Summer Undergraduate Research Fellowship

Developing Stopping Criterion for C–H Oxidation Regioselectivity Prediction

Carolyn Ruan

Project Mentor: Professor Sarah Reisman Graduate Student Mentor: Anjali Gurajapu

Introduction:

Predicting the regioselectivity of C-H oxidations of complex molecules has been an ongoing problem in the topic of target-directed synthesis.[1] Successful regioselectivity prediction models are crucial to derisk C–H activation steps in synthetic campaigns and synthesis planning, which could enable more efficient drug design and material development.[1,2]

A major obstacle in developing a successful regioselectivity prediction model is generating its dataset, especially with complex substrates, which may require significant time and laboratory resources for the purification, characterization, and precise assignment of the functionalization site.[1] Previously, the Reisman lab has developed machine learning models that predict the regioselectivity of C(sp^3)–H functionalization, with a random forest model performing the best.[1] Specifically, they used a dataset curated from the literature and active learning-based acquisition functions to reduce the number of data points needed to perform accurate prediction.[1]

The next step is establishing a successful stopping criterion to end the acquisition active learning loop aggressively, to minimize the number of required experiments.[3] This is done, without compromising on model accuracy, by capitalizing on early peaks in performance caused by the active learning strategy.[3] Preliminary work on the C(sp^3)–H functionalization regioselectivity model has shown that various thresholds on the uncertainty of the most reactive carbon prediction are ineffective as a stopping criterion.

For my proposed SURF project, I plan to work in the laboratory of Professor Sarah Reisman, under the guidance of 2nd-year graduate student Anjali Gurajapu, to make significant progress towards developing a successful stopping criterion for their previously developed C(sp^3)–H functionalization regioselectivity model. The completion of this stopping criterion would enable practical implementation of the active learning strategy and minimize experimental cost of dataset generation, without compromising accuracy.

Objectives:

The overall aim of this SURF project is to develop a successful stopping criterion for the C(sp3)–H functionalization regioselectivity model. The specific aims are to:

1. Develop a comprehensive framework for stopping criterion evaluation, which can handle molecules where accuracy is never achieved.
2. Develop and evaluate various stopping criterion functions based on properties computed from the model predictions.
3. Explore potential modern machine learning architectures for sequential data modeling to improve stopping criterion predictions, and assess the transferability to the radical arene C–H borylation dataset.

Approach:

To develop a comprehensive framework for stopping criterion evaluation, we will first establish measures to assess stopping criteria effectiveness. A possible metric is the definition from Delmas et al., for stopping criteria evaluation, reflecting the "best compromise between an early stop (consuming less than 50% of the data) and reliable accuracy at the stopping point over 10 different runs"[3] (Fig. 1), although we may also consider metrics such as area under the curve as an alternate assessment for early stopping. Since only approximately 50% of target molecules achieve accurate predictions in the model, we will also explore scenarios where accuracy is never achieved and ensure that the stopping criterion framework is comprehensive and does not stop in these cases. Success will be measured by the ability to distinguish when further iterations yield diminishing improvements.

For Aim 2, we will develop and evaluate various stopping criterion functions based on our preliminary work. This includes testing properties such as maximum and minimum distances between reactive and unreactive carbons, changes in cluster identity for these carbons, and CV accuracy on the training set, to see how these properties correlate with accuracy on the test molecule. Preliminary results have shown raw thresholds, percentage thresholds, and gradient for the uncertainty on the most reactive carbon prediction to be ineffective on their own (Fig. 2), thus necessitating exploring alternative combinations of features and thresholds to improve stopping predictions. Visualization through top-1 accuracy plots will help assess the performance of different criteria.

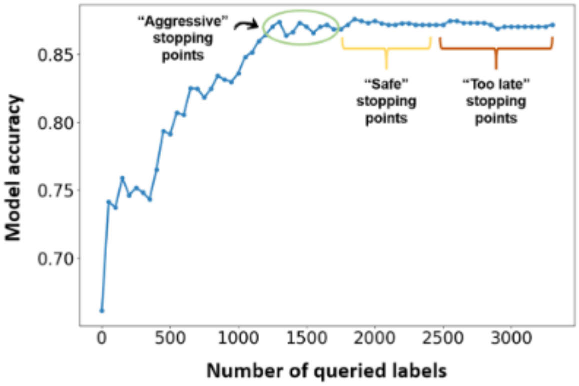

Fig. 1 Active learning curve with different natures of stopping criteria: aggressive, safe, and too late.[3]

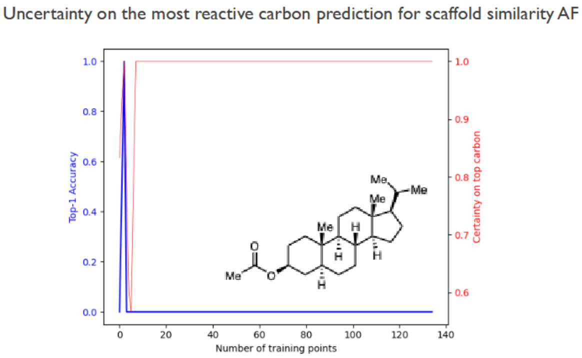

Fig. 2 From preliminary results, when active learning stops and stabilizes early on the wrong position, using uncertainty on the most reactive carbon prediction as a stopping criterion is ineffective.

For Aim 3, we will investigate modern sequential modeling techniques to refine stopping criterion predictions. This requires developing appropriate data labels to enable sequential learning and selecting suitable models for evaluation. Potential architectures, such as Gated Recurrent Units and Transformer-based models, will be explored to determine their ability to capture trends in stopping behavior. Additionally, we will assess the transferability of the stopping criteria to the radical arene C–H borylation dataset previously used as additional validation in the existing work done by the Reisman lab.[1]

The data used for the stopping criterion, specifically the accuracy and carbon predictions at each time point, across several replicate models, has already been collected. The software that will be used in the project is NumPy, a numerical computing package, along with the sklearn library in Python.[4] The step in this project that will be most difficult will be the learning curve in understanding the code base, which I plan to mitigate through weekly research meetings during Spring term. We will also collaborate with Professor Nitesh Chawla's lab at the University of North Carolina on coding and ideation for the stopping criteria.

Work Plan:

Week 0 | Set up environment and gain familiarity with the code base (with weekly research meetings during Spring term).

Week 1-3 | Establish a comprehensive evaluation framework, and script and conduct initial tests on the current system to assess baseline performance.

Week 4-6 | Explore and implement different stopping criterion functions.

Week 7-10 | Develop data labels for sequential modeling and experiment with modern machine learning architectures. Finalize results, create visualizations, and prepare a final report.

References:

1. 1. Schleinitz J, Carretero-Cerdán A, Gurajapu A, Harnik Y, Lee G, Pandey A, et al. Designing Target-specific Data Sets for Regioselectivity Predictions on Complex Substrates. Journal of the American Chemical Society. 2025 Feb 6; Available from: https://doi.org/10.1021/jacs.4c15902
2. 1. Seumer J, Ree N, Jensen JH. Enhancing Chemical Synthesis Planning: Automated Quantum Mechanics-Based Regioselectivity Prediction for C-H Activation with Directing Groups. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-vssxt This content is a preprint and has not been peer-reviewed.
3. 1. Delmas V, Jacquemin D, Aymeric Blondel, Vacher M, Laurent AD. How to actively learn chemical reaction yields in real-time using stopping criteria. Reaction Chemistry & Engineering [Internet]. 2024 Apr 30;9(5):1206–15. Available from: https://pubs.rsc.org/en/content/articlehtml/2024/re/d3re00628j
4. 1. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. arXiv:13090238 [cs] [Internet]. 2013 Sep 1; Available from: https://arxiv.org/abs/1309.0238