Yield Prediction of Organic Reactions in Biased Datasets via Positive-Unlabeled Learning

Jan Christopher Spies[*1] **Florian Boser**[*1] **Frank Glorius**[1]
[*]*Equal contribution* [1]*Organisch-Chemisches-Institut, Universität Münster, Corrensstraße 36*

Correspondence to: Frank Glorius glorius@uni-muenster.de

Introduction

Accurate *a priori* prediction of reaction yields is a long-standing objective in organic chemistry.[1] While literature and patent databases offer vast resources for Machine Learning (ML), their utility is constrained by two pervasive biases.[2–4] First, selection bias arises from the preference for familiar conditions and reagents, limiting chemical space exploration. Second, reporting bias distorts yield distributions; failed or low-yielding reactions are systematically underreported, resulting in datasets rich in 'positive' examples but critically deficient in the 'negative' examples essential for training discriminative models (see Figure 1). We address this by leveraging Positive-Unlabeled (PU) learning to compensate for the absence of negative data.[5,6]

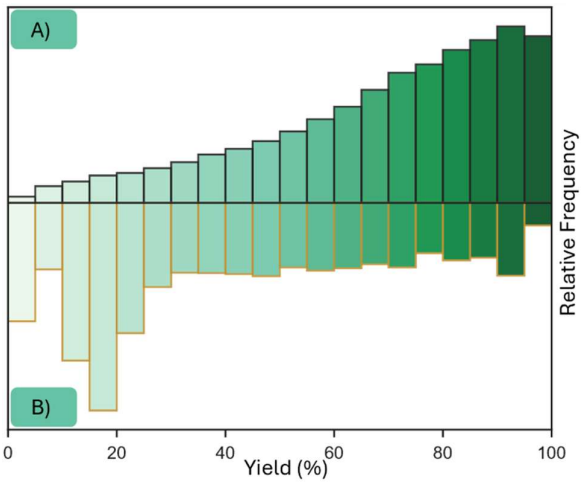

Fig. 1: Comparison of binned yield frequencies for Suzuki-Miyaura couplings. A) Literature databases such as Reaxys are severely skewed towards high-yielding examples. B) HTE Dataset by Perera et al. shows a more balanced yield distribution.[7]

Methodology

We propose PAYN (Positivity is all you need), a framework that utilizes PU learning to identify "reliable negatives" (RN) from unlabeled data to balance the labeled set. The PAYN framework employs a two-step approach:

Classification: A classifier is trained to distinguish between Positive and Unlabeled datapoints. Using this classifier, the Unlabeled datapoints are then reassigned by their probability values and a dynamically created threshold. Datapoints under the threshold are regarded as 'Reliable Negatives' (RN) and are assigned a Yield of 0%.

Regression: The combined augmented dataset (Labeled Positives + RN) is then used to train a regression model for yield prediction.

Since reaction yield is continuous, we frame the initial prediction as a binary classification problem (Positive >20% yield).

To benchmark PAYN we utilized datasets derived from High-Throughput Experimentation (HTE) datasets, which give us the ground truth. We simulated reporting bias in these HTE datasets by splitting the dataset into an Unlabeled portion (label hidden) and a labeled Positive portion (all negative examples dropped).

Results and Discussion

We validated the framework across diverse reaction types, including Ni-catalyzed borylations[8], Buchwald-Hartwig[9], and Suzuki-Miyaura couplings[7]. To contextualize the efficacy of the model trained on PAYN augmented data, two distinct benchmark models we trained in addition to our augmented regression model. First a fully labeled model was trained on the whole training data, as an upper benchmark. In addition to this, a positive only model was trained as a baseline on just the labeled positive datapoints.

As shown in Figure 2, the augmented model consistently outperforms the positives only baseline, confirming our hypothesis that the unlabeled portion of the dataset contains latent structural information critical for defining the reaction landscape. The most pronounced improvement was observed in the dataset by Ahneman et al., where the positives only model yielded a baseline MAE of 11.1%. The inclusion

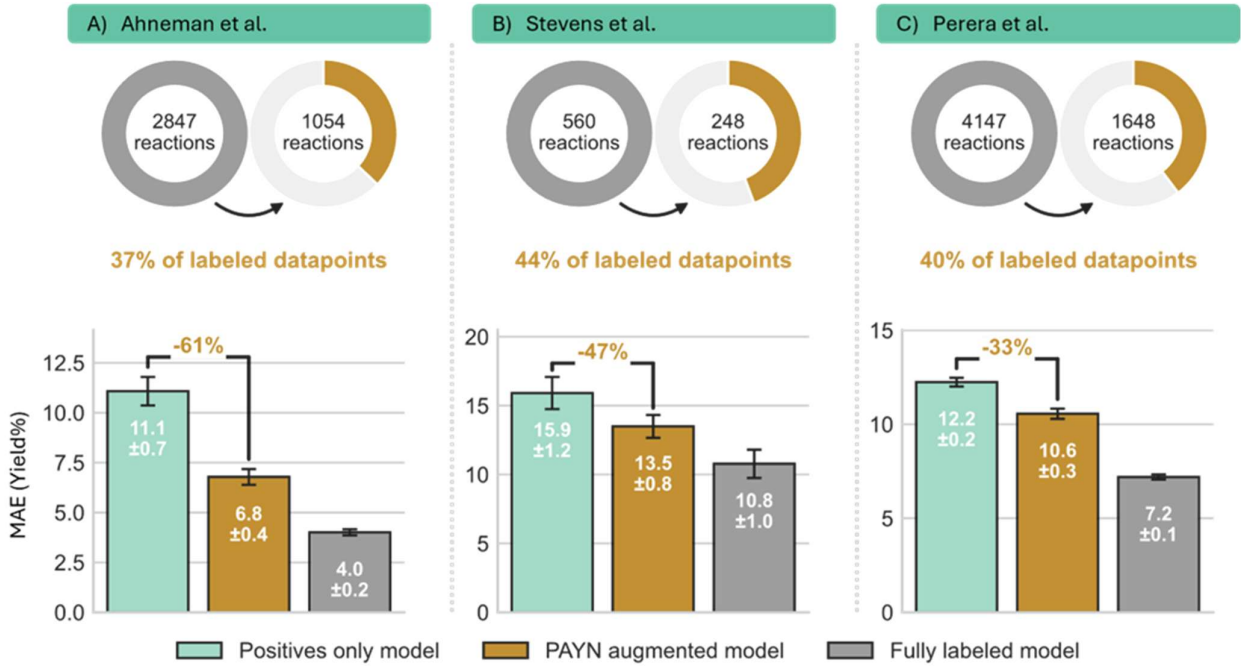

Fig. 2: Benchmarking of the PAYN augmented regression model against theoretical performance limits across three HTE datasets. Circles represent the amount of labeled datapoints used by the PAYN augmented model and positives only model in com parison to the fully labeled model. The bar chart displays the MAE for the positives only lower benchmark (light green), the proposed PAYN augmented model (gold), and the fully labeled upper benchmark (grey). Annotated percentages indicate the proportion of the performance gap between the positives only and fully labeled models that is successfully bridged by the PAYN framework. Error bars represent the standard deviation across five cross-validation folds.

of RN via the PAYN framework reduced the MAE to 6.8%. When viewed in the context of the fully labeled upper bound (MAE 4.0%), our approach successfully bridged 61% of the performance gap, without the need for any additional experiments and utilizing only 37% of the labeled data. Similar trends were observed for the Stevens et al. and Perera et al. datasets, with performance improvements of 47% and 33% (Figure 2 B-C), respectively. We attribute the variation in performance gain to the underlying topology of the chemical spaces, as well as density and size of the HTE datasets.

Conclusion

We demonstrate that PU Learning can be leveraged to train robust ML models for yield prediction even from severely biased data. Even after removing all negative datapoints to simulate extreme reporting bias, the PAYN framework successfully generates reliable negatives to construct balanced datasets. This approach allows predictive performance approaching that of models trained on fully labeled ground-truth data, unlocking the potential of literature data for predictive modelling.

Acknowledgments

Generous financial support by the European Research Council (Advanced 344 Grant Agreement No. 101098156, HighEnT.) and the Fonds der Chemischen Industrie (doctoral fellowship to J.C.S) is gratefully acknowledged. The authors thank F. Katzenburg, M. Kühnemund, D. J. Rana for proofreading and helpful discussions.

References

(1) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59* (1), 96–103. https://doi.org/10.1021/ja01280a022.

(2) Jiang, J.; Zhang, C.; Ke, L.; Hayes, N.; Zhu, Y.; Qiu, H.; Zhang, B.; Zhou, T.; Wei, G.-W. A Review of Machine Learning Methods for Imbalanced Data Challenges in Chemistry. *Chem. Sci.* **2025**, 10.1039.D5SC00270B. https://doi.org/10.1039/D5SC00270B.

(3) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J. Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis. *Nature* **2019**, *573* (7773), 251–255.

https://doi.org/10.1038/s41586-019-1540-5.

(4) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angew. Chem. Int. Ed.* **2022**, *61* (29), e202204647. https://doi.org/10.1002/anie.202204647.

(5) Bekker, J.; Davis, J. Learning from Positive and Unlabeled Data: A Survey. *Mach. Learn.* **2020**, *109* (4), 719–760. https://doi.org/10.1007/s10994-020-05877-5.

(6) Elkan, C.; Noto, K. Learning Classifiers from Only Positive and Unlabeled Data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*; ACM: Las Vegas Nevada USA, 2008; pp 213–220. https://doi.org/10.1145/1401890.1401920.

(7) Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. A Platform for Automated Nanomole-Scale Reaction Screening and Micromole-Scale Synthesis in Flow. *Science* **2018**, *359* (6374), 429–434. https://doi.org/10.1126/science.aap9112.

(8) Stevens, J. M.; Li, J.; Simmons, E. M.; Wisniewski, S. R.; DiSomma, S.; Fraunhoffer, K. J.; Geng, P.; Hao, B.; Jackson, E. W. Advancing Base Metal Catalysis through Data Science: Insight and Predictive Models for Ni-Catalyzed Borylation through Supervised Machine Learning. *Organometallics* **2022**, *41* (14), 1847–1864. https://doi.org/10.1021/acs.organomet.2c00089.

(9) 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* **2018**, *360* (6385), 186–190. https://doi.org/10.1126/science.aar5169.