Found In Translation: Using Language Models To Predict C–H Borylation Regioselectivity

Ruslan Kotlyarov and Jonathan M. Goodman

Yusuf Hamied Department of Chemistry, University of Cambridge

We investigated how encoder-decoder transformer models can be applied to predicting regioselectivity of iridium-catalysed C–H borylation using reaction SMILES as the only input. Our model performance is comparable to state of the art deep learning models trained on the same amount of data but further investigation is needed on how well it generalises to new substrates.

Can We Trust Humans To Predict Selectivity?

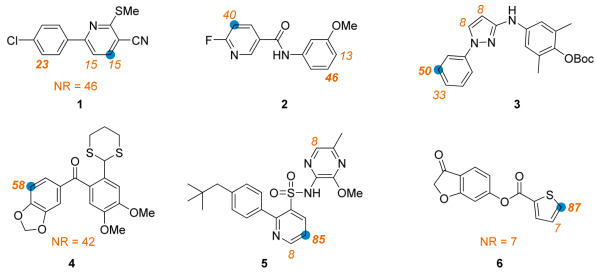
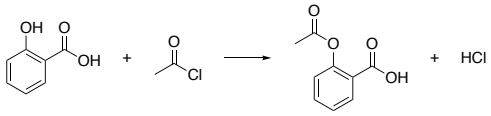


Figure 1. Borylation selectivity prediction: experts (%) vs. ground truth

How Do Transformers Predict Products?

In stark contrast with QM models[1] or graph neural networks[2], the output of the model is not a set of scores assigned to atoms, but a newly generated molecule SMILES. The model does not predict molecular structure *per se*, but the probability distribution for next token in the sequence given prompt, reaction input, and output generated so far, as illustrated by Figure 2.



Input: Product:O=C(O)c1ccccc1O.CC(=O)Cl>>
Stage 1: O
Stage 2: O=
Stage 3: O=C
Carry on until the <end> token
Output: O=C(C)c1ccccc1c(=O).Cl

Figure 2. Overview of conditional generation

This approach allows us to generate arbitrary organic molecules without need to know the underlying reaction mechanism. However, the most probable predictions are not always the most plausible ones (Figure 3).

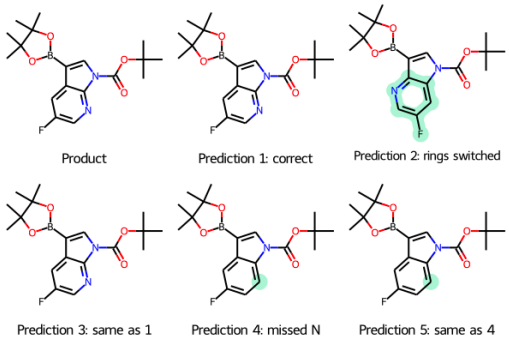


Figure 3: Conditional generation of products may not make sense chemically

The model may 'transpose' the ring substituents or generate a wrong character. In addition to that, the generated predictions may ultimately correspond to the same molecule.

What's Next?

The applicability of transformers to regioselectivity prediction is limited by the model generating predicted products from scratch, requiring it to make multiple guesses in a row. Simplifying the task to reaction site classification allowed us to achieve better performance at the cost of model universality. Interestingly, data-agnostic featurisation with a random forest classifier afforded good performance right away. We will investigate if encoder-only transformers can achieve similar results.

C–H borylation is useful for the late-stage functionalisation of organic molecules. Its products can be used in several cross-coupling reactions to rapidly explore the target chemical space in drug discovery. However, it is hard to predict the reaction site, especially with multiple aromatic rings present (Figure 1).

There are multiple approaches to modelling it, with encoder-decoder transformer architecture being the most adaptable to new chemical reactions and various relevant tasks, e.g., product generation and reaction classification.

QM-based[1]	Graph NNs[2]	Transformers[3,4,5]
+ Precise	+ Intuitive	+ SMILES only
- Slow	- Existing models	+ General in scope
- Limited	are not general	- Omits conditions

Can Transformers Predict Borylation?

We prepared a BORON1000 dataset containing aromatic iridium-catalysed C–H borylation reactions. We then investigated how the model performance depends on the task at hand (Figure 4).

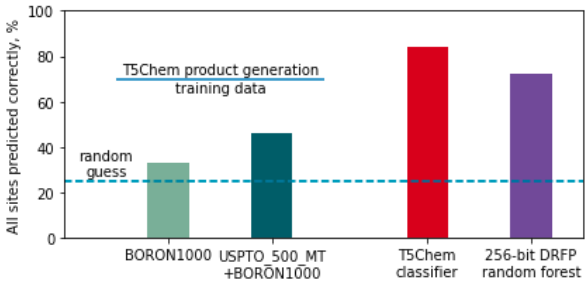


Figure 4. Prediction accuracy (%) for different T5Chem tasks trained on BORON1000.

It appears the accuracy of molecular generation does not exceed 50% but can be improved with training on unrelated reactions (see Figure 5 for successful predictions). However, using the same language model to classify the reaction sites resulted in encouraging 84% accuracy showing the model has learned the molecular context for C–H borylation. Curiously, a combination of 256-bit DRFPs[6] of the same data with a random forest classifier resulted in 72% accuracy, outperforming any molecular generation task without any hyperparameter tuning.

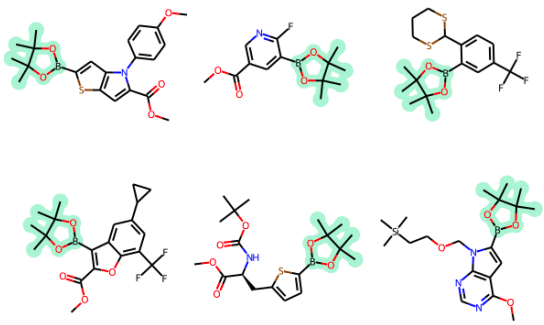


Figure 5. Examples of correctly generated products.

References

1. Caldeweyher et al. *ChemRxiv* 2022 DOI: 10.26434/chemrxiv-2022-7qw68
2. Nippa et al. *ChemRxiv* 2022 DOI: 10.26434/chemrxiv-2022-gkxm6-v2
3. Schwaller et al. *ACS Cent Sci* **5**, 1572 (2019)
4. Pesciullesi et al. *Nat Commun* **11**, 4874 (2020)
5. Lu et al. *J Chem Inf Model* **62**, 1376 (2022)
6. Probst et al. *Digital Discovery* **1**, 91-97 (2022)