Design and Application of a Screening Set for Monophosphine Ligands in Metal Catalysis

Tobias Gensch,[1]* Sleight R. Smith,[2] Thomas J. Colacot,[3] Yam Timsina,[3] Guolin Xu,[3] Ben W. Glasspoole,[3] Matthew S. Sigman[2]*

[1] Department of Chemistry, TU Berlin, Straße des 17. Juni 135, Sekr. C2, 10623 Berlin, Germany.

[2] Department of Chemistry, University of Utah, 315 South 1400 East, Salt Lake City, UT 84112, USA.

[3] MilliporeSigma (A business of Merck KGaA, Darmstadt, Germany), 6000 N. Teutonia Ave., Milwaukee, WI, 53209, USA.

ABSTRACT: In reaction discovery, the search space of discrete reaction parameters such as catalyst structure is often not explored systematically. We have developed a tool set to aid the search of optimal catalysts in the context of phosphine ligands. A virtual library, *kraken*, that is representative of the monodentate P(III)-ligand chemical space was utilized as the basis to represent the discrete ligands as continuous variables. Using dimensionality reduction and clustering techniques, we suggested a Phosphine Optimization Screening Set (PHOSS) of 32 commercially available ligands that samples this chemical space completely and evenly. We present the application of this screening set in the identification of active catalyst for various cross-coupling reactions and how well-distributed sampling of the chemical space facilitates identification of active catalysts. Furthermore, we demonstrate how proximity in ligand space can be a useful guide to further explore ligands when very few active catalysts are known.

Introduction

Reaction optimization is perhaps the most resource intensive phase of developing and applying a synthetic reaction. This process typically involves adjusting a multitude of reaction parameters that are generally interdependent. For example, the optimization search space of a standard metal-catalyzed Suzuki-Miyaura cross-coupling has been estimated to encompass 50 million reactions sampling eleven different reaction parameters.[1] Thus, in order to explore this search space efficiently and identify optimal conditions with the least number of experiments, the choice of which reactions to carry out is critical, especially at the outset when little is known about the role of the reaction parameters.[2] Typically, however, initial empirical optimization studies are selected on the basis of the experimenters intuition and compound availability. This approach is particularly common when one is exploring how a discrete variable, such as a ligand for a catalytic process, impacts the reaction outcome. The practitioner often changes one variable at a time in a linear fashion. In the case of ligand evaluation, one would select ligands with varied steric and electronic properties, and iteratively choose new sets of ligands based on the interpretation of the observed experimental results. This can also be scaled in a high-throughput experimentation (HTE) manifold to accelerate data collection. Frequently this will result in a biased ligand selection wherein the perceived properties impacting the reaction outcome are used to guide the next step.

This iterative approach contrasts how rational selection tactics in a high-dimensional search space are applied to continuous variables. It is relatively straightforward to choose experiments that sample such a parameter evenly. For example, the temperature could be sampled at even intervals of 10 °C; with 5 experiments at such intervals, identifying an optimal temperature can be simple, and one can even interpolate within the sampling interval by fitting a function to determine the optimal temperature. A statistically rigorous implementation of this principle, the Design of Experiments (DoE) approach,[3,4] has been used extensively to guide the optimization of several continuous variables, such as temperature and concentration, simultaneously. However, applying these principles to the optimization of discrete variables such as ligands remains limited as one cannot readily define the search space in order to choose experiments that sample that space effectively (Figure 1). Even with experimental knowledge, a discrete treatment does not contain information on the relationship among the other experiments. Properly defining and sampling the search space of discrete variables becomes even more relevant as data-driven tools are increasingly used to analyze and guide HTE campaigns.[5,6]

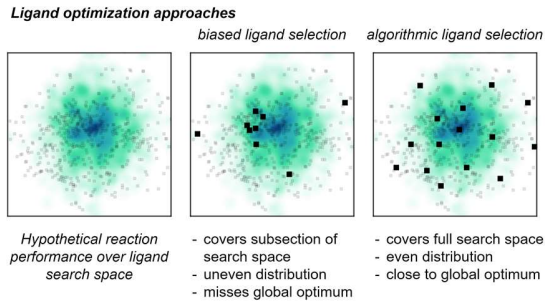


Figure 1. Comparison of ligand optimization approaches using arbitrary data. Gray points represent potential ligands defining the search space; the color map represents a hypothetical response such as yield across the search space where a darker color corresponds to better performance; the black points represent experiments that sample the search space.

It is clear that a rational selection of initial experiments could improve sustainability and shorten timelines of reaction optimization and increase the chances of achieving a global optimum.[7–9] As such, we were interested in developing a workflow to identify a screening set of monodentate phosphorus ligands, perhaps the most widely used ligand type in cross-coupling catalysis, that would quantitatively represent the chemical space that is commercially available. Using an objectively-designed set of ligands that evenly and completely covers the range of ligand properties would allow testing all property trends in a single set of ligands. We hypothesized that this type of approach has several advantages including: 1) such a set would have a representative ligand in each area of chemical space and should contain at least one or several well-performing ligands for most applications, 2) an even sampling of most ligand properties during optimization facilitates the use of data science tools to interrogate the experimental data with the aim of suggesting a next ligand set to improve performance, and 3) if the training data covers the complete property space, the search for optimal ligands based on the modeled training data will for the most part be an interpolation in the descriptor space, increasing the accuracy of such predictions.

Herein we apply our recently disclosed computationally defined library of phosphorus ligand features *kraken*[10] to define a screening set curated from those ligands commercially available at MilliporeSigma. This set of 32 ligands spans the descriptor space of the *kraken* library and was evaluated for four palladium-catalyzed cross-coupling reactions. In all four case studies, high-performing catalyst systems were identified. Finally, we showcase how proximity in the chemical space defined by the ligand features can be used to identify alternative ligands with desired performance even in scenarios with very sparse data, such as a single "hit".

Workflow

The specific workflow for the selection of a diverse molecule set adopted herein (Figure 2) represents just an example of a more general procedure where different choices can be made at every step, depending on the application. Indeed, others have used similar procedures to generate diverse sets of molecules in unique contexts,[7,9,11–16] including phosphorus ligand sets.[8,17–19]

First, selecting a set of ligands that covers the entire range of ligand properties requires a means to quantify this range in advance, thus defining the search space. To this end, we choose to utilize physicochemically relevant descriptors to quantify molecular properties that directly impact chemical reactivity and ligand behavior.[20] In this context, we have recently reported the chemical space characterization of an extensive library of ca. 1500 monodentate P-donor compounds (including phosphites, phosphoramidites and many other combinations of heteroatom-substituents at P, all collectively referred to as "phosphines" for the sake of brevity in the following) at the DFT level, taking conformational variation into account (referred to as *kraken*).[10] This incorporates many commercially available ligands, as well as examples from the academic literature and previously unsynthesized compounds and is thus representative of the structural diversity of monodentate P-donor ligands. Each conformer is characterized by 78 individual properties, which include sterics and electronics of the P-donor site as well as whole-molecule properties. To account for the conformational impact on ligand behavior, 190 "condensed" descriptors have been derived from these properties for each ligand, including the Boltzmann-weighted average and the highest and lowest value of each property across all conformers. The chemical space defined by these ligands and descriptors can be explored more conveniently and efficiently after dimensionality reduction.[17,21,22] We found that principal component analysis (PCA) yields low-dimensional spaces that are highly interpretable by chemical means (Figure 2, 1.) even just by analyzing the first four principal components (PC, 59% variance). We have hypothesized that geometric relationships in this chemical space representation correspond approximately to the relative chemical reactivity found for these ligands.[1,10,19]

After defining the entire search space, a filtering step can be necessary or useful, according to the goals of the screening set. Our goal was to create a general ligand set using commercially available ligands only, to facilitate early-stage reaction screens. Filtering to the 495 compounds we have tagged as commercially available based on SciFinder[23] reveals an even distribution of those compounds over the full search space with the exception of the "extreme corners" that mostly contain compounds that are not relevant as ligands in catalysis (Figure 2, 2.). Conceptually, the entire virtual library can be filtered by any user-defined criteria depending on prior knowledge and/or experimental requirements to arrive at more or less directed screening sets. For example, the selection procedure could be limited to a subsection of the ligand space such as that excluding heteroatom-based substitution at P if the reaction of interest is known to be catalyzed only by phosphine-derived metal complexes.[19]

We wanted to employ a sampling method that would allow us to select an arbitrary number of samples so that the screening set size can be adjusted according to experimental constraints. Furthermore, we pursued a method that would generate ranked candidates for each screening set member representing a section of the chemical space, so that further considerations such as price or stability of the ligands could be considered in screening set design without compromising the statistical validity of the approach. Both of these secondary objectives are met by clustering methods such as *k*-means clustering. While there are unsupervised learning methods that identify the natural clustering that is present in a data set to identify the optimal numbers of clusters, we specifically chose *k*-means clustering as it allows choosing the desired number of clusters via the hyperparameter *k*. The implicit assumption taken in this approach is that the phosphine chemical space representation as given by the first 4 principal components is relatively continuously populated without clusters that are separated by "empty" space. The chemical space as defined by the 190 descriptors does not fulfill this assumption because it is only sampled sparsely by the 495 commercially available compounds. In the four-dimensional space defined by the first 4 PC, the same number of samples represents a much higher density. *K*-means clustering of the commercial ligands in this 4-dimensional space into *N* clusters thus provides a natural means to select a diverse screening set (Figure 2, 3.).[15,24] The immediate output from the clustering analysis is the geometric location of the center of each cluster, rather than a specific sample. For each cluster, screening set candidates can then be sampled on the basis of increasing distance to the cluster center. This method allows for the curation of the screening set by other objectives such as price, availability or

stability, while at the same time maintaining a diverse selection fulfilling the design criteria.

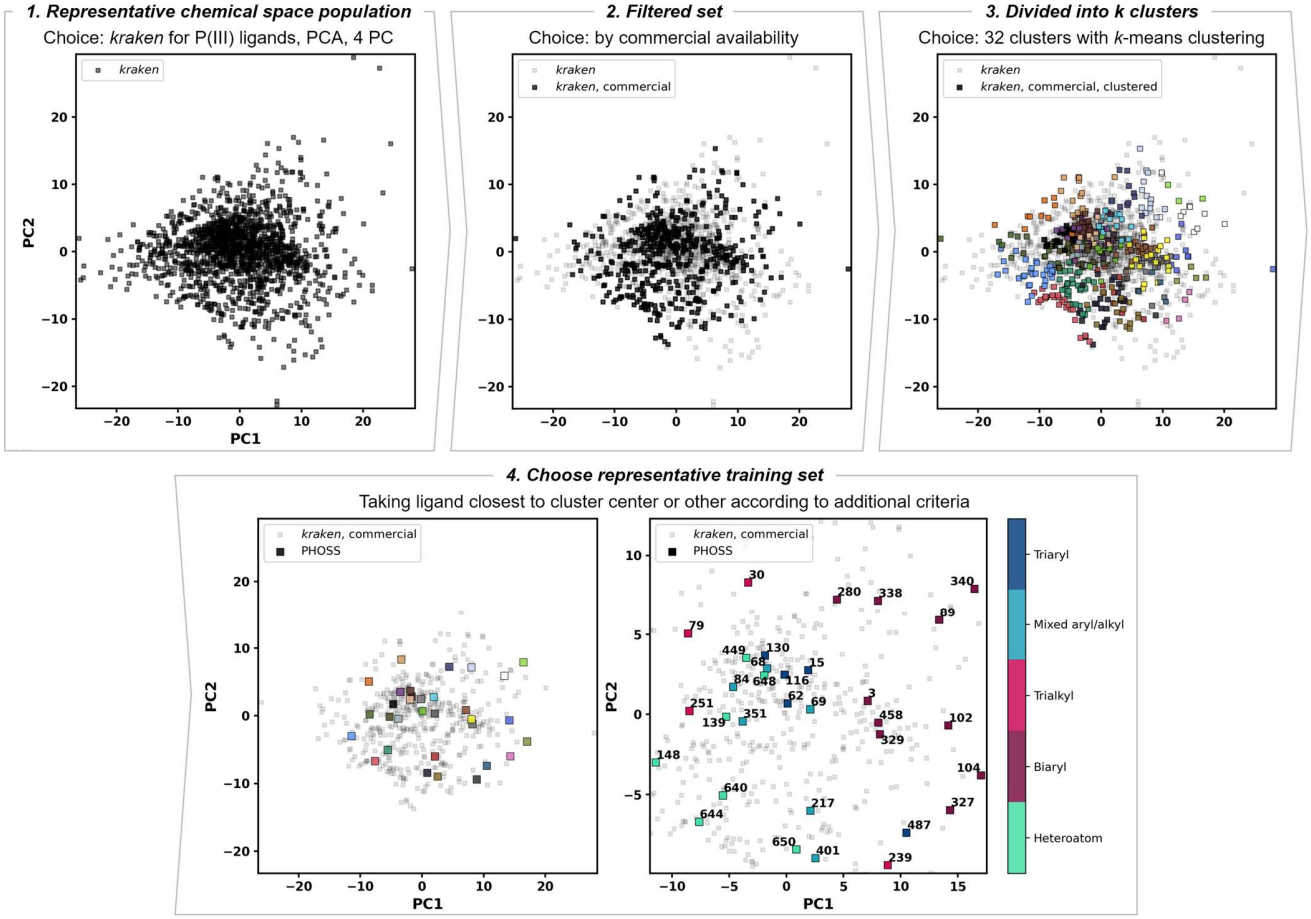


Figure 2. General workflow for the design of diverse compounds sets, illustrated with the specific exemplary choices taken in the design of the phosphorus ligand screening set described herein. The colors in 3. and 4. (left) are arbitrary, corresponding to one of each cluster. The clustering was done with the first four principle components of the *kraken* DFT feature space, thus the clusters and selected ligands appear to be overlapping in the two-dimensional projection shown. The structures of all numbered compounds in PHOSS are shown in Figure 3.

There are alternative and equally valid approaches to choosing a diverse set of ligands. For example, selections with the Kennard-Stone algorithm maximize the distances between all designated samples, thus resulting in an even and complete coverage of the space in question.[14,15,25] This will always select samples from the "outer edges" of the chemical space. For the PCA representation of the phosphine descriptor space, this results in selection of rather unusual structures that may not be feasible ligands for catalysis (e.g., by being too bulky or too electron-poor to coordinate metals). Furthermore, implementing secondary objectives or a curation step is less straightforward with this method. Ultimately, the precise algorithm used for this step likely has less of an impact on the performance of the screening set than other steps such as which compounds and descriptors are used to define the chemical space in the first place. It has been shown previously that any algorithmic sampling of a chemical space is better than random sampling, and even that is better than the biased sampling that is still common in experimental studies.[7,9,16]

With these steps in the workflow defined, we aimed to define a screening set encompassing 32 ligands (Figure 2, 4.). This was in part to achieve compatibility with typical HTE screening setups of 96 wells per plate, thus allowing space for either duplicate experiments or the testing of several reaction conditions on a single plate. Furthermore, the curation of candidates after *k*-means clustering involved selection of compounds that were available to MilliporeSigma (ca. 2018), generally selecting less expensive alternatives to facilitate adoption of the ligand set and opting for Buchwald-type ligands and against heteroatom-substituted compounds where possible. The final screening set (Figure 2, 4. and Figure 3) consists of ligands from many major classes of phosphorus ligands such as triaryl, trialkyl and mixed aryl/alkyl phosphines, Buchwald-type biaryl phosphines, and several heteroatom-substituted compounds such as a phosphinite, phosphites, phosphoramidites, and an aminophosphine. Along with very common ligands, it includes some exotic ligands that are not often tested in catalysis, which might lead to unexpected discoveries. In the following, the ligand set is referred to as the Phosphine Optimization Screening Set (PHOSS).

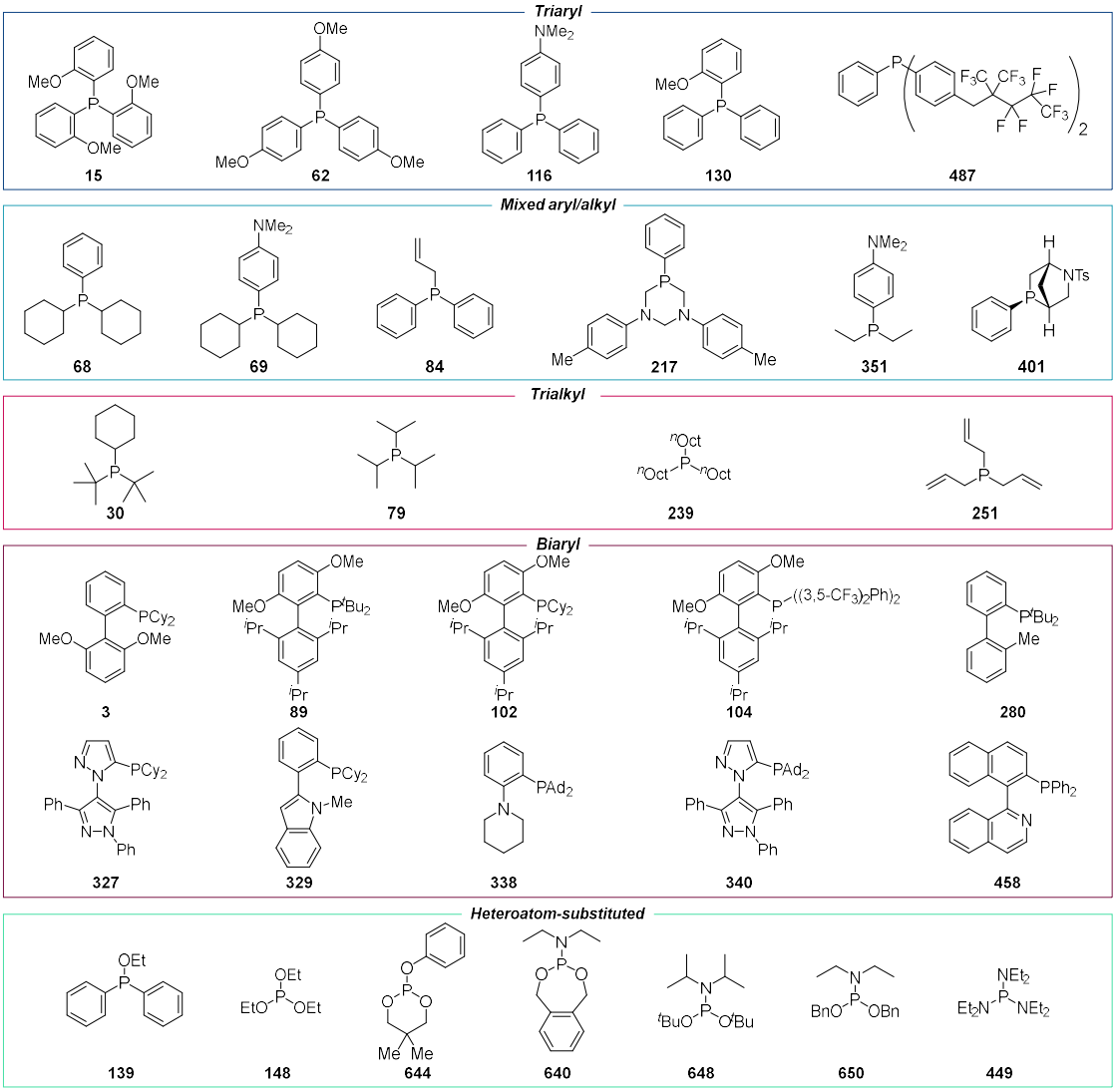


Figure 3. Screening set of monodentate P(III) ligands ("PHOSS") that was used in the following case studies. The numbers refer to the unique ID in the *kraken* database.[10]

Application of PHOSS

To test the behavior of the screening set, we experimentally performed four case studies. The goal of these case studies was to test two premises: 1) Will the screening set contain a diverse range of outputs with at least one high-performing ligand for most applications? And, where necessary, 2) will the data obtained from this screening set facilitate a means to rationally select a next set of ligands through a data driven workflow? In this context, the case studies are representative of common reaction optimization problems in catalysis, featuring two "easier" examples and two more "difficult" cases. For each case study, reactions were performed with each ligand from the screening set individually at representative, constant reaction conditions. The yield of the reactions was determined by chromatographic analysis and each reaction carried out at least in duplicate or more if there was a discrepancy between the first two results.

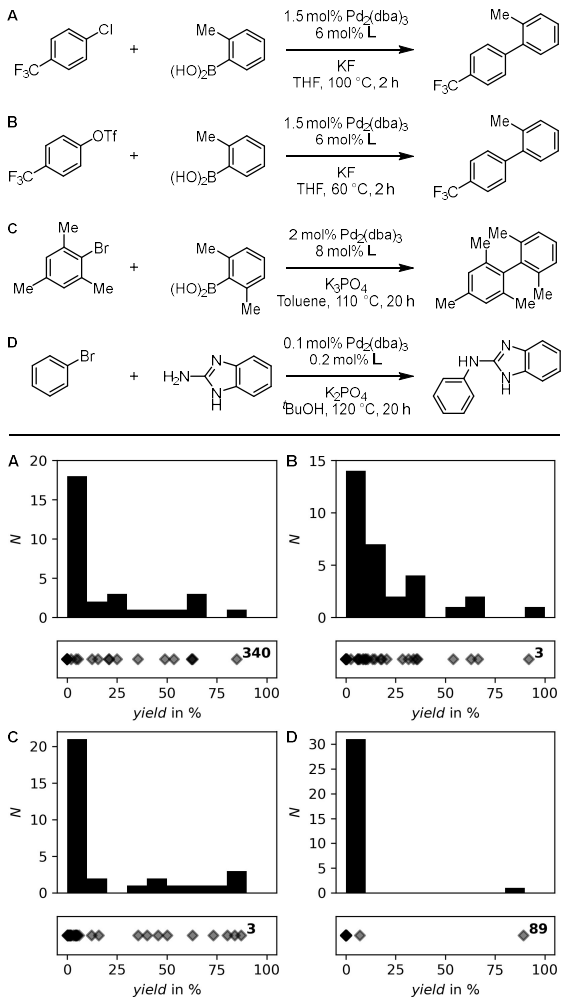
For the two easier cases, we selected prototypical Suzuki-Miyaura couplings of both an aryl chloride and aryl triflate, wherein the data was recently collected in the context of understanding how ligands effect speciation in cross-coupling reactions.[26] In the evaluation of PHOSS, excellent performance was detected, with a wide range of active catalysts found. Perhaps unsurprisingly, the best ligands from within PHOSS were Buchwald-type dialkyl biaryl phosphines in both cases. Thus, for the aryl chloride (Scheme 1, **A**), the best performing ligand was Ad-BippyPhos (**340**, 85% yield) whereas for the aryl triflate (Scheme 1, **B**), the best performing ligand was SPhos (**3**, 92% yield). Of most importance, the ligand set does provide a reasonably well-spread data response consistent with the design philosophy described above, identifying seven ligands that provided over 33% yield in both cases.

The first "more difficult" case study was the evaluation of a sterically hindered Suzuki-Miyaura reaction wherein both coupling partners present di-*ortho* substitution (Scheme 1, **C**).[27] Excitingly, the results of this screen also provide a diverse range of results as well as high performing examples. Again, SPhos (**3**, 87% yield) was the best performing ligand, but three other

structurally diverse ligands provided yields over 70%, thus demonstrating the often-non-intuitive relationships between ligand structure and reactivity: CyBippyPhos (**327**, 84%), JackiePhos (**104**, 80%) and ^{t}Bu$_2$PCy (**30**, 73%).

The final reaction is perhaps most representative of how we envision the use of this screening set in a reaction optimization setting. In this case study, we performed a chemoselective Buchwald–Hartwig amination at low catalyst loading (Scheme 1, **D**).[28] The results are qualitatively different than any of the other examples, with only one high performing ligand and only one other ligand giving a >5% yield. Clearly, this is a difficult coupling under the reaction conditions evaluated. In spite of this, the screening set does allow for the identification of a good ligand for this reaction. The best performing was tBuBrettPhos (**89**, 89% yield), which was previously reported in the optimization of this reaction and also coincidentally included in PHOSS at the outset. However, this heavily skewed distribution of reaction outputs also makes virtual screening for other high-performing catalysts by regression modeling impossible. This prompted us to investigate approaches to ligand suggestion that are applicable even with very sparse data or other scenarios where regression modeling fails.

Scheme 1. Case studies tested with PHOSS.[a]



[a]L refers to the ligands shown in Figure 3, each added individually. For further details, see the Supporting Information.

Proximity-based ligand suggestion

Based on the premise that our ligand descriptor set characterizes the chemical and physical properties of the ligands, we hypothesized that proximity in the reduced-dimensional chemical space representation provided by PCA would correspond to similar reactivity. In other words, untested ligands that are near ligands that provided high-performing catalysts have a higher chance to also provide high-performing catalysts than those that are further away. While this somewhat simplistic assumption will not always be successful, it does provide an opportunity to suggest experiments where a practitioner would like to have options for the use of a different ligand or test mechanistic hypotheses. In some cases, this type of analysis may suggest the same ligands as those that a chemist would intuitively select but it is possible that ligands with similar features may not be as obvious.[10]

To test this hypothesis, the chemical space is first classified into regions of "most active" and "less active" based on the experimentally observed yield with the nearest ligand in case study D. Here, only tBuBrettPhos (**89**) was considered in the "most active" class and all other ligands "less active" (Figure 4, A). In other scenarios, a specific yield criterion could be used instead of just the best ligand, as required by the current data and goals of the individual project. Subsequently, classes are assigned to untested ligands based on the closest PHOSS member. For example, each other ligand for which tBuBrettPhos is the closest member is also assigned the "most selective" class. Practically, this is accomplished using a *k*-nearest neighbors classifier[29] with $k = 1$ trained on all experimental results with the PHOSS ligands, and applied to all commercially available ligands in the *kraken* library.

Using this procedure, eight of the untested ligands that were commercially available at the time this case study was carried out were classified into the "most active" region (Figure 4).[30] Two of those had been tested during reaction optimization in the original study for this reaction,[28] where Me$_4$tBuXPhos (**291**) provided the desired product in 23% yield and tBuXPhos (**90**) was reported as providing < 5% yield. Encouraged by this result, we tested four of the other ligands, along with two more that are near tBuBrettPhos but just outside the region classified "most active". In the reactions using RockPhos (**103**) and Me$_3$(OMe)tBuXPhos (**534**), 65% and 64% of the desired product were observed, respectively, while all other reactions did not result in product formation. In fact, the two successful ligands are also the two ligands with the closest absolute distance to **89** of all commercially available ligands in the *kraken* database (Figure 4, X). Thus, three out of six ligands classified into the "most active" region that were tested by us or others did form active catalysts in this case study. This could be a useful starting point for further optimization of other reaction parameters, or the design of novel ligands. However, as perhaps expected, the proximity-based selection does not always guarantee finding successful ligands. Beside the simplistic representation as distance in the four-dimensional principal component space, another limiting factor can be the availability of any ligands in close proximity to a successful experimental sample. Furthermore, we do acknowledge that ligands in close proximity in the descriptor space will also be structurally similar to the point that might seem like a very straightforward variation a chemist would have suggested based on intuition alone (as is the case here). However, since the molecular structure is not an

explicit part of the descriptor set, this approach does have the potential to result in nonintuitive suggestions.

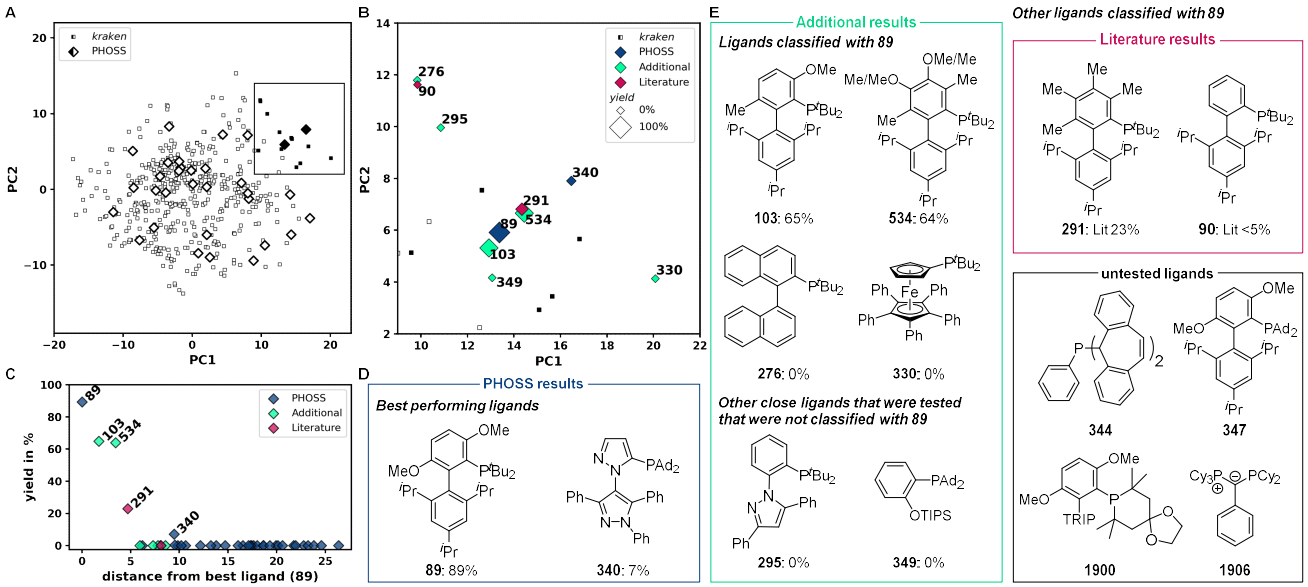


Figure 4. Ligand selection with sparse data demonstrated for case study **D**. A) Classification of the ligand space into active (black) and inactive (white) using a *k*-nearest neighbors approach trained on the experimental data (large diamonds) and applied to the commercially available part of *kraken* (small squares). B) Enlarged depiction of the chemical space around the best performing ligand, including ligands used in additional experiments and literature. Diamond size proportional to observed yield in case study **D**. Literature results refer to the yields reported in ref [28] under the same reaction conditions. C) Relationship between absolute distance from **89** in the first 4 PC and observed yield in case study **D**. D) Best ligands from PHOSS in case study **D**. E) Further ligands and results included in B).

Conclusion

In summary, we have applied a flexible workflow to aid in the screening and optimization of catalytic reactions employing monodentate phosphorus ligands in PHOSS. This approach includes the ability to select the number of ligands for a desired screening application and the ability to construct focused ligand sets based on prior knowledge, availability, or other criteria. In several of the case studies, an excellent distribution of yield outputs was observed, which could aid in future statistical modeling. We further demonstrate the ability to take a single reaction hit and explore the ligand space neighborhood even in scenarios when more traditional virtual screening approaches relying on regression models would fail.

We envision that the use of standardized ligand sets in reaction optimization could contribute to more comparable data sets, thus working towards future efforts in transfer learning across wider parts of catalysis.[31,32] Finally, one can envision integrating this ligand selection strategy into active/iterative optimization procedures such as simplex searches,[33] Bayesian optimization[34] or active learning approaches,[35] where an efficient initialization is crucial for a rapid optimization outcome.[36]

ASSOCIATED CONTENT

SUPPORTING INFORMATION.

Details on the ligand set selection process.
Experimental procedures for the case studies
Python code for the selection of ligand sets.
This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

* Tobias Gensch: tobias.gensch@tu-berlin.de
Matthew S. Sigman: matt.sigman@utah.edu

Author Contributions

M.S.S. and T.G. conceptualized the project. T.G. performed the data analysis and ligand set design. T.G., T.C. and S.R.S. designed the case studies. S.R.S., Y.T. and G.X. performed experiments. T.G. and M.S.S. wrote the manuscript with help from S.R.S. All authors have given approval to the final version of the manuscript.

ACKNOWLEDGMENT

T.G. thanks the Leopoldina Fellowship Programme of the German National Academy of Sciences Leopoldina (LPDS 2017−18). T.G. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2008/1 – 390540038, and by a Liebig Fellowship of the Fonds der Chemischen Industrie. M.S.S. acknowledges financial support from NSF under the CCI Center for Computer Assisted Synthesis (CHE-1925607).
Results from this study were used to develop a predictive analysis algorithm that is commercialized by MilliporeSigma (Phosphine Predictor, 919845).

REFERENCES

(1) Murray, P. M.; Tyler, S. N. G.; Moseley, J. D. Beyond the Numbers: Charting Chemical Reaction Space. *Org. Process Res. Dev.* **2013**, *17* (1), 40–46. https://doi.org/10.1021/op300275p.

(2) Collins, K. D.; Gensch, T.; Glorius, F. Contemporary Screening Approaches to Reaction Discovery and Development. *Nat. Chem.* **2014**, *6* (10), 859–871. https://doi.org/10.1038/nchem.2062.

(3) Carlson, R. Designs for Explorative Experiments in Organic Synthetic Chemistry. *Chemom. Intell. Lab. Syst.* **2004**, *73* (1 SPEC. ISS.), 151–166. https://doi.org/10.1016/j.chemolab.2004.04.005.

(4) Weissman, S. A.; Anderson, N. G. Design of Experiments (DoE) and Process Optimization. A Review of Recent Publications. *Org. Process Res. Dev.* **2015**, *19* (11), 1605–1633. https://doi.org/10.1021/op500169m.

(5) Renom-Carrasco, M.; Lefort, L. Ligand Libraries for High Throughput Screening of Homogeneous Catalysts. *Chem. Soc. Rev.* **2018**. https://doi.org/10.1039/C7CS00844A.

(6) Eyke, N. S.; Koscher, B. A.; Jensen, K. F. Toward Machine Learning-Enhanced High-Throughput Experimentation. *Trends Chem.* **2021**, *3* (2), 120–132. https://doi.org/10.1016/j.trechm.2020.12.001.

(7) Pötter, T.; Matter, H. Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases. *J. Med. Chem.* **1998**, *41* (4), 478–488. https://doi.org/10.1021/jm9700878.

(8) Moseley, J. D.; Murray, P. M. Ligand and Solvent Selection in Challenging Catalytic Reactions. *J. Chem. Technol. Biotechnol.* **2014**, *89* (5), 623–632. https://doi.org/10.1002/jctb.4306.

(9) Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Henle, J. J.; Denmark, S. E. Computational Methods for Training Set Selection and Error Assessment Applied to Catalyst Design: Guidelines for Deciding Which Reactions to Run First and Which to Run Next. *React. Chem. Eng.* **2021**, *6* (4), 694–708. https://doi.org/10.1039/d1re00013f.

(10) Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *ChemRxiv* **2021**. https://doi.org/10.26434/chemrxiv.12996665.v1.

(11) Carlson, R.; Carlson, J. E. Principal Properties and Designs for Discrete Variations. *Org. Process Res. Dev.* **2005**, *9* (5), 680–689. https://doi.org/10.1021/op040022w.

(12) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* **2012**, *52* (10), 2570–2578. https://doi.org/10.1021/ci300338w.

(13) Bess, E. N.; Bischoff, A. J.; Sigman, M. S. Designer Substrate Library for Quantitative, Predictive Modeling of Reaction Performance. *Proc. Natl. Acad. Sci.* **2014**, *111* (41), 14698–14703. https://doi.org/10.1073/pnas.1409522111.

(14) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363* (6424), eaau5631. https://doi.org/10.1126/science.aau5631.

(15) Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *J. Am. Chem. Soc.* **2020**, *142* (26), 11578–11592. https://doi.org/10.1021/jacs.0c04715.

(16) Kariofillis, S. K.; Jiang, S.; Żurański, A. M.; Gandhi, S. S.; Martinez Alvarado, J. I.; Doyle, A. G. Using Data Science to Guide Aryl Bromide Substrate Scope Analysis in a Ni/Photoredox-Catalyzed Cross-Coupling with Acetals as Alcohol-Derived Radical Sources. *ChemRxiv* **2021**. https://doi.org/10.33774/chemrxiv-2021-6kd0t D O I: 10.33774/chemrxiv-2021-6kd0t.

(17) Bjørsvik, H.-R.; Hansen, U. M.; Carlson, R.; Åkermark, B.; Robinson, W. T.; Wood, B. R.; Robinson, W. T.; Roos, B. O.; Vallance, C.; Wood, B. R. Principal Properties of Monodentate Phosphorus Ligands. Predictive Model for the Carbonyl Absorption Frequencies in Ni(CO)3L Complexes. *Acta Chem. Scand.* **1997**, *51*, 733–741. https://doi.org/10.3891/acta.chem.scand.51-0733.

(18) Burello, E.; Rothenberg, G. Optimal Heck Cross-Coupling Catalysis: A Pseudo-Pharmaceutical Approach. *Adv. Synth. Catal.* **2003**, *345* (12), 1334–1340. https://doi.org/10.1002/adsc.200303141.

(19) Christensen, M.; Yunker, L. P. E.; Adedeji, F.; Häse, F.; Roch, L. M.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M. S.; Aspuru-Guzik, A.; Hein, J. E. Data-Science Driven Autonomous Process Optimization. *Commun. Chem.* **2021**, *4* (1), 112. https://doi.org/10.1038/s42004-021-00550-x.

(20) Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119* (11), 6561–6594. https://doi.org/10.1021/acs.chemrev.8b00588.

(21) Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J. J.; Murray, P.; Hose, D. R. J.; Osborne, R.; Purdie, M. Expansion of the Ligand Knowledge Base for Monodentate P-Donor Ligands (LKB-P). *Organometallics* **2010**, *29* (23), 6245–6258. https://doi.org/10.1021/om100648v.

(22) For an interactive interface of the chemical space representation, see https://kraken.cs.toronto.edu. All ligand numbers used throughout this manuscript can be used to find the same ligand on the webinterface.

(23) Commercial availability was determined for all ligands in *kraken* using SciFinder in late 2018 for the compounds comprising the library at the time and June 2021 for any newer additions. This is an approximation because not all sources listed in SciFinder may be practical and availability may have changed in the meantime. An exact assignment is not necessary for our application due to the curation step later in the procedure.

(24) Burden, F. R.; Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, *42* (16), 3183–3187. https://doi.org/10.1021/jm980697n.

(25) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11* (1), 137–148. https://doi.org/10.1080/00401706.1969.10490666.

(26) Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Peters, E.; Gensch, T.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G. Linking Mechanistic Analysis of Catalytic Reactivity Cliffs to Ligand Classification. *ChemRxiv* **2021**. https://doi.org/10.26434/chemrxiv.14388557.

(27) Barder, T. E.; Walker, S. D.; Martinelli, J. R.; Buchwald, S. L. Catalysts for Suzuki-Miyaura Coupling Processes: Scope and Studies of the Effect of Ligand Structure. *J. Am. Chem. Soc.* **2005**, *127* (13), 4685–4696. https://doi.org/10.1021/ja042491j.

(28) Ueda, S.; Buchwald, S. L. Catalyst-Controlled Chemoselective Arylation of 2-Aminobenzimidazoles. *Angew. Chemie - Int. Ed.* **2012**, *51* (41), 10364–10367. https://doi.org/10.1002/anie.201204710.

(29) Zheng, W.; Tropsha, A. Novel Variable Selection Quantitative Structure--Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (1), 185–194. https://doi.org/10.1021/ci980033m.

(30) two newer ligands that are present in the *kraken* library also fall in that region and could be useful ligands for this reaction: joYPhos (**1906**) and VincePhos (**1900**), a phosphorinane analog to BrettPhos.

(31) 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.

(32) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nat. Rev. Chem.* **2019**, *3* (10), 589–604. https://doi.org/10.1038/s41570-019-0124-0.

(33) Henson, A. B.; Gromski, P. S.; Cronin, L. Designing Algorithms to Aid Discovery by Chemical Robots. *ACS Cent. Sci.* **2018**, *4* (7), 793–804. https://doi.org/10.1021/acscentsci.8b00176.

(34) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590* (7844), 89–96. https://doi.org/10.1038/s41586-021-03213-y.

(35) Eyke, N. S.; Green, W. H.; Jensen, K. F. Iterative Experimental Design Based on Active Machine Learning Reduces the Experimental Burden Associated with Reaction Screening. *React. Chem. Eng.* **2020**, *5* (10), 1963–1972. https://doi.org/10.1039/D0RE00232A.

(36) Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L. How to Explore Chemical Space Using Algorithms and Automation. *Nat. Rev. Chem.* **2019**. https://doi.org/10.1038/s41570-018-0066-y.