Finetuning Large Language Models for Prediction of C-H Functionalisation Selectivity

Ahmed M. Zaitoun, **Xacobe** C. Cambeiro, **Jiayun** Pang

School of Science, Faculty of Engineering and Science, University of Greenwich,

Af0193o@gre.ac.uk

C–H functionalisation selectivity plays a critical role in enhancing the activity of existing drugs (known as Late-Stage Functionalisation) and selectively designing new drug candidates. However, accurately predicting this selectivity remains a significant challenge due to the subtle electronic energy differences of C–H bonds within a molecule.

Transformer-based sequence-to-sequence (seq2seq) language models have shown promises in organic chemistry prediction for a variety of tasks, such as forward reaction products, reaction yields, reagents, retrosynthesis, and reaction classifications [5]. In this study, we investigate the ability of two pre-trained seq2seq language models, ByT5 and FlanT5, to predict the products of C–H functionalization reactions. These models, derived from Google's ByT5 [2] and FlanT5 [3] architectures, have been pre-trained and fine-tuned with different strategies to enhance their compatibility with chemical data [4]. We curated a dataset of 390 C–H functionalisation reactions from literature. These reactions all involve ligand-to-metal charge transfer (LMCT) with a hydrogen atom transfer (HAT) agent. We used 90% of the dataset for training and 10% for testing. After fine-tuning using our C-H dataset, FlanT5 achieved a top-1 accuracy of 41.03%, while ByT5 reached 15.38% in C-H functionalisation selectivity. To further enhance predictive accuracy, we implemented a Parameter-Efficient Fine-Tuning (PEFT) approach [4], which aims to mitigate Catastrophic Forgetting by preserving the general knowledge learned from pretraining. Interestingly, ByT5 showed a more substantial improvement, reaching 43.59% in top-1 accuracy, while FlanT5 achieved a marginal increase, also reaching 43.59% in the top-1 prediction. For comparison, a previous study using a similar approach achieved top-1 accuracy of 60.81% using a dataset of 1,041 C–H borylation reactions, [1] indicating that an increased dataset size may enhance performance.

Further analysis shows that our language models are able to predict valid structures (i.e. some incorrect predictions still exhibit chemical common sense) and can reach 60% in top-5 accuracy. We have used SHAP (Shapley Additive Explanations) analysis, an explainable AI method, to visualize how different functional groups in the reactants contribute to the predicted products. It indicates that the models have the ability to identify key reagents involved in the reactions. Our findings demonstrate that both FlanT5 and ByT5 can be effectively fine-tuned with a relatively small number of reactions to improve their predictive power for novel reactions.

References

[1] Kotlyarov, R. 2024, *Journal of Chemical Information and Modeling*, 64,10, 4286–4297

[2] Xue, L. 2022, *arXiv*:2105.13626

[3] Chung, H. W, 2022, *arXiv*:2210.11416

[4] Vulic,I, 2024, *arXiv,* 2405.10625

[5] Lu, J. 2022, *Journal of Chemical Information and Modeling*, 62,6, 1376–1387