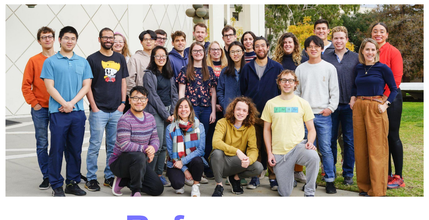
Future Work:

- Incorporating H atoms
- Inclusion of DFT
- Incorporating GNNs
- Improving Descriptors
- Larger Train Data

Open Problem: Multiple oxidation sites in one reaction

Acknowledgements

Funding: John Stauffer Trustees, Caltech SFP
Mentorship: Reisman Group



References

[1] Guiding Chemical Synthesis, Jensen J., arXiv:1710.07439
[2] Reaxys Chemical Database: https://www.reaxys.com/
[3] J. Chem. Theory Comput. 2019, 15, 3, 1652–1671
[4] RDKit: Open-source cheminformatics. https://www.rdkit.org
[5] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
[6] Morfeus - Molecular Features for Machine Learning: 10.5281/zenodo.7017599