

# Reproducing the invention of a named reaction: zero-shot prediction of unseen chemical reactions

An Su<sup>1#</sup>, Ling Wang<sup>2#</sup>, Xinqiao Wang<sup>2</sup>, Chengyun Zhang<sup>2</sup>, Yejian Wu<sup>2</sup>, Qingjie Zhao<sup>3</sup>  
& Hongliang Duan<sup>2\*</sup>

<sup>1</sup>College of Chemical Engineering, Zhejiang University of Technology, Hangzhou 310014, P. R. China

<sup>2</sup>Artificial Intelligence Aided Drug Discovery Institute, College of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou 310014, P. R. China

<sup>3</sup>Shanghai Institute of Material Medical, Chinese Academy of Sciences, Shanghai 201203, P. R. China

These authors contributed equally: An Su, Ling Wang

\*Corresponding Author:

Hongliang Duan

Email: [hduan@zjut.edu.cn](mailto:hduan@zjut.edu.cn)

## Abstract

While state-of-art models can predict reactions through the transfer learning of thousands of samples with the same reaction types as those of the reactions to predict, how to prepare such models to predict "unseen" reactions remain an unanswered question. We aim to study the Transformer model's ability to predict "unseen" reactions following "zero-shot reaction prediction (ZSRP)", a concept derived from zero-shot learning and zero-shot translation. We reproduce the human invention of the Chan-Lam coupling reaction where the inventor was inspired by the Suzuki reaction

when improving Barton's bismuth arylation reaction. After being fine-tuned with the samples from these two "existing" reactions, the USPTO-trained Transformer can predict "unseen" Chan-Lam coupling reactions with 55.7% top-1 accuracy. Our model also mimics the later stage of the history of this reaction, where the initial case of this reaction was generalized to more reactants and reagents via the "one-shot/few-shot reaction prediction (OSRP/FSRP)" approaches.

## Introduction

The history of organic synthesis began in 1828 when Friedrich Wöhler smashed the theory of vitalism by preparing urea, a naturally occurring organic compound, from an inorganic compound, ammonium cyanate<sup>1,2</sup>. In the past two centuries, a long list has been formed with more than 1,000 organic reactions named for their inventors. In the meantime, the large amount of accumulated chemical knowledge has become a suitable resource for deep learning. In the past decade, data-driven deep learning models based on the statistical learning of a vast number of existing reactions have been applied to retrosynthesis, reaction condition recommendation, generation of novel reactions, and forward-reaction prediction<sup>3,4</sup>. For forward-reaction predictions, there exist graphical convolutional neural networks<sup>5,6</sup> and simplified molecular-input line-entry system (SMILES)-based sequence-to-sequence (seq2seq) models<sup>7,8</sup> that take the SMILES of reactants and reagents as the input language and output the products as the translated language. Transformer-based models such as the Molecular Transformer<sup>9,10</sup>, adapted from Vaswani et al.'s original Transformer<sup>8</sup>, are the state-of-art SMILES-based seq2seq model for forward reaction prediction. In addition,

transfer learning has been equipped with the Transformer in the form of an additional fine-tuning step, and this combination is beneficial for the forward prediction of complex reactions that involve regioselectivity and stereoselectivity, such as carbohydrate reactions<sup>11</sup>, the Heck reaction<sup>12</sup>, and the Baeyer-Villiger reaction<sup>13</sup>. For example, in regio- and stereoselective carbohydrate reactions, a fine-tuning step with 2,000 carbohydrate reaction data yields a 30.0% increase in the prediction accuracy<sup>11</sup>. The performance of Transformer can be further improved by SMILES augmentation and beam search algorithms, and a recent state-of-art model gives top-5 accuracy higher than 96.0% in the direct reaction prediction of two USPTO-MIT datasets<sup>14</sup>.

While the Transformer equipped with transfer learning has exciting prediction performance in chemical reaction predictions, clarification and improvement are still urgently needed on two sides. First, the current transfer-learning coupled Transformers still require thousands of samples in the fine-tuning step to have the model specialized in predicting reactions in certain specific chemical space, while many chemical reactions do not have samples of this magnitude. Furthermore, the overlap of the reaction types in the training set (including fine-tuning set) and the test set in the state-of-art Transformer models, such as Schwaller et al.'s Molecular Transformer, has not been discussed in their works<sup>10,11</sup>. In other words, the capability of Transformer to predict truly "unseen" reaction is yet unclear---it might be possible that the Transformer can predict only the reaction types that it has seen in the training set.

Can we equip Transformer with the ability to predict unseen reactions through a

"zero-shot reaction prediction (ZSRP)" approach (Fig. 1c)? In the field of machine learning, there are mainly two "zero-shot" approaches including the "zero-shot learning (ZSL)" in the field of object recognition and image classification and "zero-shot translation (ZST)" in neural machine translation. ZSL means that, given training samples in seen classes  $S$ , one aims to learn a classifier for classifying test samples from "unseen" classes  $U$ <sup>15-17</sup> (Fig. 1a). One important element of ZSL is the construction of "auxiliary information" for the "unseen" classes using the instances in the feature space of the seen classes, and the use of auxiliary information is inspired by the way human beings identify new objects. A classic example is that humans can usually identify a zebra when they see it for the first time, as long as they have seen horses and stripes before and have the semantic knowledge that "a zebra is like a horse with stripes on the body" as auxiliary information<sup>16,18</sup>. On the other hand, ZST is brought up by Google's multilingual neural machine translation study in which the model learns to translate between language pairs not seen in the training set by modeling a few language pairs in a single model and setting up an "implicitly-learned bridging"<sup>19</sup> (Fig. 1b). The study gives an example that with training samples in Portuguese->English and English->Spanish provided, although this multilingual translation model has never seen a Portuguese->Spanish example in the training set, the model can translate Portuguese directly to Spanish without explicitly translating to English first<sup>19</sup>.

Unfortunately, we cannot solely follow the methodology of ZSL or ZST to implement ZSRP for Transformer---the SMILES-based reaction prediction through

Transformer does not bridge multiple languages as in ZST or do classification tasks as in ZSL. However, we can still borrow the philosophies from ZSL and ZST, that is, to get inspired by the way humans connect existing knowledge and identify unseen objects. In this work, we start by seeking inspiration from human behavior during the invention of a named reaction, a new chemical reaction named after its inventor(s).

The invention of a named reaction usually requires two fields of knowledge: the general chemistry knowledge that qualifies the inventor as a chemical scientist and the inspiration from a few existing reactions. Let us take the history of the invention of the Chan-Lam coupling reaction as an example<sup>20</sup>. When improving Barton's bismuth arylation reaction (referred to as the Barton reaction) in which N-H or O-H bonds are arylated by trivalent organobismuth compounds in the presence of copper salts<sup>21-24</sup> (Fig. 2a), Chan was inspired by the Suzuki reaction which uses arylboronic acids to cross-couple aryl bromides, iodides and triflates<sup>25</sup> (Fig. 2b). Chan used boronic acids from the Suzuki reaction to replace the bismuth compounds in the Barton reaction and successfully demonstrated a new methodology of C-O and C-N cross-coupling<sup>26</sup> (Fig. 2c), and this was the first appearance of the Chan-Lam coupling reaction. After that, Lam's group further explored the reaction through the use of alkenylboronic acids and the application of N-arylation to heterocyclic systems<sup>27</sup>, while Evans' group optimized the reaction with regard to the synthesis of biaryl ether<sup>28</sup>. After that, the scope of the reaction was further expanded by additional research groups<sup>20</sup>. To reproduce the first invention of the Chan-Lam coupling reaction using a machine learning model with the concept of ZSRP, a training set is needed to familiarize the machine learning model

with general knowledge about chemical reactions, while a fine-tuning set can teach the model the concepts of boronic acid coupling reagents, as well as the arylation of N-H and C-H bonds.

As is discussed above, besides predicting unseen reactions, the other capability that Transformer needs further development is to require fewer samples for transfer learning. With the history of the invention of the Chan-Lam coupling reaction in mind, a follow-up question is that with just one or a handful of instances of a reaction that was just invented, can machine learning models generalize the knowledge to a wide range of reactants and reagents, just like Lam, Evans, and other researchers did after seeing the first case of the Chan-Lam coupling reaction published by Chan? In this case, we again borrow the philosophy from the concept of "one-shot learning (OSL)". In 2003, Li Fei-fei et al. first proposed the idea of OSL in the area of computer vision, by which a model can learn sufficient information from one or just a handful of samples (also called "few-shot learning (FSL)") about a category.<sup>29</sup> More works regarding the concept of OSL/FSL through the use of a variety of statistical or machine learning models have been published since then<sup>30</sup>. Although being less exciting than ZSRP as samples from the seen reactions are still needed, proper one-shot/few-shot reaction prediction (OSRP/FSRP) approaches can mimic the later stages of the invention of a chemical reaction where the first case of a new reaction has been invented but has not been generalized to a wide range of reactants and reagents, and provide enlightenment for lowering the number of samples in the transfer learning of chemical reaction predictions.

In this work, we reproduce the earlier stages of the invention of the Chan-Lam coupling reaction using the concept of ZSRP. We construct a training set with USPTO, a dataset that has been used for the training of many state-of-art reaction-prediction models<sup>6,10,11</sup>. Additionally, a fine-tuning dataset that consists of 200 samples of the Suzuki reaction and 148 samples of Barton's bismuth arylation reaction is constructed. We build our model upon the Transformer based on its success in the cases that have been introduced above. The prediction performance of the model for Chan-Lam coupling reactions is evaluated using a dataset that contains 472 Chan-Lam coupling reactions (Fig. 3). After that, we reproduce the later generalization stage of the Chan-Lam coupling reaction following the concept of OSRP/FSRP by picking one or a handful of samples of Chan-Lam coupling reactions to further fine-tune the Transformer. We hope that by reproducing the whole process of the invention of the Chan-Lam coupling reaction using our adapted Transformer models, we can gain inspiration from the human discovery of new chemical knowledge to train machine learning models for predicting unseen chemical reactions as well as conducting transfer learning with fewer samples.

## Results

**Zero-shot reaction prediction (ZSRP): first invention of Chan-Lam coupling reaction.** To reproduce the first invention of the Chan-Lam coupling reaction, the Transformer model is trained on a training set (the USPTO dataset) as well as a fine-tuning dataset that includes the Suzuki reaction and Barton's bismuth arylation reaction, as described above. The USPTO dataset contains various samples of

chemical reactions, providing the model with basic knowledge of molecular representations and chemical reactions<sup>31</sup>. In our study, all Suzuki, Barton, and Chan-Lam coupling reactions are removed from the USPTO training set to ensure that there are no overlaps between the seen reactions and the "unseen" reactions (Chan-Lam coupling reactions). The performance of the model is evaluated using the top-1 accuracy metric on the test set that contains 472 Chan-Lam coupling reactions. A detailed description of the construction of the datasets can be found in the "Methods" section.

The model trained on both the USPTO training set and the fine-tuning dataset reaches a 55.7% top-1 accuracy (Table 1). Meanwhile, two other models are trained without the fine-tuning step for comparison: the model trained on the training set only yields a 4.4% top-1 accuracy, while the model trained on an alternative training set, the original USPTO dataset with only Chan-Lam reactions removed, gives a 24.8% top-1 accuracy (Table 1). The results show that the fine-tuning step using the Barton reaction and Suzuki reaction plays a significant role in improving the model's performance in predicting the Chan-Lam coupling reaction. Let us review the first part of the story of the Chan-Lam coupling reaction: Chan, a scientist in the Chemical Discovery of DuPont at the time, replaced the bismuth compounds in Barton's bismuth arylation with boric acid reagents from the Suzuki reaction without changing the main catalyst, demonstrating the first Chan-Lam coupling reaction, as described above (Fig. 2)<sup>26</sup>. Similarly, along with the basic knowledge of molecular representations and chemical reactions from the training set (like the chemical



knowledge that prepared Chan as a chemical scientist), the model needs the specific knowledge from the fine-tuning space, including the reactants for arylation, the organoboronic coupling reagents, and the chemical mechanisms of coupling reactions, to provide the first prediction of the Chan-Lam coupling reaction. As the model trained with USPTO training set and fine-tuned with Suzuki reaction and Barton reaction reproduces the first invention of Chan-Lam coupling reaction and performs significantly better than the other two models (Table 1), we refer this model as the ZSRP model in the following context.

The reactions correctly predicted by the ZSRP model are not limited to a specific type of coupling. Table 2 shows the results with respect to the different coupling types in detail. The top-1 accuracies of the two groups of C-N and C-O coupling reactions with a relatively large number of test samples are 47.4% and 65.0%, respectively. The accuracies of the C-S and C-C coupling reactions in the two groups with fewer test examples are 84.6% and 100.0%, respectively. The predictive ability of the ZSRP model for a variety of coupling types matches the fact that the Chan-Lam coupling reaction has a convenient methodology to arylate N-H-, O-H-, S-H-, and C-H-containing compounds<sup>32</sup>. Furthermore, the samples of the Barton reaction in the fine-tuning dataset have only C-N couplings and C-O couplings, which means that our model can generalize the coupling types to C-S and C-C couplings with the fundamental knowledge learned from the training set.

The Chan-Lam coupling reactions correctly predicted by the ZSRP model are not limited to a single substrate type. Table 3 shows the performance for the C-N Chan-

Lam coupling categorized by five classes of N-containing substrates. The reactions containing aliphatic amines yield the highest accuracy (65.3%), while the accuracy for the reactions involving N-aromatic heterocyclic compounds is the lowest (35.5%). The model recognizes different N-containing reactants, including amide, amine, and N-aromatic heterocyclic compounds, and this matches the fact that the Chan-Lam coupling reaction is known to accommodate a variety of substrates<sup>29</sup>.

A few correctly predicted C-N Chan-Lam coupling examples from the Table 3 are shown in Supplementary Table 1. The fourth example is the most interesting since the chlorine group competes with the amino group in the reactant to generate a Suzuki product. However, our model recognizes the difference between these two reactive groups and predicts the correct Chan-Lam coupling product under the interference of the Suzuki reaction from the fine-tuning dataset.

The ZSRP model also finds different O-H- and S-H-containing substrates that can produce Chan-Lam coupling reactions (more information can be found in Supplementary Tables 2-3), and some correctly predicted examples are shown in Supplementary Table 4. For the C-O couplings, the knowledge learned by the model during the training process determines that the substrates include aromatic alcohol, aliphatic alcohol, and amide alcohol. Again, in the first case shown in Supplementary Table 4, the model correctly identifies a Chan-Lam coupling reaction with the corresponding O-H group when competition with O-H, Cl, and Br exists. Furthermore, the model discovers that thiophenol and thiol can be used as substrates for C-S coupling, while C-C coupling can only occur when the reactant has an active

methylene structure.

**One-shot/few-shot reaction prediction (OSRP/FSRP): generalization of the first**

**Chan-Lam coupling reaction.** To reproduce the later stage of the invention of the Chan-Lam coupling reaction, where researchers expanded the ranges of the reactants and reagents, the ZSRP model is further fine-tuned via the one-shot reaction prediction (OSRP) of Chan-Lam coupling reaction samples that are outside the scope of the test set. To avoid the bias brought by the OSRP sample, the model is fine-tuned with 12 Chan-Lam coupling reaction samples, respectively (Table 4). These 12 samples are selected through stratified random sampling so that the proportions of different coupling types in the 12 reactions are close to the proportions in the test set. The testing accuracies of the 12 fine-tuned models as functions of the training steps are shown in Supplementary Tables 5-16, and the final testing accuracies range from 61.6% to 87.1% with an average of 72.3% (Table 4). We also compare OSRP with few-shot reaction prediction (FSRP), where more than one sample is used in the fine-tuning step. With the 12 selected samples applied at one time in the fine-tuning step, the testing accuracy increases to 92.2%, while fine-tuning with all 101 samples yields an additional 2.0% increase (Supplementary Table 17).

To further evaluate the effect of OSRP, the model fine-tuned with the first reaction in Table 4 (which gives the highest testing accuracy) is analyzed. The fine-tuned model correctly predicts 153 of the 209 reaction samples that are incorrectly predicted by the model without fine-tuning, and a few examples are shown in Table 5. On the other hand, the fine-tuned model correctly predicts 258 of the 263 Chan-Lam

reactions that are predicted correctly by the model without fine-tuning. Additionally, the improvement in prediction performance is not limited to reactions that have the same coupling style or the same reactant type as that of the OSRP sample. While the OSRP sample discussed above is a C-N coupling reaction that occurs for an N-heterocyclic compound, a 28.1% improvement in accuracy is found for C-O coupling reactions compared to the model without fine-tuning. Furthermore, in the C-N coupling reactions with reactants other than N-heterocyclic reactants (e.g., amides, amines), increases in accuracy ranging from 8.2% to 48.7% are found. Detailed data can be found in Supplementary Tables 18-22.

## Discussion

While the definition of "unseen classes" in ZSL and "unseen language pairs" in ZST have been clarified, the concept of "unseen reactions" in chemical reaction prediction seems difficult to define in a narrow sense. For Transformer-based models, a chemical reaction contains two main components: the SMILES of reactions and products and the rearrangement of atoms, while many chemical reactions, including different named reactions, have overlaps in SMILES or the mechanisms of atom rearrangement. Hence, we would like to define the "unseen reactions" from a different perspective, the "space" of chemical reactions. In other words, if certain reactions are clustered separately from the clusters of the reactions from the training set, they can be considered as "unseen" as they locate in the chemical reaction space that the model has never explored. A recent study by Schwaller et al. uses *rxnfp*, a BERT (Bidirectional Encoder Representations from Transformers) classifier<sup>33</sup>, to convert

SMILES-based reactions to reaction fingerprints and mapped the reaction fingerprints into TMAP<sup>34,35</sup>, a tree-like graph for the clustering of chemical reactions according to their classifications<sup>36</sup>. The *rxnfp* has been pretrained on the Pistachio database which contains 132k reactions from 792 different classes (<https://www.nextmovesoftware.com/pistachio>) and can be fine-tuned with a specific reaction classification database to follow its classification ontology<sup>36</sup>. Using *rxnfp* and TMAP, Schneider 50K, the reaction database from the work by Schneider et al.<sup>37</sup> that follows the RXNO ontology (<http://www.rsc.org/ontologies/RXNO/index.asp>), has been mapped into a chemical reaction space where reactions are well-clustered based on their reaction fingerprints and classifications<sup>36</sup>.

In this study, we use *rxnfp* and TMAP to demonstrate the clustering and the distribution of our Suzuki reactions, Barton's bismuth arylations, and Chan-Lam coupling reactions in the chemical reaction space. We use Schneider 50K instead of USPTO to form the backbone of this chemical reaction space as the Schneider 50K has been classified<sup>37</sup> and visualized<sup>36</sup> nicely as is discussed above. To achieve the best visualization effect, only the "heteroatom alkylation and arylation" and "C-C bond formation" super-classes from Schneider 50K that are closely related to our reactions are included in our TMAP along with our three reactions. Fig. 4 delivers two important messages. First, the Chan-Lam coupling reactions are clustered independently from Suzuki reactions and Barton's bismuth arylations, meaning that our fine-tuning training set and test set fall into different regions of the chemical reaction space. Meanwhile, Chan-Lam coupling reactions form their own clusters

instead of being mixed into the cluster of heteroatom alkylation and arylation. This infers that, even if there are reactions in our USPTO training set that look like Chan-Lam coupling reactions, they still fall into regions different from the Chan-Lam coupling reactions in the chemical reaction space. To confirm these two messages, two additional advanced machine learning-based visualization techniques, t-SNE<sup>38</sup>, and UMAP<sup>39</sup>, have been used to visualize the *rxnfp* fingerprints of these reactions, and similar observation to the TMAP is achieved (Supplementary Figs. 1-2). Hence, we can conclude that the Chan-Lam coupling reactions in our study are "unseen reactions" from the perspective of the space of chemical reactions.

Unlike the previous studies that focused on optimizing training processes to achieve improved prediction performances with regard to certain types of chemical reactions, our study focuses on the proof-of-concept of the fact that one can increase the possibility of correctly predicting unseen reactions via a ZSRP approach inspired by reproducing the human invention of a named reaction. In other words, we do not aim to compare the ZSRP version of Transformer prediction for "unseen" reactions to the state-of-art Transformer models that predict "seen" reactions. Instead, the 55.7% testing accuracy achieved via ZSRP approach compared to the 24.8% accuracy of the model trained with USPTO without fine-tuning shows a transition from 0 to 1---the understanding of Transformer changes from "undetermined capability for predicting unseen reactions" to "can predict unseen reactions with a proper ZSRP approach". Additionally, with additional OSRP/FSRP approaches, the model can further bring a transition from 1 to 100 by expanding one or a handful of cases of the newly

recognized reaction to various coupling types and reactant types. The results of this study show that having existing reactions as a fine-tuning training set can help the Transformer predict "unseen" reactions and that additionally providing just one or few samples of the "unseen" reaction can boost the corresponding model's generalization ability. We achieved both the transitions by reproducing the history of the invention of the Chan-Lam coupling reaction, so we infer that the training of artificial intelligence models can seek inspiration from the human history of knowledge accumulation and industrial evolution.

## Methods

**USPTO dataset.** The dataset used in our study is originally from the work of Lowe, which contains reactions extracted from the United States Patent and Trademark Office (USPTO)<sup>31</sup>. In Lowe's USPTO dataset, reagents have been eliminated so that only reactants and products are kept in the reactions. Also, the reactions with multiple products have been split into multiple reactions so that each of them has a single product. In this work, the reactions that are duplicated, incomplete, erroneous, or containing products that are the same as reactants are removed. We remove Suzuki and Chan-Lam reactions from the USPTO training set by removing all the reactions that contain boron (B) in the SMILES of the reactants. Similarly, we remove Barton's bismuth arylation reactions by removing all reactions that have bismuth (Bi) in the SMILES of the reactants. 367726 reactions are kept after data cleaning (Supplementary Table 23). All reactants and products of USPTO, Suzuki reactions,

Barton's bismuth arylations, and Chan-Lam coupling reactions are canonicalized using RDKit prior to training and testing and have had their atom mapping removed.

**Suzuki reaction.** The Suzuki reaction, the palladium-catalyzed reaction of organoboron compounds and organic halides or triflates, was first reported by Suzuki and Miyaura.<sup>25</sup> We extract the samples of the Suzuki reaction from the Reaxys database based on the name of the reaction (all entries where the "Suzuki coupling" phrase are used). The Suzuki reaction samples downloaded from Reaxys are further processed with Python scripts and RDKit. And duplicate and incorrect reactions are moved. Finally, we have 200 Suzuki reaction samples kept as the first part of the fine-tuning dataset (Supplementary Table 23).

**Barton's bismuth arylation.** Barton's bismuth arylation was first published by Barton in 1986, and the work demonstrated that arylbismuth reagents can arylate aliphatic and aromatic amines in the presence of metallic copper or a copper (II) salt<sup>21</sup>. The Barton's bismuth arylation dataset used in this work is originally derived from Barton's papers<sup>23,24,40</sup> and USPTO reactions<sup>31</sup>. We manually extract 62 Barton's bismuth arylations from the Reaxys database and 87 from the USPTO reactions. After removing the duplicates, 148 Barton's bismuth arylations are kept as the second part of the fine-tuning dataset (Supplementary Table 23).

**Chan-Lam coupling reaction.** The target reaction to predict is the Chan-Lam coupling reaction which is a class of coupling reaction between arylboronic acids and the nucleophiles that have different heteroatoms under the promotion of copper salts<sup>26-</sup>



<sup>28</sup>. We extract the Chan-Lam coupling reaction based on the name of the reaction from the Reaxys database (all entries where the "Chan-Lam coupling reaction" phrase are used). The reaction samples downloaded from Reaxys is then processed with Python scripts and RDKit. After removing the duplicates and the reactions containing multiple products, the final Chan-Lam coupling reaction dataset contains 1031 reaction samples (Supplementary Table 23).

**Model.** The model in this work is based upon the Transformer architecture, a neural machine translation model (Supplementary Fig. 3)<sup>9</sup>. The Transformer consists of two main stacks of layers: the encoder and the decoder. The encoder contains several identical layers, and each of the layers contains one multi-head self-attention sub-layer and one feed-forward sub-layer. Each layer of the decoder is composed of a multi-head self-attention sub-layer, a feed-forward sub-layer, and a masked multi-head attention corresponding to the encoder's output. The residual connection and layer normalization play a crucial role in the integration of both the encoder sublayers and decoder sublayers.

The multi-head attention consists of several scaled dot-product attention running in parallel, which is an innovative part of the Transformer architecture. With multi-head attention, the model can process different versions of queries, keys, and values simultaneously, which outperforms the models with a single-head attention. The Transformer model abandons any kind of convolutional or recurrent neural network components and is based on attention mechanism solely. Therefore, a positional encoding matrix is needed to make use of the order of the input SMILES sequence<sup>9</sup>.

**Zero-shot reaction prediction (ZSRP).** We formulate the reaction prediction as a sequence-to-sequence translation in which the reactants SMILES are translated to the products SMILES. We make a few changes to the original hyperparameters of the Transformer model: We select a batch size value of 6144 and a hidden size value of 256. We also change the vocabulary size to 64 and drop-out to 0.3. For ZSRP, the Transformer model is trained on the training set (USPTO dataset) and the fine-tuning set (Suzuki reaction and Barton's bismuth arylation). The model is then tested with the testing set that contains 472 Chan-Lam coupling reactions. Models trained on only the training set (USPTO dataset and USPTO without Chan-Lam reactions) are also tested.

**One-shot/Few-shot reaction prediction (OSRP/FSRP).** In this work, all the OSRP/FSRP is performed on the Transformer model. The hyperparameters used here are the same as in the ZSRP. 101 Chan-Lam coupling reactions outside the testing set are set aside as the Chan-Lam training set. In OSRP, the Transformer model trained on the training set (USPTO dataset) and fine-tuning dataset (Suzuki reaction and Barton's bismuth arylation) is further fine-tuned with one sample of Chan-Lam coupling reaction. 12 Chan-Lam coupling reactions are selected from the Chan-Lam training set, and each of the 12 reactions is applied to the fine-tuning process, respectively. In FSRP, the 12 selected Chan-Lam reactions and the 101 Chan-Lam coupling reactions from the Chan-Lam training set are applied to the fine-tuning step, respectively. All the models trained with OSRP/FSRP are tested with the testing set.

## **Data availability**

The training, fine-tuning, validation and testing datasets used in our study are

available from <https://github.com/hongliangduan/Reproducing-the-invention-of-a-named-reaction-Zero-shot-prediction-of-unseen-chemical-reactions>. Source data are provided with this paper.

## Code availability

The code and the trained model are available from <https://github.com/hongliangduan/Reproducing-the-invention-of-a-named-reaction-Zero-shot-prediction-of-unseen-chemical-reactions>.

## References

1. Wöhler, F. Ueber künstliche bildung des harnstoffs. *Annalen der Physik* **88**, 253-256 (1828).
2. Cello, J., Paul, A. V. & Wimmer, E. Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* **297**, 1016-1018 (2002).
3. Struble, T. J. *et al.* Current and future roles of artificial intelligence in medicinal chemistry synthesis. *J. Med. Chem.* **63**, 8667-8682 (2020).
4. Bort, W. *et al.* Discovery of novel chemical reactions by deep generative recurrent neural network. *Sci. Rep.* **11**, 3178 (2021).
5. Jin, W., Coley, C. W., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network. In *Advances in Neural Information Processing Systems*, 2607-2616 (2017).
6. Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370-377 (2019).
7. Nam, J. & Kim, J. Linking the neural machine translation and the prediction of organic chemistry reactions. Preprint at <https://arxiv.org/abs/1612.09529> (2016).

8. 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.* **9**, 6091-6098 (2018).
9. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998-6008 (2017).
10. Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572-1583 (2019).
11. Pesciullesi, G., Schwaller, P., Laino, T. & Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.* **11**, 4874 (2020).
12. Wang, L., Zhang, C., Bai, R., Li, J. & Duan, H. Heck reaction prediction using a transformer model based on a transfer learning strategy. *Chem. Commun.* **56**, 9368-9371 (2020).
13. Zhang, Y. et al. Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. *Org. Chem. Front.* **8**, 1415-1423 (2021).
14. Tetko, I. V., Karpov, P., Van Deursen, R. & Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat. Commun.* **11**, 5575 (2020).
15. Xian, Y., Lampert, C. H., Schiele, B. & Akata, Z. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**, 2251-2265 (2019).
16. Wang, W., Zheng, V. W., Yu, H. & Miao, C. A survey of zero-shot learning: settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.* **10**, 1-37 (2019).
17. Fu, Y. et al. Recent advances in zero-shot recognition: toward data-efficient understanding of

- visual content. In *IEEE Signal Processing Magazine* **35**, 112-125 (2018).
18. Fu, Z., Xiang, T., Kodirov, E. & Gong, S. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2635-2644 (2015).
  19. Johnson, M. *et al.* Google's multilingual neural machine translation system: enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* **5**, 339-351 (2017).
  20. Chan, D. M. & Lam, P. Y. *Boronic Acids* (John Wiley & Sons, Inc, New Jersey, 2005).
  21. Barton, D. H., Finet, J.-P. & Khamsi, J. Metallic copper catalysis of N-arylation of amines by triarylbi-muth diacylates. *Tetrahedron Lett.* **27**, 3615-3618 (1986).
  22. Barton, D. H., Finet, J.-P. & Khamsi, J. Copper salts catalysis of N-phenylation of amines by trivalent organobismuth compounds. *Tetrahedron Lett.* **28**, 887-890 (1987).
  23. Barton, D. H., Ozbalik, N. & Ramesh, M. Alkylation of amines using trivalent bismuth derivatives. *Tetrahedron Lett.* **29**, 857-860 (1988).
  24. Barton, D. H., Finet, J.-P. & Khamsi, J. N-phenylation of amino acid derivatives. *Tetrahedron Lett.* **30**, 937-940 (1989).
  25. Miyaura, N. & Suzuki, A. Stereoselective synthesis of arylated (E)-alkenes by the reaction of alk-1-enylboranes with aryl halides in the presence of palladium catalyst. *J. Chem. Soc., Chem. Commun.* **19**, 866-867 (1979).
  26. Chan, D. M., Monaco, K. L., Wang, R.-P. & Winters, M. P. New N-and O-arylations with phenylboronic acids and cupric acetate. *Tetrahedron Lett.* **39**, 2933-2936 (1998).
  27. Lam, P. Y. *et al.* New aryl/heteroaryl C-N bond cross-coupling reactions via arylboronic acid/cupric acetate arylation. *Tetrahedron Lett.* **39**, 2941-2944 (1998).
  28. 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.* **39**, 2937-2940 (1998).
29. Li, F.-F., Fergus & Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings Ninth IEEE International Conference on Computer Vision*, 1134-1141 (2003).
30. O'Mahony, N. *et al.* One-shot learning for custom identification tasks; a review. *Procedia Manufacturing* **38**, 186-193 (2019).
31. Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. thesis, University of Cambridge (2012).
32. Qiao, J. X. & Lam, P. Y. Copper-promoted carbon-heteroatom bond cross-coupling with boronic acids and derivatives. *Synthesis* **6**, 829-856 (2011).
33. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186 (2019).
34. Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **12**, 1-13 (2020).
35. Probst, D. & Reymond, J.-L. FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **34**, 1433-1435 (2017).
36. Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **3**, 144-152 (2021).
37. Schneider, N., Lowe, D. M., Sayle, R. A. & Landrum, G. A. Development of a novel fingerprint

- for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model.* **55**, 39-53 (2015).
38. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579-2605 (2008).
39. McInnes, L., Healy, J. & Melville, J. Umap: uniform manifold approximation and projection for dimension reduction. Preprint at <https://arxiv.org/abs/1802.03426> (2018).
40. Barton, D. H., Yadav-Bhatnagar, N., Finet, J.-P. & Khamsi, J. Phenylation of aromatic and aliphatic amines by phenyllead triacetate using copper catalysis. *Tetrahedron Lett.* **28**, 3111-3114 (1987).

## Acknowledgements

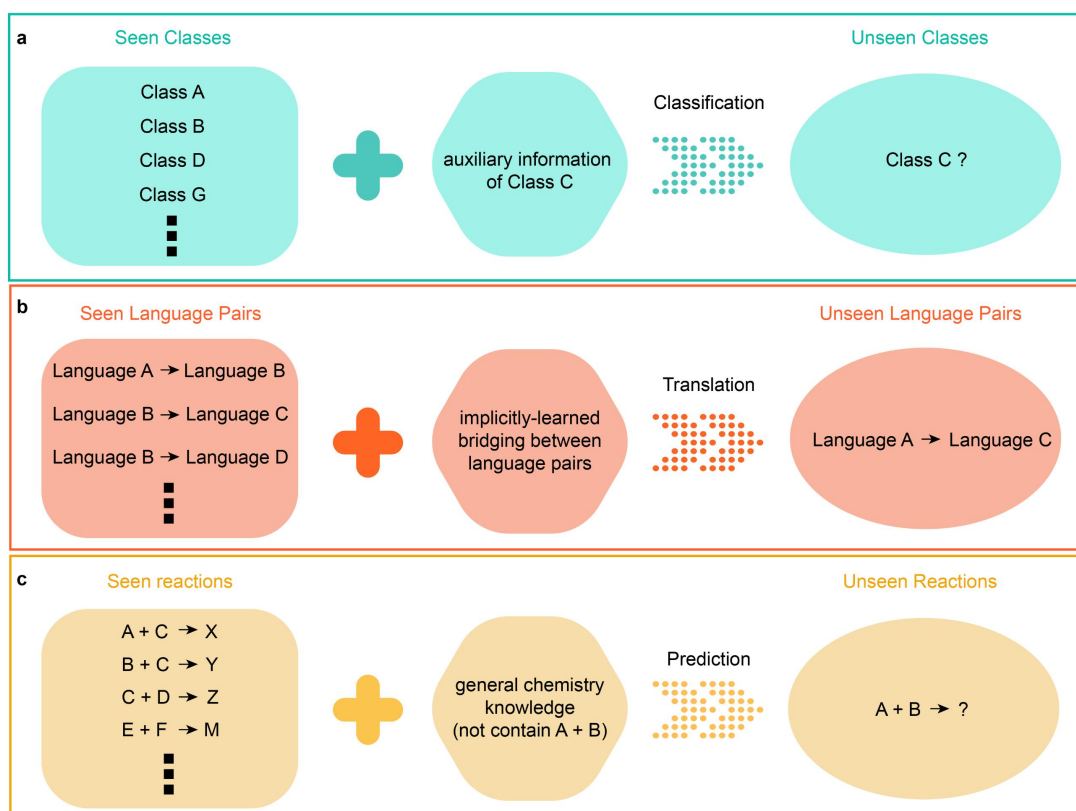
This project was supported by the National Natural Science Foundation of China (No. 81903438).

## Author contributions

These authors contributed equally: A.S. and L.W., A.S., L.W., Q.Z. and H.D. designed the research project. X.W., C.Z. and Y.W. trained models. A.S. and L.W. analyzed data and wrote the manuscript. All authors discussed the results and approved the manuscript.

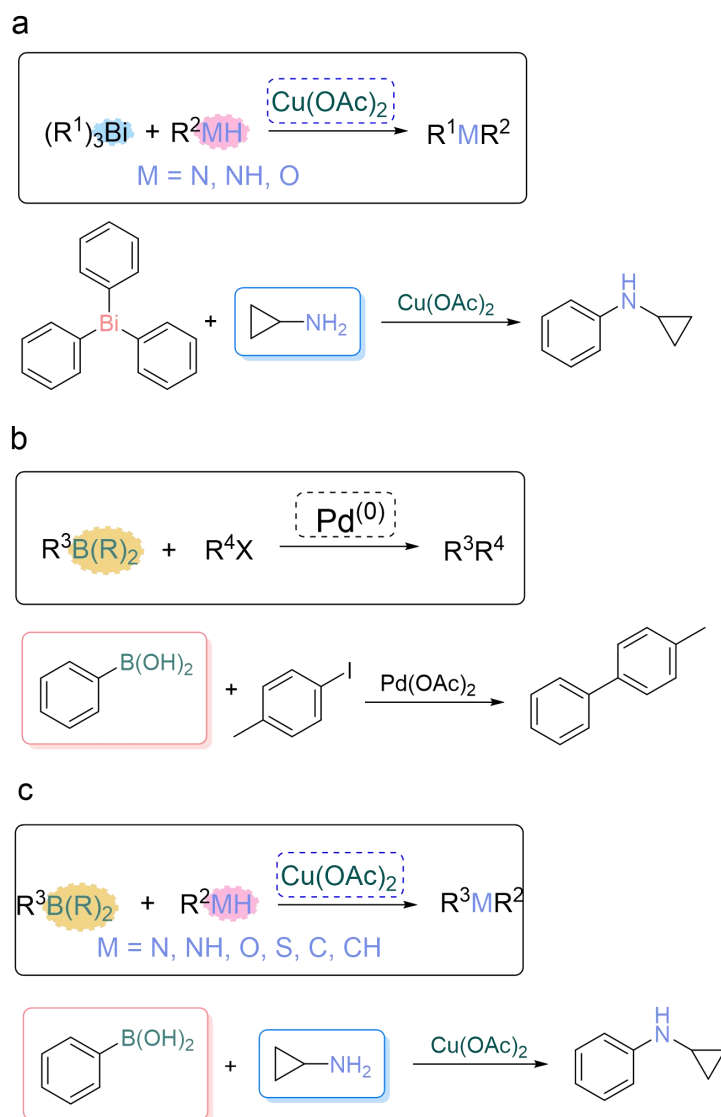
## Competing interests

The authors declare no competing interests.

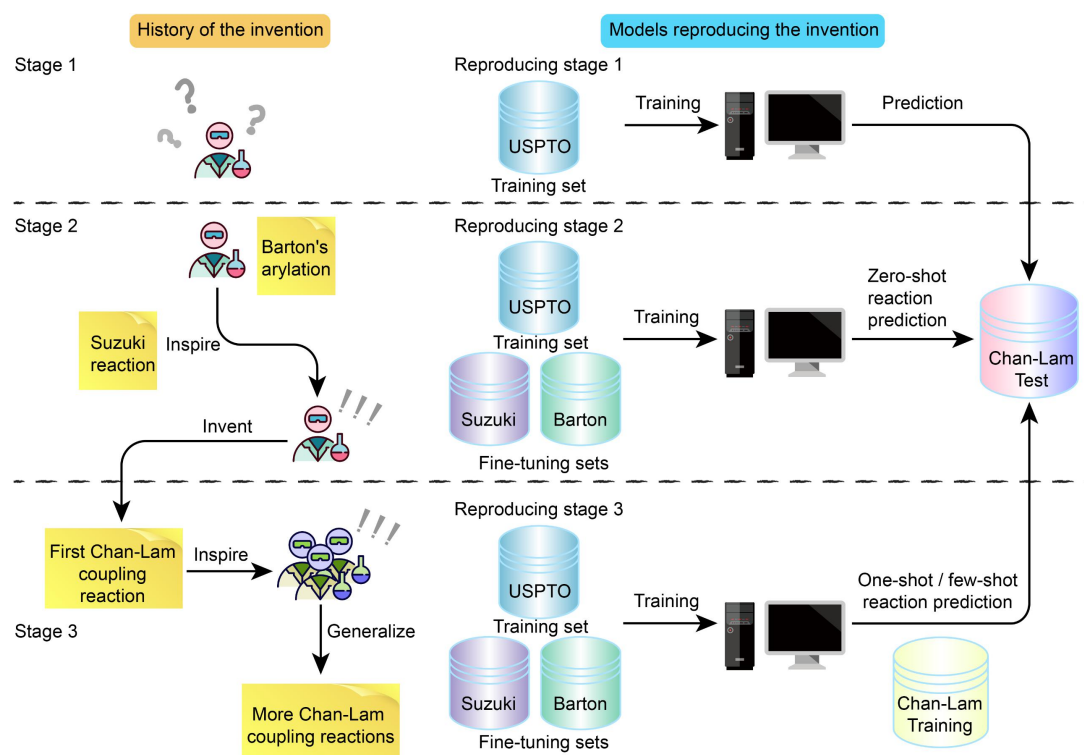


**Fig. 1** The illustration of three "zero-shot" situations. **a** Zero-shot learning (ZSL). **b** Zero-shot translation (ZST). **c** Zero-shot reaction prediction (ZSRP).

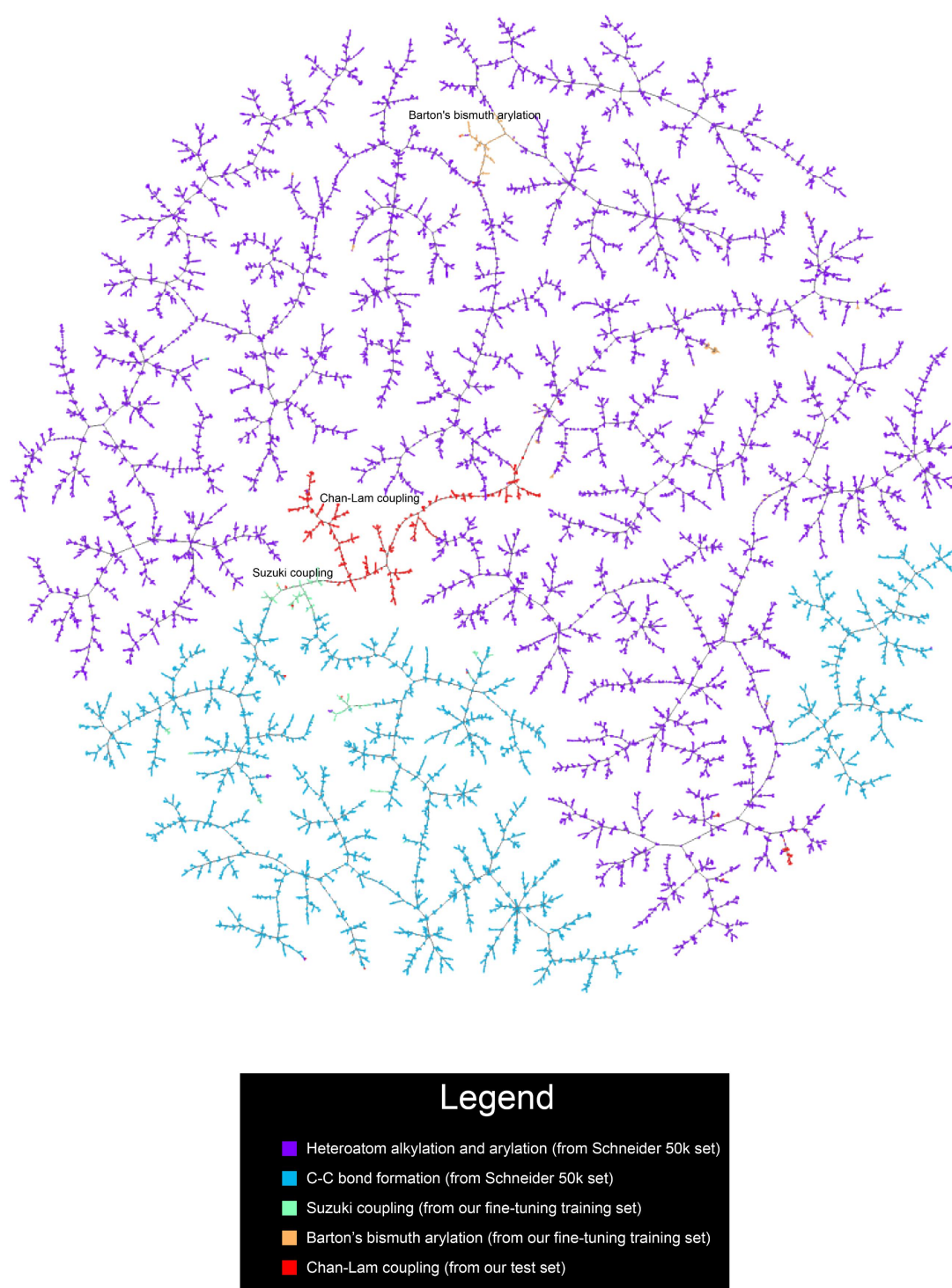




**Fig. 2** Schema of Barton's bismuth arylation, Suzuki reaction, and Chan-Lam coupling reaction.  
**a** Barton's bismuth arylation. **b** Suzuki reaction. **c** Chan-Lam coupling reaction.



**Fig. 3 Models reproducing different stages at the invention of Chan-Lam coupling reactions.** Before invention (stage 1), invention of first Chan-Lam coupling reaction (Stage 2), and generalization of more Chan-Lam coupling reactions (Stage 3).



**Fig. 4 Classification of the reactions in the fine-tuning datasets and test sets along with the two related classes of reactions from the Schneider 50K set<sup>37</sup>.** The Schneider 50K set is NOT used in the training of our model but only for the visualization purpose. The fingerprints are generated using *rxnfp*<sup>36</sup>, and the reactions are visualized using TMAP<sup>34</sup> algorithm and the Faerun<sup>35</sup> visualization library.

**Table 1 Performance of Transformer trained with different approaches.**

Training Set(s)	Top-1 Accuracy (%)
USPTO <sup>a</sup>	4.4
USPTO w/o Chan-Lam <sup>b</sup>	24.8
USPTO <sup>a</sup> + Suzuki & Barton fine-tuning (the ZSRP model)	55.7

<sup>a</sup>The "USPTO" training set is the original USPTO dataset that have all Suzuki reactions, Barton's bismuth arylations, and Chan-Lam coupling reactions removed. <sup>b</sup>The "USPTO w/o Chan-Lam" training set is the original USPTO dataset that have only Chan-Lam coupling reactions removed.

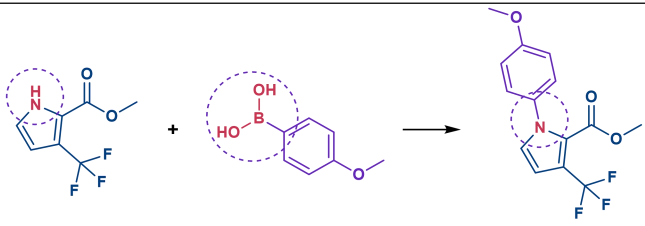
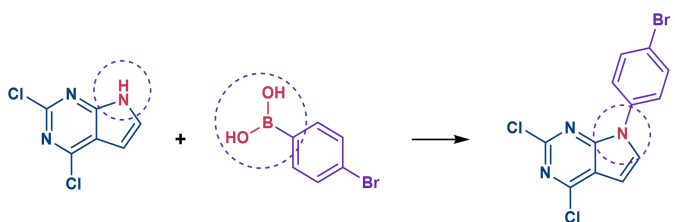
**Table 2 Performance of ZSRP model categorized by coupling types.**

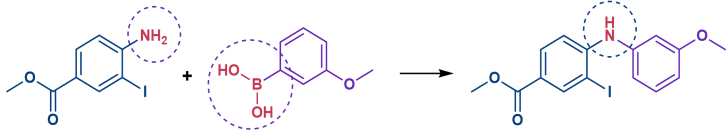
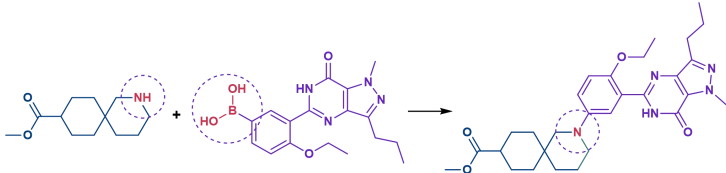
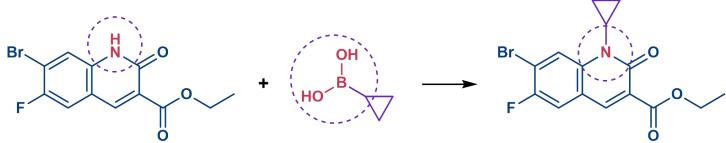
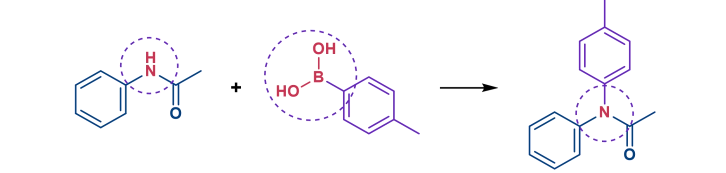
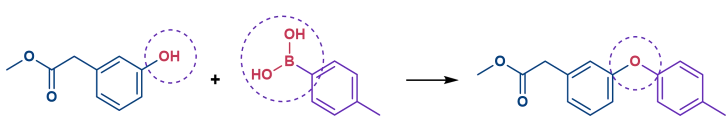
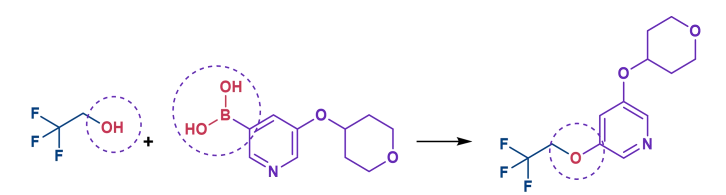
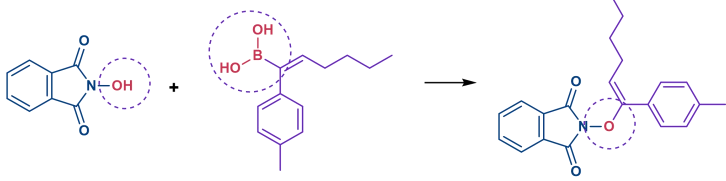
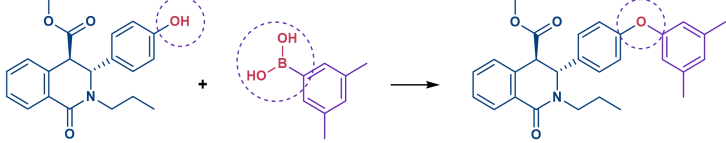
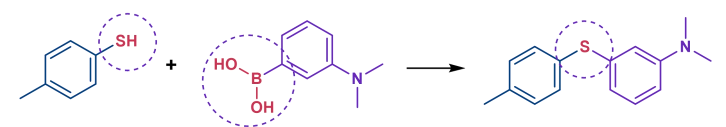
Coupling type	Test samples	Top-1 accuracy (%)
C-N	283	47.4
C-O	160	65.0
C-S	26	84.6
C-C	3	100.0
Total	472	55.7

**Table 3 Performance of ZSRP model for C-N Chan-Lam coupling categorized by reactant type.**

Reactant type	Test samples	Top-1 accuracy (%)
Amide (linear)	19	47.4
Amide (cyclic)	62	40.3
Aliphatic amine	49	65.3
Aromatic amine	77	53.2
N-aromatic heterocyclic	76	35.5
Total	283	47.4

**Table 4 12 Chan-Lam coupling reaction samples selected for OSRP and corresponding performance.**

Coupling Type	Reactant type	Reaction sample	Top-1 accuracy (%)
C-N	N-aromatic heterocyclic		87.1
C-N	N-aromatic heterocyclic		83.5

C-N	Aromatic amine		81.3
C-N	Aliphatic amine		62.3
C-N	Amide (cyclic)		69.1
C-N	Amide (linear)		79.6
C-O	Aromatic alcohol		69.5
C-O	Aliphatic alcohol		74.4
C-O	Amide alcohol		61.6
C-O	Aromatic alcohol		64.6
C-S	Thiophenol		71.0

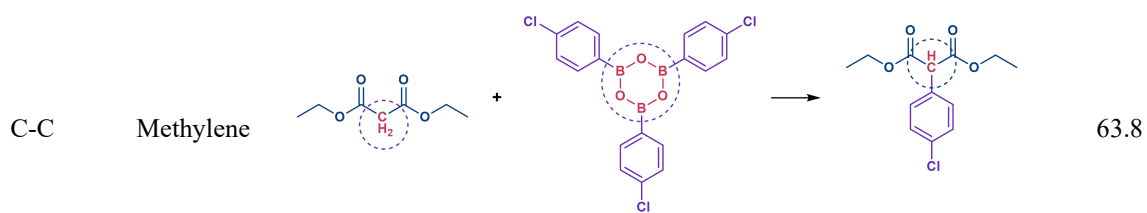


Table 5 Examples of reactions correctly predicted by OSRP but not ZSRP.

Coupling type	Reactants	ZSRP's wrong prediction	OSRP's correct prediction
C-N			
C-N			
C-N			
C-N			
C-N			
C-O			
C-O			

---

C-S

