# Guidelines for Machine Learning Yield Prediction from Small-Size Literature Dataset

J. Schleinitz*,[1,2,a] M. Langevin*,[2,3,b] Y. Smail,[4] B. Wehnert,[4] L. Grimaud,[1,c] and R. Vuilleumier[2,d]

[1] *LBM, Département de chimie, École Normale Supérieure, PSL University, Sorbonne Université, CNRS, 75005, Paris, France*

[2] *PASTEUR, Département de chimie, École Normale Supérieure, PSL University, Sorbonne Université, CNRS, 75005, Paris, France*

[3] *Molecular Design Sciences - Integrated Drug Discovery, Sanofi R&D, 94400, Vitry-Sur-Seine, France*

[4] *UPMC, PSL University, Sorbonne Université, CNRS, 75005, Paris, France*

(Dated: 18 May 2022)

Synthetic yield prediction using machine learning is intensively studied. Previous work focused on two categories of datasets: High-Throughput Experimentation data, as an ideal case study and datasets extracted from proprietary databases, which are known to have a strong reporting bias towards high yields. However, predicting yields using published reaction data remains elusive. To fill the gap, we built a dataset on nickel-catalyzed cross-couplings extracted from organic reaction publications, including scope and optimization information. We demonstrate the importance of including optimization data as a source of failed experiments and emphasize how publication constraints shape the exploration of the chemical space by the synthetic community. While machine learning models still fail to perform out-of-sample predictions, this work shows that adding chemical knowledge enables fair predictions in a low-data regime. Eventually, we hope that this unique public database will foster further improvements of machine learning methods for reaction yield prediction in a more realistic context.

Keywords: Machine learning - Dataset - Reaction yield prediction

[a] * Those authors contributed equally to this work; Electronic mail: jules.schleinitz@ens.psl.eu

[b] * Those authors contributed equally to this work; Electronic mail: maxime.langevin@sanofi.com

[c] Electronic mail: laurence.grimaud@ens.psl.eu

[d] Electronic mail: rodolphe.vuilleumier@ens.psl.eu

## I. INTRODUCTION

Machine learning (ML) algorithms learn complex functions from data. As it can leverage existing data to perform *in silico* approximations of costly experimental processes, ML applications have sparked strong interest in chemical sciences. While ML has already made a significant impact in drug development[1,2], synthetizability assessment of small molecules[3] or Computer Aided Synthesis Planning[4], the ability of ML to predict a reaction yield from its experimental conditions remains a major challenge[5] that is intensively studied[6,7]. Advances on reaction yield prediction would have a major impact on organic synthesis by significantly reducing cost, time and resources necessary to synthesize new chemicals.

Progress in ML is markedly driven by the increasing access to data. Thus, currently available datasets shape the evolution of ML for reaction yield prediction. Despite this, there are very few publicly available and easily operable datasets of chemical reactions with associated yields (Table S1). One of those few public datasets is the United State Patent and Trademark Office (USPTO) dataset,[8] that covers a wide range of chemical reactions extracted from patents. USPTO data is extremely diverse and suffers from a selection bias as only successful reactions tend to be reported in patents. ML has shown poor performance predicting yields on this dataset ($R^2 < 0.2$)[7]. In addition, two sets of High Throughput Experimentation (HTE) data, one of a Suzuki-Miyaura coupling,[9] and one of a palladium-catalyzed Buchwald-Hartwig cross-coupling,[10] are available in the literature. State-of-the-art modeling performs extremely well on those high-quality datasets ($R^2 > 0.8$)[7,10], but the extremely focused chemical reaction space covered by HTE limits the predictions to a narrow scope of experimental conditions and reactants.

While these datasets have enabled rapid progress of ML for yield prediction, there is a need for publicly available datasets[11,12] more representative of published reaction data or used by chemists in their everyday work. To the best of our knowledge, the most recent works on predicting reaction yields[5,13] rely on datasets extracted from Reaxys or Sci-Finder$^n$, which are not representative of the whole information contained in published reaction data. The main hurdle to gather a machine readable reaction database is the difficulty to automate data extraction from publications. One of the solution to overcome this issue would be a change toward a numerical data storage in the chemistry community, an option being the use of electronic laboratory notebooks[14] (ELN) interfaced with open-access database[15]. Nevertheless, the implementation of such tools requires significant time and investment. It also requires to convince the chemistry community of the merits of gathering data in a machine-readable format. Thus, we believe that showing the potentiality of a dataset derived from published reaction data to predict reaction yield would encourage chemists to embrace the new technologies available. Such a change would benefit the whole chemistry community.

To address this, we built a literature-mined, open-access reaction dataset that focuses on the Ni-catalyzed C-O bond activation to form C-C and C-N bonds: the NiCOlit dataset.[16] It gathers more than two thousand peer-reviewed reactions with detailed experimental conditions. As a singular literature representative dataset, NiCOlit stands as a benchmark for machine learning prediction of chemical yields found in published reaction data.

## II. RESULTS AND DISCUSSION

### A. Description of NiCOlit

NiCOlit was manually extracted from published reaction data cited in a recent review from Diao and co-workers[17]. In this review, the authors focus on the activation of carbon-oxygen bonds of phenol derivatives with nickel catalysts for coupling reactions. In order to reduce the size and the diversity of the dataset we arbitrarily restrained the study to challenging electrophiles towards the oxidative addition: sulfonates[18], phosphates and *in situ* activated phenols were left aside.

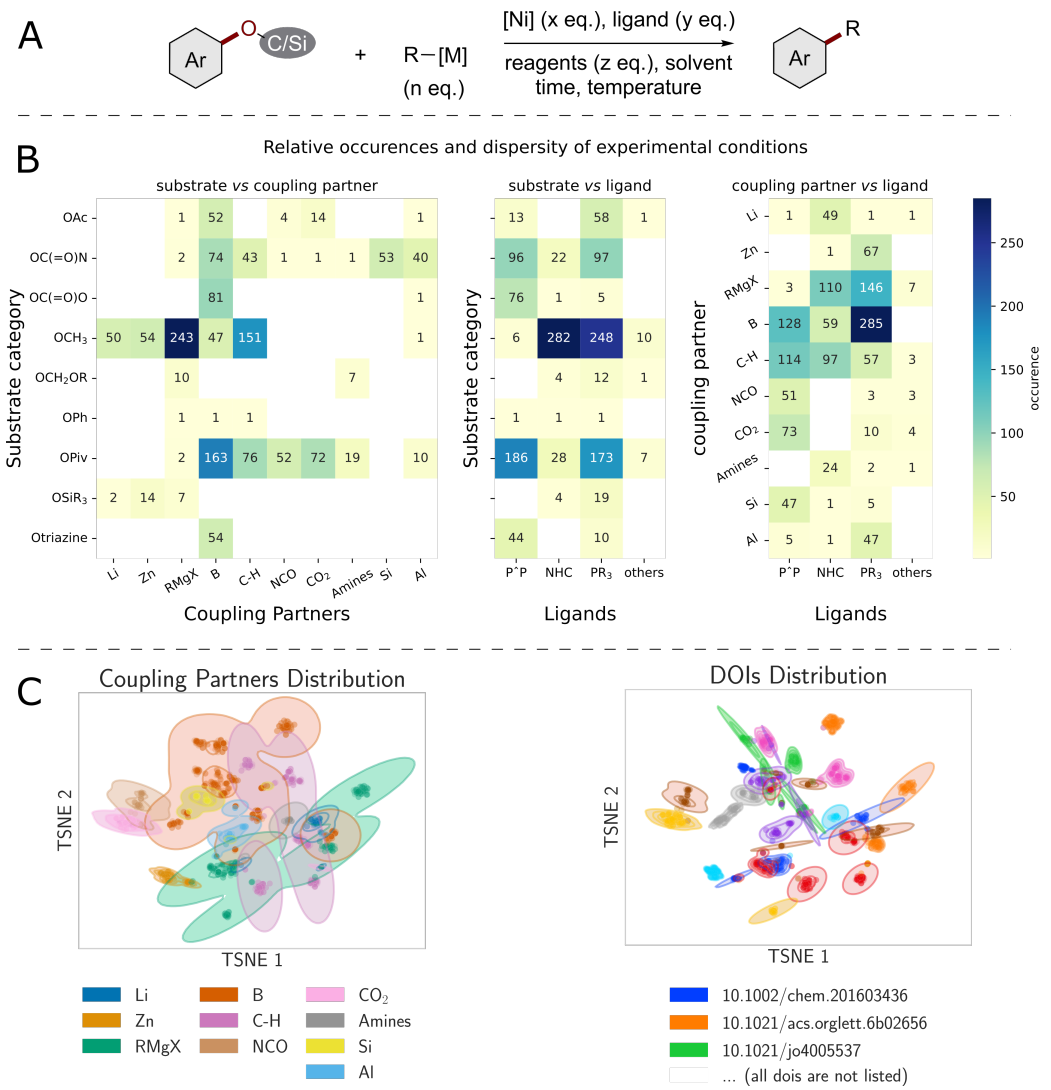


FIG. 1: (**A**) Chemical space of NiCOlit, (**B**) Diversity of NiCOlit in terms of coupling partners, substrates and ligands combinations, the color map is proportional to the number of reactions encountered for each combination. (**C**) t-SNE projection of NiCOlit, reaction data points are coloured by coupling partner category (left figure) and by publication origin (right figure).

The reactions displayed in both the main articles and their SI were extracted. For each reaction, the Simplified Molecular-Input Line-Entry System (SMILES)[19] chains of substrates, coupling partners, precursors, ligands, bases, additives, solvents, and products were gathered as well as experimental parameters: reaction time, temperature and molar ratios of the different partners (see SI Section 2). This highlighted issues when harmonizing different data sources, such as disparities in yield measurement techniques, or information being reported in prose rather than machine-readable format. The resulting database is unique, it gathers reactions from 45 publications within the chemical space illustrated Fig 1A. The different types of substrates, coupling partners and ligands reported in this dataset are summarized in Fig 1B.

The coverage of the substrate-coupling partners combinations is sparse and strongly biased toward a few reactions such as the Kumada coupling[20] (RMgX + $ArOCH_3$), which is the

most investigated with 243 reported reactions. This data-driven analysis[21] of nickel-catalyzed couplings shows that published reaction data seems to focus on specific combinations. We were surprised to see that reactions were clustered not by reaction parameters (e.g coupling partner) but almost perfectly by publications (Fig 1C). This minimal overlap between publications may be attributed to the poor reporting of reproduced experiments.

### B. Chemical Diversity: HTE versus NiCOlit

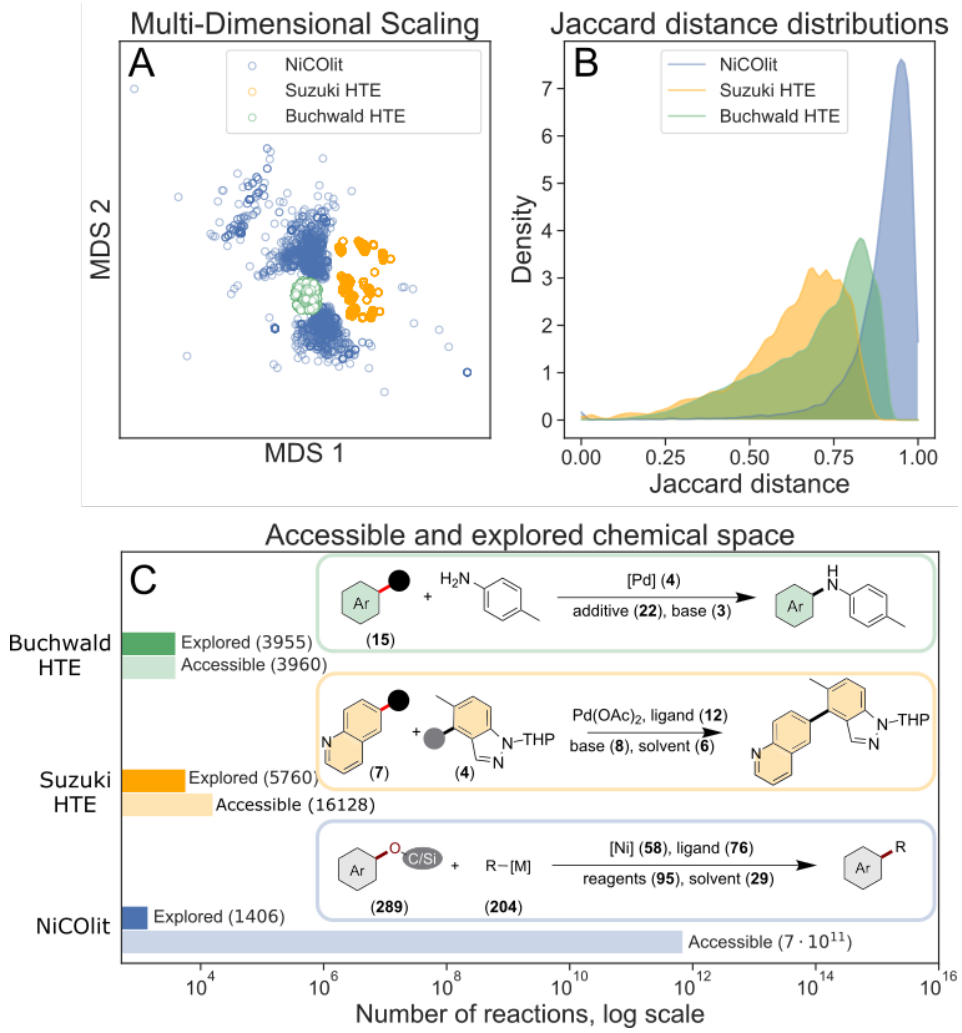


FIG. 2: Proportion of accessible chemical space observed and reaction diversities for HTE and NiCOlit data. (**A**): Multi-Dimensional Scaling (MDS) projection of the three datasets. (**B**): Distribution of Jaccard distances between reactions. (**C**): Proportion of accessible chemical space for HTE and NiCOlit data. Numbers indicated next to each parameter corresponds to the number of different choices appearing in the dataset for this parameter.

Inherent differences in terms of chemical diversity and yield distributions between NiCOlit and HTE data were laid out. This allows further understanding of how performances displayed by ML on HTE could translate to NiCOlit.

The projections of the three datasets on a common 2-dimensional space using Multi-

Dimensional Scaling (MDS)[22] (Fig. 2A, see SI for details) shows that NiCOlit is more spread out than HTE data. The distributions of the pairwise Jaccard distances between the reactions of each dataset were also computed (Fig. 2B) and corroborate the MDS analysis: distances between NiCOlit reactions are on average higher than in HTE data.

Most strikingly, we calculated the accessible chemical space as the number of all possible combinations of discrete parameters used for each category (e.g. reactants, catalysts, etc.). The proportion of accessible space explored[23] is an intensive metric that measures the ratio between the number of chemical reactions experimentally performed and the size of the accessible chemical space. Despite having roughly a similar number of reactions in the three datasets, the accessible chemical space of NiCOlit covers almost a trillion reactions versus less than 20k for both HTE datasets (Fig. 2C). This indicates that predicting yields on NiCOlit is more challenging than on HTE data. On the other hand, as most of the accessible space has been explored in HTE data (99% and 36% against only $2\text{x}10^{-7}$% for NiCOlit), developing an accurate model for NiCOlit allows to predict yields for a much larger set of unperformed reactions (almost a trillion reactions for NiCOlit).

### C. Scope/optimization structure of NiCOlit

The presence of reactions with low yields within a dataset is expected to be key for accurate predictions[11,24]. Very recently, Glorius and co-workers demonstrated that data expansion strategy with artificial low yields could boost ML predictive performances on proprietary databases[13]. However, this could introduce biases that have not yet been studied. HTE data display relatively homogeneous yield distribution, with many negative examples. Meanwhile, searching the proprietary database Sci-Finder$^n$ (Fig. 3C) for reactions matching the NiCOlit chemical reaction space returns 2,203 reactions with a clear bias toward high yields : 60% of them have a yield above 70%. NiCOlit yield distribution lays between HTE and Sci-Finder$^n$ data, with a significant amount of zero yields experiments but few reactions in the 20 to 40% yield range. This suggests a reporting bias in published reaction data, and an even stronger bias in proprietary databases. The fact that published reaction data contains significantly more negative examples than proprietary databases raises hope that such data could prove much more useful for machine learning model building.

Despite similarities in the reported yield distribution of NiCOlit and HTE data, the underlying structure of the reaction data drastically differs. In HTE, all possible combinations of reactants and reaction conditions are explored (Fig. 3A). Due to time and cost constraints, chemists tend to perform a sparser exploration of the chemical space to achieve a faster convergence. Indeed, published reaction data is reported in two categories of tables or schemes: optimization and scope. Optimization refers to the reaction conditions meaning that most parameters except substrate and coupling partner are modified in order to achieve an efficient reaction (vertical dots arrays Fig. 3B). In a complementary fashion, scope refers to reactions with various substrates and coupling partners under optimized conditions (horizontal cross arrays Fig. 3B). Scope experiments are performed in order to demonstrate the robustness of the reaction. In the case of NiCOlit, we noticed that yield distribution of optimization data is similar to the HTE yield distribution and that scope data displays a distribution reminding that of Sci-Finder$^n$ (Fig. 3C-D). Exploiting optimization tables during data extraction allows to bypass the lack of low yields reactions in proprietary datasets, and could offer improved predictive performances.

### D. Benchmark of Machine Learning Models on NiCOlit

Then, we evaluated how existing methods for yield prediction perform on NiCOlit. As the representation of chemical reactions in a machine readable format is a crucial step in statistical modeling of reaction yields[25], we selected three approaches representative of the state-of-the-art (see section III of SI). The first approach, RDKit FingerPrint (RDKit FP),

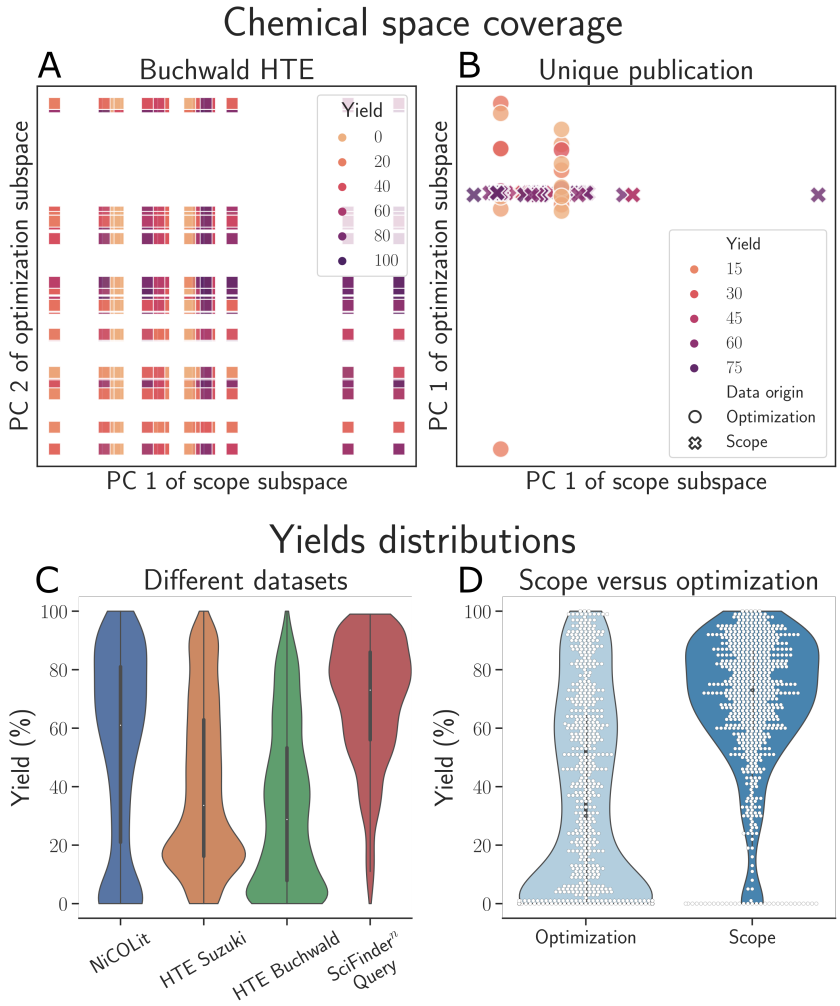


FIG. 3: Analysis of scope-optimization dataset structures and yields distributions. (**A**): Projection of the Buchwald-Hartwig dataset on the scope-optimization space, showing homogeneous coverage. The second Principal Component (PC) is displayed as the first PC is primarily driven by the 3 bases present in the dataset. (**B**): Projection of NiCOlit on the scope-optimization space, showing a biased exploration. (**C**): Yields distribution for the NiCOlit, HTE, and Sci-Finder$^n$ data. Bias towards high yields is observed on the NiCOlit and especially the Sci-Finder$^n$ datasets. (**D**): Yields distribution for the scope and optimization data on NiCOlit.

is based on the RDKit's (a cheminformatics tool) chemical reaction fingerprints[26,27] and one-hot encoding of the remaining variables. The second approach, referred to as Density Functional Theory (DFT), follows the guidelines given by the Auto-QChem framework[28] to generate DFT-based molecular descriptors adapted to the selected reaction. For the third approach, RXNFP, we featurized chemical reactions using the deep-learning RXNFP method[29]. Yield prediction models were built for each featurization with Random Forest regression models[30] that have shown excellent results on reaction yield prediction[6]. All metrics reported are averaged over 10 experiments with different random seeds. The DFT method outperforms the two others, and reaches an $R^2$ of 0.54, even though the difference with the RDKit FP is modest ($R^2$ of 0.49), while RXNFP showed the weakest performance

($R^2$ of 0.37) (Fig. S6). As expected, the performance of the model trained on optimization data performed better ($R^2$ of 0.48) than when trained on scope data ($R^2$ of 0.36) (Fig. S16).

The predictive performance on NiCOlit turns out to be far better than the one reported on the highly heterogeneous USPTO dataset ($R^2 < 0.2$)[7] and on data extracted from AstraZeneca's Electronic Lab Notebooks ($R^2 < 0.3$)[22]. This shows the potential of machine learning applied to published reaction data. Performances remain nonetheless lower than reported on the HTE data ($R^2 > 0.8$). An explanation could be the highly biased structure of published reaction data described in Fig. 3B, while HTE data cover a narrow and homogeneous chemical space (Fig. 2A and 1A-B) and are devoid of experimental and reporting bias[13,24,31,32]. Unlike published reaction data, HTE systematically reports yields for all reactions including low yields, and is comprised of reactions performed in the same experimental settings. This makes them a perfect case study for statistical learning compared to NiCOlit, at the cost of exploring a narrower chemical space. The rest of the manuscript focuses on the results obtained with the DFT model.

### E. Analysis of ML performance on out-of-sample predictions

Previous work on reaction yield prediction[7,10] focused mainly on predictive yields on random splits of the data. However, the nature of scientific discovery pushes chemists to constantly explore new reaction chemical space. Thus, the reactions for which we want to make yield prediction are not sampled from a static distribution, but undergo continuous distribution shift[33]. For instance, chemists are often interested in reactions including a novel substrate or a novel coupling partner (Fig. 4). Therefore, validation on a random split is not necessarily informative of how a model would perform when used by chemists in a prospective fashion. This problem was underlined in recent publications[7,10,34], where machine learning algorithms showed far worse predictive performances when applied to out-of-sample data (e.g. on reactions with an additive not seen in the training set). While reported models performed very well on random splits (with an $R^2$ above 0.9, see Table S1), these performances dropped significantly on some out-of-samples tests, with coefficient of correlations $R^2$ at best of 0.54 (obtained with a DFT model[6]).

We chose the prediction of reaction yield for an unseen substrate (Fig. 4) as a task of practical interest. All reactions that feature a given substrate were held-out; after training on the rest of the dataset, the model predicts the yields of the held-out reactions. Those results were aggregated over all substrates in the dataset. While the DFT model showed encouraging results on the substrate split task, the question of whether the reported predictive performance ($R^2 = 0.33$) is of practical interest is not clear. Therefore, we designed a realistic classification task, where the model classifies reactions using an unseen substrate in two classes, high yields ($> 50$) and low yields ($\leq 50$). This use case corresponds to the situation where a chemist wants to explore reactions with a new substrate, and relies on the model's prediction to discard low yield substrates, and to prioritize efficient ones. On this task, the DFT method reached high predictive performance (with a ROC-AUC, a performance metric for classifiers, of 0.74, see Fig. S18). This highlights a practical application of yield prediction models that can be achieved with existing methods.

Researchers have incentives to explore novel chemical space (Fig. 1C). Moreover, for each publication, reaction yield is biased by the chemist's skills and the way it is measured. This leads to a high heterogeneity between reactions from different publications. We evaluated how ML predicts yields on data from a new publication. A train-test split of the data, where the test set is comprised of all reactions from one publication, and the train set of all other reactions that do not appear in this publication, is used to assess the yield prediction performance of a model on a new publication (Fig. 4A - DOI Hold-out). Our results showed the inability ($R^2$ of -0.01) of ML to generalize to data from new publications.

While the reactions in NiCOlit are all extracted from publications referenced in the same review, they cover a wide range of possible mechanisms. As most of the publications extracted do not provide detailed mechanistic study of the reaction performed, a discrimination was

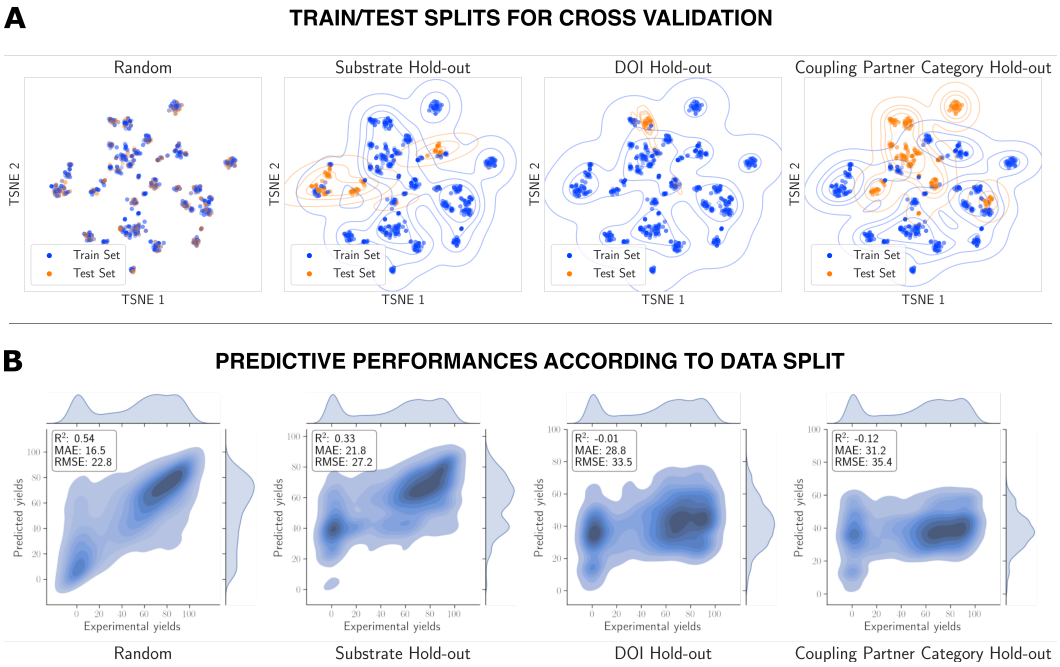


FIG. 4: (**A**): Examples of train-test splits from left to right: random split, split according to substrate (one substrate in the test set and all others in the training set), split according to publication (one Digital Object Identifier (DOI) is taken as the test set and all others as training set) and split according to coupling partner category (all reactions of one coupling partner category are taken as test set and all other coupling partners as training set). (**B**) DFT model performances for the different splits displayed in **A**. The performances displayed represent an average result over 10 random splits for the random task and all the possible substrates, DOI or coupling partner categories splits for the three remaining tasks.

made according to the nature of the coupling partner and substrate (as depicted Fig. 1B). All reactions of the same coupling partner category were held-out (e.g. boronic derivatives, see Fig. 1C). Models were trained on the rest of the dataset, and used to predict the yields of the held-out reactions (Fig. 4A - Coupling Partner Hold-out). Predictive performance reported for these experiments indicates whether the model is able to predict yields on coupling reactions using a different category of partners than the reactions of the training set. Again, the models fail to extrapolate to new coupling partners categories.

From a mechanistic point of view, the failure of the models to extrapolate is not very surprising as nickel catalysts can react through very different mechanistic pathways[35]. In this perspective, we analyzed the importance of the different features used by the models to perform predictions on more focused subsets of the NiCOlit. All reactions belonging to the same coupling partner category were regrouped and a predictive model was trained on each of these subsets. Then the importance of the features used by the models to perform the predictions were analyzed and displayed Fig. 5. The differences between the models are striking. As an example, the model trained on the RMgX coupling partner is highly dependent on the ligand, the substrate and the molar ratios used while the couplings with RLi mainly depend on substrate characteristics. Considering that widely different features are used to predict yields on the different coupling partner categories, the poor

generalizability of the models is not surprising.

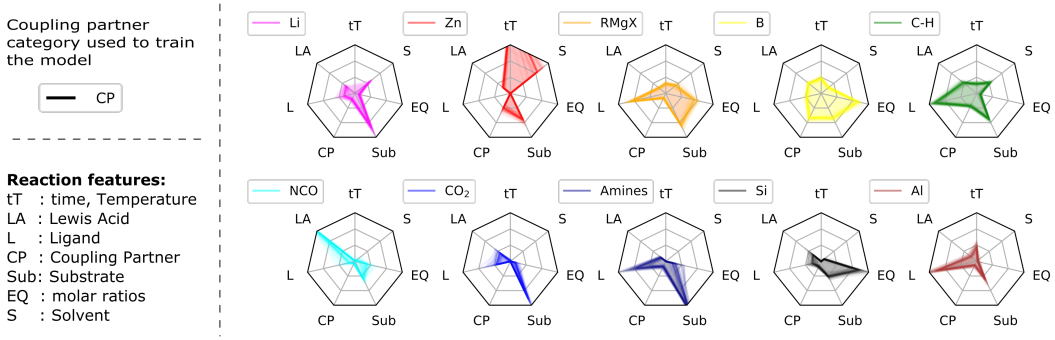


FIG. 5: Analysis of feature importance in predictive models trained on coupling partner category NiCOlit subsets. Each subset contains all the reactions belonging to a category of coupling partners.

### F. Chemist's Expertise Enables ML Predictions in a Low-Data Regime

Based on those results, we hypothesize that models trained on a restricted dataset of reactions sharing similar coupling partners would lead to equivalent predictive performances than models trained on the entire NiCOlit dataset. If true, this would give a precious guideline when gathering published reaction data in order to perform yield prediction.

To test this hypothesis, we trained the models on NiCOlit restricted to a given category of coupling partner (e.g. all reactions with a boron-based coupling partner) or substrate, and compared the results with those obtained with a model trained on the full dataset. Indeed, for most of the coupling partners or substrates, the model trained on the restricted dataset performs as well as the model trained on the full data (Table 1).

We also compared the predictive performance of the model trained on random splits of the full NiCOlit and on NiCOlit restricted to a specific coupling partner or substrate categories with two baselines. When confronted with a novel reaction with unknown yield, the Nearest Neighbor baseline (NN baseline) searches for the closest reaction in the database with a known yield and return its value as the yield prediction for the query reaction. This baseline assesses the performances of one of the natural strategy to extrapolate an unseen yield from the closest known reaction, that mimics the way a human chemist could proceed. The second baseline (scope/optimization baseline) returns the mean yield of reactions extracted from scope (resp. optimization) tables in the database if the query reaction is itself extracted from a scope (resp. optimization) table. This baseline accounts for the fact that a stratification by scope/optimization explains a part of the variability in yields observed within the database. As recent work suggests machine learning doesn't outperform simple baselines in related prediction tasks[5], our goal was to make sure that there was clear added value from using the machine learning approach. For substrates or coupling partners types with more than 70 data points, the machine learning approach systematically outperformed both baselines.

Furthermore, the performance is highly variable according to coupling partner or substrate categories. The most straightforward explanation for this behavior is the disparity between the number of reactions documented for each coupling partner category. The models exhibit poor or modest performances for coupling partners with less than 70 reactions reported. Adequate coverage of chemical space is crucial for building ML models[36]. Those results show that a dataset of a much smaller size than NiCOlit can be used to build a predictive model, provided that all reactions belong to the same coupling partner or substrate category (Fig. 6). This study also shed light on the approximate number of reactions needed to reach satisfying predictive performance (roughly one hundred reactions). We compared the performances on

| Coupling partner | $R^2$ | $R^2$ restricted | $R^2$ NN baseline | $R^2$ scope/ optimization baseline | Number of reactions | Number of publications | Accessible reactions |
|---|---|---|---|---|---|---|---|
| B | **0.47** | 0.45 | 0.17 | 0.26 | 472 | 11 | $1 \times 10^8$ |
| C-H | **0.59** | 0.56 | 0.31 | 0.11 | 271 | 3 | $3 \times 10^7$ |
| RMgX | **0.51** | 0.48 | 0.23 | 0.18 | 266 | 5 | $6 \times 10^5$ |
| $CO_2$ | **0.52** | 0.51 | 0.12 | 0.36 | 87 | 1 | $7 \times 10^3$ |
| Zn | 0.54 | **0.57** | -0.16 | 0.09 | 68 | 2 | $1 \times 10^3$ |
| NCO | **0.39** | 0.30 | -0.21 | 0.18 | 57 | 1 | $2 \times 10^4$ |
| Al | 0.20 | 0.18 | -0.35 | **0.26** | 53 | 1 | $1 \times 10^4$ |
| Si | **0.64** | 0.57 | 0.10 | 0.42 | 53 | 1 | $3 \times 10^4$ |
| Li | -0.13 | **0.05** | -0.29 | -0.03 | 52 | 1 | $2 \times 10^4$ |
| Amines | 0.17 | -0.05 | -0.58 | **0.32** | 27 | 1 | $5 \times 10^2$ |
| Substrate | $R^2$ | $R^2$ restricted | $R^2$ NN baseline | $R^2$ scope/optimization baseline | Number of reactions | Number of publications | Accessible reactions |
| OR | 0.55 | **0.57** | 0.3 | 0.13 | 546 | 11 | $5.4 \times 10^9$ |
| OPiv | 0.55 | **0.56** | 0.31 | 0.33 | 394 | 12 | $3.2 \times 10^9$ |
| OC(=O)N | **0.41** | 0.35 | -0.06 | 0.24 | 215 | 14 | $1.4 \times 10^8$ |
| OC(=O)O | **0.66** | 0.64 | 0.44 | 0.15 | 82 | 4 | $4.5 \times 10^5$ |
| OAc | 0.36 | **0.40** | -0.15 | 0.31 | 72 | 7 | $4.2 \times 10^6$ |
| Otriazine | 0.42 | **0.53** | 0.01 | 0.39 | 54 | 1 | $1.8 \times 10^5$ |
| $OSiR_3$ | -0.25 | -0.25 | -1.77 | **0.00** | 23 | 5 | $1.0 \times 10^4$ |
| OCOR | 0.06 | -0.39 | -0.74 | **0.31** | 17 | 4 | $3.2 \times 10^3$ |
| OPh | -0.30 | -1.34 | -0.41 | **-0.01** | 3 | 3 | $4.9 \times 10^2$ |

TABLE I: Performances of DFT and baselines models on each coupling partner and substrate categories subsets. $R^2$ corresponds to the predictions of a model trained on 80% of NiCOlit for a specific substrate or coupling partner category. $R^2$ restricted corresponds to the performances on the same target for a model trained on 80% of NiCOlit restricted to the targeted substrate or coupling partner. NN baseline is a nearest neighbor search based on DFT features and scope/optimization baseline returns the average yield of scope or optimization data according to whether the reaction is in a scope or optimization table. Metric reported is Pearson's coefficient of correlation $R^2$. The best values (with regard to predictive performance) are reported in bold.

the restricted sets with between 70-500 reactions and retrieved comparable performances ($R^2 \approx 0.5$) than what was obtained on Buchwald HTE data with a similar number of training points ($R^2 = 0.59$ for 150 data points)[10]. Interestingly, good performances can be attained in low-data regimes by injecting domain specific knowledge. Furthermore, for similar predictive performances, most accessible reaction spaces (last column, Table 1) are larger than their HTE counterparts.

## III. CONCLUSION

To reach its full potential, machine learning relies on high quality data. In the future, we expect that initiatives such as the Open Reaction Database[12] will provide the community

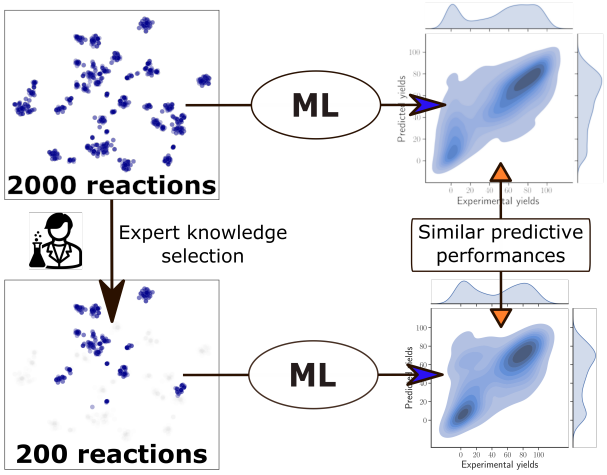


FIG. 6: ML performances when trained on full or restricted chemical space.

with the data needed. In the meantime, there is currently a lack of public unbiased datasets of chemical reactions with detailed experimental conditions. By releasing NiCOlit, we hope to foster the development of impacting data-driven approaches for yield prediction. Our results highlight the specificities of published reaction data. Comparison with HTE data shows a reporting bias towards high yields. Nonetheless, this bias is weaker in NiCOlit than in its counterpart extracted from proprietary database. Our analysis reveals that it is the extraction of optimization tables that allows the presence of negative examples and limits this reporting bias. In the light of recent results[13], this could lead to a significant improvement of machine learning models trained on published data. We also highlight a clear experiment selection bias, where reactions are explored along the two orthogonal optimization and scope directions. A benchmark of several state-of-the-art machine learning approaches confirms those observations. Indeed, predictive performance on NiCOlit lies between the one obtained on HTE data and on other available data sources of chemical reactions (proprietary databases or USPTO). We further tested the limitations of machine learning models on out-of-sample prediction tasks. While predictive models cannot currently extrapolate to new reactions or coupling partner categories, we showed that the model is able to generalize on reactions with new substrates. Leveraging this observation, we explain how domain-specific chemistry knowledge can help synthetic chemists select relevant data points to build models in low-data regimes. With only a hundred data points, building predictive models is within reach. We hope that these findings and our open-access database will stimulate the adoption of machine learning for yield predictions in the chemistry community, and encourage systematic data sharing in machine readable formats.

**Acknowledgement:**

The authors thank Maxime R. Vitale, Marc Bianciotto and Hervé Minoux for their thoughtful remarks and constructive feedback.

**Data and materials availability:**

All code and data used to produce the reported results can be found online at https://github.com/truejulosdu13/NiCOlit.

**Funding:**

CNRS and ENS are gratefully acknowledged for supporting J.S., L.G. and R.V. The French National Association of Research and Technology (ANRT) is gratefully acknowledged for supporting M.L. (contract 2019/0821). M.L. is employed by Sanofi.

**Authors contributions:**

M.L. and J.S. designed the project. Y.S. and B.W. extracted the database under the supervision of J.S.. M.L., J.S., Y.S. and B.W. conducted data analysis and developed the code. M.L. and J.S. drafted the manuscript. L.G. and R.V. supervised the project and revised the manuscript. All authors read and approved the final manuscript.

**Competing interests:**

M.L. is a Sanofi employee and may hold shares and/or stock options in the company. J.S., B.W., Y.S., R.V. and L.G. declares that they have no competing interests.

[1]S. Brogi, T. C. Ramalho, K. Kuca, J. L. Medina-Franco, M. Valko, *Frontiers in Chemistry* **8** (2020).
[2]V. Gallego, R. Naveiro, C. Roca, D. R. Insua, N. E. Campillo, *Molecular Diversity* (2021).
[3]L. Patel, T. Shukla, X. Huang, D. W. Ussery, S. Wang, *Molecules* **25**, 5277 (2020).
[4]C. W. Coley, W. H. Green, K. F. Jensen, *Accounts of Chemical Research* **51**, 1281 (2018).
[5]W. Beker, R. Roszak, A. Wołos, N. H. Angello, V. Rathore, M. D. Burke, B. A. Grzybowski, *Journal of the American Chemical Society* (2022).
[6]A. M. Żurański, J. I. Martinez Alvarado, B. J. Shields, A. G. Doyle, *Accounts of Chemical Research* **54**, 1856 (2021). PMID: 33788552.
[7]P. Schwaller, A. C. Vaucher, T. Laino, J.-L. Reymond, *Machine Learning: Science and Technology* **2**, 015016 (2021).
[8]D. M. Lowe, *Ph.D. thesis, University of Cambridge* (2012).
[9]D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson, N. W. Sach, *Science* **359**, 429 (2018).
[10]D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **360**, 186 (2018).
[11]P. M. Pflüger, F. Glorius, *Angewandte Chemie International Edition* **59**, 18860 (2020).
[12]S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen, C. W. Coley, *Journal of the American Chemical Society* **143**, 18820 (2021).
[13]F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen, F. Glorius, *Angewandte Chemie* (2022).
[14]S. G. Higgins, A. A. Nogiwa-Valdez, M. M. Stevens, *Nature Protocols* **17**, 179 (2022).
[15]K. M. Jablonka, L. Patiny, B. Smit, *Nature Chemistry* (2022).
[16]J. Schleinitz, M. Langevin, Y. Smail, B. Wehnert, truejulosdu13/nicolit: Nicolit dataset first release (2022).
[17]H. Diao, Z. Shi, F. Liu, *Synlett* **32**, 1494 (2021).
[18]S. Bajo, G. Laidlaw, A. R. Kennedy, S. Sproules, D. J. Nelson, *Organometallics* **36**, 1880 (2017).
[19]D. Weininger, *Handbook of Chemoinformatics* pp. 80 – 102 (2008).
[20]K. Tamao, K. Sumitani, M. Kumada, *Journal of the American Chemical Society* **94**, 4374 (1972).
[21]A. Pereira, C. Albornoz, O. S. Trofymchuk, *Organometallics* **0**, null (0).
[22]M. Saebi, B. Nan, J. Herr, J. Wahlers, Z. Guo, A. Zurański, T. Kogej, P.-O. Norrby, A. Doyle, O. Wiest, N. Chawla, *Chemrxiv preprint* (2021).
[23]D. Reker, E. A. Hoyt, G. J. Bernardes, T. Rodrigues, *Cell Reports Physical Science* **1**, 100247 (2020).
[24]P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **533**, 73 (2016).
[25]K. Jorner, A. Tomberg, C. Bauer, C. Sköld, P.-O. Norrby, *Nature Reviews Chemistry* **5**, 240 (2021).
[26]G. Landrum, Rdkit: Open-source cheminformatics (2020).
[27]N. Schneider, D. M. Lowe, R. A. Sayle, G. A. Landrum, *Journal of Chemical Information and Modeling* **55**, 39 (2015).
[28]A. M. Żurański, J. Y. Wang, B. J. Shields, A. G. Doyle, *Reaction Chemistry &amp Engineering* (2022).
[29]P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino, J.-L. Reymond, *Nature Machine Intelligence* **3**, 144 (2021).
[30]L. Breiman, *Machine Learning* **45**, 5 (2001).
[31]F. Strieth-Kalthoff, F. Sandfort, M. H. S. Segler, F. Glorius, *Chemical Society Reviews* **49**, 6154 (2020).
[32]W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle, E. V. Anslyn, *ACS Central Science* **7**, 1622 (2021).
[33]J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence, *Dataset Shift in Machine Learning* (The MIT Press, 2009).
[34]F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, F. Glorius, *Chem* **6**, 1379 (2020).
[35]V. M. Chernyshev, V. P. Ananikov, *ACS Catalysis* **12**, 1180 (2022).
[36]S. K. Kariofillis, S. Jiang, A. M. Żurański, S. S. Gandhi, J. I. M. Alvarado, A. G. Doyle, *Journal of the American Chemical Society* **144**, 1045 (2022).
[37]K. V. Chuang, M. J. Keiser, *Science* **362**, eaat8603 (2018).
[38]J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, Neural message passing for quantum chemistry (2017).
[39]A. Sato, T. Miyao, K. Funatsu, *Molecular Informatics* p. 2100156 (2021).
[40]A. Yada, K. Nagata, Y. Ando, T. Matsumura, S. Ichinoseki, K. Sato, *Chemistry Letters* **47**, 284 (2018).
[41]S.-J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, *IEEE Journal of Selected Topics in Signal Processing* **1**, 606 (2007).
[42]J. Friedman, T. Hastie, R. Tibshirani, *Journal of Statistic Software* **33**, 1 (2010).
[43]B. J. Reizman, Y.-M. Wang, S. L. Buchwald, K. F. Jensen, *Reaction Chemistry & Engineering* **1**, 658 (2016).
[44]Z. Fu, X. Li, Z. Wang, Z. Li, X. Liu, X. Wu, J. Zhao, X. Ding, X. Wan, F. Zhong, D. Wang, X. Luo, K. Chen, H. Liu, J. Wang, H. Jiang, M. Zheng, *Organic Chemistry Frontiers* **7**, 2269 (2020).
[45]K. Nakamura, M. Tobisu, N. Chatani, *Organic Letters* **17**, 6142 (2015).
[46]Z.-C. Cao, Q.-Y. Luo, Z.-J. Shi, *Organic Letters* **18**, 5978 (2016).
[47]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Édouard Duchesnay, *Journal of Machine Learning Research* **12**, 2825 (2011).
[48]D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Molecular Informatics* **37**, 1700153 (2018).

[49]D. Bajusz, A. Rácz, K. Héberger, *Journal of Cheminformatics* **7** (2015).
[50]L. van der Maaten, G. Hinton, *Journal of Machine Learning Research* **9**, 2579 (2008).
[51]F. Furche, R. Ahlrichs, *The Journal of Chemical Physics* **117**, 7433 (2002).
[52]M. E. Casida, C. Jamorski, K. C. Casida, D. R. Salahub, *The Journal of Chemical Physics* **108**, 4439 (1998).
[53]A. D. Becke, *The Journal of Chemical Physics* **98**, 5648 (1993).
[54]C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **37**, 785 (1988).
[55]S. H. Vosko, L. Wilk, M. Nusair, *Canadian Journal of Physics* **58**, 1200 (1980).
[56]P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, *The Journal of Physical Chemistry* **98**, 11623 (1994).
[57]M. J. Frisch, *et al.*, Gaussian09 Revision E.01. Gaussian Inc. Wallingford CT 2009.
[58]M. J. Kamlet, J. L. M. Abboud, M. H. Abraham, R. W. Taft, *The Journal of Organic Chemistry* **48**, 2877 (1983).