Title

Hybrid Machine Learning Approach to Predict the Site Selectivity of Iridium-Catalyzed Arene Borylation.

Journal

Journal of the American Chemical Society, 145(31)

Authors

Caldeweyher, Eike

Elkin, Masha

Gheibi, Golsa

et al.

Publication Date

2023-08-09

DOI

10.1021/jacs.3c04986

Peer reviewed

HHS Public Access

Author manuscript

J Am Chem Soc. Author manuscript; available in PMC 2025 January 10.

Published in final edited form as:

J Am Chem Soc. 2023 August 09; 145(31): 17367–17376. doi:10.1021/jacs.3c04986.

Hybrid Machine Learning Approach to Predict the Site Selectivity of Iridium-Catalyzed Arene Borylation

Eike Caldeweyher[#],
Data Science & Modelling, Pharmaceutical Sciences, R&D, AstraZeneca Gothenburg, SE-431 83 Mölndal, Sweden

Masha Elkin[#],
Department of Chemistry, University of California, Berkeley, California 94720, United States

Golsa Gheibi,
Department of Chemistry, University of California, Berkeley, California 94720, United States

Magnus Johansson,
Cardiovascular, Renal and Metabolism, Biopharmaceuticals R&D, AstraZeneca Gothenburg, SE-431 83 Mölndal, Sweden; Department of Organic Chemistry, Stockholm University, SE-106 91 Stockholm, Sweden

Christian Sköld,
Drug Design and Discovery, Department of Medicinal Chemistry, Uppsala University, SE-751 23 Uppsala, Sweden

Per-Ola Norrby,
Data Science & Modelling, Pharmaceutical Sciences, R&D, AstraZeneca Gothenburg, SE-431 83 Mölndal, Sweden

John F. Hartwig
Department of Chemistry, University of California, Berkeley, California 94720, United States

Abstract

The borylation of aryl and heteroaryl C–H bonds is valuable for the site-selective functionalization of C–H bonds in complex molecules. Iridium catalysts ligated by bipyridine ligands catalyze the borylation of the C–H bond that is most acidic and least sterically hindered in an arene, but predicting the site of borylation in molecules containing multiple arenes is difficult. To address

Corresponding Authors: **Per-Ola Norrby** – *Data Science & Modelling, Pharmaceutical Sciences, R&D, AstraZeneca Gothenburg, SE-431 83 Mölndal, Sweden;* per-ola.norrby@astrazeneca.com, **John F. Hartwig** – *Department of Chemistry, University of California, Berkeley, California 94720, United States;* jhartwig@berkeley.edu.

[#]Author Contributions

E.C. and M.E. contributed equally to this work. All authors conceptualized the project and designed experiments.

The authors declare no competing financial interest.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.3c04986.

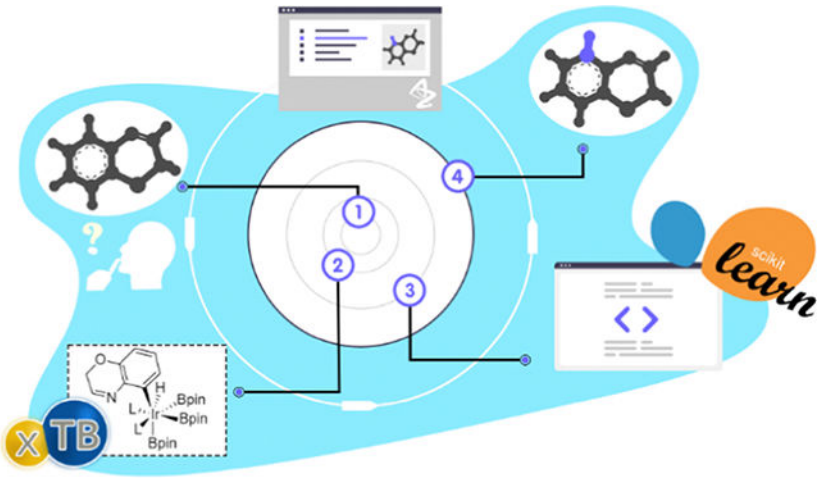
All experimental procedures, details of the computational workflow, and human survey results (PDF)

NP model statistics (XLSX)

PLS model statistics (XLSX)

this challenge, we report a hybrid computational model that predicts the Site of Borylation (SoBo) in complex molecules. The SoBo model combines density functional theory, semiempirical quantum mechanics, cheminformatics, linear regression, and machine learning to predict site selectivity and to extrapolate these predictions to new chemical space. Experimental validation of SoBo showed that the model predicts the major site of borylation of pharmaceutical intermediates with higher accuracy than prior machine-learning models or human experts, demonstrating that SoBo will be useful to guide experiments for the borylation of specific C(sp^2)–H bonds during pharmaceutical development.

Graphical Abstract



INTRODUCTION

The selective functionalization of C–H bonds in complex molecules is an emerging approach to increase the potency of lead compounds and to facilitate studies of structure-activity relationships during pharmaceutical development.[1] While reactions are being developed that occur with remarkable chemoselectivity for C–H bonds over classic functional groups, site selectivity is challenging to achieve and difficult to predict because of the ubiquity of C–H bonds and the effects of competing chemical phenomena on relative rates (Figure 1A).[2] While heuristic guidelines can help predict site selectivity, they are frequently limited to cases in which single factors dictate the reaction outcome. When multiple factors control reactivity and they oppose one another, then more sophisticated tools are necessary. However, methods to predict site selectivity have rarely been the target of modeling research, and experimental validation of the model's predictions with synthetically relevant examples is rarely reported.

One class of reaction that enables the functionalization of C–H bonds in medicinally active compounds is the undirected borylation of C–H bonds. The borylation of C–H bonds has been shown to occur on a wide range of substrates and does not require functional groups that coordinate the catalyst to direct site selectivity. The borylation of C–H bonds is especially valuable because the resulting C–B bond can be converted reliably to C–O, C–N, C–X, and C–C bonds. Given the utility of the borylation of C–H bonds, an ability to

predict the site of arene borylation in complex structures would enable the application of this reaction to discovery research.

The site selectivity of the borylation of aryl C–H bonds can be high and predicted by simple rules, but the site selectivity for the reaction of a molecule with multiple aromatic rings can be more difficult to predict. The borylation of aryl and heteroaryl C–H bonds is commonly accomplished by an iridium catalyst ligated by bidentate pyridyl ligands, such as bipyridine or phenanthroline, with stoichiometric amounts of pinacol diborane (B_2pin_2). These reactions proceed by the oxidative addition of a C–H bond, which is typically irreversible.[3] Therefore, the oxidative addition step controls site selectivity,[4] and this step occurs most rapidly at the most sterically accessible and acidic C–H bond. A series of heuristic guidelines to predict the site selectivity of borylation of various heteroarenes have been deduced from experimental studies on small heteroarenes; these guidelines indicate that borylation α to a basic nitrogen atom in a heteroarene is disfavored, that borylation of heteroarenes is faster than that of arenes, and that borylation of 5-membered heteroarenes is faster than that of 6-membered heteroarenes (Figure 1B). However, it is unclear how these competing relative rates influence site selectivity in more complex cases, including cases in which the substrate contains multiple aromatic rings, because the relative rates of borylation of multiply substituted arenes and heteroarenes are not well established, and the interplay between competing steric and electronic factors are difficult to assess. Thus, a more refined approach that builds on our current understanding of the factors influencing the site or sites of the borylation of C–H bonds is needed.[5]

Several computational approaches can be envisioned to predict site selectivity.[6] Density functional theory (DFT) has been used to rationalize experimental trends.[7,8] Several groups have developed automated tools for the generation of transition states (e.g., AARON[8,9]), but high computational costs and requisite specialized expertise continue to limit the generality and scalability of this approach. More efficient approaches to predict reaction outcomes have been developed, such as hand-coded rules,[10–14] semiempirical quantum chemical methods,[15] QSSR,[6] and related machine-learning models.[16–23] Although machine-learning methods can reveal reaction trends from experimental data, including regio-,[24] stereo-,[25] and chemoselectivity,[26] accurate predictions by these methods generally require large amounts of data over a broad scope of reactants and reaction conditions. This requirement limits the application of machine learning to synthetic chemistry because experimental data are typically available in small quantities and with varying levels of quality. In addition, accurate predictions for compounds outside the training set, such as novel chemical structures, remain an outstanding challenge.

Here, we show that combining machine learning with multiple additional computational disciplines into a hybrid model, termed SoBo (Site of Borylation), enables us to predict with high accuracy the aryl or heteroaryl C–H bond at which borylation occurs (Figure 1C). We determined site selectivities by calculating the barriers to the oxidative addition of all possible C–H bonds to a catalytically relevant iridium catalyst. To avoid computationally demanding and labor-intensive calculations of the transition-state structures for many possible reactions, we developed a streamlined, multimodal predictive system. This system combines kallisto[27] to dock the arenes at iridium, a semiempirical quantum mechanical

(SQM) method to generate approximate transition-state energies,[28,29] and two regression models to refine the predicted transition-state energies to achieve high accuracy. These regression models combine heuristics with machine learning as a function of the confidence of the model on a new substrate, thereby adapting the model to new substrates and enabling accurate predictions with small amounts of training data. High-quality experimental and computational data were used to train the models, and the precision of SoBo was demonstrated by predicting with high accuracy the site of borylation of complex molecules containing multiple aromatic units. This model outperformed predictions made by either expert synthetic chemists or previously reported machine-learning models, and these predictions are obtained within several minutes on a standard desktop computer using a command line script easily accessible to synthetic chemists (command line interface available at https://pypi.org/project/sobo).

RESULTS AND DISCUSSION

Data Collection.

High-quality training data were collected to develop a predictive model for arene borylation. From the rich body of published work on this reaction,[4,30] 86 examples of arene borylations catalyzed by iridium ligated by bipyridine or phenanthroline ligands were selected.[31] Among these examples, few gave products from borylation at more than one C–H bond. However, such examples are necessary to benchmark model performance because they allow for the direct calculation of relative transition-state energies leading to isomeric products. Thus, we conducted reactions of 15 additional substrates that undergo borylation at two positions to augment the literature examples and experimentally determined the ratio of products formed (Figure 2A). The subsequent combined dataset of 101 examples comprised the training and testing set for model development.

This dataset primarily consists of simple arenes, but the envisioned application of a predictive model is the borylation of substrates containing multiple aromatic subunits. To test the feasibility of this type of extrapolation—training a model on isolated arenes to predict the reactivity of substrates containing multiple substituted arene—we assessed the site selectivity of borylation of a series of arenes in separate reactants and within one reactant. These experiments were designed to test if the site selectivity of borylation in one arene is affected by the presence of another arene in the same molecule. To this end, we compare the site selectivity of arenes as distinct units and of arenes tethered as one molecule.

Figure 2B demonstrates that the relative reactivity of two arenes in the same reaction vessel mimics the relative reactivity of one substrate that contains both arenes. For example, the borylation of *N*-methyl pyrrole occurs to a greater extent than that of toluene in both intermolecular (84:16) and intramolecular (85:15) competition experiments. Within each arene, the site selectivity is conserved, both for the borylation of *N*-methyl pyrrole (C2:C3 selectivity; 88:12 intermolecular; 89:11 intramolecular) and toluene (C3:C4 selectivity; 63:17 intermolecular; 67:13 intramolecular). Good agreement between intramolecular and intermolecular reactivity was observed for several other competition experiments (see Supporting Information Sections C.2 and C.3 for details), demonstrating that the rate of

functionalization of one C–H bond is independent of the rate of functionalization of another. As a result, a model trained on the reactions of isolated arenes could predict the site selectivity for reactions of a substrate containing multiple arenes.

Hybrid Computational Workflow.

The workflow for the computational model was developed by combining several distinct predictive methodologies to leverage the capabilities and compensate for the deficiencies of each approach. We termed this model SoBo for Site of Borylation, and the workflow by which it predicts the site of borylation is shown in Figure 3. In Step 1, a user provides a substrate of interest in the form of a Simplified Molecular Input Line Entry System (SMILES)[32] string. Three-dimensional coordinates were constructed from this one-dimensional representation using RDKit.[33] In Step 2, the transition state for the oxidative addition of the C–H bond in benzene to the iridium catalyst ligated by tert-butyl bipyridine was calculated by DFT (B3LYP[34,35] D3(BJ)[36]/LACVP**/PB(THF)). The benzene in this structure was replaced by a heteroarene or substituted arene of interest using kallisto,[27] and the structure of the transition state for the addition of the C–H bond was optimized using a constrained SQM method (GFN2-xTB). Exchanging arenes for benzene in DFT-optimized structures provides an approach to generate the initial structures for calculations of the transition state containing a substituted arene, enabling relatively accurate transition-state structures to be calculated in orders of magnitude less time by SQM than by DFT (ca. minutes vs. hours), and obviating the need for user intervention when identifying the transition state. In this way, kallisto leverages the accuracy of DFT with the efficiency of SQM. We repeated this calculation for each aryl C–H bond in the molecule of interest. The resulting relative transition-state barriers alone did not accurately predict the product distribution. Thus, we layered additional computations involving machine learning and linear regression to calculate finer differences between the barriers for the addition of various C–H bonds.

To improve the accuracy of the energies predicted by SQM, we applied two regression models (Step 3). In Step 3a, extended connectivity molecular fingerprints (ECFP)[37] were constructed from the various aryl C–H bonds of the substrate. This representation was used to train a series of machine-learning models, and the 10-fold, cross-validated predictive performance was analyzed by the mean absolute error (MAE) and root-mean-square error (RMSE). We defined two dummy regressors as computational baselines, one that predicts the mean (mean regressor) and one that predicts the median (median regressor). A series of machine-learning architectures, such as random forest, Bayesian ridge, k-nearest neighbors, kernel ridge regression, Gaussian processes, and partial least squares (PLS) regression, with different kinds of feature preprocessing, were implemented using scikit learn.[38]

The most accurate model was a PLS regression model (n components = 13) with a polynomial combination feature preprocessing (degree = 2), with MAE = 3.1 kJ mol^{-1} and RMSE = 4.7 kJ mol^{-1}. Both the mean regressor (MAE = 6.4 kJ mol^{-1}, RMSE = 6.5 kJ mol^{-1}) and median regressor (MAE = 6.5 kJ mol^{-1}, RMSE = 9.0 kJ mol^{-1}) predicted the relative energy barriers of the oxidative addition of different C–H bonds in a molecule, derived from experimental ratios of borylation, with approximately half the accuracy

of the regressor trained using chemically meaningful data. While this PLS regressor is more accurate than the baseline models, it alone does not provide sufficient accuracy for synthetic purposes. In addition, fingerprint-based models are unable to extrapolate beyond the scope of chemical space represented by the encoded fingerprints. Thus, we combined PLS regression with additional computational approaches to create a more accurate and robust predictor.

To augment the energies from SQM calculations and predictions from the PLS model, we introduced a Neighbor Penalty (NP, Step 3b) to capture the deactivating effect of large substituents *ortho* to C–H bonds. For each C–H bond, Sterimol descriptors were calculated for all *ortho* substituents (L, B_{min}, and B_{max}).[39] These descriptors were fit to hybrid DFT energies in a multivariate regression model (coefficient of determination, R^2 = 0.76, see Supporting Information Section A.2 for details). This approach represents a quantification of the experimental trend that borylation frequently occurs at the most sterically accessible C–H bond. The resulting intuitive regression model, termed Neighbor Penalty (NP), complements the less-readily interpretable, fingerprint-based PLS model.

To combine the two correlation models (Step 3c), we calculated the binary Rogers–Tanimoto similarity[40] scores for the C–H bond of interest against all C–H bonds in the training set used to construct the PLS model. The similarity score was used to construct a mixing function between the PLS and NP regressors. By this mixing function, the PLS prediction is weighted more heavily when the environment of a C–H bond is like those in the training set, but the NP prediction is prioritized when the C–H bond is in a position less like those in the training set. This dynamic mixing enables ML predictions to be used when they are most applicable, and NP predictions, which are more extrapolative, to be used when the C–H bond is in a chemical environment that lies outside the chemical space of the training set. This approach effectively quantifies the trustworthiness of the model's predictions and adapts the model architecture to each substrate of interest. This adaptation allows for the extrapolation of predictions to new chemical space that is not represented by the training data, thereby circumventing a common challenge of models based purely on ML.

Finally, in Step 4, the predictions from regression models were combined with the lowest SQM-calculated barrier to generate absolute activation barriers, which were used to calculate Boltzmann populations for the isomeric products. The SQM barrier value is necessary to allow a reproduction of absolute barriers, but the relative energy barriers and, thus, intramolecular regioselectivity are calculated from the PLS and NP models. The absolute barrier is critical for determining intermolecular selectivity (i.e., competition reactions) and for detecting cases in which the barrier to C–H activation is prohibitively high.

Across the training set, we obtained an accuracy of 97.1% for the site selectivity of arene borylation. This entire workflow uses open-source software, and the prediction of site selectivity for a new substrate using SoBo is complete within minutes on a standard desktop computer. More information on computational techniques is available in the SI, and all code and data are available on GitHub.[41]

Results of SoBo on Simple Arenes.

Figure 4 shows the SoBo predictions and experimental product ratios for sample substrates from the original data set, depicted as product ratios out of 100, with the major site of borylation highlighted by a circle. In each case, the predictions agree well with the experimental outcome. The entire list of predicted and reported sites of borylation is available on GitHub.[41] Having trained a model to predict the site of arene borylation for members of a dataset comprising mostly substrates containing a single aromatic unit, we next investigated the ability of the model to accurately predict the reactivity of more complex, polyaromatic systems that are chemically distinct from the training set. In particular, we sought to create a predictive model for the late-stage borylation of compounds relevant to pharmaceutical programs. Thus, we tested the workflow on molecules that represent the chemical space of pharmaceuticals.

Experimental Validation with Pharmaceutically Related Substrates.

We assembled an out-of-sample validation set of compounds from the AstraZeneca collection that contain at least two aromatic rings, possess a range of functional groups, and are publicly available. These molecules differ significantly from the ones used to develop the model, which consists predominantly of substrates containing a single arene. The higher level of structural complexity and substitution pattern in the validation set evaluates the ability of SoBo to extrapolate to new chemical space. To ensure rigorous validation, no modification to the model was allowed during the work with this validation set.

Figure 5 shows the results of this validation with experimental data obtained under the standard reaction conditions used to obtain the dataset on small arenes. Despite the number of potentially reactive C–H bonds in each compound, one product was observed in all but one case. In every case, SoBo correctly predicted the major site of borylation. The model quantitatively predicted the major product (entries 4 and 5) and correctly identified both the major and minor products when two products formed (entry 3). In some cases, the model predicted a minor isomer that was not observed (entries 1, 2, and 6). These data demonstrate the accuracy of the model for synthetic applications, and future work will expand the model to predict when no reaction or side reactions will be observed during arene borylation.

Comparison to Other Approaches to Predict Site Selectivity.

To understand the extent to which SoBo can augment human intuition, a series of alternative approaches to predict reaction outcomes were tested against the validation set. A reported computational model for predicting site selectivity of arene borylation is a multitask Weisfeiler–Lehman neural network (WLN),[16] which predicts site selectivity for a range of C–H functionalization reactions, including borylation and electrophilic aromatic substitutions. The WLN model was trained using reactions that primarily follow an electrophilic aromatic substitution mechanism, which results in different site selectivity than does a model based on the iridium-catalyzed borylation of C–H bonds. For this reason, the WLN model incorrectly predicted the major site of borylation for all six substrates in our validation set, resulting in an accuracy of predicting the major product of 0% for compounds **1–6**. This result highlights the importance of pairing computational models with mechanistically sound approaches to achieve high predictive power.

An alternative predictor for the site selectivity of arene borylation is human knowledge and intuition.[42] To assess the predictive power of chemists, relative to computational models, 15 chemists from AstraZeneca and UC Berkeley predicted the site at which borylation occurred in the molecules of the validation set. Each chemist classified themselves as an expert in the borylation of C–H bonds (5–6 respondents) or an experienced synthetic chemist lacking specific expertise in borylation (7–10 respondents). Each respondent was allowed to specify one or more sites of borylation, or no reaction, and could consult the literature to inform their predictions.

The ability of these chemists to predict the major site of borylation of molecules in the validation set was compared to that of the SoBo model (Figure 6A). In general, the experts in borylation predicted the major site of reaction more accurately than did general synthetic chemists, but they did so less accurately than did the SoBo model. The chemists accurately predicted the major products for some of the substrates (>80% for **5** and **6**), but they did so less accurately for other substrates (<40% for **1** and **2**). While the predictions of some chemists were more accurate than those of others, none of the chemists correctly predicted the major product for all six molecules, while SoBo predicted the major site of borylation for each. In some cases, the chemists incorrectly predicted the borylation product when the reaction occurred ortho to a functional group (**1** and **2**), highlighting the difficulty of balancing the effect of steric and electronic influences on selectivity.

We also compared the precision of the two approaches by the distribution of predicted sites of reaction. The distribution of predictions made by all chemists is shown in Figure 6B, with the major site of borylation (and SoBo's prediction) highlighted by a blue circle. In general, the chemists predicted a wider range of site selectivities than did SoBo. Across the validation set, the chemists predicted borylations to occur at 16 of the 34 aromatic C–H bonds (47%), while SoBo predicted borylations to occur at only 10 of the C–H bonds (29%). This comparison indicates that a computational approach can yield predictions of both higher accuracy and a greater level of precision than human knowledge alone. While the human chemists surveyed herein cannot speak for all possible experimentalists, their responses illustrate the difficulty of predicting the site of arene borylation in complex systems, even for chemists highly experienced with this transformation. The precision and accuracy of SoBo are high enough to guide those seeking to conduct the borylation of specific C–H bonds in complex substrates.

CONCLUSIONS

The value of late-stage functionalization of C–H bonds relies on accurate predictions of site selectivity and the degree of selectivity. We have shown that a predictive model, created by combining a series of computational tools to leverage the strengths and supplement the weaknesses of each, identifies the site of borylation of arenes and heteroarenes catalyzed by iridium complexes ligated by bipyridine ligands. DFT was used to create approximate transition-state geometries, SQM was used to optimize these structures for new substrates, and ML, in combination with cheminformatics was used to refine the predictions of site selectivities. By comparing the similarity of a new substrate to the training data, the model calculates the applicability of machine-learned predictions and supplements

these predictions with those arising from empirical-based rules derived from mechanistic understanding. This mixing results in a model (SoBo) that can adapt to new substrates and make meaningful predictions from data-limited sources. SoBo accurately predicted the major site of borylation of substrates in the training set (97.1%) and out-of-sample validation set (100%), demonstrating the strong ability of the model to extrapolate to new chemical space and to be valuable for designing experiments for late-stage functionalizations of C–H bonds in pharmaceutically relevant molecules. SoBo proved to be more accurate than a collection of expert chemists or prior machine-learning models and should complement chemical intuition during synthetic planning. Future efforts will expand predictive models to capture reactivity trends, using mechanistic information to predict catalyst poisoning and side reactivity.

The prediction of site selectivity for a new substrate using SoBo requires no specialized computational experience and is complete within minutes on a standard desktop computer; a user simply enters the substrate as a SMILES string to a submission script, and the fully automated workflow returns the predicted likelihood of borylation at each aryl C–H position, as is described in the GitHub repository.[43] The computational approach developed herein can be applied to any reaction for which the transition state of the product-determining step is well defined and, therefore, constitutes a general platform to predict the outcome of many different chemical reactions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors thank the NIH (1R35GM130387) for support of this work at UC Berkeley. E.C. is a fellow of the AstraZeneca postdoc programme. M.E. was supported in part by an NIH Kirschstein-NRSA postdoctoral fellowship (F32-GM134579). G.G. was supported in part by a Pfizer La Jolla Academic Industrial Relations (AIR) Diversity Research Fellowship. This work was supported by an NIH instrumentation grant S10OD024998 for an NMR spectrometer.

Data Availability Statement

The training set for machine-learning is shared in a separate GitHub repository.[41] Within this repository, we list all GFN2-xTB optimized catalyst structures, all training labels, the database used to calculate Rogers–Tanimoto similarities, helper scripts that enable the creation of the PLS machinelearning model, and the final trained model. The validation set is shared in a separate GitHub repository,[44] which includes structures from the AstraZeneca database that are publicly available. We queried this database for structures with at least two different substituted aromatic rings that have more than two aryl C–H bonds and made sure that no iridium-catalyzed borylation of the molecules was previously reported. In addition, we created a chemist’s survey to create a human predictor system for C–H borylation. The results of the chemist’s survey and all experimental information are found in the Supporting Information. The workflow implementation is shared in a separate GitHub repository.[43] This workflow was implemented in python using the dask library to create a directed acyclic graph workflow[45] together with the slurm[46] high-performance computing

job scheduler. Main dependencies are numpy,[47] scipy,[48] rdkit,[33] scikit learn,[38] kallisto,[27] and GFN2-xTB.[28] Command line interface is available at https://pypi.org/project/sobo.

REFERENCES

(1). Guillemard L; Kaplaneris N; Ackermann L; Johansson MJ Late-Stage C–H Functionalization Offers New Opportunities in Drug Discovery. Nat. Rev. Chem. 2021, 5, 522–545. [PubMed: 37117588]

(2). Nippa DF; Hohler R; Stepan AF; Grether U; Konrad DB; Martin RE Late-Stage Functionalization and Its Impact on Modern Drug Discovery: Medicinal Chemistry and Chemical Biology Highlights. CHIMIA 2022, 76, 258.

(3). Boller TM; Murphy JM; Hapke M; Ishiyama T; Miyaura N; Hartwig JF Mechanism of the Mild Functionalization of Arenes by Diboron Reagents Catalyzed by Iridium Complexes. Intermediacy and Chemistry of Bipyridine-Ligated Iridium Trisboryl Complexes. J. Am. Chem. Soc. 2005, 127, 14263–14278. [PubMed: 16218621]

(4). Larsen MA; Hartwig JF Iridium-Catalyzed C–H Borylation of Heteroarenes: Scope, Regioselectivity, Application to Late-Stage Functionalization, and Mechanism. J. Am. Chem. Soc. 2014, 136, 4287–4299. [PubMed: 24506058]

(5). During preparation of this manuscript, a related study was disclosed: Nippa DF.; Atz K.; Hohler R.; Müller AT.; Marx A.; Bartelmus C.; Wuitschik G.; Marzuoli I.; Jost V.; Wolfard J.; Binder M.; Stepan AF.; Konrad DB.; Grether U.; Martin RE.; Schneider, G.Enabling Late-Stage Drug Diversification by High-Throughput Experimentation with Geometric Deep Learning ChemRxiv 2022 DOI: 10.26434/chemrxiv-2022-gkxm6-v2.

(6). Zahrt AF; Athavale SV; Denmark SE Quantitative Structure–Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. Chem. Rev. 2020, 120, 1620–1689. [PubMed: 31886649]

(7). Andersson T; Broo A; Evertsson E Prediction of Drug Candidates' Sensitivity Toward Autoxidation: Computational Estimation of C–H Dissociation Energies of Carbon-Centered Radicals. J. Pharm. Sci. 2014, 103, 1949–1955. [PubMed: 24823496]

(8). Wheeler SE; Seguin TJ; Guan Y; Doney AC Noncovalent Interactions in Organocatalysis and the Prospect of Computational Catalyst Design. Acc. Chem. Res. 2016, 49, 1061–1069. [PubMed: 27110641]

(9). Guan Y; Ingman VM; Rooks BJ; Wheeler SE AARON: An Automated Reaction Optimizer for New Catalysts. J. Chem. Theory Comput. 2018, 14, 5249–5261. [PubMed: 30095903]

(10). Salatin TD; Jorgensen WL Computer-Assisted Mechanistic Evaluation of Organic Reactions. 1. Overview. J. Org. Chem. 1980, 45, 2043–2051.

(11). Gasteiger J; Hutchings MG; Christoph B; Gann L; Hiller C; Löw P; Marsili M; Saller H; Yuki K A New Treatment of Chemical Reactivity: Development of EROS, an Expert System for Reaction Prediction and Synthesis Design. In Organic Synthesis, Reactions and Mechanisms, Topics in Current Chemistry; Springer: Berlin, Heidelberg, 1987; pp 19–73 DOI: 10.1007/3-540-16904-0_14.

(12). Ugi I; Bauer J; Bley K; Dengler A; Dietz A; Fontain E; Gruber B; Herges R; Knauer M; Reitsam K; Stein N Computer-Assisted Solution of Chemical ProblemsThe Historical Development and the Present State of the Art of a New Discipline of Chemistry. Angew. Chem., Int. Ed. 1993, 32, 201–227.

(13). Satoh H; Funatsu K SOPHIA, a Knowledge Base-Guided Reaction Prediction System - Utilization of a Knowledge Base Derived from a Reaction Database. J. Chem. Inf. Comput. Sci. 1995, 35, 34–44.

(14). Socorro IM; Taylor K; Goodman JM ROBIA: A Reaction Prediction Program. Org. Lett. 2005, 7, 3541–3544. [PubMed: 16048337]

(15). Ree N; Göller AH; Jensen JH RegioSQM20: Improved Prediction of the Regioselectivity of Electrophilic Aromatic Substitutions. J. Cheminformatics 2021, 13, 10.

(16). Struble TJ; Coley CW; Jensen KF Multitask Prediction of Site Selectivity in Aromatic C–H Functionalization Reactions. React. Chem. Eng. 2020, 5, 896–902.

(17). Jorner K; Tomberg A; Bauer C; Sköld C; Norrby P-O Organic Reactivity from Mechanism to Machine Learning. Nat. Rev. Chem. 2021, 5, 240–255. [PubMed: 37117288]

(18). Ree N; Göller AH; Jensen JH RegioML: Predicting the Regioselectivity of Electrophilic Aromatic Substitution Reactions Using Machine Learning. Digital Discovery 2022, 1, 108–114.

(19). Dotson JJ; van Dijk L; Timmerman JC; Grosslight S; Walroth RC; Gosselin F; ntener K.; Mack KA.; Sigman MS. Data-Driven Multi-Objective Optimization Tactics for Catalytic Asymmetric Reactions Using Bisphosphine Ligands. J. Am. Chem. Soc. 2023, 145, 110–121. [PubMed: 36574729]

(20). Hoque A; Sunoj RB Deep Learning for Enantioselectivity Predictions in Catalytic Asymmetric β-C–H Bond Activation Reactions. Digital Discovery 2022, 1, 926–940.

(21). Boni YT; Cammarota RC; Liao K; Sigman MS; Davies HML Leveraging Regio- and Stereoselective C(Sp3)–H Functionalization of Silyl Ethers to Train a Logistic Regression Classification Model for Predicting Site-Selectivity Bias. J. Am. Chem. Soc. 2022, 144, 15549–15561. [PubMed: 35977100]

(22). Qiu J; Xie J; Su S; Gao Y; Meng H; Yang Y; Liao K Selective Functionalization of Hindered Meta-C–H Bond of o-Alkylaryl Ketones Promoted by Automation and Deep Learning. Chem 2022, 8, 3275–3287.

(23). Xu L-C; Frey J; Hou X; Zhang S-Q; Li Y-Y; Oliveira JCA; Li S-W; Ackermann L; Hong X Enantioselectivity Prediction of Pallada-Electrocatalysed C–H Activation Using Transition State Knowledge in Machine Learning. Nat. Synth. 2023, 2, 321–330.

(24). Li X; Zhang S-Q; Xu L-C; Hong X Predicting Regioselectivity in Radical C–H Functionalization of Heterocycles through Machine Learning. Angew. Chem., Int. Ed. 2020, 59, 13253–13259.

(25). Moon S; Chatterjee S; Seeberger PH; Gilmore K Predicting Glycosylation Stereoselectivity Using Machine Learning. Chem. Sci. 2021, 12, 2931–2939.

(26). Jorner K; Brinck T; Norrby P-O; Buttar D Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies. Chem. Sci. 2021, 12, 1163–1175. [PubMed: 36299676]

(27). Caldeweyher E Kallisto: A Command-Line Interface to Simplify Computational Modelling and the Generation of Atomic Features. J. Open Source Software 2021, 6, 3050.

(28). Bannwarth C; Caldeweyher E; Ehlert S; Hansen A; Pracht P; Seibert J; Spicher S; Grimme S Extended Tight-Binding Quantum Chemistry Methods. Wiley Interdiscip. Rev.: omput. Mol. Sci. 2021, 11, No. e1493.

(29). Ehlert S; Stahn M; Spicher S; Grimme S Robust and Efficient Implicit Solvation Model for Fast Semiempirical Methods. J. Chem. Theory Comput. 2021, 17, 4250–4261. [PubMed: 34185531]

(30). Wright JS; Scott PJH; Steel PG Iridium-Catalysed C–H Borylation of Heteroarenes: Balancing Steric and Electronic Regiocontrol. Angew. Chem., Int. Ed. 2021, 60, 2796–2821.

(31). Caldeweyher E ICB_db_regioselectivity: A Database of Real Compounds and Their Regioselective Products 2022 https://github.com/C-H-activation/ICB_db_regioselectivity (accessed April 26, 2023).

(32). Daylight Theory: SMILES. https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed April 26, 2023).

(33). Landrum GA RDKit: Open-Source Cheminformatics Software 2023 https://github.com/rdkit/rdkit (accessed April 26, 2023).

(34). Becke AD Density-functional Thermochemistry. III. The Role of Exact Exchange. J. Chem. Phys. 1993, 98, 5648–5652.

(35). Stephens PJ; Devlin FJ; Chabalowski CF; Frisch MJ Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. J. Phys. Chem. A 1994, 98, 11623–11627.

(36). Grimme S; Antony J; Ehrlich S; Krieg H A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. J. Chem. Phys. 2010, 132, No. 154104. [PubMed: 20423165]

(37). Rogers D; Hahn M Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50, 742–754. [PubMed: 20426451]

(38). Pedregosa F; Varoquaux G; Gramfort A; Michel V; Thirion B; Grisel O; Blondel M; Prettenhofer P; Weiss R; Dubourg V; Vanderplas J; Passos A; Cournapeau D; Brucher M; Perrot M; Duchesnay É Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.

(39). Brethomé AV; Fletcher SP; Paton RS Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. ACS Catal. 2019, 9, 2313–2323.

(40). Hassan M; Brown RD; Varma-O'Brien S; Rogers D Cheminformatics Analysis and Learning in a Data Pipelining Environment. Mol. Divers. 2006, 10, 283–299. [PubMed: 17031533]

(41). Caldeweyher E ICB-Ml-Training: Machine Learning Training Set for the Iridium-Catalyzed Borylation 2022 https://github.com/C-H-activation/ICB-ml-training (accessed April 26, 2023).

(42). Shields BJ; Stevens J; Li J; Parasram M; Damani F; Alvarado JIM; Janey JM; Adams RP; Doyle AG Bayesian Reaction Optimization as a Tool for Chemical Synthesis. Nature 2021, 590, 89–96. [PubMed: 33536653]

(43). Caldeweyher E ICB-Workflow: Iridium-Catalyzed Borylation Dask Workflow 2022 https://github.com/C-H-activation/ICB-workflow (accessed April 26, 2023).

(44). Caldeweyher E ICB-Validation 2022 https://github.com/C-H-activation/ICB-validation (accessed April 26, 2023).

(45). Dask Development Team. Dask: Library for dynamic task scheduling.. https://dask.org (accessed April 26, 2023).

(46). Yoo AB; Jette MA; Grondona M SLURM: Simple Linux Utility for Resource Management. In Job Scheduling Strategies for Parallel Processing; Feitelson D.; Rudolph L.; Schwiegelshohn U., Eds.; Springer: Berlin, Heidelberg, 2003; pp 44–60 DOI: 10.1007/10968987_3.

(47). Harris CR; Millman KJ; van der Walt SJ; Gommers R; Virtanen P; Cournapeau D; Wieser E; Taylor J; Berg S; Smith NJ; Kern R; Picus M; Hoyer S; van Kerkwijk MH; Brett M; Haldane A; del Río JF; Wiebe M; Peterson P; Gérard-Marchant P; Sheppard K; Reddy T; Weckesser W; Abbasi H; Gohlke C; Oliphant TE Array Programming with NumPy. Nature 2020, 585, 357–362. [PubMed: 32939066]

(48). Virtanen P; Gommers R; Oliphant TE; Haberland M; Reddy T; Cournapeau D; Burovski E; Peterson P; Weckesser W; Bright J; van der Walt SJ; Brett M; Wilson J; Millman KJ; Mayorov N; Nelson ARJ; Jones E; Kern R; Larson E; Carey CJ; Polat İ; Feng Y.; Moore EW.; VanderPlas J.; Laxalde D.; Perktold J.; Cimrman R.; Henriksen I.; Quintero EA.; Harris CR.; Archibald AM.; Ribeiro AH.; Pedregosa F.; van Mulbregt P.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat. Methods 2020, 17, 261–272. [PubMed: 32015543]

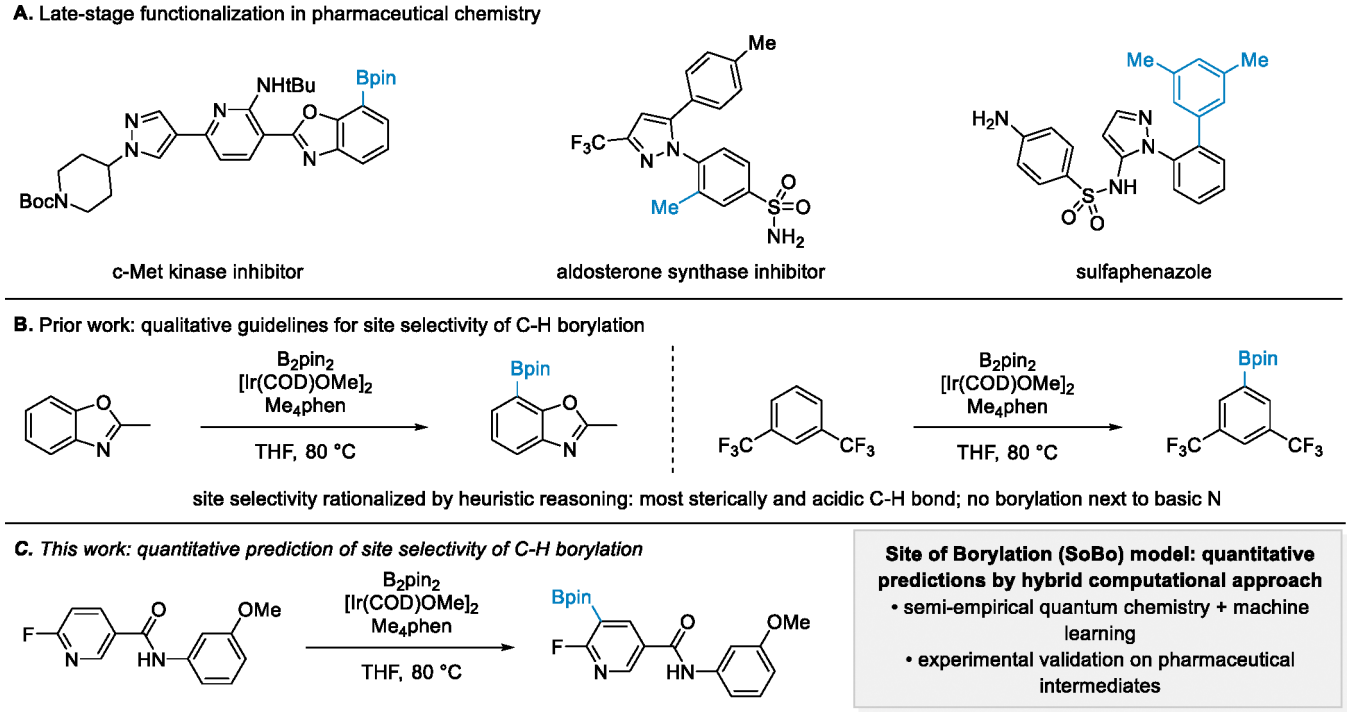


Figure 1.

(A) Sample compounds synthesized by late-stage C–H functionalization of pharmaceutical intermediates. (B) Site selectivity of the borylation of C–H bonds rationalized in simple arenes by heuristic guidelines. (C) Hybrid computational model enabling accurate prediction of the site of borylation in complex substrates containing multiple arenes and relevant to medicinal chemistry.

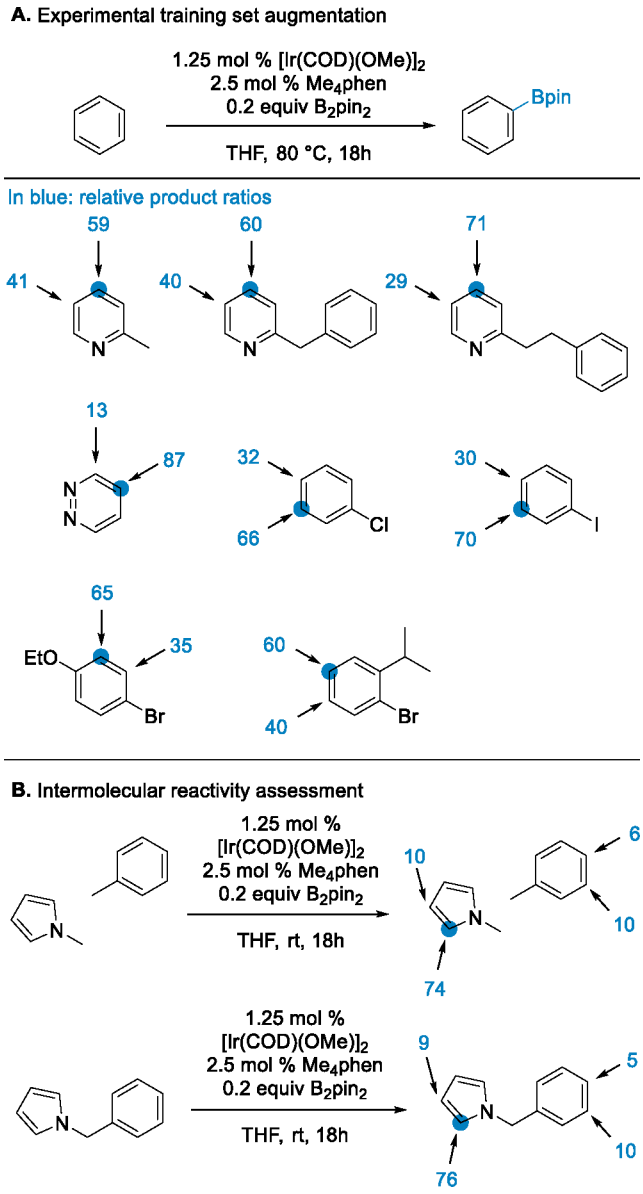


Figure 2.

(A) Experimental training set of arenes that undergo borylation at two positions augments a literature-based dataset. Product distributions are normalized to 100, and the major site of borylation is highlighted. (B) Intermolecular competition experiment mirrors the selectivity of the borylation of two arenes in the same molecule, demonstrating the feasibility of training a model on isolated arenes to predict the reactivity of a substrate containing multiple arenes.

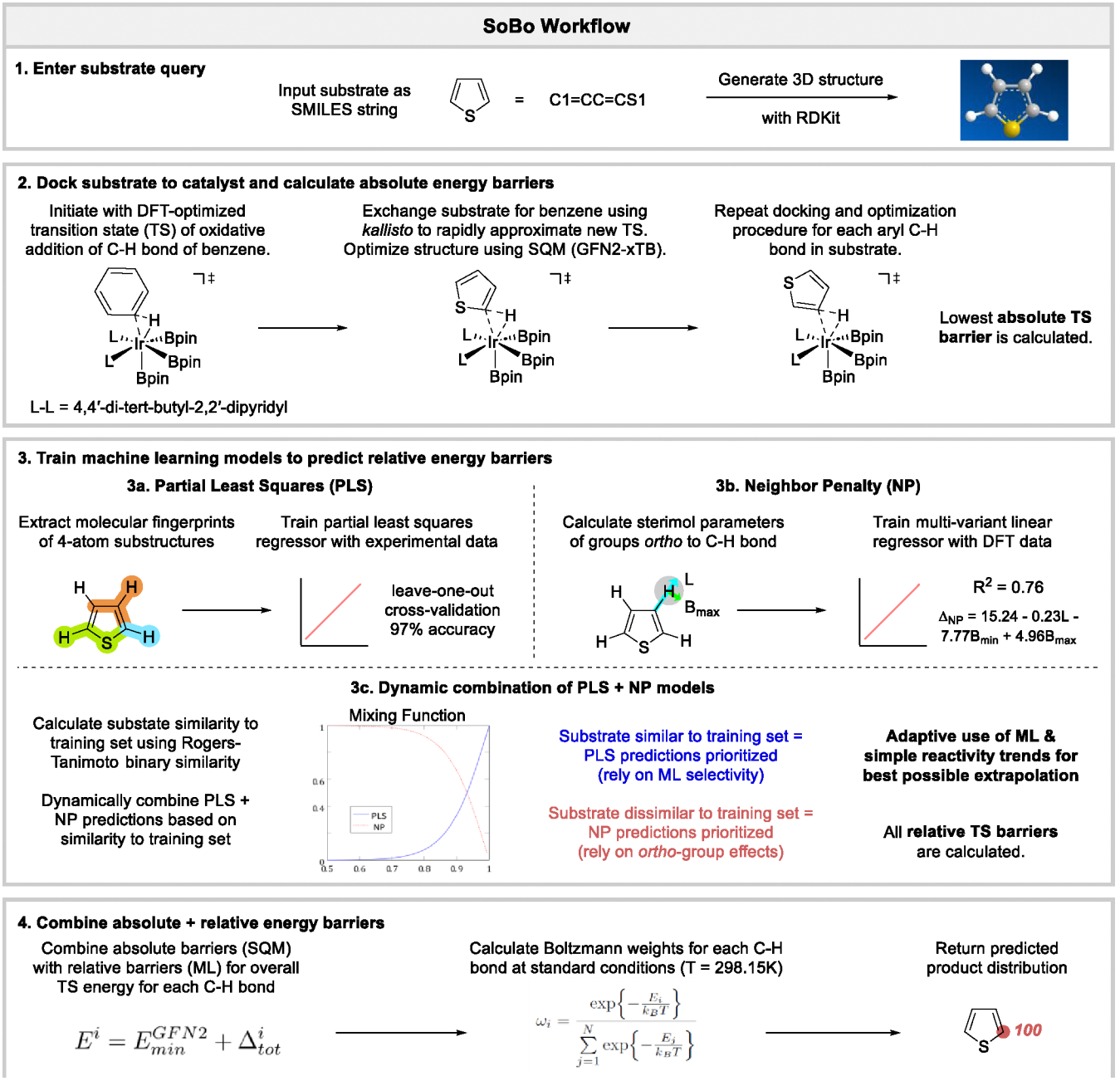


Figure 3.

Computational workflow to train SoBo to predict the site of borylation. Starting from a molecular representation, three-dimensional structures are generated (1) and activation barriers for the oxidative addition of each substrate C–H bond to the iridium catalyst are calculated (2). Next, a partial least squares regressor (3a) and sterimol-based steric approximator (3b) are trained to predict site selectivity. These regressors are combined (3c), and the absolute energy barriers for oxidative addition of all C–H bonds are adjusted (4). The workflow outputs Boltzmann weights in percentages as calculated from activation barriers at standard conditions.

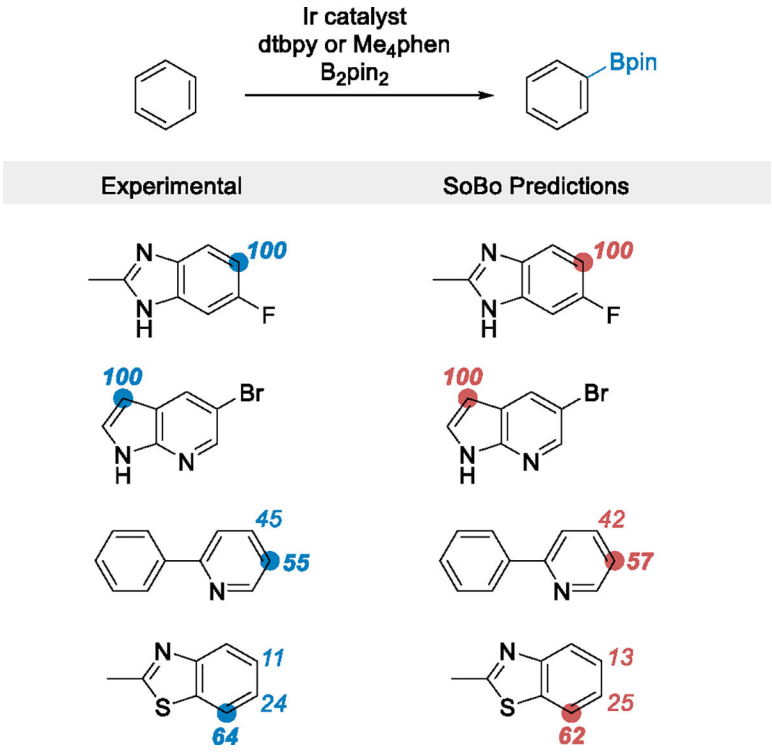


Figure 4.

Comparison of SoBo predictions and experimental results from model training. Starting materials are shown along with the relative ratio of borylation at various positions, normalized out of 100. The major site of borylation is indicated by a colored circle.

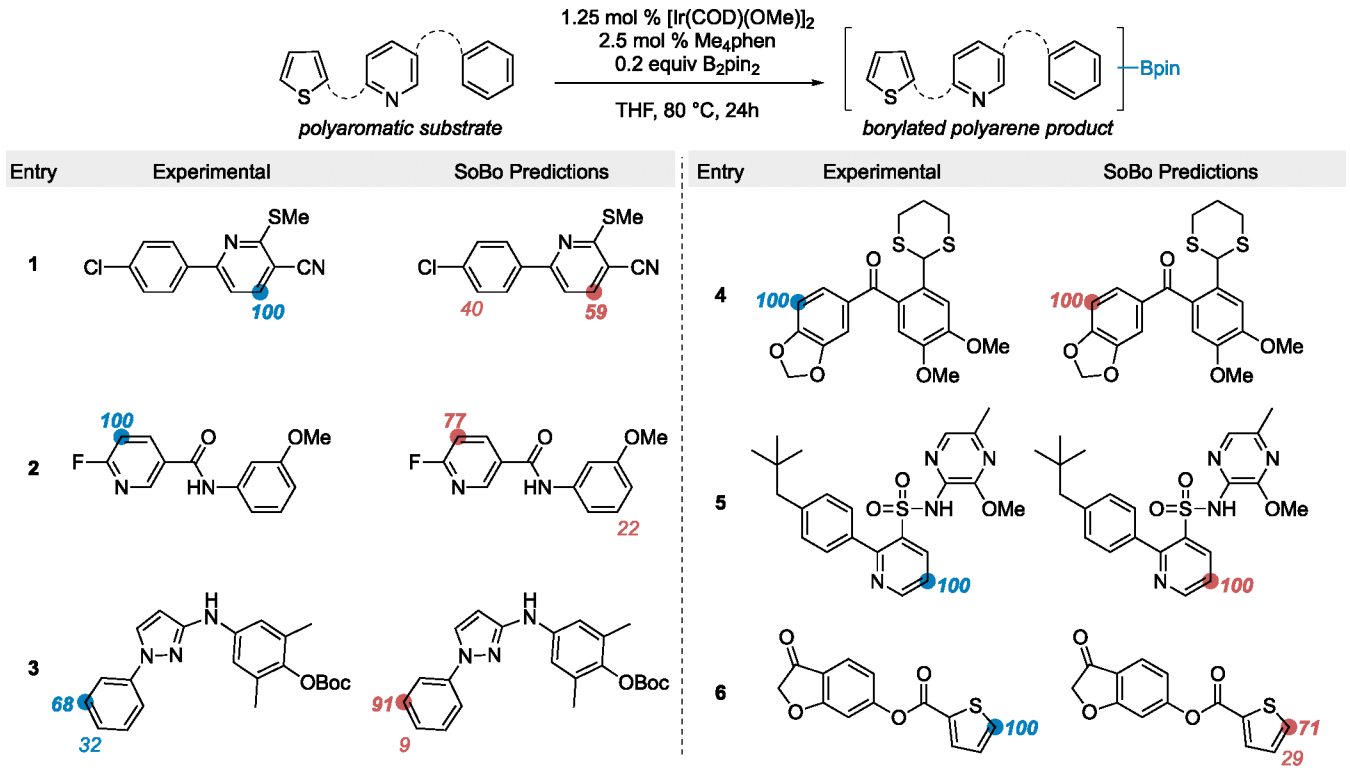


Figure 5.

Experimental validation using intermediates from medicinal chemistry programs compares SoBo predictions and experimentally determined sites of borylation. Product distributions are normalized out of 100 with the major site of borylation highlighted with a circle.

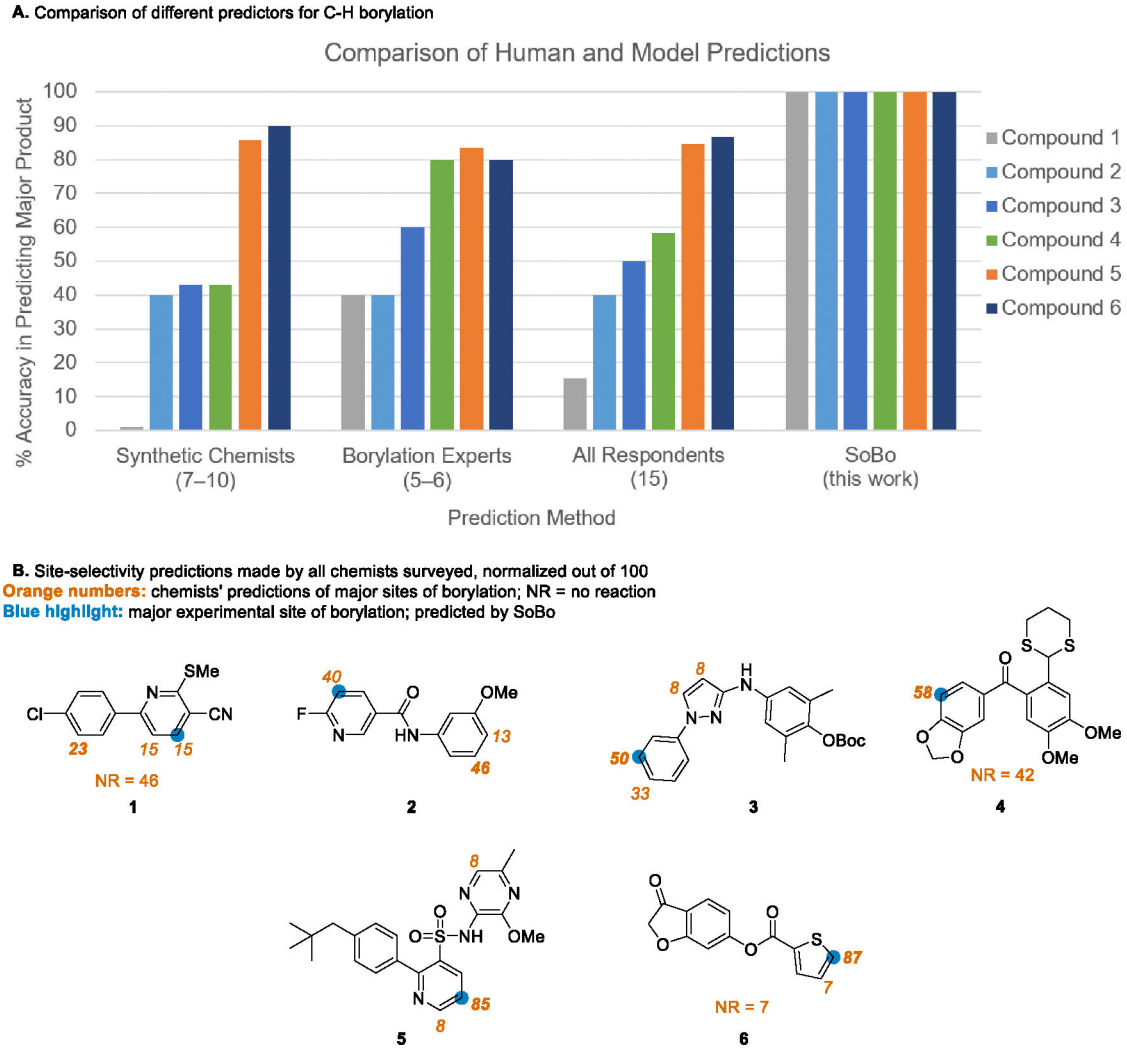


Figure 6.

(A) Predictions by human chemists possessing various levels of familiarity with C–H borylation and by SoBo for the major site of arene and heteroarene borylation across the validation set. (B) Predicted major sites of borylation made by all chemists surveyed, normalized out of 100, with experimental sites of borylation highlighted with a circle.