Supporting Information for 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 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

Details of the DFT descriptors

The atomic charges used here are Hirshfeld partial charge as it has a small dependency on basis sets, and are able to reproduce the electrostatic potential, which are important to reactivity and molecular properties.[1,2]

The Fukui index were used to estimate the most electrophilic or nucleophilic sites of a molecule.[3,4] Charge distribution among a molecule in each atoms is changed when a netural molecule accept or losing an electron. The electrophilicity (susceptibility of an nucleophilic attack) of an atom in a molecule with N electrons is described as:

$$f_{elec} = \rho_{N+1}(r) - \rho_N(r) \tag{1}$$

, and the nucleophilicity (susceptibility of an electrophilic attack) of an atom in a molecule with N electrons is described as:

$$f_{nuclei} = \rho_N(r) - \rho_{N-1}(r) \tag{2}$$

The NMR chemical shift is the third atomic descriptors, which indicates the local atomic environment by encoding informaiton about attached or adjacent atoms.[5] The NMR shift will be presented as shielding constants and calculated using the Gauge-Independent Atomic Orbital (GIAO) method.[6]

Comparison of qmdesc and Fqmdesc

We first apply qmdesc model introduced by ml-QM-GNN[7] to predicting Hirshfeld partial charge, and Fukui index of reactants in metal-catalyzed cross-coupling reactions (MCCCRs). The predictions made by qmdesc is compared with DFT calculated results, as shown in Figure S1 a to c. The electrophilicity of an atom, labeled as f_{elec}, and the nucleophilicity of an atom, labeled as f_{nuclei}, can not be correctly predicted via qmdesc. Considering qmdesc model is trained on 137k molecules curated from PubChem and Pistachio, specific atoms or functional groups of reactant molecules in MCCCRs may barely involved in qmdesc's training set. We treat the qmdesc model as a pre-trained model and use the calculated DFT descriptors as new data to update the weights in qmdesc, resulting in a fintuned qmdesc model (Fqmdesc). The predictions made by Fqmdesc is also compared with DFT calculated results, as shown in Figure S1 d to f. Predictions made by Fqmdesc are closer to DT calculated results than those made by qmdesc.

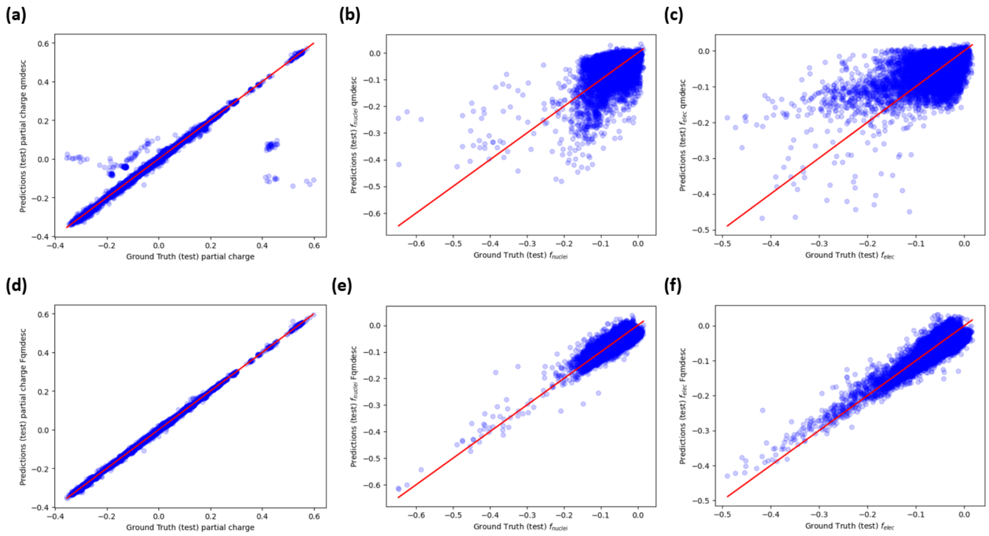

Figure S1: Prediction of target descriptors using qmdesc and Fqmdesc compared to DFT calculated results (Ground Truth) on test set. (a) Hirshfeld partial charge predicted by qmdesc versus DFT calculated. (b) f_{nuclei} predicted by qmdesc versus DFT calculated. (c) f_{elec} predicted by qmdesc versus DFT calculated. (d) Hirshfeld partial charge predicted by Fqmdesc versus DFT calculated. (e) f_{nuclei} predicted by Fqmdesc versus DFT calculated. (f) f_{elec} predicted by Fqmdesc versus DFT calculated.

Statistics of Metal Elements used as Catalyst in CAS Content Collection data

Table S1: Statistics of Metal Element used as Catalyst in CAS Content Collection data

Reaction Type	Pd	Ni	Co	Fe	Cu	Ru	Mn	Au	Rh	Pb	In
Heck	405	6	4	3	0	5	0	3	0	1	1
Hiyama	46	1	0	0	12	0	0	1	0	0	0
Kumada	77	47	21	25	3	0	4	0	0	0	0
Negishi	74	53	5	0	13	0	0	0	1	0	0

Details for the alternating reactant example

The yield of the original reaction for an industrial process is 89 %[8] as reported by Karlsson *et al.*. The hetero-substituted starting material, with chlorine and iodine groups, can be replaced by homo-substituted starting material, with two chlorine groups, without shifting the regioselectivity site. Regioselectivity for the homo-substituted starting material can be verified by reaction as shown in Figure S2 and the yield of this reaction is 64 %[9] as reported by Shah *et al.*. Considering the prices of starting material and reaction yields, the homo-substituted starting material would be a good substitution for the original starting material with different halogen groups.

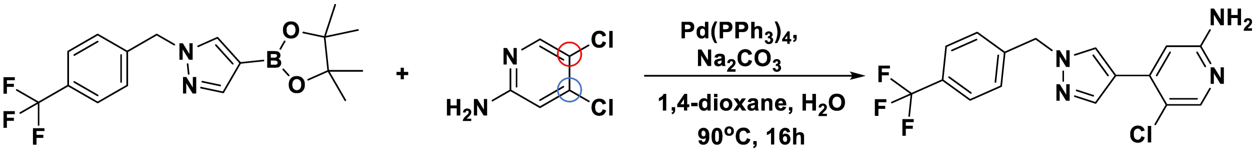

Figure S2: Regioselectivity result of the alternative reactant with the homo-substituted group in Figure 1 (e) for a Suzuki-Miyaura reaction.[9]

Input Features

Table S2: Extracted Atom and Bond Features for Molecule Graph

Atom Features	one-hot	values
Atom Symbol	yes	C, N, O, F, etc.
Atom Degree	yes	0, 1, 2, 3, 4, 5
Explicit Valence	yes	1, 2, 3, 4, 5, 6
Implicit Valence	yes	0, 1, 2, 3, 4, 5
Aromatic	no	0, 1

Bond Features	one-hot	values
Bond Type	yes	Single, Double, Triple, Aromatic
Conjugated Bond	no	0, 1
Bond in Ring	no	0, 1

Details for implementation of other models

Except for Regio-MPNN and random guess, Extended-Connectivity Fingerprints (ECFPs) based, chemical descriptors based, and sequence based multilayer perceptron (MLP) models are also used to accomplish the regioselectivity prediction task. Detailed implementation of these models is as follows.

For ECFPs based MLP model, ECFPs for reactants and product in a reaction are computed with a length of 2,048 and a radius of 2. ECFPs for reactions are calculated by the subtraction of reactants fingerprints and product fingerprints. Then, concatenation of reactants fingerprints and reaction fingerprints are fed into a fully-connected neural network to generate the predicted "score" for candidates.

For chemical descriptors based MLP model, Hirshfeld partial charges, Fukui index, and NMR chemical shifts of reacting center atoms are used as model inputs. Reacting center atoms are atoms whose neighbor atoms or bonds connected to neighbor atoms are different between reactants and products. Then, these chemical descriptors are fed into a fully connected neural network to generate the predicted "score" for candidates.

For sequence based model + MLP, the Simplified Molecular Input Line Entry System (SMILES) of a reaction is vectorized by the encoder of rxnfp generation with a length of 256.[10] Concatenation of this generated vector and chemical descriptors of reacting center atoms, as shown in the aforementioned chemical descriptors based model, is used as model inputs. Then, the input vectors are fed into a fully connected neural network to generate the predicted "score" for candidates.

Experiment on under-sampling training set

To achieve a more balanced training set on the Pistachio, we decrease the number of data points for the majority reaction type in the Pistachio training set. The total number of the remaining training data is 3,271 and reaction type distribution for this under-sampled training set is shown in Figure S3 a. The overall accuracy on the test set using MPNN + Fqmdesc is 87.98% and the percentage of incorrect prediction for each type is shown in Figure S3 b. The results indicate that applying under-sampling strategy on our unbalanced training set reduces the ability of generalization of the MPNN+Fqmdesc model.

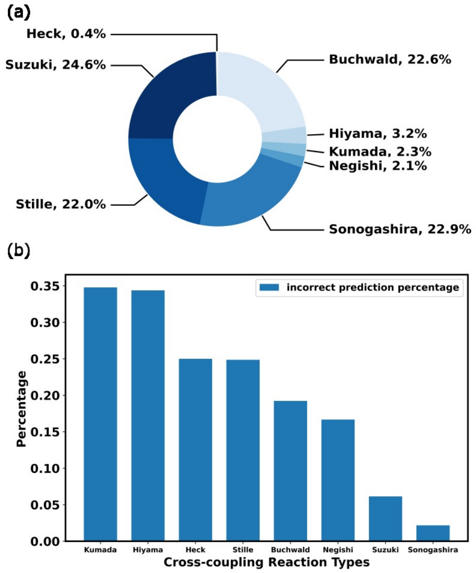

Figure S3: (a) Reaction Type Distribution for Metal-Catalyzed Cross-Coupling Reactions in under-sampled Pistachio training data. (b) The Percentage of Incorrect Predictions in Different Cross-Coupling Types in Test Set by Regio-MPNN with Fqmdesc trained by under-sampled training set.

Regioselectivity sites for different reaction types

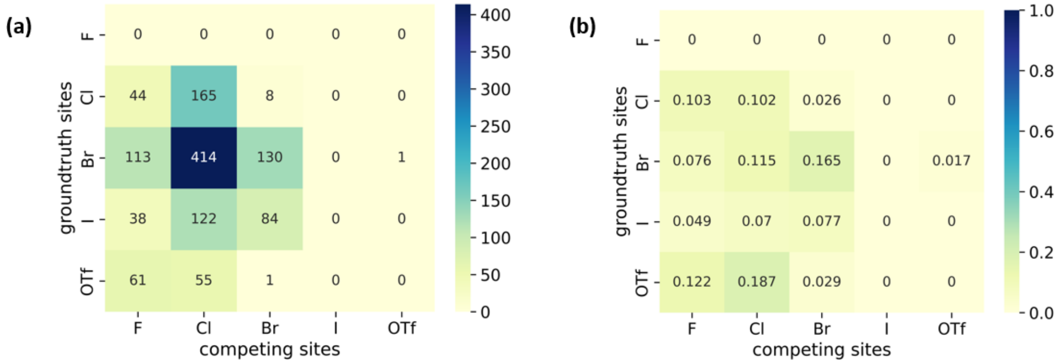

Figure S4: (a) Pistachio + CAS Content Collection Buchwald-Hartwig reaction competing sites summary and (b) The posterior probability of Buchwald-Hartwig reactions for different competing site pairs.

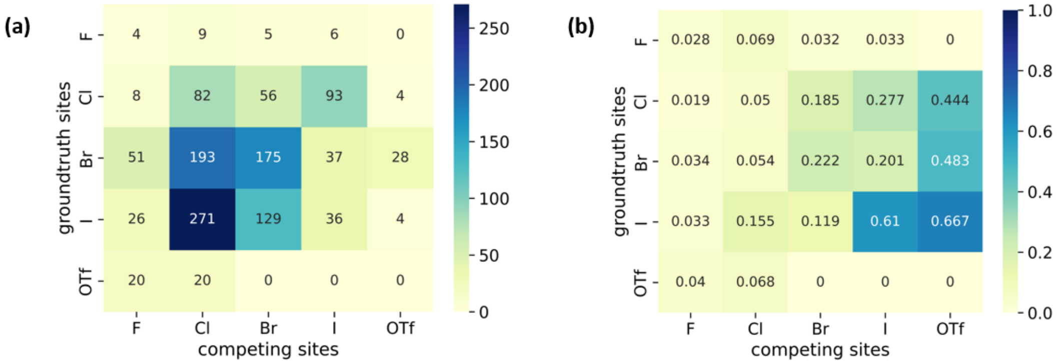

Figure S5: (a) Pistachio + CAS Content Collection Heck reaction competing sites summary and (b) The posterior probability of Heck reactions for different competing site pairs.

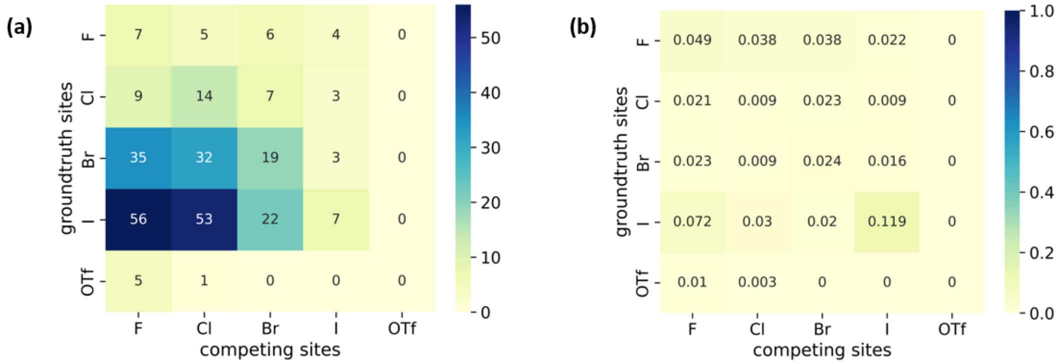

Figure S6: (a) Pistachio + CAS Content Collection Hiyama reaction competing sites summary and (b) The posterior probability of Hiyama reactions for different competing site pairs.

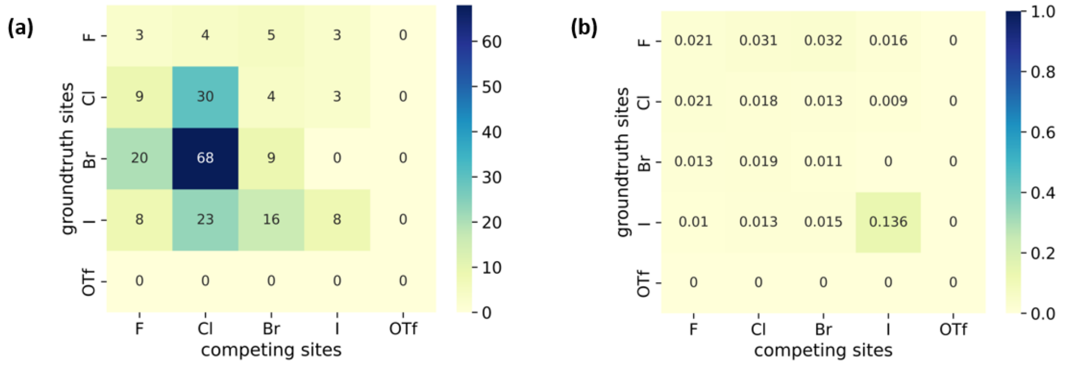

Figure S7: (a) Pistachio + CAS Content Collection Negishi reaction competing sites summary and (b) The posterior probability of Negishi reactions for different competing site pairs.

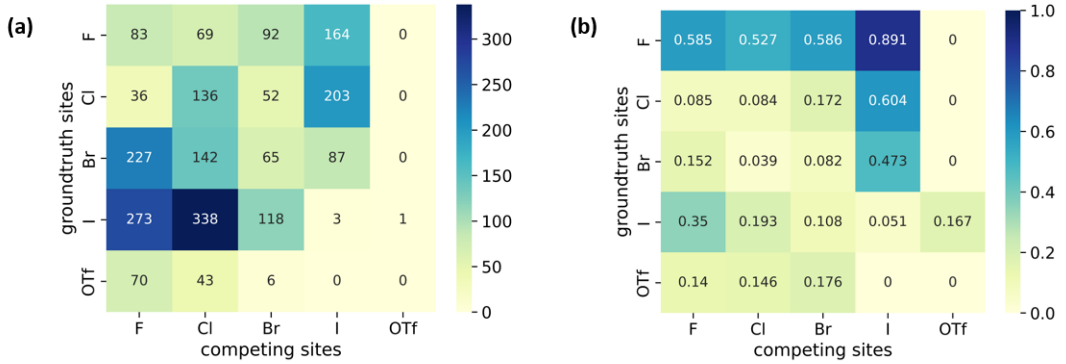

Figure S8: (a) Pistachio + CAS Content Collection Sonogashira reaction competing sites summary and (b) The posterior probability of Sonogashira reactions for different competing site pairs.

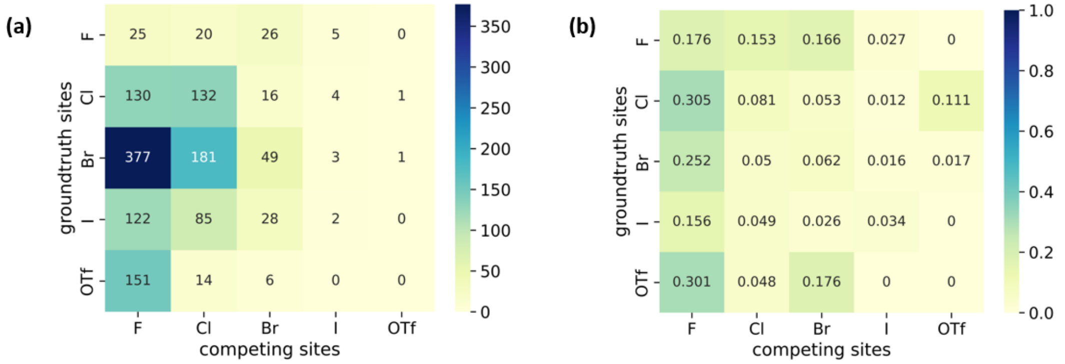

Figure S9: (a) Pistachio + CAS Content Collection Stille reaction competing sites summary and (b) The posterior probability of Stille reactions for different competing site pairs.

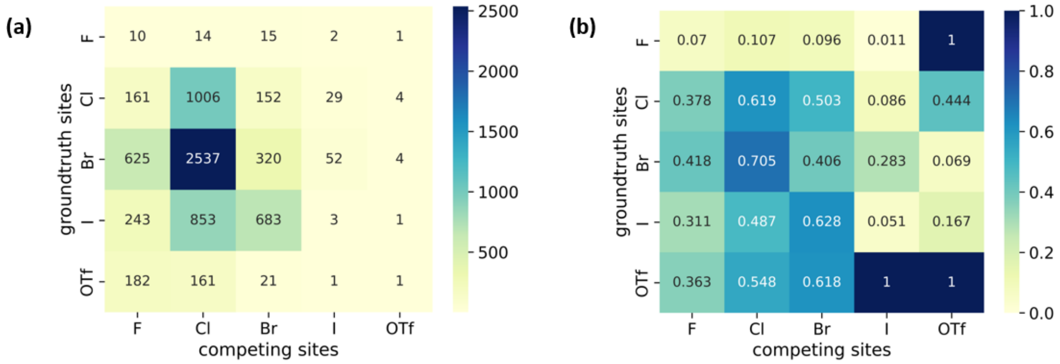

Figure S10: (a) Pistachio + CAS Content Collection Suzuki-Miyaura reaction competing sites summary and (b) The posterior probability of Suzuki-Miyaura reactions for different competing site pairs.

Human Chemist Test and Results

We randomly picked 100 reactions with regioselectivity issue from test set and invited six senior chemists to identify the main product for these reactions independently. The amount of reactions for each reaction type kept the same percentage as they are in the whole database. Namely, 10 entries are Buchwald-Hartwig reaction, 4 for Heck reaction, 2 for Hiyama reaction, 3 for Kumada reaction, 2 for Negishi reaction, 10 for Sonogashira reaction, 10 for Stille reaction, 58 for Suzuki-Miyaura reaction. Among these 8 reaction types, Kumada, Sonogashira, and Stille reactions have Fluorine bonded carbon as possible reacting site. Accuracy of predictions made by the chemists varies from 73.7 % to 93.9%, while accuracy of predictions made by model is 100 %. Reactants with regioselectivity issue for reactions on which at least one human made incorrect predictions are shown from Figure S11 to S15 for all the above mentioned reaction types. These reactants share similar substructures of hetero-aromatic rings with substitution halogen atoms. Full reactions can be found from the original patents indicated on the figures.

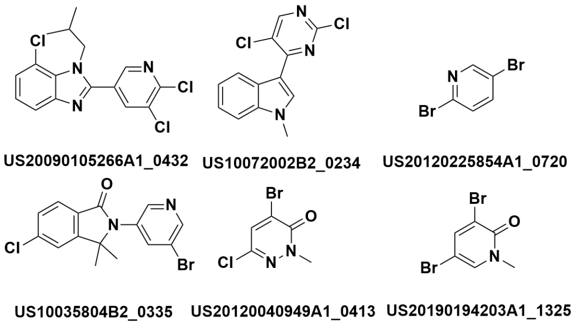

Figure S11: Reactants in regioselective reactions on which at least one human made incorrect predictions for Buchwald-Hartwig reaction type.

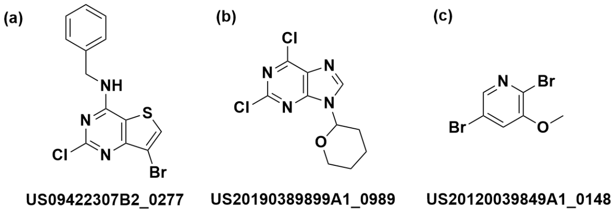

Figure S12: Reactants in regioselective reactions on which at least one human chemist made incorrect predictions. (a) Hiyama reaction; (b) Negishi reaction; (c) Kumada reaction.

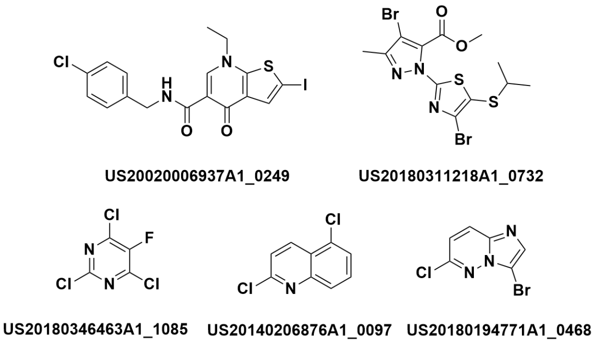

Figure S13: Reactants in regioselective reactions on which at least one human made incorrect predictions for Sonogashira reaction type.

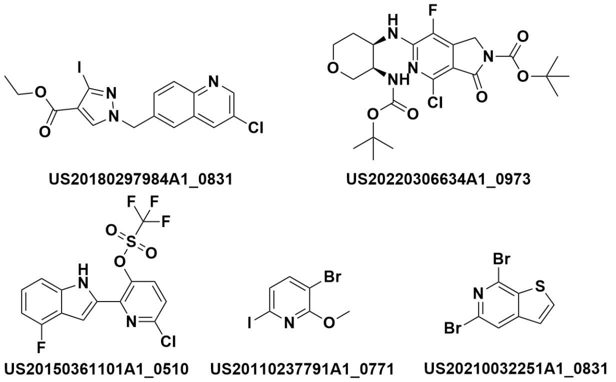

Figure S14: Reactants in regioselective reactions on which at least one human made incorrect predictions for Stille reaction type.

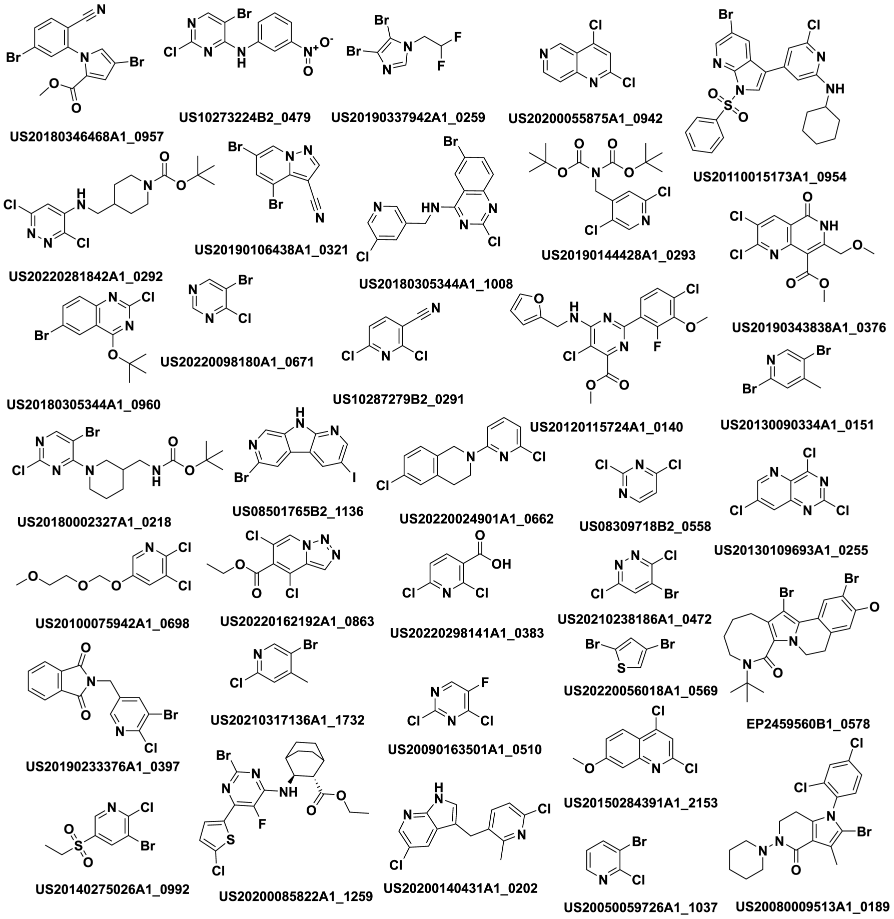

Figure S15: Reactants in regioselective reactions on which at least one human made incorrect predictions for Suzuki-Miyaura reaction type.

References

(1) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theoretica chimica acta* **1977**, *44*, 129–138.

(2) Wiberg, K. B.; Rablen, P. R. Comparison of atomic charges derived via different procedures. *Journal of Computational Chemistry* **1993**, *14*, 1504–1518.

(3) Yang, W.; Mortier, W. J. The use of global and local molecular parameters for the analysis of the gas-phase basicity of amines. *Journal of the American Chemical Society* **1986**, *108*, 5708–5711.

(4) Fuentealba, P.; Pérez, P.; Contreras, R. On the condensed Fukui function. *The Journal of Chemical Physics* **2000**, *113*, 2544–2551.

(5) Verma, R. P.; Hansch, C. Use of 13C NMR Chemical Shift as QSAR/QSPR Descriptor. *Chemical Reviews* **2011**, *111*, 2865–2899.

(6) Wolinski, K.; Hinton, J. F.; Pulay, P. Efficient implementation of the gauge-independent atomic orbital method for NMR chemical shift calculations. *Journal of the American Chemical Society* **1990**, *112*, 8251–8260.

(7) 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.

(8) Karlsson, S.; Benson, H.; Cook, C.; Currie, G.; Dubiez, J.; Emtenäs, 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.

(9) 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/patent/search/family/084323512/publication/WO2022256075A1?q=WO2022256075A1`.

(10) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence* **2021**, *3*, 144–152.