Using supervised learning methods to predict Buchwald-Hartwig amination yields from high throughput experimentation data

Project Proposal

Nathan M. Lui

Introduction

The world is made of chemistry; from the materials that form your everyday electronics, to the petroleum products that power the global supply chain, to the hundreds of thousands of pharmaceuticals that are prescribed by doctors every day across the world, molecules are an inextricable part of our lives. A lucky handful of these compounds (e.g., latex, crude oil, food, etc.) are created by mother nature herself; for everything else (e.g., synthetic plastics, industrial dyes, most medicines, etc.) there's a large-scale manufacturing process. The more complex the compound the more complicated the process. This aphorism is exemplified nowhere better than in the manufacturing of small-molecule active pharmaceutical ingredients (APIs).

Pharmaceuticals are typically small, but incredibly complex molecules. Designing large-scale syntheses for small molecules requires efficient and incredibly robust chemical reactions. Consider a 10-step plant-scale process for a particular pharmaceutical drug. If every step of that process causes just 5% loss (95% efficiency), over 40% of the product will be lost to production. Given the cost of chemical feedstocks and potential losses to purification and transport it becomes clear that industrial process chemistry requires nearly perfect reactions. The *academic* synthetic chemistry community has begun to understand this fact and in the design of more efficient transformations has turned to machine learning methods for the optimization of reaction conditions and even the design of reagents and new reactions.[1–4]

Synthesis of aryl- and heteroaryl-amines

There has been considerable research effort directed at the development of synthetic routes to aryl- and heteroaryl-amines (**Chart 1**) because these are important building blocks for pharmaceuticals, agrochemicals, organic dyes, and functional polymers, etc.[5] A glance at the world's best-selling pharmaceuticals from the past several years reveals the importance of this structural motif (**Chart 2**). Prior to 1995 the best way of forming this structure without highly toxic reagents was a fairly expensive reaction with limited substrate scope called the S_NAr. Today, while the S_NAr remains a significant tool for aryl C-N bond formation it has been largely replaced in the academic synthetic community by a reaction known as the Buchwald-Hartwig amination.

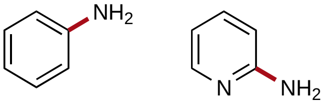


Chart 1 | Aniline (left) and 2-aminopyridine (right), the simplest aryl- and heteroaryl-amine, respectively. The carbon-nitrogen (C-N) bond formed from the Buchwald-Hartwig amination.

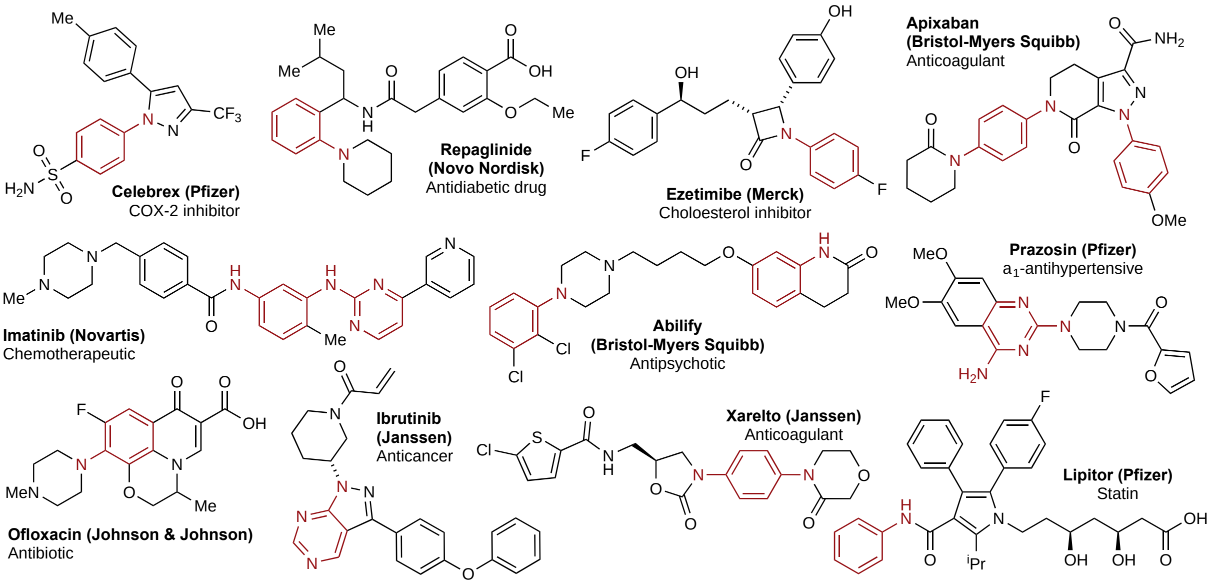
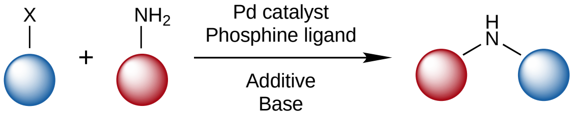


Chart 2 | Examples of top-selling active pharmaceutical ingredients (APIs) containing the arylamine and herteroaryl motif (substructure highlighted in red). A list of chemical abbreviations is included after the references.

The Buchwald-Hartwig amination

The seminal reports of what would become the Buchwald-Hartwig amination was first published by Migita and coworkers who developed the palladium (Pd)-catalyzed substitution of aryl halides using incredibly toxic amino-tin reagents.[6] In 1995, a decade later the groups of Stephen Buchwald (MIT) and John Hartwig (Berkeley) simultaneous published tin-free Pd-catalyzed couplings of aryl halides (blue) and amines (red) (**Scheme 1**).[7] In the twenty years since, what has come to be known as the Buchwald-Hartwig C-N cross-coupling (because the reaction generates the carbon-nitrogen (C-N) bond shown in red, **Chart 1**) or the Buchwald-Hartwig amination, has become one of the most widely used organometallic reactions in synthetic chemistry.[7,8] Owing to the utility of the Buchwald-Hartwig reaction a veritable library of ligands, bases, and additives have been developed to facilitate reactions between almost every possible aryl halide and amine (**Chart 3**).



Scheme 1 | General model of the palladium (Pd) catalyzed Buchwald-Hartwig C-N cross coupling reaction.[8]

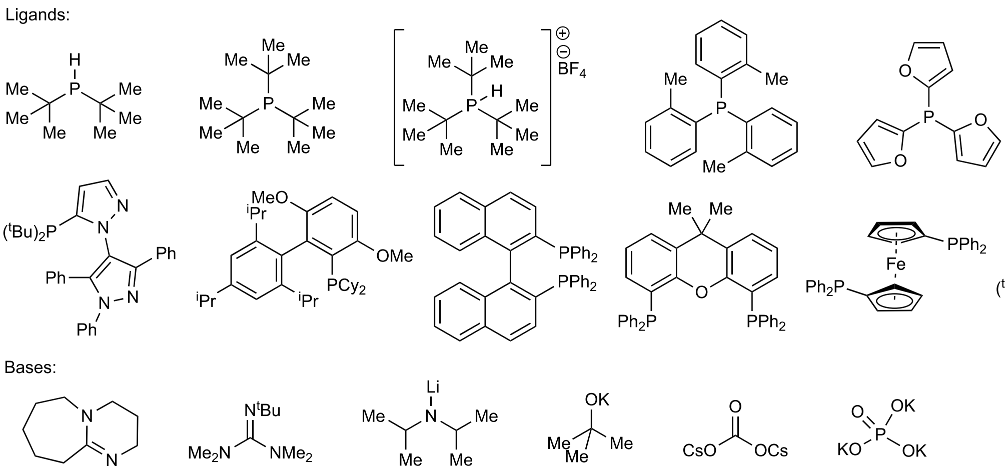


Chart 3 | Examples of commercially available phosphine ligands and bases used for Buchwald-Hartwig C-N cross coupling.[8] A list of chemical abbreviations is included after the references.

Previous work

In 2018, the group of Prof. Abigail Doyle (Princeton) set out to further optimize the Buchwald-Hartwig amination.[1] They noticed the poor performance of the cross-coupling reaction in the presence of more complex "drug-like" molecules, especially those containing 5-membered 1,2-heterocycles (see additives in **Figure 1** in the next section). In collaboration with Merck Research Labs they screened over 4000 reaction combinations generating models to determine the effects of these additives on the reaction yield. Their model and other efforts to predict reaction yields (both using Doyle's Buchwald-Hartwig dataset and other common reaction databases) is described in later sections. In this project we attempt to improve on the model determined by Doyle and coworkers on the Buchwald-Hartwig dataset, by incorporating not only the original computed features, but also high- and low-level static representations of the molecules.

Background

Choice of dataset

The original Buchwald-Hartwig amination dataset developed by Doyle and coworkers in collaboration with the high throughput screening facility at Merck Research Labs consists of over 4600 reactions (**Figure 1a**).[1] The published dataset maps five compounds (a palladium catalyst, a base, an aryl halide, an isoxazole additive, and a coupled product) to the percent yield of the Buchwald-Hartwig amination. These reactions were run simultaneously on several 1536-well plates under identical nanomolar reaction conditions. In total 4 catalysts, 3 bases, 15 aryl halides, and 23 isoxazoles were screened (**Figure 1b**).[1] Approximately 30% of the reactions failed to produce any product, however the remaining 70% are relatively uniformly distributed (**Figure 2**).[1]

Doyle's Buchwald-Hartwig dataset was chosen because it is chemically complete. The sampled reaction space is fully covered by the dataset. Every aryl halide is screened with every palladium catalyst, every additive, and every base, giving us a full view of the sampled reaction landscape. One must note that just because the sample space is fully covered does not imply that the *global reaction space* is adequately covered. Generalizability to *out-of-set* reactions (i.e., Buchwald-Hartwig aminations involving coupling partners that are *not* studied) is, of course, largely dependent on their similarity to the compounds sampled. Nevertheless, the likelihood of successful modeling of the *in-set* reaction landscape (and yield prediction) is greatly improved by the completeness of the dataset. Note that a complete sample space necessitates special generalization test set treatment; this will be further discussed in the methods section below.[2,4,9]

Other well-explored reaction datasets have been used to predict reaction yields; two of the most popular are the USPTO and its subset USPTO-50k, which contain 424,621 and 50,036 (respectively) atom mapped reactions from organic chemistry patents from the US Patent and Trademark Office.[10] While these would be more than suitable for modeling reaction yields, their vast generality (over 10 general reaction classes) makes the modeling task much more complicated. Furthermore, the USPTO datasets are mined from published patents resulting in approximately 70% accuracy in product yields and 89% accuracy in chemical identity.[10] These factors lead us to use the Buchwald-Hartwig dataset despite its smaller size and scope.

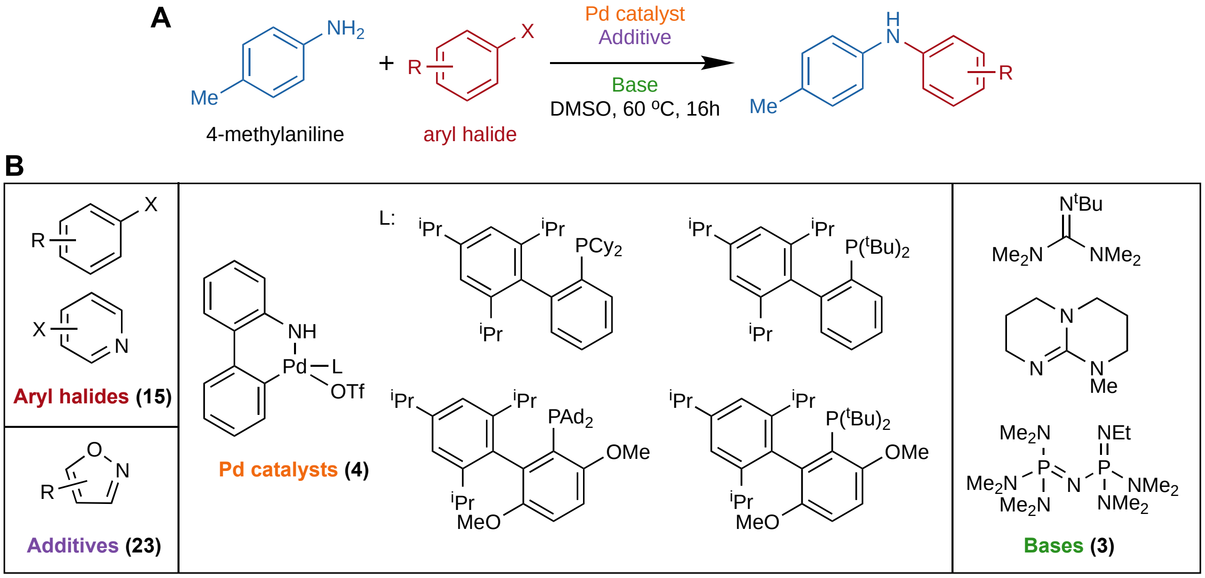


Figure 1 | (a) The Buchwald Hartwig amination screened in the dataset. (b) Four chemical variables (15 aryl halides, 4 palladium catalysts, 3 bases, and 23 isoxazole additives) were studied simultaneously using high throughput screening in 1536-well reaction plates. Each well was run under the same conditions (nanomolar-scale) and the yield of each combination of reactions was determined spectroscopically. Figure adapted from Ahneman et al.[1]

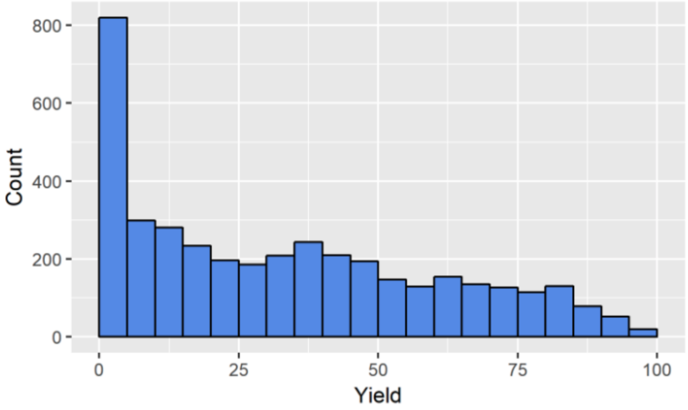


Figure 2 | The distribution of reaction yields (labels) from Doyle's Buchwald-Hartwig dataset. The control reactions have been removed. Approximately 30% of the reactions failed to produce any product, however the remaining 70% are relatively uniformly distributed. Figure reproduced from Ahneman et al.[1]

The challenge of molecular representation

One of the largest obstacles to the application of machine learning methods to chemistry is the conversion of molecules to meaningful input data. A sufficiently learned student can be taught general rules/models to characterize the differences in, for example, the three bases or four catalysts in the dataset, however the rules of undergraduate organic chemistry can hardly be used to translate complex 3-dimensional molecules into machine-readable features. The question of how one translates molecules to feature vectors began with quantitative structure–activity relationship (QSAR) and quantitative structure–property relationship (QSPR) modeling in the pharmaceutical industry where QSAR modeling has been used to screen libraries of compounds for "druglikeness," reducing the possible chemical space by orders of magnitude.[2,11] In these QSAR/QSPR studies important molecular descriptors such as a compound's molecular weight, dipole moment, solubility, or "linearity" are computed for each molecule in the dataset.[12] There have been recent efforts to apply the methods of natural language processing and graph theory to chemical structures; while these methods have shown considerable promise they are still in their infancy.[2,9,11,13] The "computable descriptor" method is the workhorse for modern chemical machine learning.

In the original paper Doyle and coworkers[1] used density function theory (DFT) methods to generate features for the 46 compounds. A total of 120 various chemical descriptor (e.g. intramolecular vibrational modes/intensities, dipole moment, electrostatic charges, etc.) were calculated for the compounds. These data are included in the original dataset, bringing the total size of the dataset to 4600 samples in $\mathbb{R}^{120}$. The authors note that, in addition to the very high computational cost, DFT functionalization is sensitive to basis set and functional selection.[2] The dataset's molecular features were generated using the B3LYP/6-31G* level of theory, a general method used largely out of tradition.[1] Recent reports have demonstrated the significant shortcomings of this method.[14,15] To avoid the biases of this particular choice without recomputing every one of these features, the 120 DFT calculated features will be supplemented with several high (e.g. polar surface area, H-bond donor/acceptor ability, etc.) and low-level (i.e. molecular fingerprints) static representations using the integrated molecular transformers from Therapeutics Data Commons (TDC)[16] and DeepChem,[17] methodology which has recently been shown promising results.[13,18]

Methods

Model selection

A plethora of machine learning models will be explored in this project. The original work of Ahneman et al.[1] tested linear regression (with various regularization), *k*-nearest neighbors regression, support vector regression, Bayes generalized linear model, feed-forward neural network regression, and random forest regression. We will examine all of these models in addition to gaussian process regression, gradient boosted regression trees, as well as XGBoost regression trees.

Special considerations for cross-validation and holdout testing of HTS data

Candidate models are trained for generalizability using a hidden test set. Typically, the dataset is split randomly into a training and test set (commonly, 70%/30% or 80%/20%). The model is then trained on the training set and then evaluated on the test set to determine how well the model generalizes to unseen data.[19] This approach is well accepted, however it was recently reported that using random split of high throughput data provides optimistically biased estimates of model performance on *out-of-sample* molecules since the dataset is combinatorially complete.[2,4] Random splitting of the dataset creates a training set in which some combinations of molecules are hidden, but *every* molecule is eventually seen, allowing the members of the test set to infiltrate the training set.[4] An analogous argument is applicable for the process of hyperparameter tuning through cross-validation. We follow the suggestions of Żurański et al.[2] using *leave-one-molecule-out* cross-validation and testing instead of random or stratified cross-validation; formally this is similar to *leave-p-out* cross-validation.[4,19]

Performance evaluation

Generalizability of the candidate models will be tested on a holdout test set following the methodology described above. To evaluate the individual model performance each model will be tested against the baseline models on a sequestered test set following the method described by Chuang and Keiser[9] and Żurański et al.[2]: a learned model using one-hot encoded features and a non-learned model derived by averaging yields across the dataset. Performance of the candidate models above the baseline indicates predictive ability on out-of-sample molecules.[1,2]

Alternative representations

In addition to sequence representations of molecules described through this proposal these inherently 3-dimensional molecules also have graph representations.[11] In recent years the application of graph neural networks and message passing neural networks to organic chemistry has developed rapidly.[11,16] If time permits and results necessitate the dataset can be converted into graph representation and used to train messaging passing and graph neural network using a preexisting architecture such as Chemprop[20] or MolNet.[21] As this analysis represents a significant change in the proposed methodology and would require significant training resources it is only proposed as an option of last resort and a potential future direction for this project.

References

1. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science (80-.).* **360**, 186–190 (2018).
2. Żurański, A. M., Martinez Alvarado, J. I., Shields, B. J. & Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **54**, 1856–1865 (2021).
3. Gao, H. *et al.* Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **4**, 1465–1476 (2018).
4. Zahrt, A. F., Henle, J. J. & Denmark, S. E. Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets. *ACS Comb. Sci.* **22**, 586–591 (2020).
5. Lawrence, S. A. *Amines: synthesis, properties and applications*. (Cambridge University Press, 2004).
6. Kosugi, M., Kameyama, M. & Migita, T. Palladium-catalyzed aromatic amination of aryl bromides with N,N-diethylamino-tributyltin. *Chem. Lett.* **12**, 927–928 (1983).
7. Dorel, R., Grugel, C. P. & Haydl, A. M. The Buchwald–Hartwig Amination After 25 Years. *Angew. Chemie Int. Ed.* **58**, 17118–17129 (2019).
8. Ruiz-Castillo, P. & Buchwald, S. L. Applications of Palladium-Catalyzed C–N Cross-Coupling Reactions. *Chem. Rev.* **116**, 12564–12649 (2016).
9. Chuang, K. V. & Keiser, M. J. Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science (80-.).* **362**, eaat8603 (2018).
10. Lowe, D. M. Extraction of chemical structures and reactions from the literature. (2012).

doi:10.17863/CAM.16293.

11. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688-702.e13 (2020).
12. Dehmer, M., Varmuza, K., Bonchev, D. & Emmert-Streib, F. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*. (Wiley, 2012).
13. Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach. Learn. Sci. Technol.* **2**, 15016 (2021).
14. Morgante, P. & Peverati, R. The devil in the details: A tutorial review on some undervalued aspects of density functional theory calculations. *Int. J. Quantum Chem.* **120**, e26332 (2020).
15. Kruse, H., Goerigk, L. & Grimme, S. Why the Standard B3LYP/6-31G* Model Chemistry Should Not Be Used in DFT Calculations of Molecular Thermochemistry: Understanding and Correcting the Problem. *J. Org. Chem.* **77**, 10824–10834 (2012).
16. Huang, K. *et al.* Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. (2021) doi:10.48550/ARXIV.2102.09548.
17. Ramsundar, B. *et al. Deep Learning for the Life Sciences*. (O'Reilly Media, 2019).
18. Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **6**, 1379–1390 (2020).
19. Hastie, T., Tibshirani, R. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Springer, 2009).
20. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
21. Kim, Y. *et al.* MolNet: A Chemically Intuitive Graph Neural Network for Prediction of Molecular Properties. (2022) doi:10.48550/ARXIV.2203.09456.

Abbreviations

Ad: adamantyl (–$C_{10}H_{15}$); API: active pharmaceutical ingredient; ^{t}Bu: *tert*-butyl (–$C(CH_3)_3$); Cy: cyclohexyl (–C_6H_{11}); DFT: density functional theory; HTS: high throughput screening; Me: methyl (–CH_3); Ph: phenyl (–C_6H_5); ^{i}Pr: iso-propyl (–$CH(CH_3)_2$); QSAR: quantitative structure–activity relationship; QSPR: quantitative structure–property relationship; TDC: Therapeutics Data Commons