

Leveraging Language Model Multi-Tasking to Predict C–H Borylation Selectivity

Ruslan Kotlyarov¹, Konstantinos Papachristos², Geoffrey P. F. Wood², Jonathan M. Goodman^{1*}

¹ Yusuf Hamied Department of Chemistry, Lensfield Road, Cambridge CB2 1EW

² Exscientia plc, The Schrödinger Building, Oxford Science Park, Oxford OX4 4GE

* Email: jmg11@cam.ac.uk

C–H borylation is a high-value transformation in the synthesis of lead candidates for the pharmaceutical industry thanks to a wide array of downstream coupling reactions available. However, predicting its regioselectivity, especially in drug-like molecules that may contain multiple heterocycles, is not a trivial task. Using a dataset of borylation reactions from Reaxys® we explored how a language model originally trained on USPTO_500_MT, a broad-scope set of patent data, can be used to predict the C–H borylation reaction product in different modes: product generation and site reactivity classification.

Our fine-tuned T5Chem multi-task language model can generate the correct product in 79% of cases. It can also classify the reactive aromatic C–H bonds with 95% accuracy and 88% positive predictive value, exceeding purpose-developed graph-based neural networks.

Introduction

Late-stage functionalisation (LSF) of C–H bonds is an important approach to lead compound development in the pharmaceutical industry^{1–3}. LSF can be used both for fine-tuning the structure of a lead and supporting extensive structure-activity relationship studies. The preference for C–H bonds is both an advantage, due to their ubiquity, and a drawback, due to the difficulty of differentiating similar bonds. While for simple cases it is possible to derive a set of heuristics for site-selectivity, the presence of multiple competing factors necessitates more complex models.

Iridium-catalysed C–H borylation is an example of such reaction. Its products are safe to handle and can undergo a wide range of cross-coupling reactions, forming C–C bonds via Suzuki-Miyaura reaction^{4,5} or connecting to heteroatoms via Chan-Lam-Evans coupling^{6–8}. This makes organoboronates ideal candidates for streamlined modular drug candidate synthesis.

The mechanism of the reaction has been determined (Figure 1) and can be used to explain the selectivity.⁹ For aromatic compounds, the rate-determining step is an irreversible oxidative addition¹⁰ to the C–H bond in a substrate, and the reaction proceeds faster at sterically unencumbered acidic C–H bonds. Studies of heterocyclic substrates¹¹ helped to derive a set of guidelines for borylation selectivity: the reaction avoids taking place next to ortho substituents or basic nitrogen atoms, electron-deficient heteroarenes react faster than arenes, and 5-membered heterocycles are preferred to 6-membered heterocycles, probably due to steric factors.

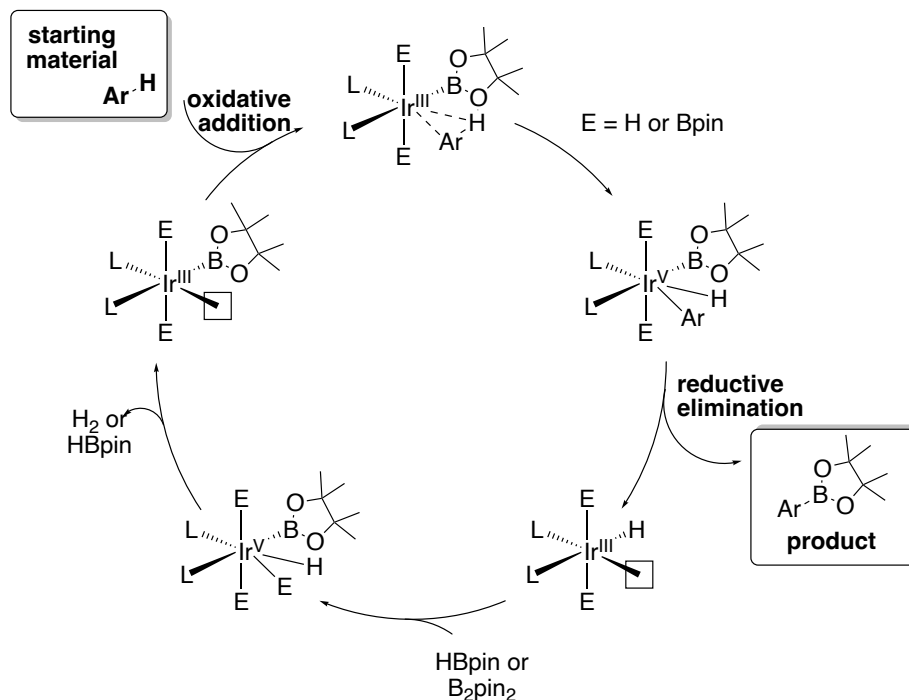


Figure 1. Mechanism of undirected iridium-catalysed borylation (resting state omitted)

The complexity of the competition between electronic and steric factors rapidly increases with the size of the molecule. For example, if there are multiple aromatic rings, which one would react, and which bond of this ring would be preferred? For complex systems, simple heuristics are insufficient and a different approach is needed.

Prior art

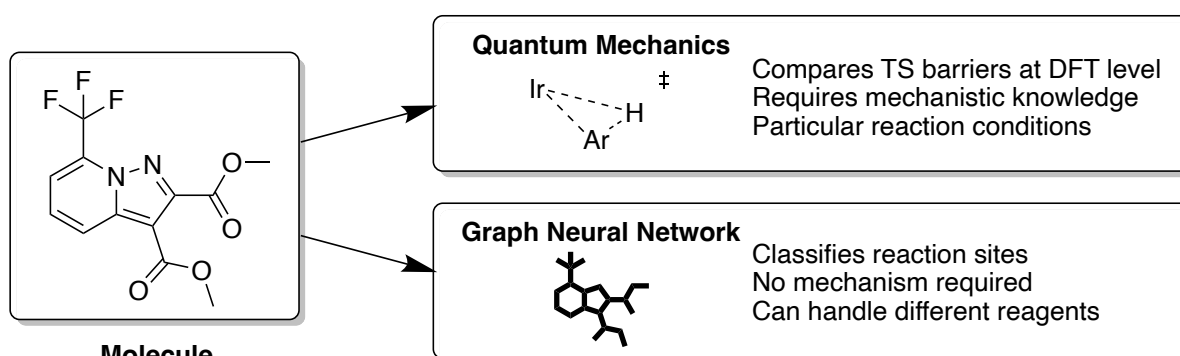


Figure 2. Overview of existing approaches for regioselectivity prediction

Quantum Mechanics-based models

Ab initio quantum-mechanics modelling has been used in elucidation of the catalytic cycle of C–H borylation^{10,11}. Calculating the reaction barriers for each position is an effective way to determine the selectivity, including the stereoselectivity, but it is also computationally expensive because of the variety of catalytic pathways, active species, and solvent effects coupled with the effort required to find each transition state¹². Performing such calculations for every possible reaction site may end up uncompetitive to running an experiment. It is, therefore, necessary to consider approximations, simplifications and alternative methods to speed up selectivity prediction.

Noting that the selectivity for iridium-catalysed borylation is controlled at the oxidative addition step (Figure 1), researchers from AstraZeneca and UC Berkeley have built the hybrid SoBo model¹³ to predict relative barriers for each position (Figure 2, top right). The model uses a transition state for benzene pre-optimised at the density functional theory (DFT) level and uses a semi-empirical quantum mechanical method to get the approximate barrier heights. The predictions are refined further using a combination of two correction terms. The first term is the neighbour penalty, which estimates steric bulk in the *ortho* position next to the reaction site. The second term is a partial least squares regressor which models the local chemical environment. Depending on the similarity to the training set, these two correction terms are combined dynamically. Using the SoBo model, the authors found the prediction can be generated in minutes using a desktop computer as opposed to hours on a high-performance cluster for a traditional DFT

calculation. The model does not take the absolute barrier height into account, which means it cannot predict if the reaction is fast or slow.

Graph Neural Networks

The connectivity of a molecule can be represented as a 2D graph, and additional features, such as bond lengths, can be incorporated into 3D graphs. Within the network, the atoms and their connections are associated with a set of features including atom types, ring membership, aromaticity, and atom hybridization, which are commonly known as embedding vectors. These features can be updated with features of neighbouring atoms or connections. After several iterations, each expanding the number of atoms which influence each feature, the model should be able to take long-range interactions between the atoms into account. A graph-based approach for C–H borylation has been developed by researchers from Roche, LMU, and ETH¹⁴ (Figure 2, bottom right). To model regioselectivity, an atomistic graph neural network (aGNN) architecture was employed, which represents the borylation substrate as a molecular graph.

This model was applied to a selectivity task: which non-quaternary carbon atoms are reactive? While the lowest accuracy model, aGNN2D, which used only 2D information, gave 88% accuracy, the F-score was only 38% with true positive at 30%, demonstrating how the class imbalance distorts the metrics. The use of 3D structures to initialise the graphs (aGNN3D) improved the accuracy to 90%, and the true positive rate improved to 56%, demonstrating that the model was now much more effective. Augmenting the graphs with DFT-accuracy¹⁵ Mulliken partial charges for each atom (aGNN2DQM, aGNN3DQM) had a negligible impact on the metrics.

The approach was restricted to carbon (C), hydrogen (H), oxygen (O), nitrogen (N), sulfur (S), phosphorus (P), and the halogens. The site-level accuracy metrics (whether the reactivity of a C–H bond is predicted correctly) do not reflect the accuracy of an overall molecular-level prediction: what is the major reaction site for the molecule. The molecular-level prediction is probably the standard use case and so it should be addressed while evaluating site classification models.

Transformer models

Molecular transformer models leverage developments in natural language processing which make it possible to translate one language into another.¹⁶ This technology has been applied to ‘translating’ reactants into products. Molecules can be represented as lines of text using SMILES¹⁷, DeepSMILES¹⁸, SELFIES¹⁹, or other methods, and the relationships between the structures are modelled using the self-attention mechanism which is described in detail elsewhere^{20,21}. Other applications of these models include predicting yields²² or reaction class²³, keeping track of atoms in a chemical reaction²³ (atom mapping), and identifying the active sites in enzymes²⁴.

Transformers may contain an encoder module, a decoder module, or both. The encoder converts the text input into a context-dependent embedding, i.e., an internal vector representation which takes the relationships between the neighbouring tokens into account. The decoder module generates new tokens from this embedding. If the process has been trained on the SMILES representation of a molecule (see Figure 3) then the output should be the SMILES string of a new molecule.

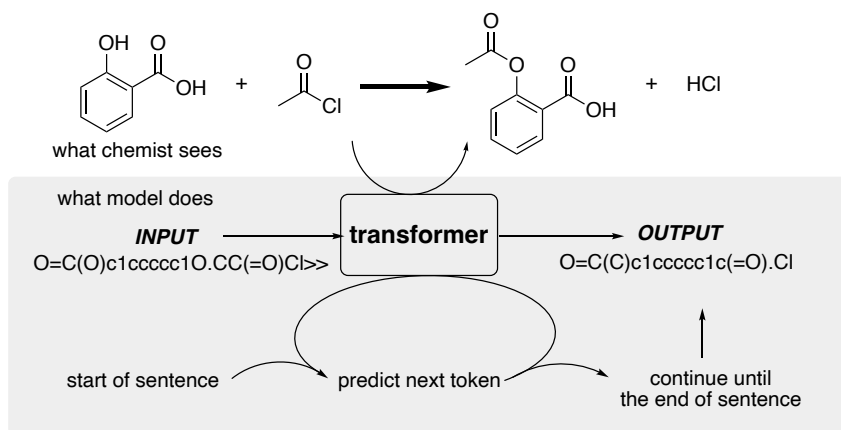


Figure 3. Autoregressive conditional generation of the reaction product by encoder-decoder transformers using character-level tokenisation

The development of the encoder-decoder transformer architecture¹⁶ allows us to treat reaction prediction as a translation task, generating products based on reactants and reagents provided. Molecular Transformer²⁵ by Schwaller et al. was the first model of this kind to predict reaction products. By reversing the translation direction, the model was successfully repurposed to retrosynthesis tasks²⁶. Subsequent development²⁷ improved the prediction quality both for forward and retrosynthesis, enhancing performance for scarce data^{28,29} and increasing the diversity of the possible retrosynthetic disconnections^{30,31}.

For our study, we chose to use T5Chem³², a multi-task encoder-decoder model (Figure 4). In addition to SMILES generation as in Molecular Transformer, it can also assign a reaction class and predict a yield using task-specific output layers, known as “heads”, on top of a common encoder-decoder module. In contrast with other encoder-decoder transformers, the authors chose to use primitive character-level tokenisation (e.g., ‘Cl’ corresponds to two tokens and ‘[C@@H]’ corresponds to six tokens) rather than the regular expression-based atom-level tokenisation

proposed by Schwaller³³. The reduction in vocabulary size led to a higher prediction accuracy despite the increased number of tokens in a sentence.^{28,32}

The T5Chem model is available pre-trained on SMILES for molecules encountered in PubChem and tokenized at the character level. This improves the model performance on the downstream prediction tasks despite using less task-specific data, as the model appears to have learned the representation of a molecule. This allows us to use the model for several purposes, thus saving on computational resources.

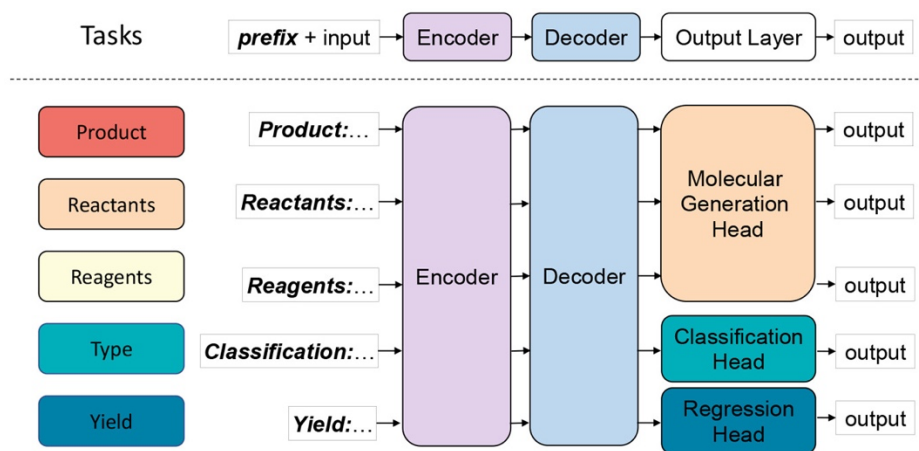


Figure 4. Overview of T5Chem model architecture. Reprinted with permission from *J. Chem. Inf. Model.* 2022, 62, 6, 1376-1387. Copyright 2022 American Chemical Society.

Data Curation

Data was compiled from Reaxys^{®34} and used in this study as provided by Elsevier Limited under license. A naïve search for a C–H bond in a reactant and a C–B bond in a product resulted in roughly 500 000 transformations. However, many of these are not associated with the C–H borylation of interest (see SI Figure 1). To further curate our dataset, we fragmented the products along the C–B bonds and checked if the fragment structures matched the reactant. That left us with around 20 000 reactions, out of which only about 12 000 had all the reaction species identified by PubChem. This is necessary, as conversion from structure to SMILES representation is done using

the PubChem record. Among these, only 4105 had associated yield data and only 1041 involved iridium-catalysed aromatic borylations. This is comparable to the 1300 reaction set assembled from the literature keyword search of SciFinder reported by Nippa et al¹⁴. The resulting set was termed **BORON1000**.

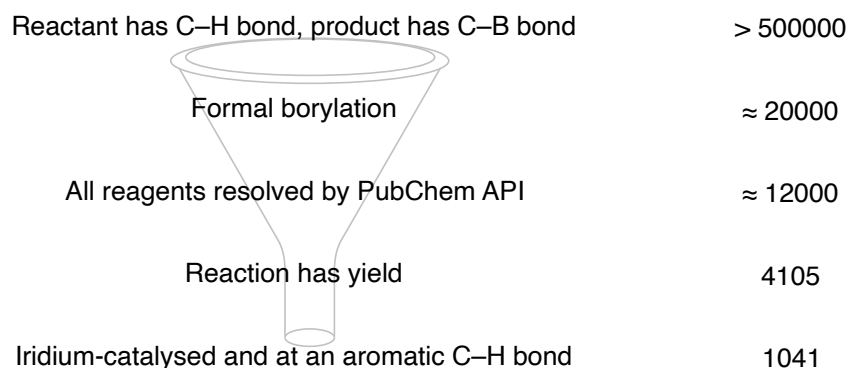


Figure 5. Overview of filtering for the **BORON1000** dataset generation

BORON1000 is limited to aromatic borylations and, while aromatic motifs are common in drugs, further refinement and expansion of the dataset are required to capture advances in other borylation classes, including sp^3 -rich substrates.

Enumerating rings in the substrates demonstrates the prevalence of benzenes, with thiophenes, pyridines, indoles, and quinolines also abundant (Figure 6). Such motifs are also found in drug molecules which should make the model applicable to lead development³⁵.

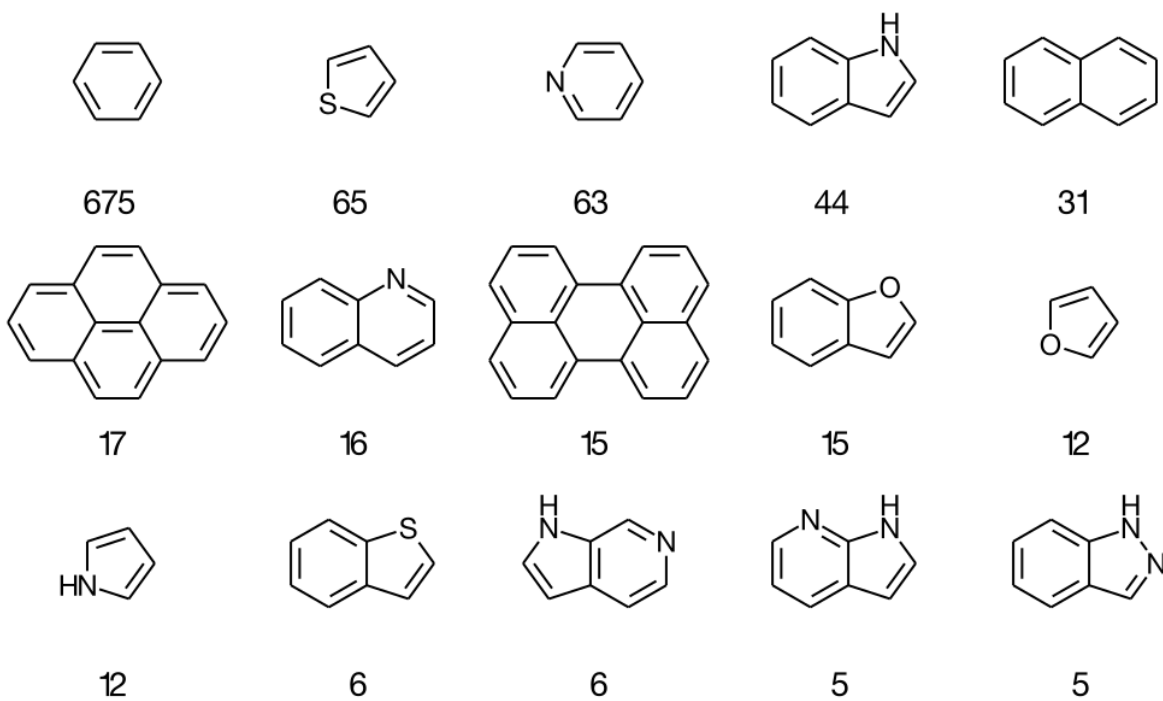


Figure 6. Fifteen most common reacting aromatic systems in **BORON1000**.

Table 1. Breakdown of **BORON1000** data by type of the reactive aromatic system and number of aromatic rings in a substrate

BORON1000	1 aromatic ring	>1 aromatic ring	sum
reactions at carbocycles	530	315	845
reactions at heterocycles	114	82	196
sum	644	397	1041

Training details

The models were trained on GeForce RTX 3080 for 100 epochs unless specified otherwise. From the original publication, the batch size was reduced from 32 to 16, and the initial learning rate was reduced from $5e-4$ to $2.5e-4$ accordingly. Character-level tokenisation was employed to take

advantage of available pre-trained models. For molecular generation, the model was set to return five highest-probability predictions with a beam search of width ten, so that the ten most probable predictions so far are kept during the prediction with the five most probable retained for later.

This study investigates three approaches to the C-H borylation selectivity problem: (A) product SMILES prediction; (B) reaction site classification; (C) yield prediction. The next three sections of the paper goes through these in order.

(A) C-H borylation selectivity analysis by product SMILES prediction

Model development and evaluation: T5Chem models for product SMILES prediction task

We set out to study how well the T5Chem³² model predicts the borylation products. The metric T5Chem molecular generation employs is a top- k accuracy which reflects if the first k predictions contain a correct answer. In this model, we use RDKit³⁶ to turn the SMILES representations of the predicted product into canonical SMILES. For the correct answer, these must be identical to the experimental results. For **BORON1000**, we found the median number of aromatic C–H bonds is equal to four so random guessing of the reaction site has a 25% chance of being correct.

We took advantage of the existing pre-trained models³⁷ which were supplied alongside the GitHub repository for the T5Chem model. The first model, **pretrain 1** (denoted as *simple* in the original manuscript), was pretrained on the SMILES representations of molecular structures encountered in PubChem, using masked language modelling.³⁸ The model was given SMILES with one character randomly masked and trained to predict the missing token. This helps the model learn the syntax of SMILES.

The **pretrain 1** model was trained further on USPTO_500_MT^{32,39} reaction SMILES data as **pretrain 2** (denoted as *USPTO_500_MT* in the original manuscript) in mixed mode⁴⁰, so that it could perform product, reagents, and reactants prediction.

USPTO_500_MT is a subset of USPTO 1k TPL dataset²³ containing reactions corresponding to 500 most frequent reaction templates and was developed by Lu and Zhang to test how well the T5Chem architecture would handle training for multiple tasks.

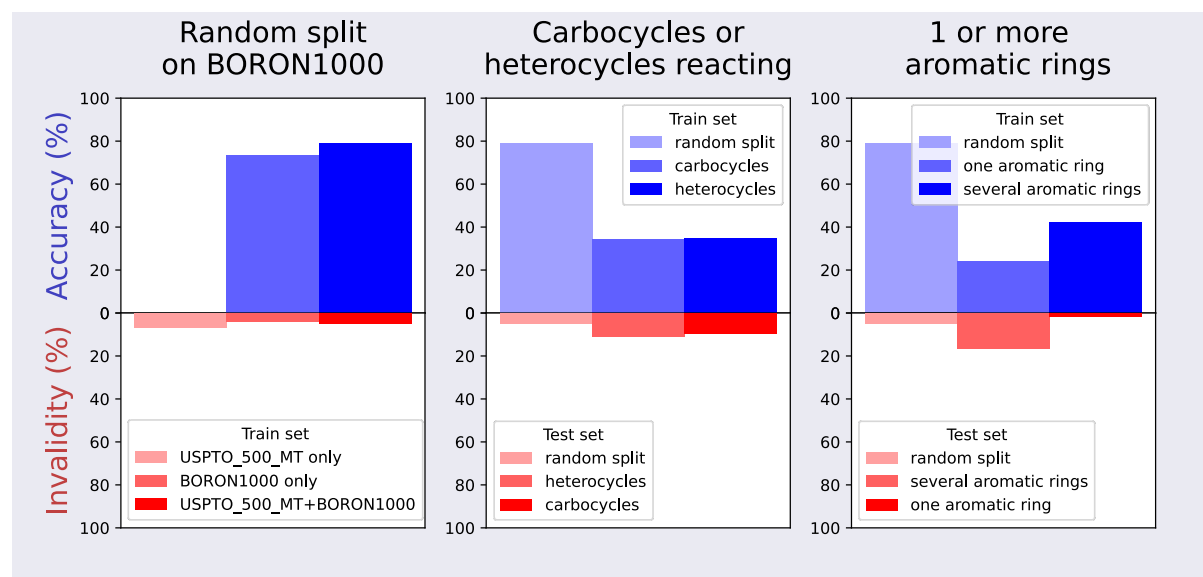


Figure 7. Top-k prediction accuracy for the models pre-trained from **pretrain 2**.

To establish a baseline, we tested **pretrain 2**, which is pre-trained on patent data on the **BORON1000** dataset (Figure 7, left). 93% of predictions were syntactically valid, demonstrating that **pretrain 2** has sufficient information about SMILES to generate reasonable molecules. This model was pre-trained on the USPTO_500_MT dataset which contains no iridium-mediated borylations. As a result, none of the top predictions corresponded to the products in the test set.

Encouraged by the level of valid molecules that were generated, we further trained the model using borylation data.

The **pretrain 1** model was further trained for 100 epochs on borylation data to generate a new model, **finetune 1**. This showed 73% top-1 accuracy (the top molecule was correct), which is a major improvement over the base case. Training the **pretrain 2** model for 100 epochs on borylation data **BORON1000** allowed the model **finetune 2** to generate the correct product structure as the most probable in 79% of cases. The model appears to benefit from further training on translation tasks as it is getting better conditioned for output generation through exposure to common structural changes in the reactions.

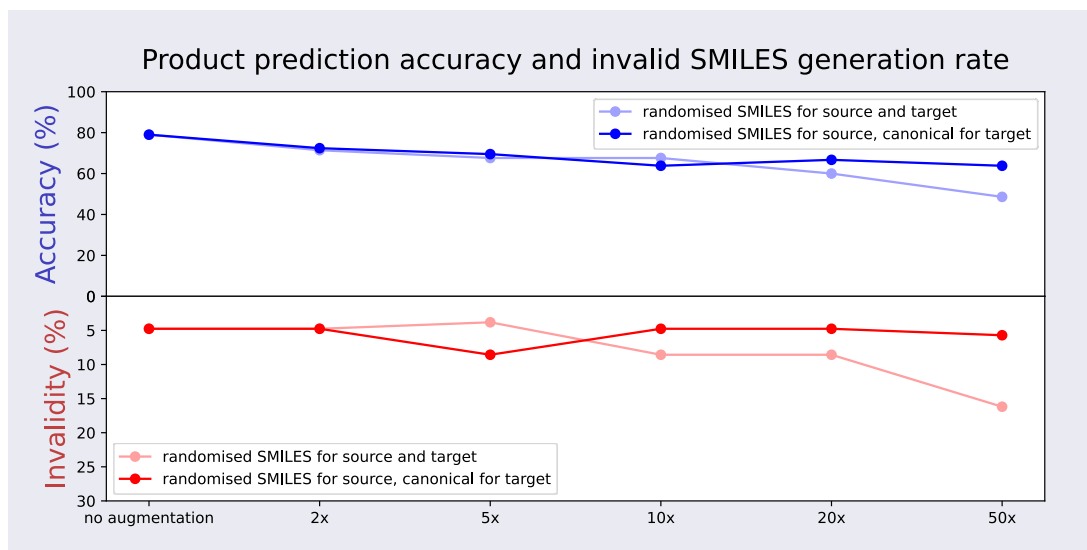


Figure 8. Impact of **BORON1000** dataset augmentation on Top-1 product prediction accuracy. The accuracy and the proportion of syntactically valid SMILES both decrease with the extent of reaction SMILES augmentation in the training set.

Studies into generative language models have shown training on multiple SMILES representations of the same molecule improves the quality of generated SMILES^{41,42}. However, when trained from

pretrain 2, we found the accuracy of borylation product prediction decreases upon augmentation of the **BORON1000** dataset (Figure 8), especially when the target SMILES sequence was made non-canonical as well. We did not add augmentation to the training data for our models, therefore.

Scaffold-based cross-validation

In general, random splitting of the dataset for model evaluation is not a reliable method for assessing its performance, as it may lead to an overestimation of the model's accuracy.

Table 2. Subsets of **BORON1000** used in cross-validation

Subset name	substrate features
BORON1000_HET	reacts at a heterocyclic ring
BORON1000_CARB	reacts at a carbocyclic ring
BORON1000_ONE	has one aromatic ring
BORON1000_MULT	has several aromatic rings

To gauge the extrapolative power of the T5Chem model, we split **BORON1000** into sections, summarized in Table 2. **BORON1000_HET** contains reactions of heterocycles only, and **BORON1000_CARB** is only carbocycles. We then trained the **pretrain 2** model on reactions at heterocycles and evaluated on reactions of carbocycles, and vice versa to yield **trained_on_heterocycles** and **trained_on_carbocycles**.

The **trained_on_heterocycles** and **trained_on_carbocycles** models still generate syntactically correct SMILES, albeit at a lower rate, but the Top-1 accuracy plummets from 79% to 34% for either split (Figure 7, centre).

To assess how well the architecture performed in predicting molecules with different levels of structural complexity than those in the training set, we re-split the dataset into reactions of molecules with one aromatic ring **BORON1000_ONE** and multiple aromatic rings **BORON1000_MULT**. We then trained the **pretrain 2** model on these sets in the same manner as above to obtain **trained_on_1** and **trained_on_multiple**.

As the model **trained_on_1** extrapolated to the molecules containing multiple rings, the results showed a 24% top-1 accuracy with 16% of the predictions being syntactically incorrect. Testing the model trained on polyaromatic molecules **trained_on_multiple** on molecules with one aromatic ring resulted in 42% accuracy with just 1.5% syntactically incorrect predictions (Figure 7, right). All of these new models are listed in Table 3.

Table 3. Summary of T5Chem models trained for product SMILES generation.

Model	Tokenisation	Trained from	trained on
pretrain 1	character-level		PubChem SMILES (masked LM)
pretrain 2	character-level	pretrain 1	USPTO_500_MT
finetune 1	character-level	pretrain 1	BORON1000, random split
finetune 2	character-level	pretrain 2	BORON1000, random split
trained_on_heterocycles	character-level	pretrain 2	BORON1000_HET

trained_on_carbocycles	character-level	pretrain 2	BORON1000_CARB
trained_on_1	character-level	pretrain 2	BORON1000_ONE
trained_on_multiple	character-level	pretrain 2	BORON1000_MULT

Comparison of product generation by T5Chem methods and by mechanistic analysis

The **finetune 2** model's lower probability predictions illustrate how the internal representation is gathering the key features of the transformations through the lens of the data that trained it. Changes that look quite dramatic to a chemist with a knowledge of organic synthesis appear to be less significant to the model, as Figure 9 illustrates.

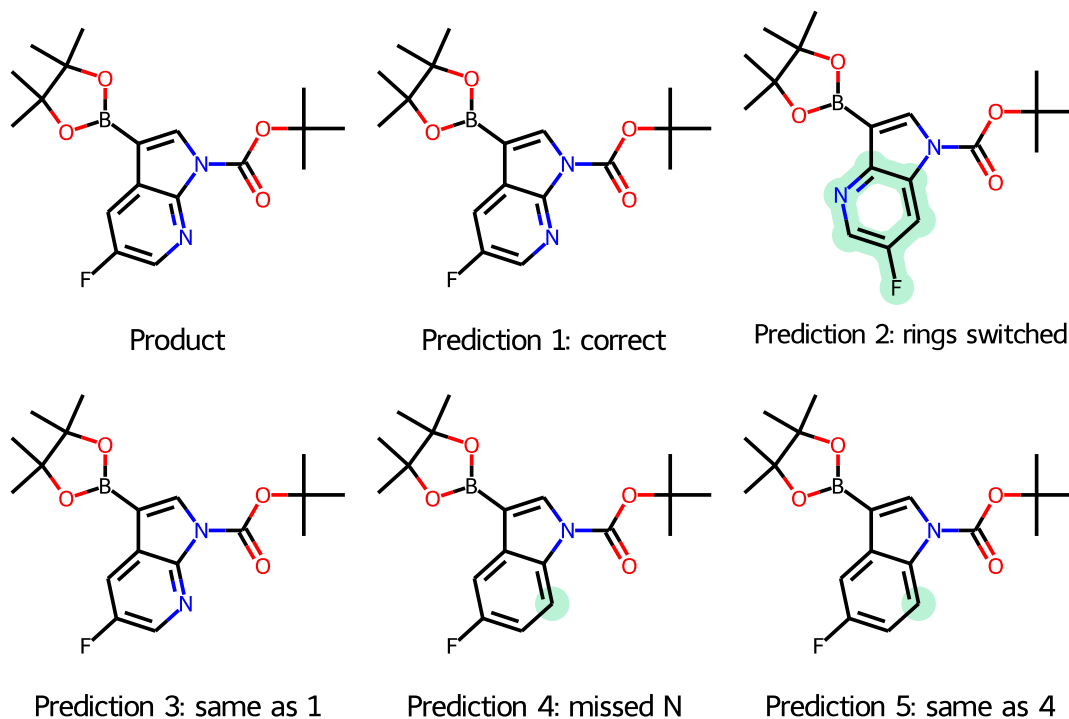


Figure 9. An example of predictions for a borylation product.

For example, moving a nitrogen within a pyridine ring, or changing a pyridine to a carbocycle, are difficult synthetically, but are small changes in terms of a SMILES string. Branch transposition also appears probable to the model and may well be synthetically challenging. This information, which is "obvious" to humans, is not part of the training set for the model. Fortunately, this extra information can readily be added at the post-processing stage by checking for substructures in the predicted products. The molecules with major structural changes may be readily filtered out using the starting material as an RDKit substructural filter, because any substrate of C–H activation is considered a substructure of a product.

We suggest that reduction in performance for complex structures as shown by application of model **trained_on_1** to the **BORON1000_MULT** molecules may be a consequence of autoregressive generation since the probabilities for the next token are dictated by input and output generated so far. For example, a model trained exclusively on structures with one ring, when extrapolating to molecules with multiple rings, having generated one ring, would assign a low probability of generating a token to open another ring, let alone a matching character to close it.

It appears the model **trained_on_multiple**, that was trained on more complex substrates, can extrapolate to simpler molecules of **BORON1000_ONE** despite fewer training points available. Exposure to complex examples of **BORON1000_MULT** allows for a more robust generation of SMILES strings but not greater accuracy. This demonstrates the importance of a representative training set as the models did not extrapolate out of training data distribution well.

The unusual negative impact of augmentation may be caused by the deterioration of the generative capacity of the model. While augmentation by generating multiple distinct SMILES for each molecule can make molecular representation more robust, as the model is exposed to several representations of the same structure, it may also erode the model confidence during generation, since during the training the T5Chem model text generation is evaluated using the cross-entropy loss function. The model loss is minimal if the generated SMILES is identical to the target SMILES, but the generation of different SMILES representing the same molecule is penalized. Further augmentation increases this problem.

The compiled T5Chem model only supports atom-wise and character-wise SMILES tokenizer. Whilst it would be interesting to explore larger token sizes, which could encode functional groups and other common molecular patterns in a single token⁴³, it was not practical to implement this, as models require complete re-training with each new tokenizer. Atoms with two-letter symbols, such as chlorine, are represented by two tokens rather than by one, which may be counterintuitive. However, in an independent study, the use of data-driven tokenisation was shown not to bring about an improvement in molecular generation accuracy⁴⁴. We decided, therefore, to focus on testing the character-wise tokenizer.

Comparing the generative model with quantum mechanics calculations

For a comparison with the prior art, the model **finetune 2** was tested on the validation set of six pharmaceutical intermediates from the SoBo model paper¹³. The **finetune 2** model correctly

predicted the products for two molecules out of six (Figure 10). For the four erroneous predictions, the model has either predicted a wrong site or no reaction. The correct answer was in the top five predictions for all six molecules, and in the top two for four out of six. The survey undertaken by the authors of the SoBo paper suggests our model performs at least on par with an average synthetic chemist. Considering the complexity of the mechanism (Figure 1) and the potential for diverse features of the process to control the outcome, it is remarkable that **finetune 2**, which has no direct knowledge of the mechanism, can be so effective.

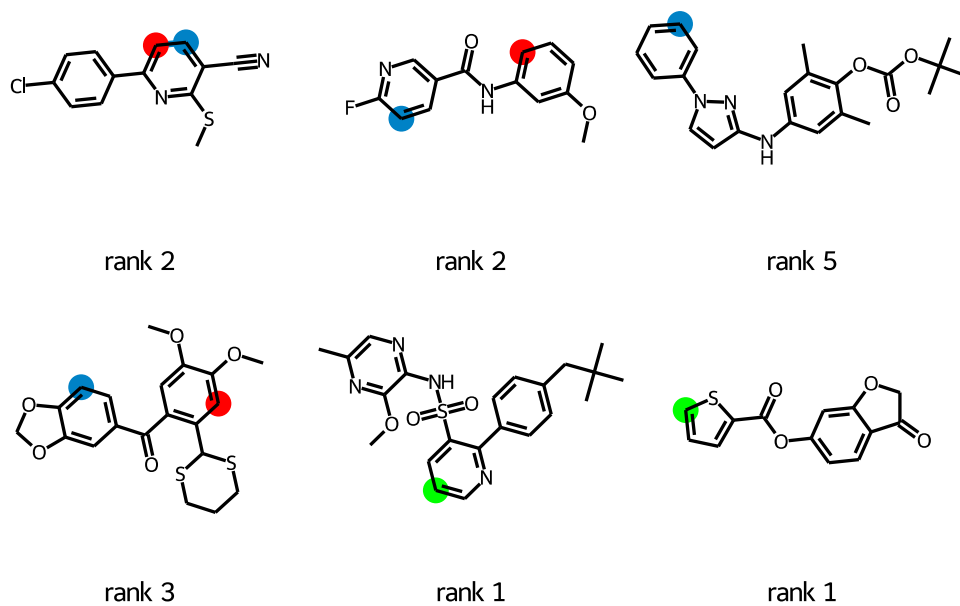


Figure 10. T5Chem is generating predictions for the SoBo validation set of molecules¹³. The T5Chem predictions are shown in red, while the experimental results are shown in blue. The green colour indicates a match between the two. For each molecule, the rank of the experimental outcome (ground truth) returned as a prediction by the T5Chem model is also displayed.

(B) C-H borylation selectivity analysis by reaction site classification

Are we asking the right question in the right way?

The generative model in the previous section is doing two different things: first it generates an internal representation of the input molecule in a form which may be suitable for chemistry-related tasks; second it generates a new molecule based on this information. Even if the first step were performed perfectly, the second step could introduce uncertainty and inaccuracy into the model prediction. If the T5Chem language model has an accurate, or reasonably accurate, internal representation of SMILES suitable for chemistry-related tasks, it should be possible to get useful information from this without going through the process of generating a new molecule as the output. Token classification of molecular SMILES should be able to point out the reactive atoms based on the SMILES of the molecule alone using encoder-only models⁴⁵. However, the encoder-decoder T5Chem model is typically used to predict a singular output such as reaction class or yield and requires adaptation to predict atomic properties such as regioselectivity. Therefore, if we want to predict reacting atoms rather than whole molecules, we must re-formulate the question.

Site selectivity via classification

We can treat the borylation reaction as an ensemble of reactions for each aromatic C–H bond. All possible monoborylation products can be enumerated and compared to the experimental outcome. Each of these possible reactions are either put into class 1 (reactive) or class 0 (unreactive, Figure 11). This approach has the advantage that negative reactions are specified explicitly, which should improve learning effectiveness since adversarial examples are now available. In addition, if all sites are classified as non-reactive, we can conclude that the reaction does not happen at all. The output of the model is list of all the reactive sites of the input molecule.

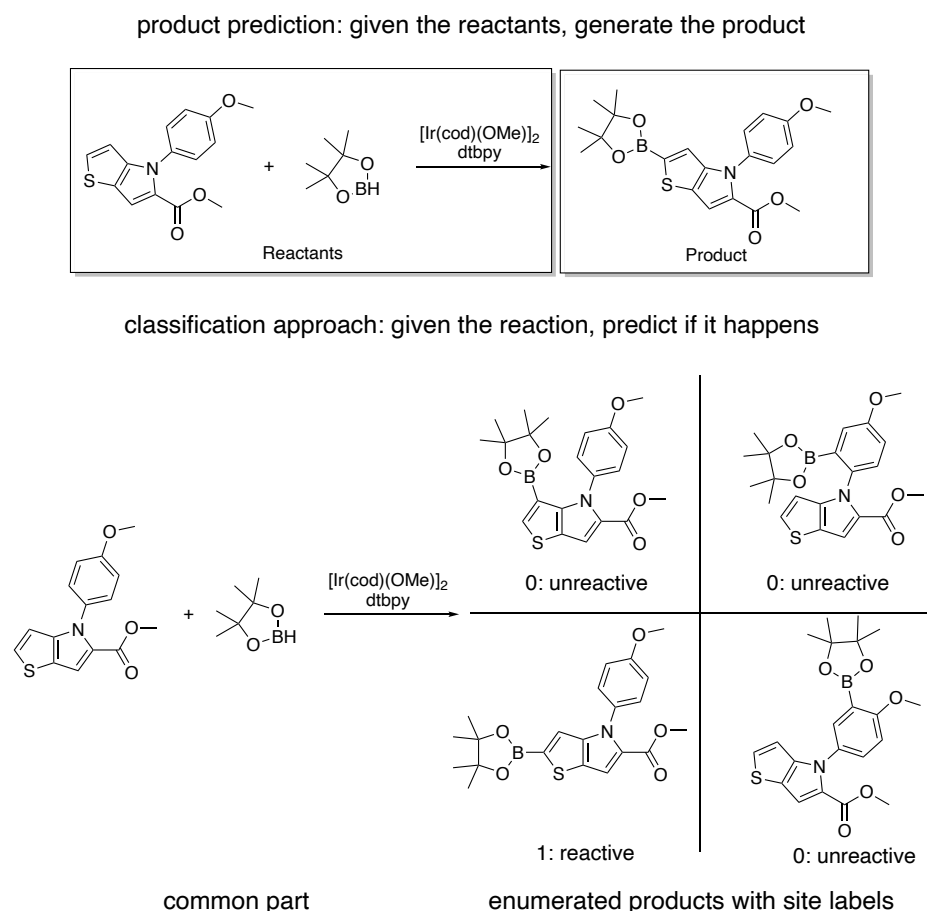


Figure 11. For C–H activation, product prediction may be expressed as a site-classification problem. For the T5Chem model, this can be implemented by enumerating all possible monoborylation reactions and determining whether they take place.

The classification task shares most of its hidden states with the autoregressive molecular generation which we have explored previously. However, instead of producing a probability distribution for the next character across the vocabulary space (i.e., ‘C’, ‘O’, ‘1’), it outputs the probability distribution over the two classes (class 0 if the site is not reactive, class 1 otherwise). The model inference only runs once per reaction site, circumventing the demanding task of molecular generation.

Now that the problem has been reduced to binary site classification, it is possible to make meaningful comparisons with other site-classifying models. Due to an imbalance between reactive and unreactive sites, simple accuracy becomes an unreliable metric. We, therefore, choose to use Matthews' correlation coefficient (MCC)^{46,47} which has been successfully employed as a binary classification metric.

We have used the **pretrain 1** model (see Table 3), which was pretrained only on molecular SMILES from PubChem and no reaction data as a baseline. Using the same split as product prediction, we have obtained 95% accuracy in classification, with Matthews' Correlation Coefficient at 82%. This means the hidden representation of the molecules in the T5Chem model is sufficient for predicting the reactivity for each aromatic C–H bond in the molecule. Interestingly, using **pretrain 2** as a starting point brought no improvement to the classification accuracy, even though this model had been trained on reactions as well as on molecular structures. We did not try **finetune 1** or **finetune 2** because they had already been trained on the borylation data. Overall, the model predicted correct selectivity for 84% of **BORON1000** validation set molecules.

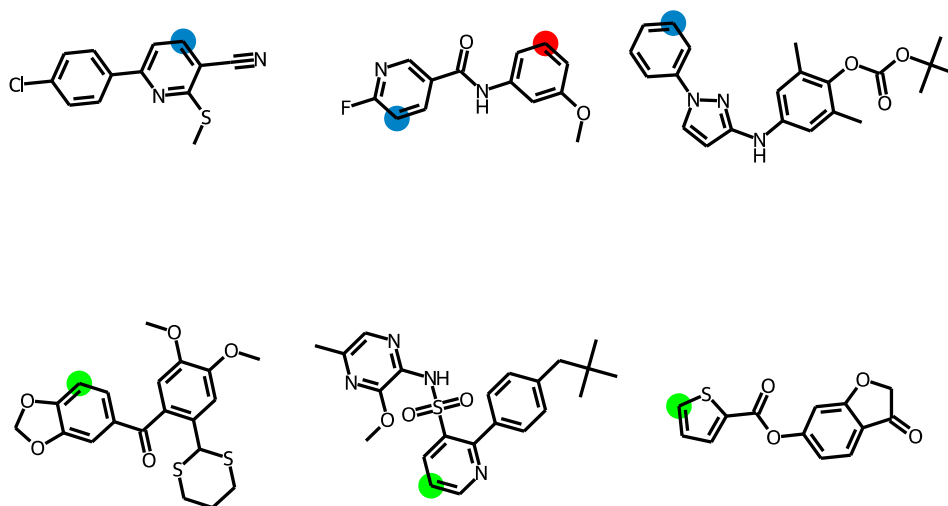


Figure 12. T5Chem Classifier is classifying reaction sites for the SoBo validation set of molecules¹³. The T5Chem predictions are shown in red, while the experimental results are shown in blue. The green colour indicates a match between the two.

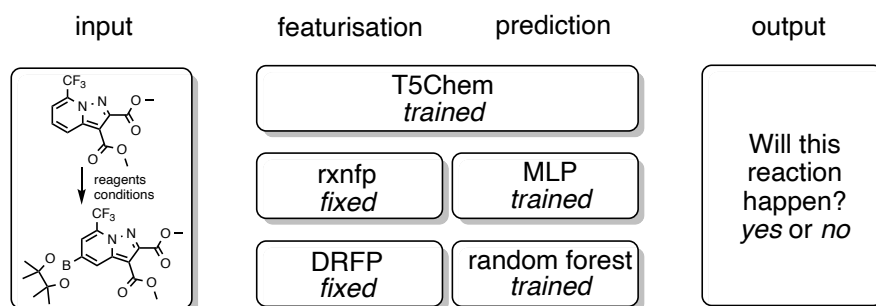


Figure 13. Simpler models were used to estimate the trade-off between T5Chem classifier complexity and performance.

Transformers are comparatively computationally expensive to train and evaluate. Therefore, a comparison with less sophisticated methods is required to justify their use (Figure 13). For our baseline, we selected the pretrained encoder-only transformer RXNFP²² model, which converts reaction SMILES into a feature vector of floating-point numbers and fitted a random-forest

classifier on top of it to translate this reaction encoding into a reaction site classification. The MCC was 44%, which suggested further fine-tuning was required. A multilayer perceptron-based classifier improved the Matthews' correlation coefficient to 61%. We also investigated knowledge-agnostic differential reaction fingerprints (DRFPs) which are based on a symmetric set difference of the SMILES representation of molecular features extracted by extended connectivity fingerprints of products and reactants.⁴⁸ The fingerprint is a bit vector (i.e., contains only 0s and 1s) and can be directly matched to the structural features it encodes. We generated 256-bit DRFPs of the same data and fit a random forest classifier using default hyperparameters. We achieved a 93% site classification accuracy with MCC at 79% despite the simplicity of the model. However, the correct reactivity pattern, i.e., all sites in the molecule are classified correctly, was reproduced only in 72% of cases. This demonstrates how a small reduction in the quality of the site classifier may dramatically affect the correctness of the prediction for the entire molecule, highlighting the need for improved accuracy.

We expected that the RXNFPs should be able to achieve better performance because they are more complex than DRFPs. However, the opposite is true. We suspect the discrepancy is due to the way they were obtained. RXNFPs were trained on a USPTO data set that does not contain iridium-catalysed borylations, while DRFPs are data agnostic. The embedding generated for an out-of-scope reaction might be of poor quality. On the molecular level, we got a 72% accuracy for the DRFP-based classifier, performing on par with the molecular generation by T5Chem using a simple model with features that directly map onto the structure of a molecule.

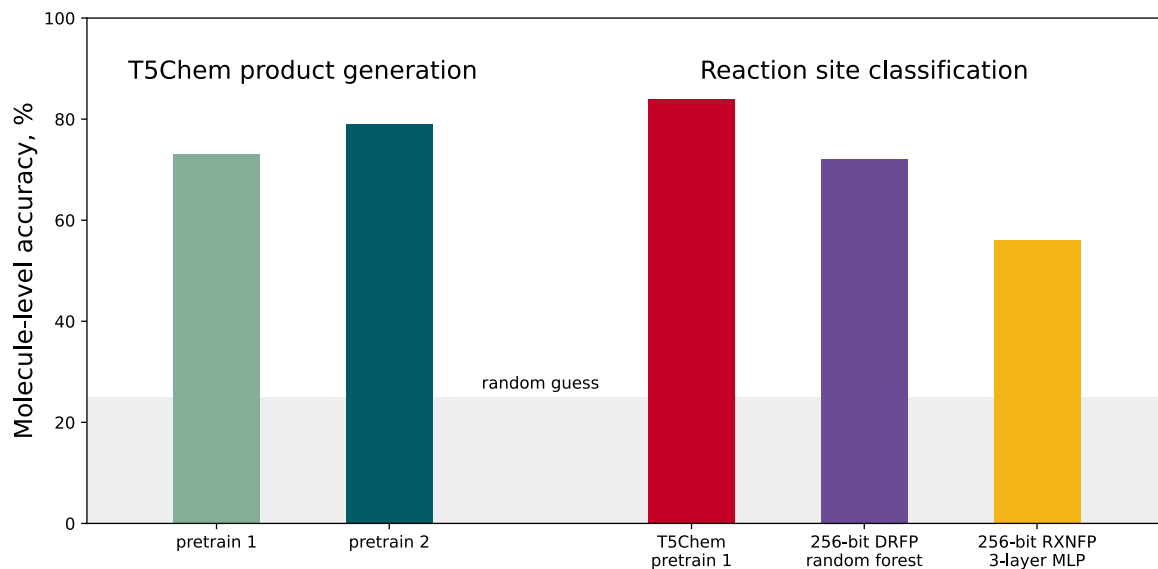


Figure 14. Molecule-level model performance after training and testing on **BORON1000** data. For the T5Chem product generation task, accuracy is the proportion of correct first predictions; for classification tasks, it requires all sites to be classified correctly.

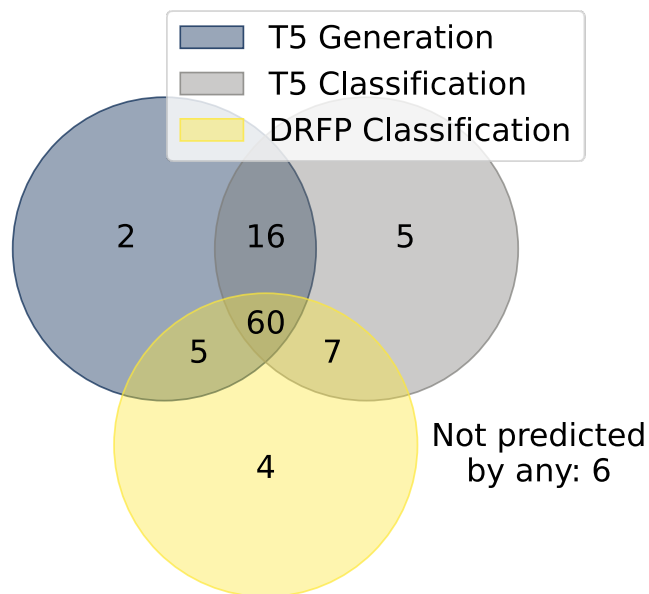


Figure 15. Venn diagram comparing number of the successful selectivity predictions for molecules of various methods on a held-out validation set of **BORON1000**. The set contains 105 molecules in total.

We compared the DRFP-based model and T5Chem approaches using the same validation dataset. We wondered whether the differences in model performance might be due to characteristics of the molecule, i.e., there would be molecules all models would predict correctly and molecules no model would predict successfully. The Venn diagram comparing these predictions (Figure 15) shows there are only six ‘hard’ molecules that no model could predict (Figure 16).

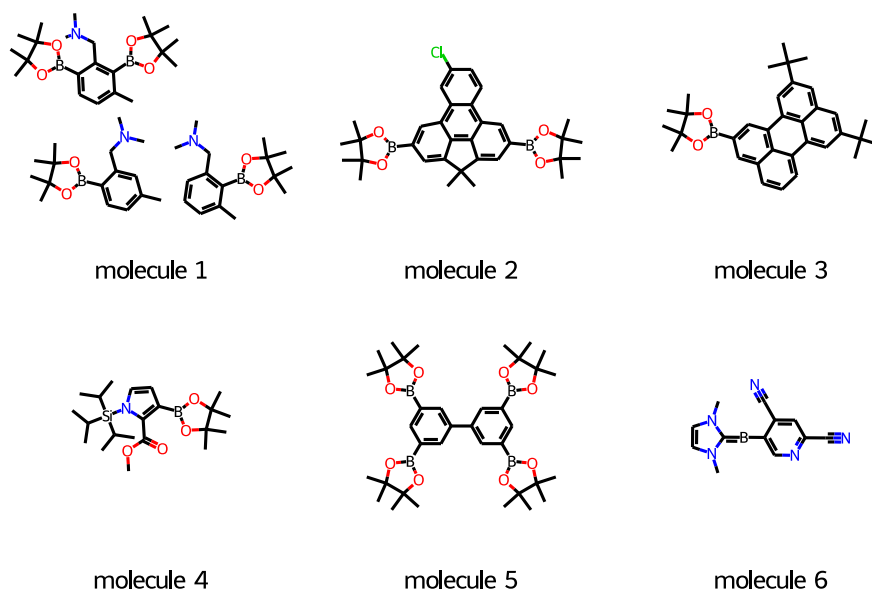


Figure 16. The six hard molecules that T5Chem Generator, T5Chem Classifier, and DRFP Classifier failed to predict correctly.

The models appear to struggle with mixtures of products (molecule 1), molecules containing condensed aromatic rings (molecules 2, 3, 5) or unconventional motifs like N-heterocyclic carbene fragment (molecule 6).

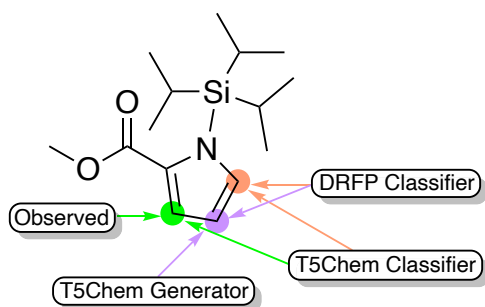


Figure 17. Comparison of site selectivity prediction using different approaches. The experimental borylation site is highlighted in green.

While Figure 18, molecule 4 looks easy to predict using expert-derived rules¹¹, the three models did not arrive at a consensus on which positions would react (Figure 17). While T5Chem Generator proposed a 4-position so that borylation happens at the least hindered site away from both ester and bulky *N*-triisopropylsilyl protecting group, the actual reaction has taken place at 3-position, presumably due to ester acting as a Lewis base for iridium. This suggests the T5Chem Generator may have learned the importance of common steric factors but not chelation.

The classifier models proposed two reactive sites each. The T5Chem Classifier suggested 3 and 5 positions, which matches taking chelation and selectivity for 2 and 5 positions of pyrrole. The DRFP Classifier suggested 4 and 5 positions, reflecting general steric trend and selectivity in pyrroles.

Table 4. Selection of site classification methods applied to **BORON1000** data.

	Site accuracy / %	PPV / %	MCC / %
T5Chem Classifier	95	87	87
DRFP + RF Classifier	93	94	80
RXNFP + RF Classifier	81	91	44
RXNFP + MLP	85	68	61

Comparison with existing site classification models

It would be interesting to compare the approach with other site-level regioselectivity predictors, namely graph neural networks developed by Nippa et al.¹⁴. We trained the **pretrain 1** model for site classification ten times using randomly split data from their study and found the site

assignment accuracy was $94\pm 1\%$, positive predictive value (precision) was $84\pm 5\%$, and F1 score was $82\pm 5\%$. This is a meaningful improvement over the aGNN3DQM performance reported by Nippa et al.: $90\pm 1\%$, $62\pm 2\%$, and $60\pm 4\%$ respectively. Despite seemingly close accuracy numbers, the multi-task T5Chem architecture demonstrates a greater precision than a purpose-built model, enabling greater confidence in its predictions. However, a simple combination of 256-bit DRFP and a random forest achieves the same result with considerably less resource.

Table 5. Predictions on Roche dataset. For consistency, the models were trained on a dataset prepared by Roche to classify all non-quaternary carbons, average of 10 random splits listed.

	Site accuracy / %	PPV / %	F1-score, %
aGNN3DQM ¹⁴	90 ± 1	62 ± 2	60 ± 4
T5Chem Classifier	94 ± 1	84 ± 5	79 ± 5
DRFP+RF Classifier	94 ± 1	95 ± 3	80 ± 3

(C) C-H borylation selectivity analysis by yield prediction

Yield prediction for site selectivity analysis

We modified our T5Chem classification approach for yield prediction (Figure 18). The R^2 of the regression was 0.75, the mean absolute error was 6, and the root mean square error was 17. The error metrics may not be accurate as the data are skewed due to the prevalence of non-reactive sites, and misclassified C–H bonds (e.g., yield estimated to be 0 % instead of 70 %). These may drive a dramatic error increase, especially for RMSE.

Using Kullback-Leibler divergence as the loss function, the T5Chem-based yield prediction reproduced the distribution of yields in the test set (Figure 20). To compare the model with the baselines, we have used the same sets of RXNFPs and DRFPs and fit a random forest regressor on top of those. We found that the RXNFP-based model does not reproduce the distribution well, with R^2 at 0.25. The DRFP-based model has achieved an R^2 of 0.67, outperforming an RXNFP-based model and reproducing the target distribution more accurately.

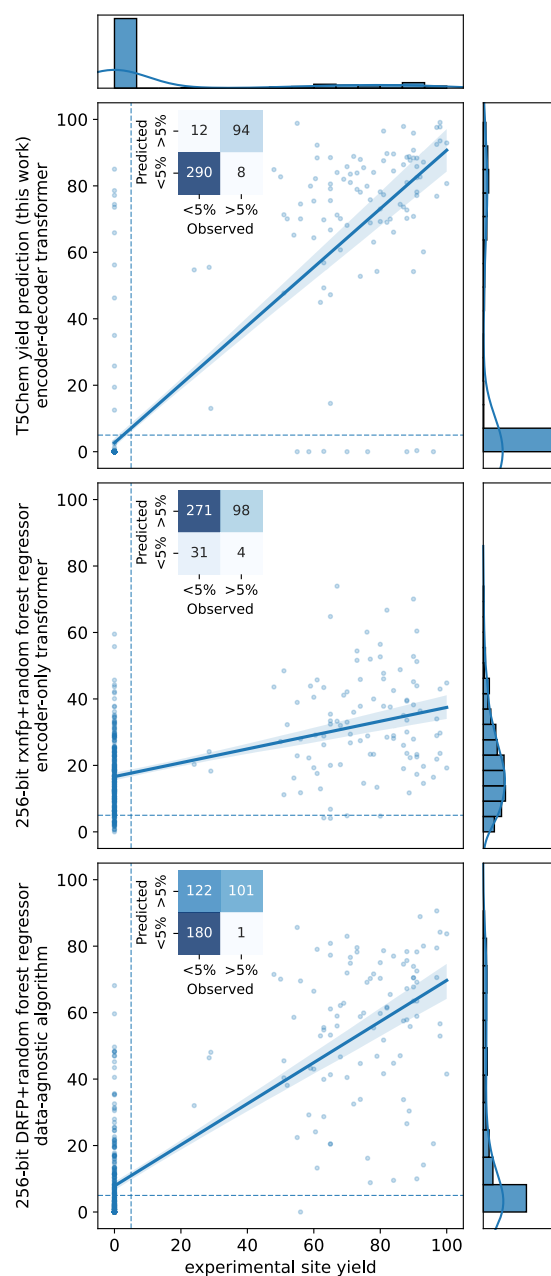


Figure 18. Parity plot for site-level yield predictions by T5Chem (this work), 256-bit RXNFP with random forest regressor and 256-bit DRFP with random forest regressor as two baseline models. The confusion matrix (top left of each plot) shows the high performance of T5Chem.

We set a classification threshold at the 5% yield and obtained an MCC value of 87% which is comparable to the classification approach. The consistent performance is reasonable to expect since the model shares all hidden states, except the task-specific heads, which both constitute a linear transform of the same feature vector (see Figure 4). Aggregating those predictions by molecule has shown 85% accuracy for a major product.

The classification and yield prediction heads both comprise a linear transform and output layer but produce the final output differently. The classification head assigns an arbitrary score to each class (in this case, class 0 if there is no reaction and class 1 otherwise) and returns the class with a higher score. The yield prediction head uses a soft label approach so that the outputs correspond to probability distribution between minimum label (0) and maximum (100) and are then normalized to produce the yield. We hope this will prevent model overconfidence by forcing it to consider the likelihood of an alternative outcome as well.

Table 6. Yield prediction accuracy by site-level regressor models.

model	R ²	Reaction outcome, MCC	Molecule-level accuracy, %
T5Chem	0.75	0.87	85
RXNFP+random forest	0.25	0.10	2
DRFP+random forest	0.67	0.51	39

While this approach can quantify the regioselectivity, we may be limited by reporting bias as reactions resulting in mixtures may not have all its products listed and yields omitted, as the lack of reactions with yields under 50% would suggest. Having a representative set of reactions would hopefully improve model performance on novel molecules.

Conclusions

Using a model based on T5Chem, we can treat reaction selectivity prediction within the same architecture in three distinct ways: generation of a product SMILES, reaction site classification, and site-wise reaction yield prediction.

SMILES generation is the most challenging, because molecules are generated from scratch and there is the potential to generate products which are completely unlike the starting materials. Whilst some distant products are generated, sensible products which show the expected reactivity are generated in 79% of cases. When tested on a validation set for SoBo¹³ model, it predicted selectivity for two out of six substrates correctly on par with a synthetic chemist with no expertise in borylation. Reaction site classification with T5Chem is a more straightforward task because it selects between the possible reactive sites of the starting material rather than generating a completely new molecule. These restrictions lead to it being more effective with 84% molecular-level accuracy at cost of universal reaction applicability. The model also performed better on the SoBo validation set, predicting selectivity correctly for 3 molecules out of 6, putting it above an average synthetic chemist. Predicting selectivity from T5Chem yield calculations also fits the data well with an R² score of 0.75 and can predict the reaction success (yield \geq 5%) on par with the T5Chem classifier.

The best model for predicting C-H borylation selectivity is the T5Chem site classification model. This works in the absence of detailed knowledge of the reaction mechanism (Figure 1). The model can be readily configured for use by people having no computational experience and trained in one command. Another advantage of fine-tuning the existing model is low resource demand: it only

takes about 20 minutes of consumer-grade GPU time to train the model on 1000 reactions, with predictions returned in seconds. When trained on the same dataset, the T5Chem classifier outperforms existing purpose-built graph neural networks¹⁴, with a higher F-score (78% vs. 55%) despite encoding no steric and electronic information about the substrate. Moreover, unlike quantum mechanics-based methods¹³, we are not restricted to a single reacting system and a singular reaction mechanism, allowing for greater flexibility in applying the model across a broad range of chemical systems and in a mechanism-agnostic fashion.

Overall, we believe it is now possible to predict selectivity for a complex reaction well enough to be helpful to many synthetic chemists, without any mechanistic knowledge or need for purpose-built models.

DATA AND SOFTWARE AVAILABILITY STATEMENT

A GitHub repository with the scripts employed for data processing is available at <https://github.com/ruslankotl/rxn-data-proc>

The repository also contains borylation regioselectivity dataset as prepared by Nippa et al.¹⁴ and Reaxys IDs for the reactions that went into **BORON1000** dataset.

Code to run the T5Chem model is available at <https://github.com/HelloJocelynLu/t5chem/tree/main>

AUTHOR INFORMATION

Corresponding Author

*Jonathan M. Goodman, jmg11@cam.ac.uk

Author Contributions

R.K. has trained the models and wrote the manuscript. J.M.G. edited the manuscript. K.P and G.P.F.W. have advised on project direction.

Funding Sources

Exscientia and EPSRC via SynTech CDT

ACKNOWLEDGMENT

We thank Exscientia and EPSRC for funding; R.K thanks Margarita Kotlyarova for support.

References

- (1) Moir, M.; Danon, J. J.; Reekie, T. A.; Kassiou, M. An Overview of Late-Stage Functionalization in Today's Drug Discovery. *Expert Opin. Drug Discov.* **2019**, *14* (11), 1137–1149. <https://doi.org/10.1080/17460441.2019.1653850>.
- (2) Börgel, J.; Ritter, T. Late-Stage Functionalization. *Chem* **2020**, *6* (8), 1877–1887. <https://doi.org/10.1016/j.chempr.2020.07.007>.
- (3) Hassan, M. M. M.; Guria, S.; Dey, S.; Das, J.; Chattopadhyay, B. Transition Metal–Catalyzed Remote C–H Borylation: An Emerging Synthetic Tool. *Sci. Adv.* **2023**, *9* (16), eadg3311. <https://doi.org/10.1126/sciadv.adg3311>.
- (4) Miyaura, N.; Yanagi, T.; Suzuki, A. The Palladium-Catalyzed Cross-Coupling Reaction of Phenylboronic Acid with Haloarenes in the Presence of Bases. *Synth. Commun.* **1981**, *11* (7), 513–519. <https://doi.org/10.1080/00397918108063618>.
- (5) Miyaura, Norio.; Suzuki, Akira. Palladium-Catalyzed Cross-Coupling Reactions of Organoboron Compounds. *Chem. Rev.* **1995**, *95* (7), 2457–2483. <https://doi.org/10.1021/cr00039a007>.
- (6) Chan, D. M. T.; Monaco, K. L.; Wang, R.-P.; Winters, M. P. New N- and O-Arylations with Phenylboronic Acids and Cupric Acetate. *Tetrahedron Lett.* **1998**, *39* (19), 2933–2936. [https://doi.org/10.1016/S0040-4039\(98\)00503-6](https://doi.org/10.1016/S0040-4039(98)00503-6).
- (7) Lam, P. Y. S.; Clark, C. G.; Saubern, S.; Adams, J.; Winters, M. P.; Chan, D. M. T.; Combs, A. New Aryl/Heteroaryl C–N Bond Cross-Coupling Reactions via Arylboronic Acid/Cupric Acetate Arylation. *Tetrahedron Lett.* **1998**, *39* (19), 2941–2944. [https://doi.org/10.1016/S0040-4039\(98\)00504-8](https://doi.org/10.1016/S0040-4039(98)00504-8).
- (8) Evans, D. A.; Katz, J. L.; West, T. R. Synthesis of Diaryl Ethers through the Copper-Promoted Arylation of Phenols with Arylboronic Acids. An Expedient Synthesis of Thyroxine. *Tetrahedron Lett.* **1998**, *39* (19), 2937–2940. [https://doi.org/10.1016/S0040-4039\(98\)00502-4](https://doi.org/10.1016/S0040-4039(98)00502-4).

- (9) Mkhaliid, I. A. I.; Barnard, J. H.; Marder, T. B.; Murphy, J. M.; Hartwig, J. F. C–H Activation for the Construction of C–B Bonds. *Chem. Rev.* **2010**, *110* (2), 890–931. <https://doi.org/10.1021/cr900206p>.
- (10) Hartwig, J. F. Regioselectivity of the Borylation of Alkanes and Arenes. *Chem. Soc. Rev.* **2011**, *40* (4), 1992–2002. <https://doi.org/10.1039/C0CS00156B>.
- (11) Larsen, M. A.; Hartwig, J. F. Iridium-Catalyzed C–H Borylation of Heteroarenes: Scope, Regioselectivity, Application to Late-Stage Functionalization, and Mechanism. *J. Am. Chem. Soc.* **2014**, *136* (11), 4287–4299. <https://doi.org/10.1021/ja412563e>.
- (12) Sperger, T.; Sanhueza, I. A.; Kalvet, I.; Schoenebeck, F. Computational Studies of Synthetically Relevant Homogeneous Organometallic Catalysis Involving Ni, Pd, Ir, and Rh: An Overview of Commonly Employed DFT Methods and Mechanistic Insights. *Chem. Rev.* **2015**, *115* (17), 9532–9586. <https://doi.org/10.1021/acs.chemrev.5b00163>.
- (13) Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Sköld, C.; Norrby, P.-O.; Hartwig, J. F. Hybrid Machine Learning Approach to Predict the Site Selectivity of Iridium-Catalyzed Arene Borylation. *J. Am. Chem. Soc.* **2023**. <https://doi.org/10.1021/jacs.3c04986>.
- (14) Nippa, D. F.; Atz, K.; Hohler, R.; Müller, A. T.; Marx, A.; Bartelmus, C.; Wuitschik, G.; Marzuoli, I.; Jost, V.; Wolfard, J.; Binder, M.; Stepan, A. F.; Konrad, D. B.; Grether, U.; Martin, R. E.; Schneider, G. Enabling Late-Stage Drug Diversification by High-Throughput Experimentation with Geometric Deep Learning. *Nat. Chem.* **2023**, 1–10. <https://doi.org/10.1038/s41557-023-01360-5>.
- (15) Atz, K.; Isert, C.; A. Böcker, M. N.; Jiménez-Luna, J.; Schneider, G. Δ -Quantum Machine-Learning for Medicinal Chemistry. *Phys. Chem. Chem. Phys.* **2022**, *24* (18), 10775–10783. <https://doi.org/10.1039/D2CP00834C>.
- (16) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. arXiv December 5, 2017. <https://doi.org/10.48550/arXiv.1706.03762>.
- (17) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. <https://doi.org/10.1021/ci00057a005>.
- (18) O’Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. ChemRxiv September 19, 2018. <https://doi.org/10.26434/chemrxiv.7097960.v1>.
- (19) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn. Sci. Technol.* **2020**, *1* (4), 045024. <https://doi.org/10.1088/2632-2153/aba947>.
- (20) *The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time.* <https://jalammar.github.io/illustrated-transformer/> (accessed 2022-11-22).

- (21) Phuong, M.; Hutter, M. Formal Algorithms for Transformers. arXiv July 19, 2022. <https://doi.org/10.48550/arXiv.2207.09238>.
- (22) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn. Sci. Technol.* **2021**, *2* (1), 015016. <https://doi.org/10.1088/2632-2153/abc81d>.
- (23) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the Space of Chemical Reactions Using Attention-Based Neural Networks. *Nat. Mach. Intell.* **2021**, *3* (2), 144–152. <https://doi.org/10.1038/s42256-020-00284-w>.
- (24) Nana Teukam, Y. G.; Kwate Dassi, L.; Manica, M.; Probst, D.; Schwaller, P.; Laino, T. *Language Models Can Identify Enzymatic Active Sites in Protein Sequences*; preprint; Chemistry, 2023. <https://doi.org/10.26434/chemrxiv-2021-m20gg-v3>.
- (25) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583. <https://doi.org/10.1021/acscentsci.9b00576>.
- (26) Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeod, J. L.; Butler, C. R. Molecular Transformer Unifies Reaction Prediction and Retrosynthesis across Pharma Chemical Space. *Chem. Commun.* **2019**, *55* (81), 12152–12155. <https://doi.org/10.1039/C9CC05122H>.
- (27) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: A Pre-Trained Transformer for Computational Chemistry. *Mach. Learn. Sci. Technol.* **2022**, *3* (1), 015022. <https://doi.org/10.1088/2632-2153/ac3ffb>.
- (28) Zhang, Y.; Wang, L.; Wang, X.; Zhang, C.; Ge, J.; Tang, J.; Su, A.; Duan, H. Data Augmentation and Transfer Learning Strategies for Reaction Prediction in Low Chemical Data Regimes. *Org. Chem. Front.* **2021**, *8* (7), 1415–1423. <https://doi.org/10.1039/D0QO01636E>.
- (29) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Is GPT-3 All You Need for Low-Data Discovery in Chemistry? ChemRxiv February 14, 2023. <https://doi.org/10.26434/chemrxiv-2023-fw8n4>.
- (30) Toniato, A.; Vaucher, A. C.; Schwaller, P.; Laino, T. Enhancing Diversity in Language Based Models for Single-Step Retrosynthesis. 25.
- (31) Thakkar, A.; Vaucher, A.; Byekwaso, A.; Schwaller, P.; Toniato, A.; Laino, T. *Unbiasing Retrosynthesis Language Models with Disconnection Prompts*; preprint; Chemistry, 2022. <https://doi.org/10.26434/chemrxiv-2022-gx9gb>.
- (32) Lu, J.; Zhang, Y. Unified Deep Learning Model for Multitask Reaction Predictions with Explanation. *J. Chem. Inf. Model.* **2022**, *62* (6), 1376–1387. <https://doi.org/10.1021/acs.jcim.1c01467>.
- (33) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. “Found in Translation”: Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-to-Sequence Models. *Chem. Sci.* **2018**, *9* (28), 6091–6098. <https://doi.org/10.1039/C8SC02339E>.

- (34) Reaxys. <https://www.reaxys.com/#!/search/quick> (accessed 2021-11-05).
- (35) Shearer, J.; Castro, J. L.; Lawson, A. D. G.; MacCoss, M.; Taylor, R. D. Rings in Clinical Trials and Drugs: Present and Future. *J. Med. Chem.* **2022**, *65* (13), 8699–8712. <https://doi.org/10.1021/acs.jmedchem.2c00473>.
- (36) RDKit: Open-Source Cheminformatics. <https://www.rdkit.org>.
- (37) *T5Chem model files*. <https://yzhang.hpc.nyu.edu/T5Chem/index.html>.
- (38) *T5Chem model pretrained on 97 million PubChem Molecules with BERT-like self-supervised mask-filling scheme*. https://yzhang.hpc.nyu.edu/T5Chem/models/simple_pretrain.tar.bz2 (accessed 2023-09-25).
- (39) Lu, J.; Zhang, Y. USPTO_500_MT. https://yzhang.hpc.nyu.edu/T5Chem/data/USPTO_500_MT.tar.bz2 (accessed 2023-09-25).
- (40) *T5Chem model pretrained on 97 million PubChem Molecules with BERT-like self-supervised mask-filling scheme and fine-tuned on USPTO_500_MT*. https://yzhang.hpc.nyu.edu/T5Chem/models/USPTO_MT_model.tar.bz2.
- (41) Grisoni, F. Chemical Language Models for de Novo Drug Design: Challenges and Opportunities. ChemRxiv December 30, 2022. <https://doi.org/10.26434/chemrxiv-2022-5f14w>.
- (42) *Randomized SMILES strings improve the quality of molecular generative models | Journal of Cheminformatics | Full Text*. <https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0393-0> (accessed 2023-03-06).
- (43) Li, X.; Fourches, D. SMILES Pair Encoding: A Data-Driven Substructure Tokenization Algorithm for Deep Learning. *J. Chem. Inf. Model.* **2021**, *61* (4), 1560–1569. <https://doi.org/10.1021/acs.jcim.0c01127>.
- (44) Jaume-Santero, F.; Bornet, A.; Valery, A.; Naderi, N.; Vicente Alvarez, D.; Proios, D.; Yazdani, A.; Bournez, C.; Fessard, T.; Teodoro, D. Transformer Performance for Chemical Reactions: Analysis of Different Predictive and Evaluation Scenarios. *J. Chem. Inf. Model.* **2023**, *63* (7), 1914–1924. <https://doi.org/10.1021/acs.jcim.2c01407>.
- (45) Fu, N.; Hu, J.; Feng, Y.; Morrison, G.; Loye, H.-C. zur; Hu, J. Composition Based Oxidation State Prediction of Materials Using Deep Learning. arXiv November 28, 2022. <https://doi.org/10.48550/arXiv.2211.15895>.
- (46) Chicco, D.; Jurman, G. The Matthews Correlation Coefficient (MCC) Should Replace the ROC AUC as the Standard Metric for Assessing Binary Classification. *BioData Min.* **2023**, *16* (1), 4. <https://doi.org/10.1186/s13040-023-00322-4>.
- (47) Chicco, D. Ten Quick Tips for Machine Learning in Computational Biology. *BioData Min.* **2017**, *10* (1), 35. <https://doi.org/10.1186/s13040-017-0155-3>.

- (48) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction Classification and Yield Prediction Using the Differential Reaction Fingerprint DRFP. *Digit. Discov.* **2022**, *1* (2), 91–97. <https://doi.org/10.1039/D1DD00006C>.

For Table Of Contents Only

