

Regio-MPNN: Predicting Regioselectivity for General Metal-Catalyzed Cross-Coupling Reactions using Chemical Knowledge Informed Message Passing Neural Network

Baochen Li,^{†,§} Yuru Liu,^{†,§} Haibin Sun,[†] Rentao Zhang,[†] Yongli Xie,[†] Klement Foo,[‡] Frankie S. Mak,[‡] Ruimao Zhang,[¶] Tianshu Yu,[¶] Sen Lin,[†] Peng Wang,[†] and Xiaoxue Wang^{*,†}

[†]*ChemLex Technology Co., Ltd., 1976 Gaoke Mid. Rd., Shanghai, China, 201210*

[‡]*Experimental Drug Development Centre (EDDC), Agency for Science, Technology and Research (A*STAR), 10 Biopolis Road, Chromos, Singapore, 138670*

[¶]*School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-SZ), 2001 Longxiang Boulevard, Shenzhen, China, 518172*

[§]*Contributed equally to this work*

E-mail: wxx@chemlex.tech

Abstract

As a fundamental problem in organic chemistry, synthesis planning aims at designing energy and cost-efficient reaction pathways for target compounds. In synthesis planning, it is crucial to understand regioselectivity, or the preference of a reaction over competing reaction sites. Precisely predicting regioselectivity enables early exclusion of unproductive reactions and paves the way to designing high-yielding synthetic

routes with minimal separation and material costs. However, it is still at emerging state to combine chemical knowledge and data-driven methods to make practical predictions for regioselectivity. At the same time, metal-catalyzed cross-coupling reactions have profoundly transformed medicinal chemistry, and thus become one of the most frequently encountered types of reactions in synthesis planning. In this work, we for the first time introduce a chemical knowledge informed message passing neural network(MPNN) framework that directly identifies the intrinsic major products for metal-catalyzed cross-coupling reactions with regioselective ambiguity. Integrating both first principle methods and data-driven methods, our model achieves an overall accuracy of 95.24% on the test set of eight typical metal-catalyzed cross-coupling reaction types, including Suzuki-Miyaura, Stille, Sonogashira, Buchwald-Hartwig, Hiyama, Kumada, Negishi, and Heck reactions, outperforming other commonly used model types. To integrate electronic effects with steric effects in regioselectivity prediction, we propose a quantitative method to measure the steric hindrance effect. Our steric hindrance checker can successfully identify regioselectivity induced solely by steric hindrance. Notably under practical scenarios, our model outperforms 6 experimental organic chemists with an average working experience of over 10 years in the organic synthesis industry in terms of predicting major products in regioselective cases. We have also exemplified the practical usage of our model by fixing routes designed by open-access synthesis planning software and improving reactions by identifying low-cost starting materials. To assist general chemists in making prompt decisions about regioselectivity, we have developed a free web-based AI-empowered tool. Our code and web tool have been made available at <https://github.com/Chemlex-AI/regioselectivity> and <https://ai.tools.chemlex.com/region-choose>, respectively.

Introduction

Regioselectivity describes the preference of a reaction to occur at a specific site within a molecule when multiple sites are available (Figure 1a). Accurately predicting regioselectivity

for chemical reactions is crucial for designing feasible and high-yielding synthetic routes with minimal separation and material costs.¹ Though human experts have accumulated rich experience during practice, predicting regioselectivity is still highly challenging for most experts, especially for molecules where competing reactive centers have subtle differences in intrinsic stereoelectronic reactivity. Conventionally, when given a new reaction with undetermined regioselectivity concerns, human experts need to study the mechanism of the reaction and conduct small-scale experiments to determine the results. Given such challenges often arise when the target molecule reaches a level of substantial complexity, any chemical material generated would often be considered precious, thus making such experimental approaches expensive. Moreover, this single-step product optimization may take several days, significantly hindering the process of synthetic route determination. Therefore, a fast computational tool to assist human experts during this process would have an invaluable impact and be in high demand.

Concurrently, the recent integration of artificial intelligence(AI) into computer-aided synthesis planning(CASP) has profoundly revolutionized early-stage drug discovery and pre-clinical manufacturing process development.⁵⁻¹³ Not only useful for saving opportunity and tangible costs for bench scientists, fast and accurate computational methods to determine regioselectivity are also critical for designing green and efficient synthetic routes in CASP.¹⁴⁻¹⁶ Aligning with such unmet needs, computational methods to predict regioselectivity for various organic chemistry reactions have been long pursued. However, despite relentless endeavors over decades, developing high-performing predictive computational tools for regioselectivity remains a significant challenge. In the 1990s, with limited computational power, scientists focused on using feature engineering to learn the correlation between molecular descriptors and regioselectivity.¹⁷ Oslob and co-workers¹⁸ used Quantitative Structure-Activity Relationship (QSAR) to predict the regioselectivity for palladium-catalyzed allylations. They calculated the energy for steric probing, the bond length from palladium to the reacting center carbon, and two dihedral angles, to fit the experimental selectivity data. Nonethe-

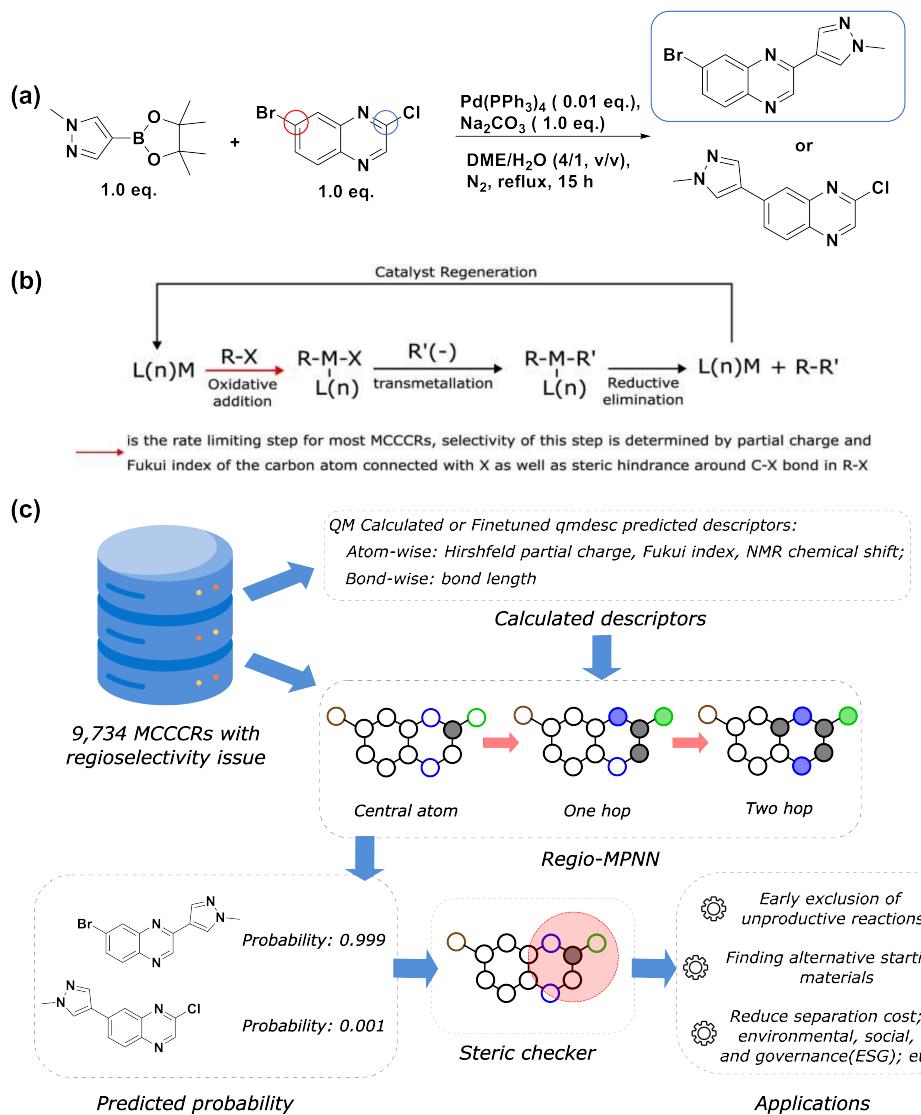


Figure 1: Regioselectivity in Metal-Catalyzed Cross-Coupling Reactions. (a) An example of regioselectivity challenge in a Suzuki-Miyaura reaction. In the heteroaryl halide, there are two competing coupling sites, marked by blue and red circles. As reported by Saxty *et al.*,² the blue circle denotes the major coupling site with the corresponding major product (with a yield of 82 %) highlighted in the blue rectangle; (b) The general mechanism for Metal-Catalyzed Cross-Coupling reactions(MCCCRs); (c) An overview of this work. The dataset comprising 9,734 MCCCRs with regioselectivity ambiguity is licensed from Pistachio³ and CAS Content Collection.⁴ Calculated descriptors, together with atom and bond features are passed through Regio-MPNN to get the predicted probability for each candidate product. A steric hindrance checker guarantees that the predicted major reaction site is within the “safe” steric hindrance range. Black circles stand for carbon atoms, blue circles stand for nitrogen atoms, green circles stand for chlorine atoms, brown circles stand for bromine atoms, and filled circles stand for neighboring atoms considered at each message-passing step.

less, to ensure the performance, data points with relatively high activation energy were removed from their data set, limiting their method's application and imposing unresolved bias. Banerjee and co-workers¹⁹ used various Machine Learning(ML) tools to analyze the outcome of catalytic regioselective difluorination reactions of alkenes and decipher the complex interplay of various molecular parameters and their non-linear dependencies. In their work, they used Density functional theory (DFT) to compute molecular parameters for 66 alkenes and further discovered the dependencies between these molecular parameters and regiochemical outcomes. However, computing these parameters via DFT is relatively time-consuming, underscoring the need for more efficient approaches. To enhance computational efficiency, Hong *et al.*²⁰ used ML models trained on DFT results to predict descriptors of interest. They combined 32 key physical organic descriptors and developed a regression algorithm to predict regioselectivity for radical C-H functionalization of heterocycles. Their work achieved a rapid and reliable prediction of descriptors through ML models successfully. However, the approach of featurizing reactants solely with physical organic descriptors may lead to missing critical structural information about molecules. Ree and co-workers²¹ calculated Charge Model 5 (CM5)²² atomic charges and used them to predict the regioselectivity of electrophilic aromatic substitution reactions with a light gradient boosting machine. Similarly, Tomberg *et al.*²³ also focused on electrophilic aromatic substitution reactions. They computed Fukui coefficient, partial charge, bond orders, and atomic solvent accessible surface for each aromatic carbon and found random forest models achieved the best performance for classifying carbons as active or inactive in electrophilic aromatic substitution. Guan *et al.*¹⁷ introduced a new method that combines graph-based molecular representation with simulated chosen quantum mechanical descriptors to predict regioselectivity in substitution reactions. After that, they focused on the regioselectivity problem in nucleophilic aromatic substitution (SNAr) reactions and attached a checker at the end of their model to distinguish the possible products with similar predictive scores.²⁴ In Ree's, Tomberg's, and Guan's work, they all treated the electronic properties as the major factors for regioselectivity in

their interested reactions and did not take steric effects into account explicitly. Currently, there is no predictive model available that combines both electronic and steric effects with data-driven deep learning methods.

On the other hand, medicinal chemistry has been profoundly reshaped by metal-catalyzed cross-coupling reactions(MCCCRs) due to their impressive ability to forge carbon-carbon/carbon-heteroatom bonds between diverse chemical moieties, enabling the creation of compounds that are otherwise challenging to obtain.²⁵ Palladium-catalyzed cross-coupling reactions, such as the Suzuki-Miyaura, Buchwald-Hartwig, Heck, and Stille reactions, have stood among the most popular reaction types in medicinal chemistry.²⁵ As shown in Figure 1b, the mechanism of MCCCRs generally involves:²⁶ 1. Oxidative addition converting an organic halide (RX) to $\text{L}(\text{n})\text{MR}(\text{X})$ at the presence of catalyst $\text{L}(\text{n})\text{M}$ (L = spectator ligand). 2. Transmetalation converting $\text{L}(\text{n})\text{MR}(\text{X})$ to $\text{L}(\text{n})\text{MR}(\text{R}')$, where the source of $\text{R}'(-)$ varies in different metal coupling reactions. For Suzuki, $\text{R}'(-)$ comes from a boronic acid or the corresponding ester. For Buchwald, Heck, Sonogashira, Negishi, Stille, Hiyama, and Kumada reactions, the sources are amine, alkene, alkyne, organozinc, -tin, -silicon reagent and Grignard reagent, respectively. 3. Reductive elimination of $\text{L}(\text{n})\text{MR}(\text{R}')$ to regenerate catalyst $\text{L}(\text{n})\text{M}$ and to release the resulting product $\text{R}-\text{R}'$. Among the three steps, oxidative addition is generally considered the rate-limiting step for most MCCCRs.²⁷ The extent of this step is heavily influenced by the leaving group ability of X and the steric hindrance around the $\text{C}-\text{X}$ bond.²⁶ Therefore, it is possible to use quantitative descriptors, *e.g.* partial charge, Fukui index, and volumetric measures from conformational analysis, to describe the properties of the rate-limiting step, and further characterize the different kinetics between competing reaction sites where regioselectivity is considered. Additionally, regioselectivity-related experimental results have been widely reported in the literature, enabling data-driven methods to mine statistical rules behind the literature data. As a result, a computational model that accurately predicts the regioselectivity is thus made possible by both theoretical analysis and data-driven machine learning.

In this work, we propose Regio-MPNN, Message Passing Neural Network(MPNN) backbone combined with chemical descriptors, to directly predict intrinsic major products for MC-CCRs with regioselectivity risks, as shown in Figure 1c. We use computed atomic charges, Fukui index, Nuclear Magnetic Resonance(NMR) chemical shift, and bond length through DFT calculation to train an MPNN model based on the graph representation of a molecule, and examine the steric hindrance effect on possible reactive sites through a steric hindrance checker. It is worth noting that in practice, regioselectivity can also be affected by reaction conditions.^{28–30} However, literature has limited reported records on the relationship between regioselectivity and reaction conditions, making it difficult to capture this effect using data-driven methods. Therefore, in this work, we only focus on the intrinsic properties of reactants and ignore the differences caused by reaction conditions. Our model exhibits an overall accuracy of 95.24 % on a test set comprising eight types of metal-catalyzed cross-coupling reactions, demonstrating capability for practical usage. Our model architecture outperforms other frequently used model types, including Extended-Connectivity Fingerprints (ECFPs)³¹ based multilayer perceptron (MLP), descriptors based MLP, and sequence based model + MLP, regarding prediction accuracy and robustness. We have also demonstrated that our steric hindrance checker is able to detect regioselectivity solely induced by steric hindrance regardless of electronic effects. Furthermore, we invited 6 experimental organic chemists with an average working experience of over 10 years in industry to compete with our model on regioselectivity tests. In this test, we randomly picked 100 reactions from the test set and collected the predictions of major products from the chemists and our model. Our model significantly outperformed human chemists, demonstrating the advantage of using machine learning methods with first-principle results. In the end, we show the practical usage cases of our model on synthesis planning and material cost saving by finding more accessible starting materials.

Methods

Data preparation

Data used in this manuscript came from two commercial datasets: Pistachio³ and Chemical Abstracts Service (CAS) Content Collection.⁴ For the Pistachio dataset, we first used NameRxn numbers provided together with the data to pick out each cross-coupling reaction, the Buchwald, Heck, Hiyama, Kumada, Negishi, Sonogashira, Stille, and Suzuki reactions, respectively. Then, for each reaction type, we applied several filters to get a dataset with 8,923 data points. The filters added are discussed in detail in the following paragraph. To enhance the data balance among different metal-catalyzed cross-coupling reactions, we also licensed 811 curated data points from CAS Content Collection, producing a dataset with 9,734 data points.

The data cleaning process is as follows. Here we take the Suzuki-Miyaura cross-coupling reactions as examples. First, we removed the duplicate reactions. Then we filtered out reactions with yields less than 40% to ensure the product in the dataset is indeed the preferred reacting site. Next, we used SMARTS to match each reactant and classify reactants into two types: organohalides or organoboron.³² The SMARTS of the reactant organohalides is defined as “[F,Cl,Br,I,\$(OS(=O)(=O)C(F)(F)F)][#6]”, which covers aryl halides and allyl halides. Here, the “generalized” halides are not limited only to halogens but also include halogen-like functional groups, *e.g.* trifluoromethane sulfonate (OTf). SMARTS of reactant organoboron is defined as “B(O)O”, which matches boronic acid or boronate ester. In order to be identified as Suzuki reactions, all reactions should have at least one organohalide and one boronic substance. Then another filter was applied to pick only organohalides with more than 2 halogen and halogen-like leaving groups. Finally, we use rdkit.Chem.rdChemReactions module to enumerate all possible products.³³ The SMIRKS we used for the Suzuki reactions is: “B(O)(O)([#6:1]).[Cl,Br,I, \$(OS(=O)(=O)C(F)(F)F)][#6:2]>>[#6:1]-[#6:2]”. For all the obtained products after applying the reaction SMIRKS, the product that

is the same as in the reaction dataset is labeled with 1, which is the ground truth, and all the rest of the possible products are labeled as 0. Other types of MCCR_s other than Suzuki reactions were treated similarly with their own chemical definitions respectively. The corresponding SMIRKS for each reaction type can be found in our GitHub repository.³⁴

Density Functional Theory(DFT) calculation

Reaction data were extracted from the aforementioned commercial datasets according to the reaction SMARTS templates. Then the reactants were extracted from those reactions for further DFT calculations. The DFT descriptors used in the model were calculated by an automated computational workflow developed as in Guan *et al.*^{17,35} The input for the workflow is a list of SMILES. The output for the workflow is the atom descriptors of each molecule. Firstly, each of the SMILES strings was processed by RDKit³³ using the Merck Molecular Force Field(MMFF94s) to obtain initial structure coordinates for each molecule.³⁶ Then the molecular structure was optimized at the GFN2-xtb level of theory.^{37–39} We made sure there were no imaginary frequencies to guarantee the structure was correctly optimized. Next, the optimized structure was used to calculate the Fukui index with Density functional theory(DFT). The DFT calculations were performed using Gaussian 16,⁴⁰ with the B3LYP functional and def2svp basis set. Then, the target descriptors, Hirschfeld partial charge, Fukui index, and bond lengths were extracted from the Gaussian outputs. Details of the DFT descriptors are included in supporting information(SI).

MPNN for regioselectivity prediction

MPNN is a type of graph neural network suitable to learn molecular features wherein atoms are vertices and bonds are edges.^{41–43} Concretely, a molecule can be represented as a graph $\mathcal{G} = (\mathbf{A} \in \mathbb{R}^{l \times a}, \mathbf{B} \in \mathbb{R}^{l \times l \times b})$, where \mathbf{A} is the matrix of atom features and \mathbf{B} is the adjacency tensor of bond features. Here, l is the number of atoms in the molecule, a is the dimension of atom features, and b is the dimension of bond features. Such a graph \mathcal{G} serves

as the input for the MPNN framework. In an MPNN layer, there are two fundamental steps for a forward pass: a message-passing step and an update step. The number of stacked MPNN layers typically depends on the number of connected bonds to be considered around the central atom.⁴¹

A message-passing step reads

$$\mathbf{m}_v^{t+1} = \sum_{n \in N(v)} M(\mathbf{h}_v^t \parallel \mathbf{h}_n^t \parallel \mathbf{B}^{(v,n)}) \quad (1)$$

where, t denotes the index of step; $\mathbf{m}_v^{t+1} \in \mathbb{R}^{l_m}$ denotes the new message of atom v at the next time step $t + 1$ with dimension l_m ; $N(v)$ denotes the neighboring atom indices of v ; $\mathbf{h}_v^t \in \mathbb{R}^{l_h}$ denotes the hidden state for atom v at time step t with dimension l_h ; $\mathbf{h}_n^t \in \mathbb{R}^{l_h}$ denotes the corresponding atom features for atom n at time step t where $n \in N(v)$; the initial hidden states $\mathbf{h}_v^0 \in \mathbb{R}^a$ and $\mathbf{h}_n^0 \in \mathbb{R}^a$ are vectors of length a that denotes the v th and n th element of atom features matrix \mathbf{A} , correspondingly; $\mathbf{B}^{(v,n)} \in \mathbb{R}^b$ is a vector of length b that denotes the (v, n) element of the adjacency tensor \mathbf{B} ; $\cdot \parallel \cdot$ denotes the operation of vector concatenation; M is the message neural network as a mapping, $M(\mathbf{x}_M) = \text{ReLU}(\mathbf{W}_M \cdot \mathbf{x}_M + \mathbf{b}_M)$, ReLU is the rectified linear unit activation function, $\mathbf{W}_M \in \mathbb{R}^{l_m \times (2l_h+b)}$ and $\mathbf{b}_M \in \mathbb{R}^{l_m}$ are the weights and bias of M respectively.

In the update step,

$$\mathbf{h}_v^{t+1} = U(\mathbf{h}_v^t \parallel \mathbf{m}_v^{t+1}) \quad (2)$$

where, U denotes the update neural network as a mapping, $U(\mathbf{x}_U) = \text{ReLU}(\mathbf{W}_U \cdot \mathbf{x}_U + \mathbf{b}_U)$, $\mathbf{W}_U \in \mathbb{R}^{l_h \times (l_h+l_m)}$ and $\mathbf{b}_U \in \mathbb{R}^{l_h}$ are the weights and bias of the update neural network U respectively.⁴¹

An overall pipeline of our model is shown in Figure 2. At the beginning, input reactant pairs were fed into DFT computation module or finetuned qmdesc model(Fqmdesc)¹⁷ to obtain atom-wise descriptors (Hirshfeld partial charge, nucleophilicity, and electrophilicity) and bond-wise descriptors (bond length). At the same time, atom and bond features of

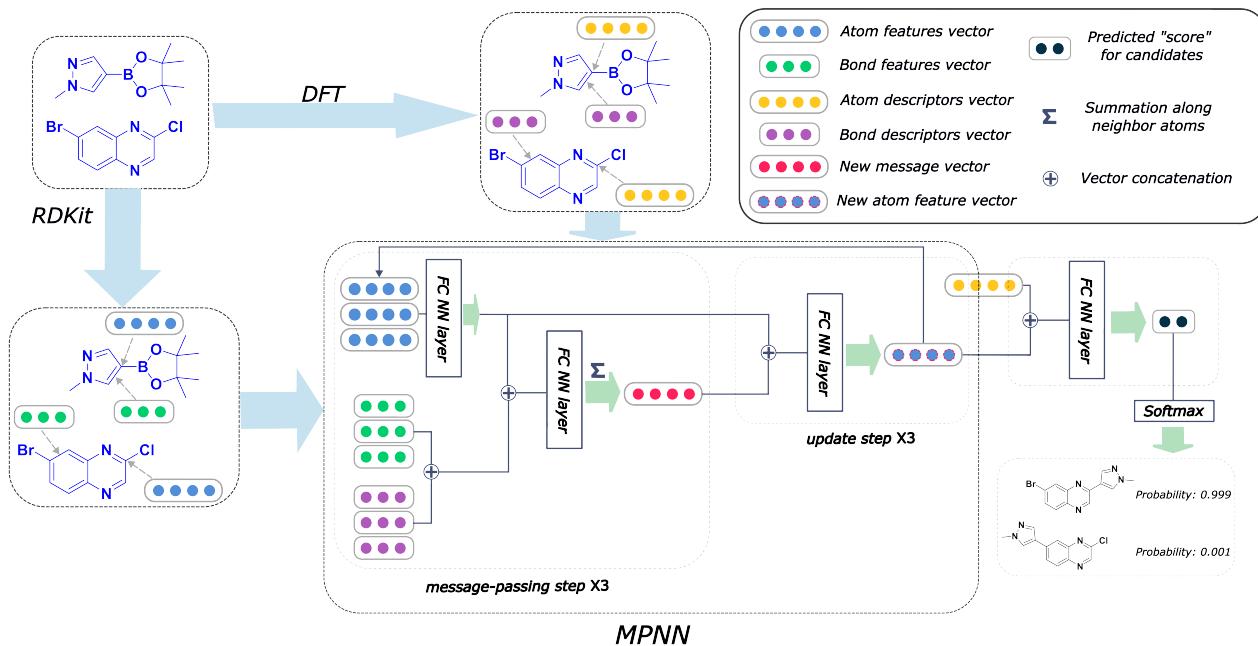


Figure 2: Architecture of Regio-MPNN. We calculate the atom and bond descriptors using density functional theory (DFT) or a finetuned qmdesc model(Fqmdesc),¹⁷ at the same time the atom features and bond features are computed using RDKit. We combine the quantum mechanics (QM) descriptors with the features and feed the overall input into the MPNN network, which outputs a learned embedding of the reactants. We then use this embedding together with the atom descriptors to predict the probabilities for candidate products. A steric hindrance checker (shown in Figure 1c) is used after Regio-MPNN to filter out steric hindered results.

input reactant pairs were extracted using RDKit.³³ Extracted atom and bond features are described in SI Table S2. Atom features were concatenated with atom descriptors and bond features were concatenated with bond descriptors. Together, the concatenated atom and bond features were fed into the MPNN module as the input **A** and **B**, respectively, to perform representation learning for atoms. This new representation, together with the descriptors, was used to compute the probability of being the main product for each candidate product. This probability was estimated as follows. We first calculated a score for each candidate product using Equation 3,

$$s = S\left(\sum_c R(\mathbf{h}_c^T \parallel \mathbf{d}_c)\right) \quad (3)$$

where $s \in \mathbb{R}$ is the resulting score; $\mathbf{h}_c^T \in \mathbb{R}^{l_h}$ is the hidden state of atom c learned by the

MPNN at final time step T , c belongs to reaction center atoms, reaction center atoms are the atoms which undergo changes in their bonding pattern during the course of MCCCRs, for example, the carbon atom attached to the boronic acid or the corresponding ester group and the carbon atom attached to the reacting halogen or OTf group in a Suzuki-Miyaura reaction are the reaction center atoms; $\mathbf{d}_c \in \mathbb{R}^{l_d}$ is the QM computed or Fqmdesc computed atom descriptors with a dimension of l_d for atom c ; R is a neural network as a mapping, $R(\mathbf{x}_R) = \mathbf{W}_R \cdot \mathbf{x}_R$, $\mathbf{W}_R \in \mathbb{R}^{l_h \times (l_h + l_d)}$ is the weights of the neural network; S is a neural network as a mapping, $S(\mathbf{x}_S) = \mathbf{W}_S \cdot \mathbf{x}_S + b_S$, $\mathbf{W}_S \in \mathbb{R}^{1 \times l_h}$ and $b_S \in \mathbb{R}$ are the weights and bias of the neural network. After computing the scores for each candidate product, the probabilities of being the main product for each candidate product were calculated by feeding the concatenation of the scores through a softmax activation function.

Steric hindrance checker

As described in Oslob *et al.*,¹⁸ steric effect has a significant impact on selectivity results in the oxidative addition step of MCCCRs. Therefore, a steric hindrance checker was attached to filter out heavily steric-hindered reactive sites in the inference stage. To compute the steric hindrance effect of the reactive sites, a conformer of the molecule was generated by the ETKDG method from RDKit.⁴⁴ We located the reaction center carbon with the carbon-halogen bond or carbon-triflate bond, placed this carbon atom as the center, and considered a sphere with a radius equal to 5 Å. We defined the steric hindrance value as the occupied volume by other atoms (calculated using van der Waals radii) in the sphere divided by the total volume of the sphere. We then analyzed the steric hindrance values for all metal-catalyzed cross-coupling reaction entries in Pistachio³ and CAS Content Collection,⁴ chose 95 % percentile of the steric hindrance values and got 0.60 as the maximum “safe” steric hindrance value. After removing carbon-halogen or carbon-triflate sites with steric hindrance values greater than the threshold, we determined the product with the highest probability in the remaining candidates as the predicted main product.

Results and Discussions

Analysis of reaction data

We first analyze the reaction data in our cleaned commercially available dataset. Figure 3a shows the distribution of data from each reaction type extracted from the Pistachio dataset. Suzuki coupling reactions overweight all the other seven coupling reactions in terms of volume. The phenomenon is not unexpected as the Suzuki coupling reaction has been extensively used in the synthesis of various compounds in synthetic organic chemistry.^{45,46} In order to improve the data balance, we supplement the data for Heck, Hiyama, Kumada, and Negishi reactions using data from the CAS Content Collection.⁴ Figure 3b shows the distribution of data from each reaction type in the dataset combining both Pistachio and CAS Content Collection datasets. After adding high-quality data from CAS Content Collection, there is an increase in the data portion of Heck, Hiyama, Negishi and Kumada coupling reactions. Especially, the portion of Heck reaction increased from 0.2% to 4.6%. Considering the various catalysts used in MCCR_s, we also analyze the data distribution of metals used in the datasets. Most reactions extracted from Pistachio use Palladium-based catalysts and only Sonogashira reactions use Zinc-based catalysts. In total, 11 different metals are used as catalysts in selected CAS Content Collection data. A summary of different metal elements used as catalysts in CAS Content Collection data is provided in Table S1.

To get a deeper understanding of our dataset, we plot the competing reaction sites with regioselectivity risks in the overall cross-coupling reaction dataset in Figure 3c. The vertical axis denotes the ground-truth halogens or halogen-like group corresponding to the major product, and the horizontal axis denotes other potential reactive sites present in the same reactant. For instance, the deepest blue section in the heatmap represents that the most commonly seen competing site pair in MCCR_s is Br vs. Cl (3,599 data points),

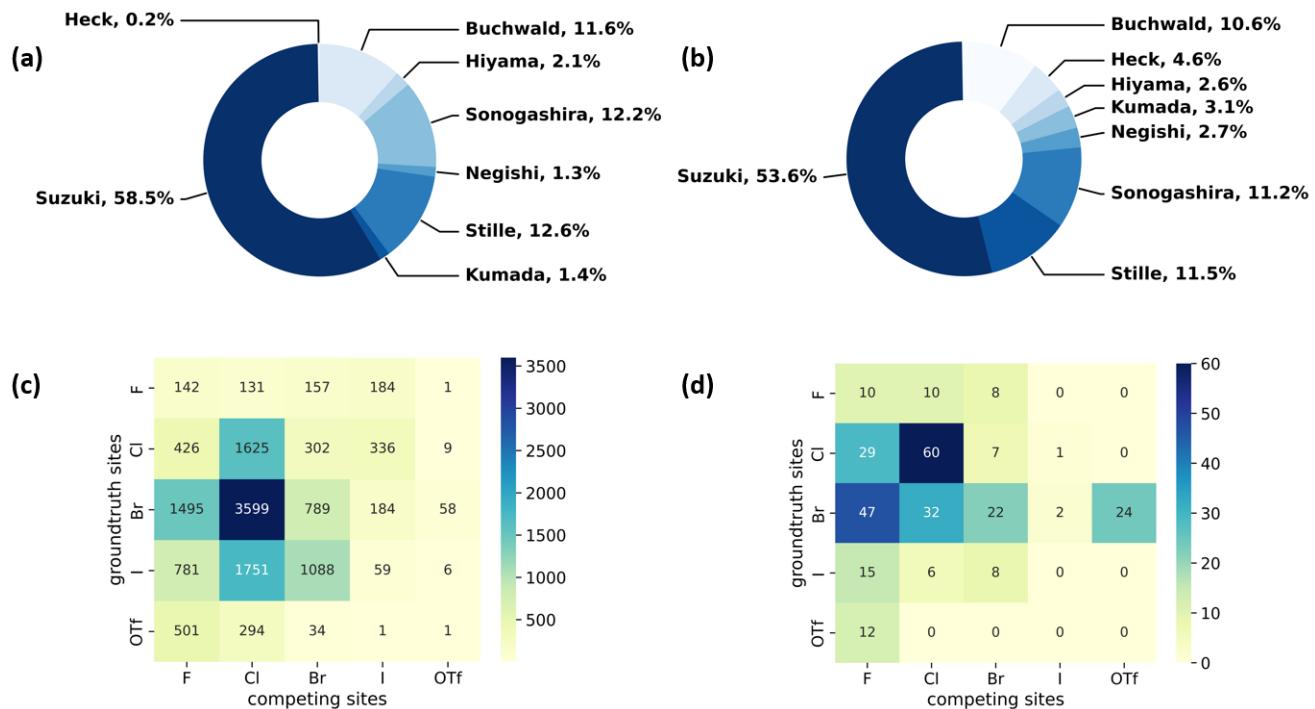


Figure 3: Reaction Type Distribution for Metal-Catalyzed Cross-Coupling Reactions in (a) Pistachio and (b) Pistachio + CAS Content Collection; The analysis of competing reaction site pairs in the Pistachio + CAS Content Collection metal-catalyzed cross-coupling reaction dataset (c) for all reactions in the dataset and (d) for Kumada reactions.

where the actual reactions occur at the bromine-substituted sites. The total number in the heatmap exceeds the total number of our dataset (9,734) because there may be more than two competing sites in one molecule.

Analysis of model performance

We split the combined dataset comprising 9,734 data points from Pistachio³ and CAS Content Collection⁴ into training/validation/test by a ratio of 8:1:1. To ensure the fidelity of the estimation of our model's generalization ability, we incorporated a stratified sampling strategy.⁴⁷ In other words, reactions in the test set do not contain any reactants that are seen in the training or validation sets. In addition, we made sure that each reaction type is distributed among the three sets according to their distribution in the overall dataset shown in Figure 3b.

Table 1: Performance of Regio-MPNN and other model architectures on the stratified sampled test set. MLP stands for multilayer perceptron.

Model	Molecular Structure Info	Descriptors Used	Accuracy %
Random guess	✗	✗	41.74 ± 4.83
ECFPs based MLP	✓	✗	46.34 ± 2.56
Descriptors based MLP	✗ ✗	DFT Fqmdesc	60.52 ± 5.98 61.40 ± 1.93
Sequence based model + MLP	✗ ✗ ✗	✗ DFT Fqmdesc	62.53 ± 5.32 71.65 ± 3.05 73.51 ± 2.63
Regio-MPNN	✓ ✓ ✓	✗ DFT Fqmdesc	61.46 ± 1.57 95.03 ± 0.56 95.24 ± 0.57

We implement different model architectures to predict the main product for MCCCRs with regioselectivity risks (Table 1). The random guess refers to randomly picking a product from the candidate products and treating it as the main product. Besides the aforementioned graphic representation in the MPNN framework, other frequently used representations for chemical compounds or reactions, such as Extended-Connectivity Fingerprints (ECFPs),³¹ Simplified Molecular Input Line Entry System (SMILES) sequence,⁴⁸ and chemical descrip-

tors, are used to do the same prediction task. Detailed implementation of each representation is illustrated in SI. Robustness of each model is evaluated by a five-fold cross-validation in which the data splitting also follows the stratified sampling strategy.⁴⁷ Accuracy for each model is the average accuracy of the model running on different data splits for five times and robustness for each model is the maximum deviation between the average accuracy and a one-time running accuracy. Accuracy and robustness results for different models are shown in Table 1. The overall accuracy on the test set using only MPNN is 61.46 %, while the overall accuracy can increase to 95.24 % when integrating Fqmdesc-calculated descriptors into the MPNN, emphasizing the necessity of introducing chemical descriptors to the model. The accurate results of MPNN + DFT model and MPNN + Fqmdesc model indicate that the difference between using DFT- vs Fqmdesc- calculated descriptors is negligible. Therefore, Fqmdesc can be an effective alternative to the DFT calculation for inference, reducing the total inference time by ~20000 times. Putting this into context, calculating descriptors with DFT takes 1.5 days vs 6 seconds with Fqmdesc for the same data in the test set.

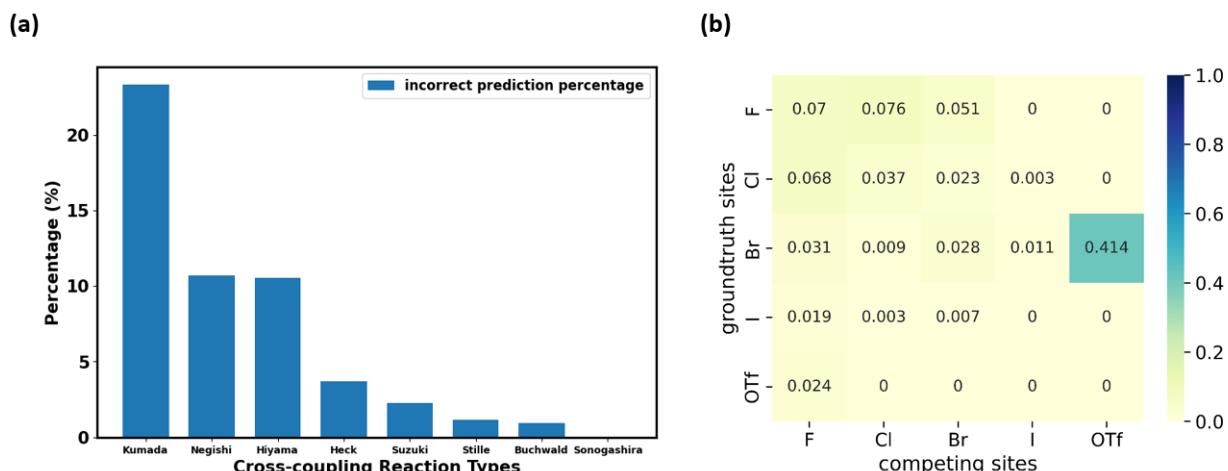


Figure 4: The analysis of incorrect predictions. (a) The percentage of incorrect predictions in different cross-coupling types in the test set by Regio-MPNN with Fqmdesc computed chemical descriptors; (b) The posterior probability of a reaction being categorized into Kumada coupling given fixed competing site pairs.

We also evaluated the accuracy of prediction among different cross-coupling reaction types in the test set. As shown in Figure 4a, four reaction types with relatively fewer data

points(as shown in Figure 3b), Kumada, Negishi, Hiyama, and Heck, have relatively higher percentage of erroneous regioselectivity prediction. This could be rationalized by the lack of data in the data set for these reaction types (Figure 3b). Even though we have combined all the relevant data from pistachio and CAS Content Collection⁴ datasets, the extent of data imbalance is still significant. To mitigate this, data scientists typically use measures such as data augmentation,⁴⁹ under-sampling,⁵⁰ and weighted loss.⁵¹ However, data augmentation may not be a suitable solution for the task at hand, because regioselectivity may vary even with small structural changes, which means the generation of new data points from scratch or from existing data points for rare reaction types would not be feasible. Under-sampling the majority reaction type leads to a small training set and would significantly reduce the ability to generalize our model (Detailed under-sampling experiments are shown in SI). Unlike classification tasks, the scaling strategy in the loss function can not be applied to our model either. Thus, in the future, the best solution to the data imbalance is to intentionally acquire more wet lab data from high throughput experiments, as proposed by recent literature.⁵²⁻⁵⁴ However, this approach is out of the scope of this work.

The reaction type that challenged our prediction model the most was Kumada coupling reactions (Figure 4a). Surprisingly, as shown in Figure 3b, Kumada reactions are not the rarest reaction type in the dataset. In order to understand the uniqueness of Kumada reactions, we conduct a fine-grained analysis of the data distributions. We notice the obvious difference in competing site distribution between all reactions in the overall data set and Kumada reactions as shown in Figure 3c-d. To quantify this difference, Kullback–Leibler(KL) divergence,⁵⁵ Jensen–Shannon(JS) divergence,⁵⁶ and Bhattacharyya coefficient⁵⁷ are computed with Equation 4 and the difference of competing site pairs among different reaction types is shown in Table 2.

$$KL^k = \sum_i \sum_j P(i, j | k) \log \frac{P(i, j | k)}{P(i, j)}$$

$$JS^k = \sum_i \sum_j \left(\frac{1}{2} P(i, j|k) \log \frac{P(i, j|k)}{\frac{1}{2} P(i, j|k) + \frac{1}{2} P(i, j)} + \frac{1}{2} P(i, j) \log \frac{P(i, j)}{\frac{1}{2} P(i, j|k) + \frac{1}{2} P(i, j)} \right)$$

$$BC^k = \sum_i \sum_j \sqrt{P(i, j|k) \times P(i, j)} \quad (4)$$

where, KL is short for KL divergence; JS is short for JS divergence; BC is short for Bhattacharyya coefficient; superscript k denotes the specific reaction type; $P(i, j|k)$ denotes the percentage of regioselectivity sites for ground-truth site i and competitor site j in reaction type k ; $P(i, j)$ denotes the percentage of regioselectivity sites for ground-truth site i and competitor site j in the whole data set.

Among the 8 reaction types, Kumada reactions exhibit the highest KL and JS divergence as well as the lowest Bhattacharyya coefficient compared to the entire dataset, which indicates the competing site pairs in Kumada reactions are significantly different from those in the entire data set. To better understand this difference, we calculate an approximation of the posterior probability of reaction type k given competing site pairs i, j using Equation 5, and this posterior probability for Kumada reactions is shown in Figure 4b.

$$P(k|i, j) = \frac{P(i, j|k) \times P(k)}{P(i, j)} \quad (5)$$

where, $P(k|i, j)$ denotes the approximated posterior probability of reaction type k given ground-truth site i and competing site j ; $P(k)$ denotes the percentage of reaction type k in the whole data set.

An interesting fact is discovered that the competition between a Br vs. OTf with Br as the preferred site is a signature to Kumada reactions. This is unexpected since the reactivity for Bromine sites and that for OTf sites in MCCRAs are normally considered as closely similar and hard to distinguish,⁵⁸ which could be attributable to the relatively low accuracy for predicting regioselectivity for Kumada reactions. Moreover, among all the reaction types we are considering, the Kumada reactions require strict experimental condition control owing to the high reactivity of Grignard reagent with water or with functional groups

in other reactants,⁵⁹ which also makes Kumada reactions unique among MCCCCRs. Figures of competing site pair distribution and approximated posterior probability for other reactions are available in SI Figure S3 to Figure S9.

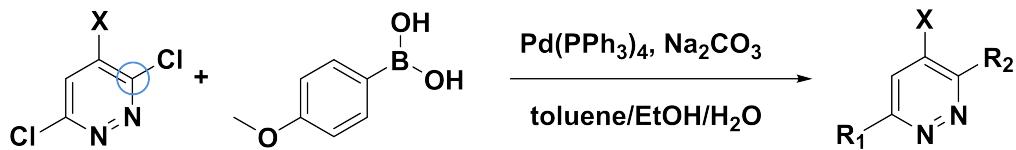
Table 2: The distribution deviation of competing site pairs in each reaction category compared to the overall dataset.

Reaction Type	KL divergence	JS divergence	Bhattacharyya coefficient
Buchwald	0.065	0.027	0.941
Heck	0.127	0.052	0.930
Hiyama	0.114	0.054	0.936
Kumada	0.219	0.101	0.876
Negishi	0.068	0.033	0.945
Sonogashira	0.164	0.058	0.909
Stille	0.142	0.058	0.919
Suzuki	0.035	0.011	0.975

Steric hindrance checker case study

To examine the effectiveness of our steric hindrance checker, we conduct a case study of a couple of MCCCCRs whose regioselectivity is dominantly affected by the steric hindrance. In Figure 5, four reported Suzuki-Miyaura reactions with regioselectivity risks are shown.⁶⁰ These four reactions undergo the same reaction condition but result in different primary reacting sites. The main products for the first three reactions correspond to the reaction site at the *ortho*- position with respect to the X group (circled in blue), while the main product for the last reaction corresponds to the *meta*- chlorine. Spivey *et al.* attribute this difference to steric hindrance effect.⁶⁰ Based on their analysis, we extract the steric hindrance values of carbon atoms highlighted by the blue circle in Figure 5 for these four reactions. The steric hindrance values for the first three reactions are 0.456, 0.507, 0.477, respectively, while the steric hindrance value for the last reaction is 0.603, above our steric hindrance threshold (0.6) determined by the statistics on the MCCCCRs in Pistachio³ and CAS Content Collection.⁴ The first three reactants are "safe" in terms of steric hindrance at the *ortho*- chlorine, but the last reaction is under steric hindrance "risk" at the same reaction site. Therefore, our

model predicts that the main products of the first three reactants are products with R_2 as reacting site and the main product of the last reactant is product with R_1 as reacting site, which agrees with experimental results.



X	Steric value	R ₁	R ₂	yield (%)
-NH ₂	0.456	-Cl	4-OMebenzyl	85
-NHBn	0.507	-Cl	4-OMebenzyl	77
-NMe ₂	0.477	-Cl	4-OMebenzyl	84
-NMeBoc	0.603	4-OMebenzyl	-Cl	63

Figure 5: Case study of steric impact on a series of Suzuki-Miyaura reactions. Regioselectivity may vary due to the crowded chemical environment around the reacting halogen group. Our steric hindrance checker can successfully capture this change. The yield reported here is from the literature.⁶⁰

Contest between human experts and Regio-MPNN

To compare model performance with predictions made by senior chemists, we randomly picked 100 reactions with regioselectivity risks from the test set and invited six senior chemists to identify the main products for these reactions independently. All of the human experts are from organic chemistry educational backgrounds and have worked in the chemical synthesis contract research organization (CRO) industry for more than 10 years. The accuracy of predictions made by the chemists varies from 74 % to 94 %, while the accuracy of predictions made by our model is 100 %, as shown in Table 3. Figure 6 shows some examples in which at least three senior chemists have made incorrect predictions. These examples indicate that predicting regioselectivity between reaction sites substituted by the same kind of halogen atoms remains a challenge for some human experts. In contrast, our

model can successfully determine the regioselectivity for these situations. We have developed a web tool⁶¹ empowered by Regio-MPNN to assist organic chemists when they encounter regioselectivity problems.

Table 3: Comparison of the performance of human experts and our model on 100 reactions randomly selected from the test set

Test Taker	Accuracy
Expert_1	94 %
Expert_2	91 %
Expert_3	81 %
Expert_4	80 %
Expert_5	78 %
Expert_6	74 %
Regio-MPNN	100 %

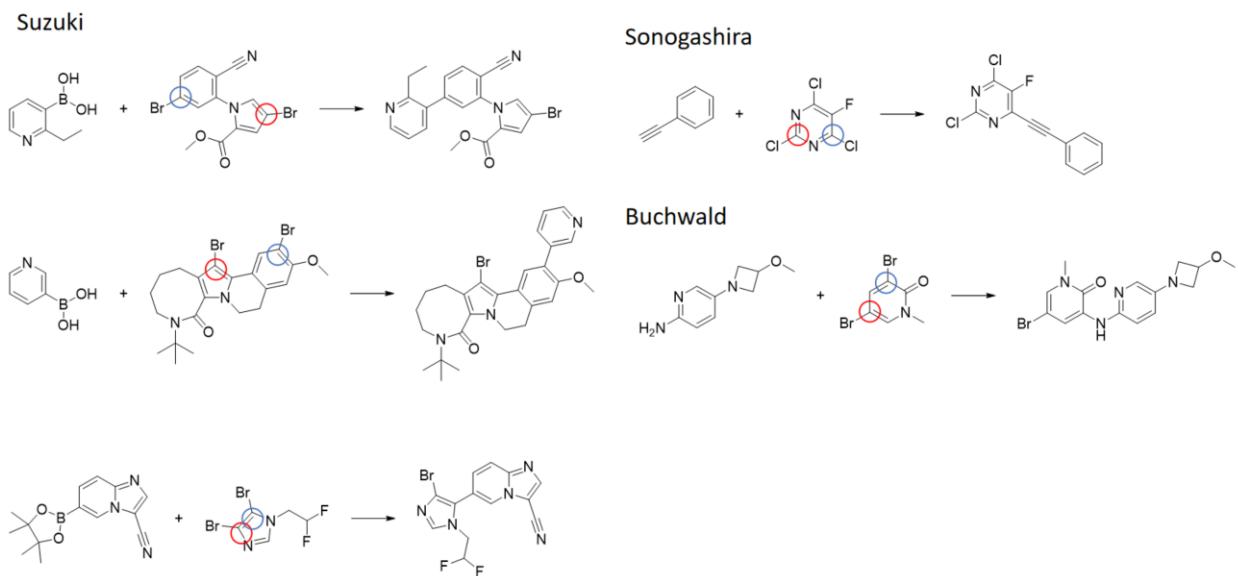
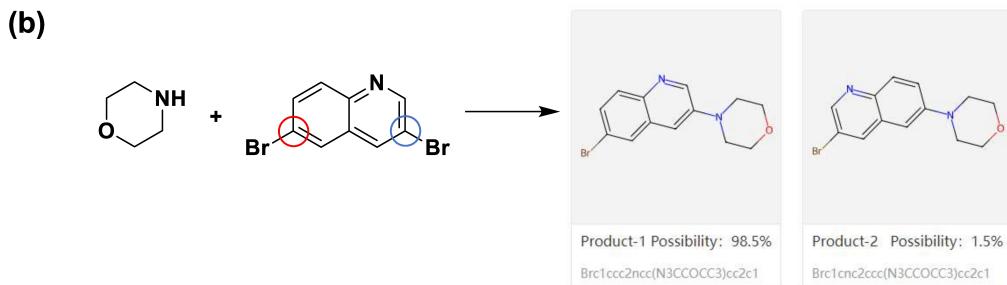
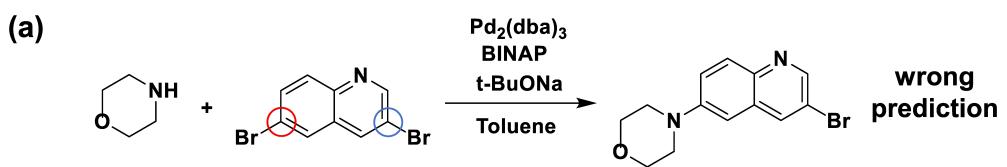
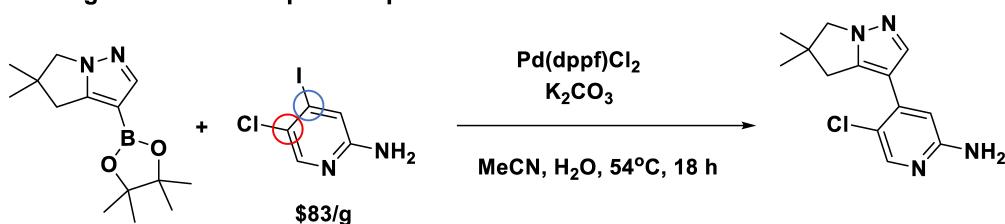


Figure 6: Examples of erroneous predictions made by at least half of the human experts. The reactions are from the 100 randomly sampled reactions with regioselectivity issues from test set. The ground-truths are shown as the product;^{62–66} The sites predicted by our model are marked by blue circles; The sites erroneously predicted by human experts are marked by red circles.



(c) an original reaction for process phase:



alternating reactant can reduce the cost:

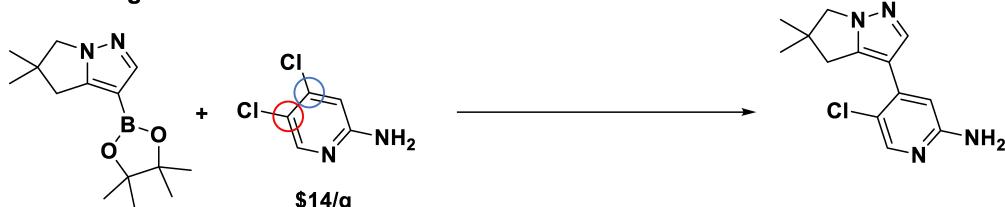


Figure 7: Applications of Regio-MPNN model. (a) An example of a regioselectivity mistake made by an AI-driven CASP system. This Buchwald-Hartwig reaction is used in a route designed by a CASP system with open online access. However, the ground truth⁶⁷ demonstrates that the blue-circled site is more active than the mistakenly chosen red-circled one; (b) A snapshot of the prediction result from our Regio-MPNN web tool for the Buchwald-Hartwig reaction in (a), showing that our model can fix the above regioselectivity mistake; (c) An example of how regioselectivity models can potentially save material costs by replacing expensive starting materials (upper reaction scheme)⁶⁸ with economical ones (lower) without significantly affecting the yield.

Model application case study

To further exemplify the practical application of our proposed model, we demonstrate two cases where Regio-MPNN can facilitate synthetic route planning for bench chemists in Figure 7a-c. Figure 7a-b shows an example where Regio-MPNN can potentially improve the performance of existing CASP systems. Recent open-source CASP software, *e.g.* ASKCOS,⁵ AiZynthFinder,¹² and RXN for Chemistry,⁷ have made a significant impact on the synthesis planning industry and have facilitated the process of organic synthesis profoundly. However, mistakes can still happen in regioselectivity prediction for AI-embedded CASP systems. For example, Figure 7a shows an erroneously planned Buchwald-Hartwig reaction with regioselectivity issue, recommended during synthesis planning by an open source CASP software. Integration of a regioselectivity model in such systems will be useful. This is exemplified when the correct prediction of the same Buchwald-Hartwig reaction was performed by our model as shown in Figure 7b, demonstrating the ability of our model to mitigate the regioslectivity issue for CASP systems. The second case is shown in Figure 7c, demonstrating the determination of regioselectivity can be useful in industrial process development. In Figure 7c, the original reaction from an industrial process⁶⁸ involves a reactant containing different halogen groups, chlorine and iodine, with a price of \$ 83 /g.⁶⁹ Typically such hetero-substituted starting materials are expensive and short in stock. Our regioselectivity model suggests an alternative starting material substituted with two chlorine atoms. The price of the latter compound is \$ 14 /g⁶⁹ and the yield is not significantly affected according to literature⁷⁰ as shown in SI. Thus, our model provides a potentially more economical method to develop new process routes with more accessible starting materials by leveraging fast and accurate predictions of regioselective preferences over competing reaction sites. Based on our model, a web tool⁶¹ (Figure 7b) has been made available to assist general bench chemists in making quick decisions about regioselectivity.

Conclusions

We analyzed metal-catalyzed cross-coupling reactions with regioselectivity issues from the pistachio and CAS Content Collection and applied MPNN together with DFT calculated descriptors or Fqmdesc computed descriptors to identify the correct main product. We also developed a statistical checker to take steric hindrance effect into consideration. The overall accuracy on a stratified sampled test set for our model is 61.46 % and the accuracy is significantly boosted to > 95% by adding DFT or Fqmdesc calculated descriptors. Our model outperformed other commonly used model architectures in terms of accuracy and robustness. We have also demonstrated that our steric hindrance checker is able to identify the cases where regioselectivity is solely induced by steric hindrance effects. On the basis of the superior performance of our model, we showed that intrinsic weaknesses in literature datasets, *e.g.* insufficient training data for rare competition pairs of reaction sites and data imbalance, are major reasons for relatively weaker predictions for certain types of reactions, proposing the potential approaches to collect experimental data in order to further advance the current method. In addition, we conducted a fine-grained analysis of the data distributions and demonstrated that Kumada reactions might pose unique challenges to regioselectivity prediction tasks compared to other types of metal-catalyzed cross-coupling reactions. We also collected the results of a set of regioselectivity challenges from 6 senior level organic chemists, and found that our model outperformed the human experts in this test. Based on the high accuracy of our model and the results of the competition with human experts, we demonstrate that our model can help senior chemists determine the regioselectivity in metal-catalyzed cross-coupling reactions in synthesis planning efficiently and accurately. We have also made a web based regioselectivity prediction tool for general chemists to use.

Acknowledgement

The authors thank members of the organic chemistry team at ChemLex for their insights on regioselectivity issues in Metal-Catalyzed Cross-Coupling Reactions and feedback on the 100-example test. We acknowledge the software development team at Chemlex for making our model available on the website. The authors would like to acknowledge that the data from the CAS Content Collection was provided under license for this study by CAS. CAS is a division of the American Chemical Society and provider of SciFinder[®], STN[®], and CAS Custom Services. www.CAS.org.

Data and Software Availability

The source code is available at <https://github.com/Chemlex-AI/regioselectivity> upon request. A Regio-MPNN based wet tool is also available at <https://ai.tools.chemlex.com/region-choose>. All reaction data used in the paper are available from commercial databases Pista-chio³ and CAS Content Collection.⁴ To obtain more information about the data, we encourage readers to directly contact the original authors.

Author Contributions

B.L. and X.W. conceived the project. B.L. and H.S. carried out the experiments. Y.L. processed the data and performed the DFT calculations. R.T.Z. and Y.X. organized the human expert test under the supervision of P.W. and S.L.. K.F. and F.M. proposed the model applications and reviewed the medicinal chemistry part. R.M.Z. and T.Y. discussed and guided the machine learning studies with B.L. and X.W.. X.W. supervised the project. B.L., Y.L. and X.W. prepared the manuscript. All authors discussed the results and contributed to the final manuscript.

Supporting Information Available

The following files are available free of charge.

- Supporting information file.
- Code is available at <https://github.com/Chemlex-AI/regioselectivity>.
- Web tool is available at <https://ai.tools.chemlex.com/region-choose>.

References

- (1) Li, J.; Eastgate, M. D. Making better decisions during synthetic route design: leveraging prediction to achieve greenness-by-design. *Reaction Chemistry & Engineering* **2019**, *4*, 1595–1607.
- (2) Saxty, G.; Murray, C. W.; Berdini, V.; Besong, G. E.; Hamlett, C. C. F.; Woodhead, S. J.; Ligny, Y. A. E.; Angibaud, P. R. Substituted quinoxalines as FGFR kinase inhibitors. 2016; <https://patents.google.com/patent/US9290478B2>.
- (3) Sayle, R. A.; Mayfield, J. W.; Lagerstedt, I.; Pirie, R. Nextmove Software Pistachio. 2022; <http://www.nextmovesoftware.com/pistachio.html>.
- (4) CAS Content Collection. CAS Content Collection through CAS Custom Services, <https://www.cas.org/cas-data>.
- (5) Coley, C. connorcoley/ASKCOS: First public release of ASKCOS. 2019; <https://zenodo.org/record/3261361>.
- (6) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research* **2018**, *51*, 1281–1289.
- (7) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science* **2020**, *11*, 3316–3325.
- (8) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology* **2022**, *3*, 015022.
- (9) Lin, K.; Xu, Y.; Pei, J.; Lai, L. Automatic retrosynthetic route planning using template-free models. *Chemical Science* **2020**, *11*, 3355–3364.

- (10) Shi, C.; Xu, M.; Guo, H.; Zhang, M.; Tang, J. A Graph to Graphs Framework for Retrosynthesis Prediction. 2020; <https://arxiv.org/abs/2003.12725>.
- (11) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.
- (12) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics* **2020**, *12*.
- (13) Klucznik, T. et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, *4*, 522–532.
- (14) Wang, X.; Qian, Y.; Gao, H.; Coley, C. W.; Mo, Y.; Barzilay, R.; Jensen, K. F. Towards efficient discovery of green synthetic pathways with Monte Carlo tree search and reinforcement learning. *Chemical Science* **2020**, *11*, 10959–10972.
- (15) Wang, P.-S.; Gong, L.-Z. Palladium-Catalyzed Asymmetric Allylic C–H Functionalization: Mechanism, Stereo- and Regioselectivities, and Synthetic Applications. *Accounts of Chemical Research* **2020**, *53*, 2841–2854.
- (16) Beak, P.; Basu, A.; Gallagher, D. J.; Park, Y. S.; Thayumanavan, S. Regioselective, Diastereoselective, and Enantioselective Lithiation-Substitution Sequences: Reaction Pathways and Synthetic Applications. *Accounts of Chemical Research* **1996**, *29*, 552–560.
- (17) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chemical Science* **2021**, *12*, 2198–2208.

- (18) Oslob, J. D.; Åkermark, B.; Helquist, P.; Norrby, P.-O. Steric Influences on the Selectivity in Palladium-Catalyzed Allylation. *Organometallics* **1997**, *16*, 3015–3021.
- (19) Banerjee, S.; Sreenithya, A.; Sunoj, R. B. Machine learning for predicting product distributions in catalytic regioselective reactions. *Physical Chemistry Chemical Physics* **2018**, *20*, 18311–18318.
- (20) Li, X.; Zhang, S.-Q.; Xu, L.-C.; Hong, X. Predicting Regioselectivity in Radical C–H Functionalization of Heterocycles through Machine Learning. *Angewandte Chemie International Edition* **2020**, *59*, 13253–13259.
- (21) Ree, N.; Göller, A. H.; Jensen, J. H. RegioML: predicting the regioselectivity of electrophilic aromatic substitution reactions using machine learning. *Digital Discovery* **2022**, *1*, 108–114.
- (22) Marenich, A. V.; Jerome, S. V.; Cramer, C. J.; Truhlar, D. G. Charge Model 5: An Extension of Hirshfeld Population Analysis for the Accurate Description of Molecular Interactions in Gaseous and Condensed Phases. *Journal of Chemical Theory and Computation* **2012**, *8*, 527–541.
- (23) Tomberg, A.; Johansson, M. J.; Norrby, P.-O. A Predictive Tool for Electrophilic Aromatic Substitutions Using Machine Learning. *The Journal of Organic Chemistry* **2018**, *84*, 4695–4703.
- (24) Guan, Y.; Lee, T.; Wang, K.; Yu, S.; McWilliams, J. C. SNAr Regioselectivity Predictions: Machine Learning Triggering DFT Reaction Modeling through Statistical Threshold. *Journal of Chemical Information and Modeling* **2023**, *63*, 3751–3760.
- (25) Buskes, M. J.; Blanco, M.-J. Impact of Cross-Coupling Reactions in Drug Discovery and Development. *Molecules* **2020**, *25*, 3493.

- (26) Busch, M.; Wodrich, M. D.; Corminboeuf, C. A Generalized Picture of C–C Cross-Coupling. *ACS Catalysis* **2017**, *7*, 5643–5653.
- (27) Kurti, L.; Czako, B. *Strategic Applications of Named Reactions in Organic Synthesis*; Elsevier Academic Press, 2005.
- (28) Houpis, I. N.; Liu, R.; Wu, Y.; Yuan, Y.; Wang, Y.; Nettekoven, U. Regioselective Cross-Coupling Reactions of Boronic Acids with Dihalo Heterocycles. *The Journal of Organic Chemistry* **2010**, *75*, 6965–6968.
- (29) Norman, J. P.; Larson, N. G.; Entz, E. D.; Neufeldt, S. R. Unconventional Site Selectivity in Palladium-Catalyzed Cross-Couplings of Dichloroheteroarenes under Ligand-Controlled and Ligand-Free Systems. *The Journal of Organic Chemistry* **2022**, *87*, 7414–7421.
- (30) Cai, C.; Chung, J. Y. L.; McWilliams, J. C.; Sun, Y.; Shultz, C. S.; Palucki, M. From High-Throughput Catalyst Screening to Reaction Optimization: Detailed Investigation of Regioselective Suzuki Coupling of 1, 6-Naphthyridone Dichloride. *Organic Process Research & Development* **2007**, *11*, 328–335.
- (31) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.
- (32) Daylight Theory Manual, Chapter 4: SMARTS - A Language for Describing Molecular Patterns. <https://daylight.com/dayhtml/doc/theory/theory.smarts.html> [Accessed: 2023-10-01].
- (33) Landrum, G. et al. rdkit/rdkit: 2022_03_1 (Q1 2022) Release. 2022; <https://doi.org/10.5281/zenodo.6388425>.
- (34) Regioselectivity code on Github. <https://github.com/Chemlex-AI/regioselectivity>.

- (35) Yanfei Guan, O. W., Duminda Ranasinghe QM descriptors calculation. https://github.com/yanfeiguan/QM_descriptors_calculation, 2020.
- (36) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *Journal of Cheminformatics* **2014**, *6*.
- (37) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements ($Z = 1\text{--}86$). *Journal of Chemical Theory and Computation* **2017**, *13*, 1989–2009.
- (38) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019**, *15*, 1652–1671.
- (39) Pracht, P.; Caldeweyher, E.; Ehlert, S.; Grimme, S. A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for large Molecules. **2019**,
- (40) Frisch, M. J. et al. Gaussian~16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.
- (41) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. **2017**,
- (42) Withnall, M.; Lindelöf, E.; Engkvist, O.; Chen, H. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. *Journal of Cheminformatics* **2020**, *12*.
- (43) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **2019**, *59*, 3370–3388.

- (44) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling* **2015**, *55*, 2562–2574.
- (45) Beletskaya, I. P.; Alonso, F.; Tyurin, V. The Suzuki-Miyaura reaction after the Nobel prize. *Coordination Chemistry Reviews* **2019**, *385*, 137–173.
- (46) Farhang, M.; Akbarzadeh, A. R.; Rabbani, M.; Ghadiri, A. M. A retrospective-prospective review of Suzuki–Miyaura reaction: From cross-coupling reaction to pharmaceutical industry applications. *Polyhedron* **2022**, *227*, 116124.
- (47) Xu, J.; Kalyani, D.; Struble, T.; Dreher, S.; Krska, S.; Buchwald, S. L.; Jensen, K. F. Roadmap to Pharmaceutically Relevant Reactivity Models Leveraging High-Throughput Experimentation. **2022**,
- (48) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.
- (49) XU, X.; CHEN, W.; SUN, Y. Over-sampling algorithm for imbalanced data classification. *JSEE* **2019**, *30*, 1182–1191.
- (50) Seiffert, C.; Khoshgoftaar, T. M.; Hulse, J. V.; Napolitano, A. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* **2010**, *40*, 185–197.
- (51) Khan, S. H.; Hayat, M.; Bennamoun, M.; Sohel, F. A.; Togneri, R. Cost-Sensitive Learning of Deep Feature Representations From Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems* **2018**, *29*, 3573–3587.
- (52) Eyke, N. S.; Koscher, B. A.; Jensen, K. F. Toward Machine Learning-Enhanced High-Throughput Experimentation. *Trends in Chemistry* **2021**, *3*, 120–132.

- (53) Nippa, D. F. et al. Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning. *Nature Chemistry* **2023**,
- (54) Biyani, S. A.; Moriuchi, Y. W.; Thompson, D. H. Advancement in Organic Synthesis Through High Throughput Experimentation. *Chemistry—Methods* **2021**, *1*, 323–339.
- (55) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* **1951**, *22*, 79–86.
- (56) Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* **1991**, *37*, 145–151.
- (57) Bhattacharyya, A. On a Measure of Divergence between Two Multinomial Populations. *Sankhyā: The Indian Journal of Statistics* **1946**, *7*, 401–406.
- (58) Chen, Z.; Gu, C.; Yuen, O. Y.; So, C. M. Palladium-catalyzed chemoselective direct α -arylation of carbonyl compounds with chloroaryl triflates at the C–Cl site. *Chemical Science* **2022**, *13*, 4762–4769.
- (59) Denmark, S. E. *Organic Reactions*; Wiley, 2019.
- (60) Almond-Thynne, J.; Blakemore, D. C.; Pryde, D. C.; Spivey, A. C. Site-selective Suzuki–Miyaura coupling of heteroaryl halides – understanding the trends for pharmaceutically important classes. *Chemical Science* **2017**, *8*, 40–62.
- (61) Regioselectivity Analyzer web tool. <https://ai.tools.chemlex.com/region-choose>.
- (62) Schann, S.; Mayer, S.; Manteau, B. Substituted tricyclic 1,4-benzodiazepinone derivatives as allosteric modulators of group ii metabotropic glutamate receptors. 2021; <https://patents.google.com/patent/US20180346468A1>.
- (63) Loozen, H.; Stock, H.; Timmers, C. Ring-annulated dihydropyrrolo[2,l-a]isoquinolines. 2015; <https://patents.google.com/patent/EP2459560B1>.

- (64) Seshadri, B.; Darne, C.; Rahaman, H.; Warrier, J.; Kamble, M.; Liu, P.; Mannoori, R.; Borzilleri, R.; Velaparthi, U. Tgf beta receptor antagonists. 2019; <https://patents.google.com/patent/US20190337942A1>.
- (65) Tang, C.; Yin, J.; Yi, K.; Ren, Q.; Lin, X.; Zhang, Y. Inhibitors of influenza virus replication, application methods and uses thereof. 2018; <https://patents.google.com/patent/US20180346463A1>.
- (66) Wei, B.; Ortwine, D.; Crawford, J.; Young, W. Heteroaryl pyridone and aza-pyrodine compounds. 2019; <https://patents.google.com/patent/US20190194203A1>.
- (67) Choi, D. S.; Cho, Y. H. Preparation of aromatic amine compounds for organic light emitting device comprising organic solar cell, electronic paper, an organic photoconductor or an organic transistor. 2015; https://scifinder-n.cas.org/patent-viewer?docuri=gWep6ap-cBSz8swSpe1GUAI18baFQvVxG5F0ETKig5I&markedFullTextKey=Ea8sHV9JXE8n-p1FFGaRhEr9f21HT06emPlpj8oMv7A.pdf&fullTextKey=0keY4Y8TltkYQTyRmPQqbAOTxGN1_Co3lXnr6ndqwtQ.pdf.
- (68) Karlsson, S.; Benson, H.; Cook, C.; Currie, G.; Dubiez, J.; Emténäs, H.; Hawkins, J.; Meadows, R.; Smith, P. D.; Varnes, J. From Milligram to Kilogram Manufacture of AZD4573: Making It Possible by Application of Enzyme-, Iridium-, and Palladium-Catalyzed Key Transformations. *Organic Process Research & Development* **2021**, *26*, 601–615.
- (69) SciFinder. <https://scifinder-n.cas.org/?referrer=scifinder.cas.org>.
- (70) Beaton, G.; Ravula, S. B.; Tucci, F.; Lee, S. J.; Shah, C. R. Pyrimidine and pyridine amine compounds and usage thereof in disease treatment. 2022; <https://worldwide.espacenet.com/search/family/084323512/publication/WO2022256075A1?q=W02022256075A1>.