ORGANIC PROCESS RESEARCH & DEVELOPMENT
OPR&D

pubs.acs.org/OPRD

Article

Impact of Phosphine Featurization Methods in Process Development

Jennifer M. Crawford, Tobias Gensch, Matthew S. Sigman,* Jennifer M. Elward,* and Janelle E. Steves*

Cite This: https://doi.org/10.1021/acs.oprd.1c00357

ABSTRACT: Modern high-throughput experimentation (HTE) has enabled the rapid exploration of large expanses of chemical reaction space to accelerate the development of key synthetic steps in pharmaceutical processes. However, the dimensionality of reaction parameters, the desire to use minimal starting material, and the need to thoroughly analyze reaction outcomes still require the judicious selection of which experiments to perform in which order. Therefore, the development of a capability to quantify reagent diversity and analyze reaction outcomes in HTE holistically is paramount. A method to address this goal would combine computational featurization of key reaction components with the use of multivariate linear regression modeling to correlate the reaction performance outputs. In this context, we describe a process of establishing a computational featurization platform for monodentate phosphine ligands and considerations for its implementation at GSK. We demonstrate that the choice of computational method has an impact on phosphine descriptor values, ligand selection for experiments, and the development of linear regression models of reaction outcomes.

KEYWORDS: *high-throughput experimentation, phosphine, computation, features*

Many perspectives have been published[1] on approaches to HTE design that showcase diverse strategies that pharmaceutical companies use to optimize reaction conditions in a plate-based format. In our experience, most experimental designs originate from variations of literature-precedented reaction conditions, application of chemists' intuition and expert knowledge, and best-guess approaches to exploring chemical space. Based on demands for an acceleration of API delivery and the prospect of minimal material in early-phase small-molecule development, we were compelled to examine and redesign our platform for Pd-catalyzed cross-coupling reactions. Specifically, we envisioned a data science-informed method to efficiently explore chemical space to maximize the information obtained, with the goal of conducting fewer, more information-rich experiments to optimize a particular reaction.

Toward this goal, we were inspired by recent work on phosphine featurization,[2] as this ligand class is commonly used within our HTE platform in various cross-coupling reactions. We were specifically interested in exploring the chemical space of our internal phosphine library by quantifying steric and electronic differences between ligands and the application of this technology to reaction development. Therefore, we initiated an academic−industrial collaboration between our groups at GSK and the University of Utah to probe how to integrate phosphine featurization and linear regression modeling with HTE reaction outcomes to maximize the information gained from GSK Pd-catalyzed cross-coupling experiments. A key goal of our collaboration was to determine how the choice of computational featurization method impacts experimental design and analysis. Based on our related recent work on comprehensive phosphine featurization[3a] and selection,[3b,c] we also wanted to develop interactive resources to enable GSK chemists to customize their experimental designs while exploring phosphine space in a more quantitative manner. Herein, we describe our process to featurize phosphines and utilize these features for experimental design at GSK with a specific focus on how the choice of computational method may impact experimental design and analysis. To begin our investigation, we computed monodentate phosphine features for our internal ligand library using two methods recently reported by the Sigman group: featurization of the ligands as phosphine oxides[2c] and as phosphines.[2d] Bidentate phosphines were excluded from this endeavor as methods for their featurization have not yet been established to the level achieved for monodentate phosphines. In the first step, we conducted a molecular mechanics-based conformational search of each molecule (Figure 1). While rigid structures produced 10 or fewer conformers, more flexible ligands that resulted in greater than 10 conformers were assessed by clustering structures with similar conformations into groups based on root-mean-square deviation (RMSD), a common metric for analyzing conformer similarity.[4] One conformer from each cluster was then selected for further calculation, resulting in a collection of 10 diverse conformers for each ligand. Steric and electronic features were then computed and extracted via quantum mechanical (QM) optimization of molecular geometries obtained from our conformational search, computation of the steric and electronic features of each geometry, and rapid feature extraction with a scripted workflow. For certain features

Received: September 10, 2021

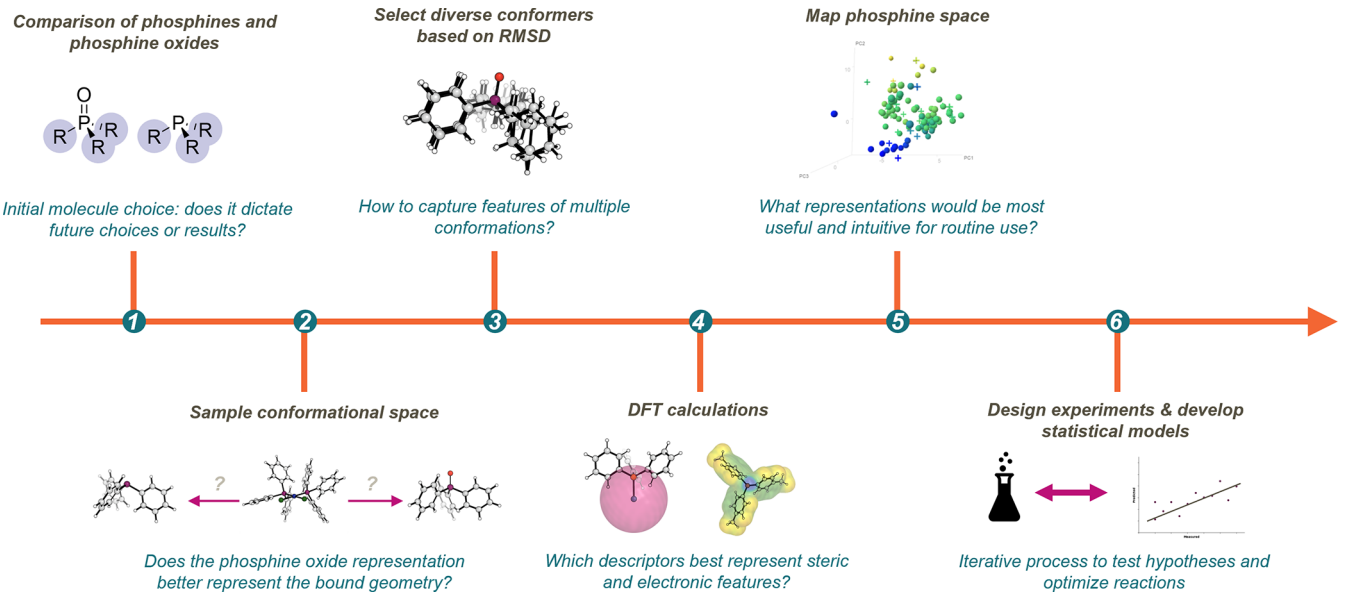


Figure 1. Workflow for featurization and mapping of phosphines and phosphine oxide space.

describing electron density at phosphorus, the oxygen atom was removed from the phosphine oxide prior to descriptor calculation (see the Supporting Information for further details).

After completing the extraction and curation of descriptors, we sought to create maps of phosphine space to qualitatively analyze differences between the two computational featurization methods and to build a tool that our chemists would use to design their experiments. In creating this tool, we considered the plot types that were available in software we typically use in GSK Chemical Development (e.g., JMP, Spotfire[5]) as well as feedback from our colleagues about the tool's ease of use and consistency of ligand arrangement with their chemical intuition. We then chose to develop dendrograms from the first three principal components of each set of features (Figure 2).[6] The dendrogram enables chemists to explore phosphine space by selecting ligands from different branches or exploit phosphine space by focusing experiments on ligands from a smaller branch.

We immediately observed that ligands were arranged differently across the two dendrograms created via the same statistical methods and that this ligand arrangement could impact chemists' phosphine selection for experiments. Considering that the values of steric and electronic features were likely responsible for this shift in ligand arrangement, we analyzed two key features for an array of ligands across the dendrograms, percent buried volume ($\%V_{bur}$)[7,8] and the minimum electrostatic potential in the phosphorus lone pair region (V_{min}).[9] In this context, buried volume was especially relevant as it has recently been shown to correlate with reactivity and ligand speciation.[10] To facilitate comparison, we analyzed the Boltzmann-weighted average of the $\%V_{bur}$ and V_{min} of our computed phosphine to phosphine oxide structures and included feature values of crystal structures to represent a ligand geometry on a Pd complex.[11] This analysis revealed that though most ligands exhibit buried volume similar to that of their reference crystal structure, there is a subset of ligands that do not follow this trend (Figure 3).

Most notably, a series of biaryl ligands including RuPhos, XPhos, SPhos, and CyvBRIDP exhibited crystal structure-derived buried volumes far lower than both our phosphine and phosphine oxide computed Boltzmann-weighted averages. To better understand the observed differences in this ligand series, the QM minimum energy structures for the phosphines and phosphine oxides were compared to their corresponding minimum energy reference crystal structures. We determined that while both the reference crystal structure and biaryl phosphines in this series exhibited a flipped geometry (Figure 3B), our phosphine oxide ensembles did not include these flipped geometries. In this case, the oxide moiety presumably serves as a steric blocking group (see the Supporting Information for further information). Our initial interest in phosphine oxide featurization originated from the assumption that an oxide could substitute for a Pd atom to enable efficient calculation of ligand features while producing geometries assumed to be relevant to catalysis.[2c] Theoretical and experimental work conducted by Buchwald[12] and Espinet,[13] however, demonstrates the relevance of flipped biaryl ligand geometries in potential mechanisms for phosphine oxidation and catalyst deactivation. Our recent work with collaborators[3a] captures these biaryl ligand geometries at the initial stages of a related computational workflow by intentionally selecting conformations that represent the extremes of steric descriptors. Work is ongoing to understand how to appropriately sample the complex conformational landscape to address different chemistry challenges.

To further understand the origin of variations in ligand arrangement in Figure 2, we anticipated that V_{min} would be influential. However, the comparison between phosphines, phosphine oxides, and the corresponding reference crystal structures did not result in a clear trend (see the Supporting Information for comparison figure). While a thorough analysis of all feature values is out of scope of this work, we conclude based on our initial analysis that the computational method can impact feature values, leading to clear changes in ligand arrangement. We reasoned that any method that provides a realistic feature space could serve equally well for exploratory purposes. Nevertheless, we sought to further understand the impact of computational method on experimental design and models of reaction outcomes (vide infra).

To qualitatively assess how phosphine selection from across the dendrograms could be impacted by changes in arrangement,

https://doi.org/10.1021/acs.oprd.1c00357
Org. Process Res. Dev. XXXX, XXX, XXX−XXX

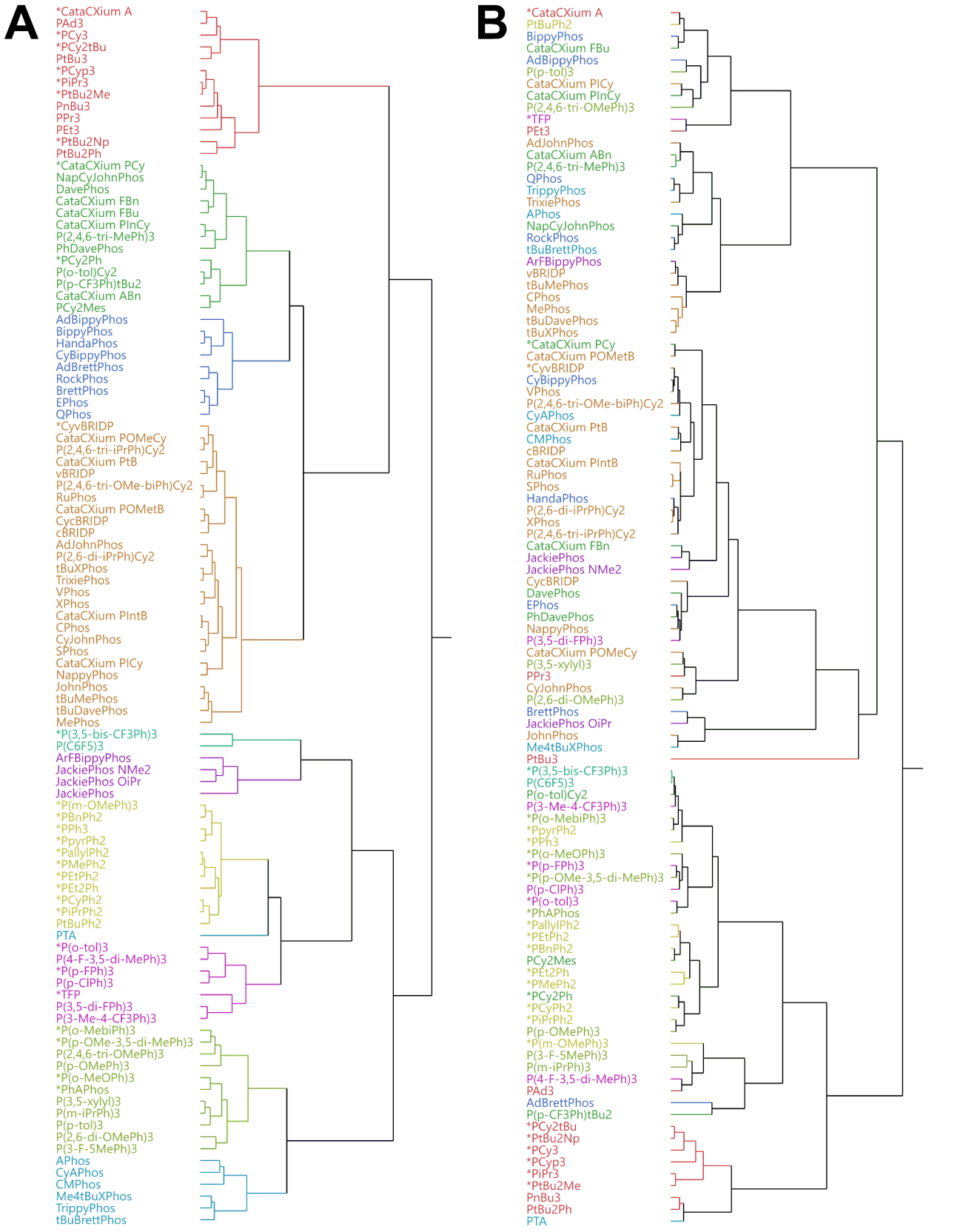


Figure 2. (A) Dendrogram using phosphine oxide features. (B) Dendrogram using phosphine features.

we first performed *k*-means clustering to identify up to 24 diverse ligands from each dendrogram. This phosphine set size is relevant to performing experiments in 96-well HTE plates; a set of 24 ligands can accompany two bases and two solvents to survey Pd-catalyzed cross-coupling reactions. When we used the first three principal components of each feature set to conduct *k*-means clustering for the phosphines and phosphine oxides

https://doi.org/10.1021/acs.oprd.1c00357
Org. Process Res. Dev. XXXX, XXX, XXX−XXX

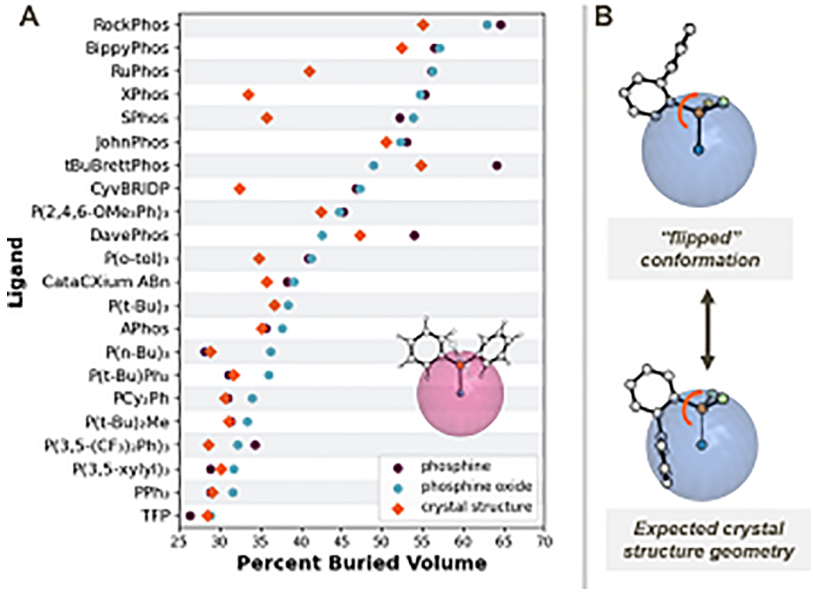


Figure 3. (A) Differences in Boltzmann-averaged $\%V_{bur}$ across phosphine oxides, phosphines, and metal−ligand crystal structures. (B) Flipped vs expected conformations of phosphine ligands.

separately,[14] we found six ligands closest to the cluster centroids in common (Figure 4).

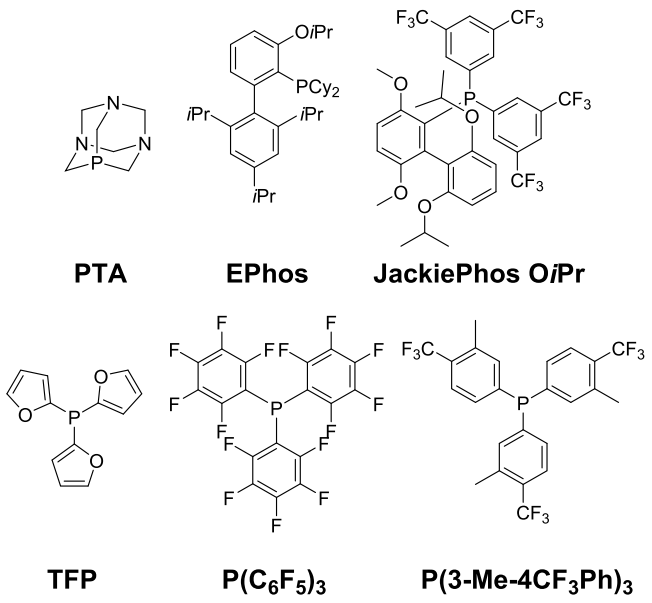


Figure 4. Ligands identified from separate *k*-means clustering of phosphines and phosphine oxides that are shared between the two *k*-means cluster sets.

Since the majority of these shared ligands are not among the most commercially available options on a scale required by pharmaceutical process chemists, we manually substituted many of them when selecting the larger set by choosing alternatives from the same dendrogram branch when possible (Figure 5). For example, we replaced JackiePhos OiPr with JackiePhos. In our manual selection from across each dendrogram, we also considered the perceived prevalence of the ligands in literature reports.

Overall, 70% of our independent selections from the phosphine oxide and phosphine dendrograms were in common, demonstrating that the two computational methods that we examined produced comparable representations of ligand space when used to identify sterically and electronically diverse ligands. Although exploring ligand space from each dendrogram can produce similar selections, focusing on smaller dendrogram branches reveals distinct differences between our maps of ligand space. These local changes in arrangement could impact a chemist's experimental design when seeking to exploit steric and electronic features that previously led to a successful reaction. An analogous comparison using a dendrogram created from the "*kraken*" features[3a] resulted in the independent selection of a ligand set similar to those chosen in Figure 5 (see the Supporting Information for further details). While a quantitative assessment of the impact of ligand arrangement on experimental design is out of the scope of this work, this qualitative analysis enables us to guide our colleagues when designing and analyzing cross-coupling experiments.

APPLICATION OF FEATURES TO HTE DATA: CHEMOSELECTIVE MIYAURA BORYLATION

While global trends in our maps of ligand space led to the selection of similar ligands, we also wanted to understand how models of HTE reaction outcomes might be influenced by different computational featurization methods. We envisioned using such models to help our chemists understand how ligand features influenced reactivity and selectivity to better inform their design of subsequent experiments. A chemoselective Miyaura borylation previously developed by our colleagues[15] proved to be an apt case study for proof of concept; HTE outcomes revealed that the yield of the product (Scheme 1) significantly varied as a function of the phosphine ligand identity.

To determine the potential impact of ligand descriptors on the reaction outcome, we first compared the desired product yield to each of our computed features. Similar to recent work exploring the implications of single-node decision trees to $\%V_{bur}$ (and other phosphine descriptors) within the context of nickel and palladium catalysis,[10] a plot of product yield vs $\%V_{bur}$ revealed a distinct reactivity cliff between ligands over or under approximately 36% V_{bur} (Figure 6A). This cliff was observed

https://doi.org/10.1021/acs.oprd.1c00357
Org. Process Res. Dev. XXXX, XXX, XXX−XXX

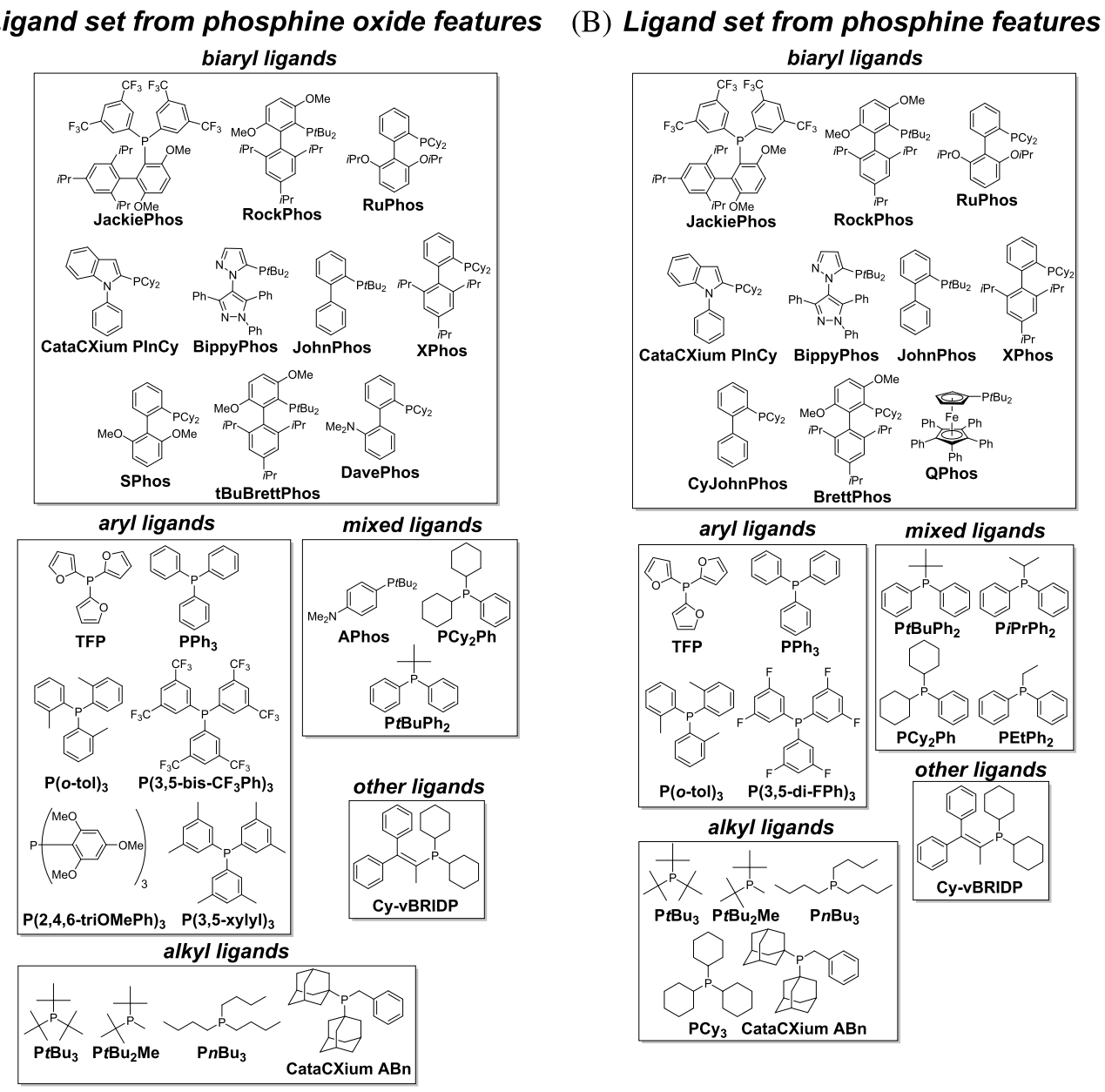


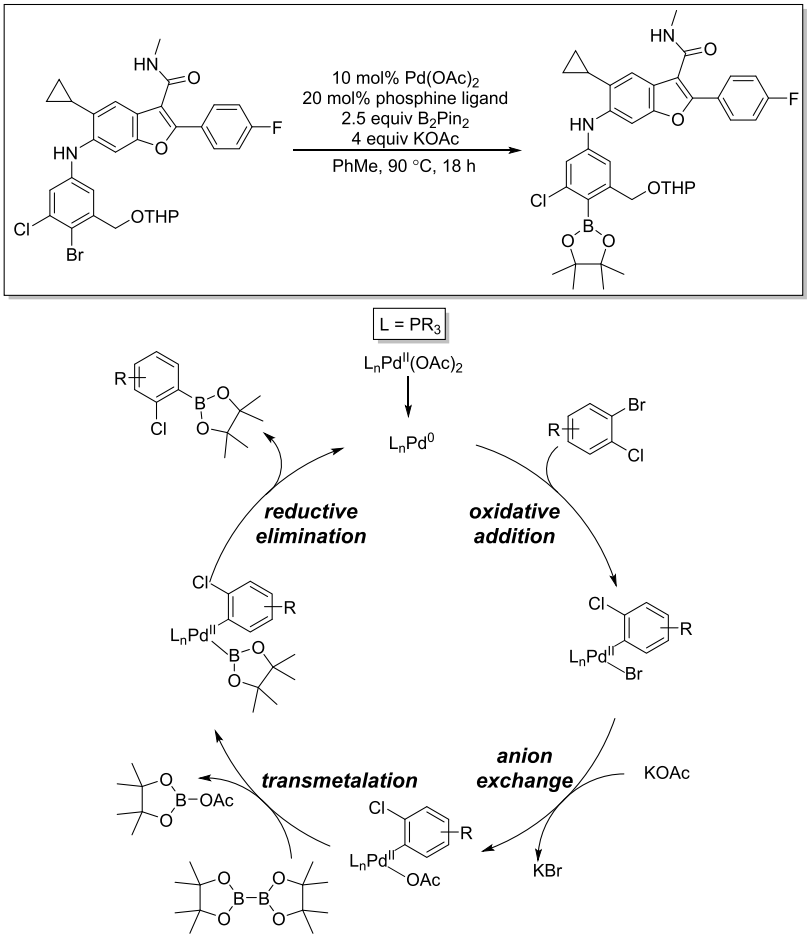
Figure 5. (A) Ligand set selected from phosphine oxide dendrogram. (B) Ligand set selected from phosphine dendrogram.

in plots with either the phosphine oxide or the phosphine features. This result led us to consider the possible role of Pd speciation in the formation of the borylation product.

To gain more insight into key phosphine features beyond this steric-driven reactivity classification, we first initiated statistical modeling to determine whether the yield is related to multiple computed molecular descriptors.[17] However, preliminary modeling efforts to account for all available data points were unsuccessful. Consequently, examples wherein the palladium complex did not complete one turnover were removed, as this poor activity could be attributed to either on- or off-cycle failure. Additionally, the classification process wherein six of these seven removed data points were above the defined reactivity threshold of 36% V_{bur} supported this curation step. Ultimately, two statistical models were developed for the yield using the phosphine oxide descriptors. A bivariate model (Figure 6B, validation R^2 = 0.77, Q^2 = 0.76, *k*-fold = 0.74) incorporates the Boltzmann average of the Fukui nucleophilicity index (Boltz f_{-P}) and CM5 partial charge on phosphorus (Boltz P CM5) from the phosphine oxide.[18,19] Both the f_{-P} function and CM5 represent phosphine σ-donor capacity, and the signs of both terms indicate that phosphines with moderate σ-donor capacity are optimal. Larger positive f_{-P} values coupled with a negative coefficient (−0.31) reflect that excellent σ-donor ligands result in diminished reaction performance. In contrast, a large phosphorus CM5 charge coupled with a negative coefficient (−0.26) indicates the importance of good σ-donors; a high CM5 value is indicative of less electron-rich phosphines. Our colleagues' analysis, however, also suggested that the aryl bromide bond is sterically encumbered compared to the adjacent aryl chloride bond, necessitating a relatively unhindered Pd(0) species to promote oxidative addition. Indeed, similar statistics and improved interpretability can be obtained with the inclusion of a third term (validation R^2 = 0.80, Q^2 = 0.74, *k*-fold = 0.71), the maximum value of $\%V_{bur}$ (minimum quadrant) (see Supporting Information Figure S2). This term

https://doi.org/10.1021/acs.oprd.1c00357
Org. Process Res. Dev. XXXX, XXX, XXX−XXX

Scheme 1. Miyaura Borylation Reaction Conditions and General Mechanism[16]



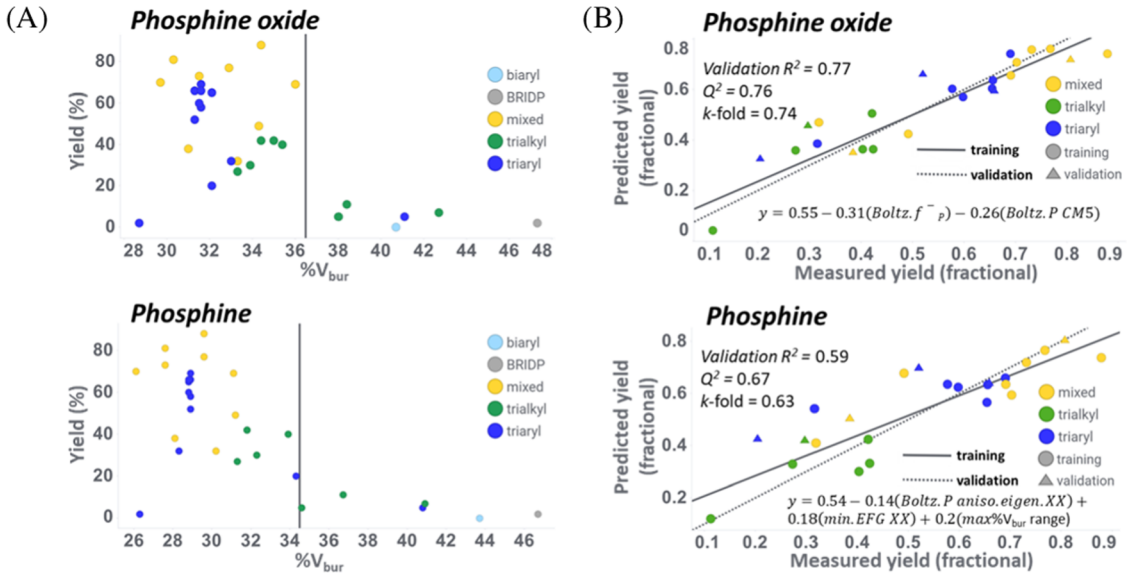


Figure 6. (A) Univariate relationship of phosphine and phosphine oxide $\%V_{bur}$ to the yield of the desired product. (B) Multivariate relationship between the yield and ligand features using phosphine and phosphine oxide feature sets.

reflects the ability of a ligand to provide at least one free quadrant in the coordination sphere to accommodate additional ligands at the metal. This inclusion and the cutoff observed in the threshold analysis emphasize the potential importance of complex speciation in this particular Miyaura borylation; a mono- or bis-ligated Pd species may be required to facilitate different steps in the catalytic cycle. The best multivariate linear regression model identified when employing features from the phosphine ligands, however, is less robust than the ones identified from the phosphine oxide structures. The three

https://doi.org/10.1021/acs.oprd.1c00357
Org. Process Res. Dev. XXXX, XXX, XXX−XXX

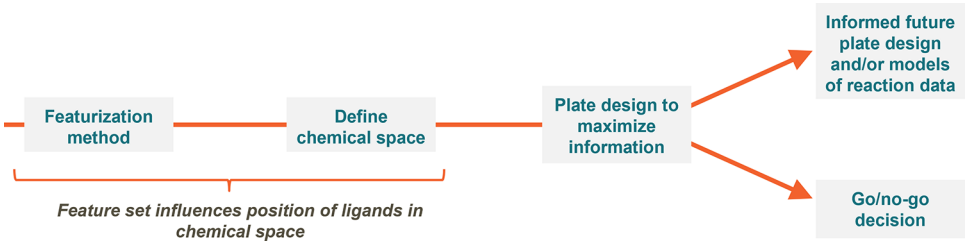


Figure 7. Process for implementing a computationally informed workflow for reaction optimization.

parameters included in our model generated from phosphine features are the Boltzmann-weighted average of the ^{31}P chemical shift anisotropy (11 eigenvalues at phosphorus (Boltz aniso. Eigen. XX)), the minimum value of the smallest eigenvalue of the electric field gradient (EFG) tensor (min. EFG XX), and the buried volume range (max. %V_{bur} range).[20] Compared to both models generated from phosphine oxide features (Figures 6B and S2), this model contains slightly inferior statistics for each measure (validation R^2 = 0.59, Q^2 = 0.67, k-fold = 0.63). Moreover, although the inclusion of a buried volume term again highlights speciation, it is less clear what phosphorus chemical shift anisotropy and EFG tensor value may imply. The magnitude and sign of the EFG tensor, as it describes the local electric fields at a given nucleus, has previously been rationalized by comparison to computed atomic orbitals.[21] Similarly, Copéret and co-workers have shown through natural chemical shift analysis that chemical shift tensors can be linked to frontier molecular orbitals in the context of metathesis.[22] Together, these terms are more challenging to relate to key characteristics of the phosphine ligands that promote high product yield. While models with interpretable features may be intellectually satisfying, models with less interpretable features may still be useful to predict the performance of additional phosphines. We believe that both performance prediction and interpretability have distinct advantages; predictive models may enable virtual ligand screening to save time and material, while interpretability may provide insight into mechanistic factors driving reactivity or selectivity. At this point, these models do not fully answer whether employing phosphine- or phosphine oxide-derived molecular features is more beneficial in a global sense. Nevertheless, the differences in these model terms show that there is a potential impact on experimental prioritization; one model may lead a scientist to choose different ligands (e.g., more expensive, less available on a large scale) for subsequent experiments than another, leading to different experimental outcomes.[23]

CONSIDERATIONS FOR THE IMPLEMENTATION OF PHOSPHINE FEATURES IN EXPERIMENTAL DESIGN AND ANALYSIS

Recent work on phosphine featurization[2,3,10,24] offers industrial chemists the opportunity to both streamline experimental design and maximize information gained from each experiment. As we have shown, integration of any reported tools necessitates thoughtful consideration of experimental goals and the potential impact of the computational method on scientific choices (Figure 7). Based on a chemist's goal, we recommend using ligand features to accomplish any or all of the following:

(i) design of an HTE plate that incorporates ligands that occupy diverse regions of chemical space or a specific region based on prior experimental knowledge and
(ii) development of regression models that inform mechanism-driven hypotheses and/or predict ligand performance through analysis of the descriptors contained in the top model.

As we have demonstrated, the choice of computational method has an impact on ligand conformation, ligand feature values, and the makeup of ligand clusters in maps. By incorporating computed features into a hierarchical map of chemical space and highlighting a set of diverse ligands for general purpose screening in the absence of prior knowledge, we have enabled the incorporation of two discrete strategies into process development. First, an experienced process chemist can supplement prior knowledge of likely ligand performance with monophosphine ligands that occupy different regions of chemical space using the dendrogram. Mapping the aforementioned Miyaura borylation ligands to the dendrograms (Figure 2 ligands with asterisks, vide supra) reveals that the ligands in our colleagues' study primarily occupy one region in the dendrogram. If a high-performing ligand had not been identified from that HTE plate, these dendrograms would enable a chemist to quickly and strategically target new regions of chemical space by exploring alternate branches. Next, by choosing diverse ligands, further statistical modeling is feasible. The adoption of one or both of these strategies can maximize the information derived from a single plate through the selection of key monophosphine ligands (supplementing other experimental variables such as solvent, base, or precatalyst), leading to a faster decision to pursue or abandon a synthetic strategy.

CONCLUSIONS AND OUTLOOK

Although statistical modeling tools have helped to accelerate reaction optimization and enhance mechanistic understanding, the specific demands of process development warranted closer analysis of the holistic workflow first reported by the academic community. While developing this workflow to guide experimental design in process development, we have shown that the choice of which molecule to compute and the subsequent featurization influences chemical space, redefining which phosphine ligands are neighbors. This direct comparison of the maps of chemical space and the selection of linear regression models warrants further exploration to understand the practical implications of seemingly modest computational choices. To democratize a computationally driven workflow, it is necessary to design a system that can be adapted to include expert knowledge. By computing phosphine ligand descriptors and developing a map of this feature space, we have implemented a strategy to transform the optimization of Pd-catalyzed cross-coupling reactions in our labs. Although this is a start, a broader comparison of models obtained from HTE data is necessary to conclude if any set of ligand features leads to superior models by

https://doi.org/10.1021/acs.oprd.1c00357
Org. Process Res. Dev. XXXX, XXX, XXX−XXX

a variety of statistical measures. Based on our own subjective criteria, the phosphine oxide descriptors provided a more interpretable linear regression model. However, different criteria could be reasonably employed throughout this workflow, wherein the phosphine representation could be preferred. Future decisions to employ a specific feature set in HTE design efforts and/or modeling of reaction outcomes may depend on reaction context. Since the identification of an optimal ligand feature set requires one to obtain reaction outcomes and regression models, any set that provides a realistic feature space serves equally well for exploratory experimental design. Regardless of which descriptor set is more effective for a particular reaction, building a platform amenable to chemists with diverse experience requires evaluating choices within the workflow that could have consequences for reaction development decisions. The specific constraints of process chemistry (e.g., scalability and availability of raw materials) offer an exciting opportunity to develop creative and innovative approaches using data science-derived tools. The flexibility of our approach suggests that computation can be harnessed for data-rich experimentation, facilitating experimental design and the identification of key connections between data.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.oprd.1c00357.

Computational and statistical methods; list of crystal structures for Figure 3 (PDF)

List of computed feature values for phosphines and phosphine oxides and XYZ coordinates for ligands in Figure 5 (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Matthew S. Sigman − *Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States;* orcid.org/0000-0002-5746-8830; Email: sigman@chem.utah.edu

Jennifer M. Elward − *Molecular Design, GlaxoSmithKline, Collegeville, Pennsylvania 19426, United States;* orcid.org/0000-0002-3786-3411; Email: jennifer.m.elward@gsk.com

Janelle E. Steves − *Chemical Development, GlaxoSmithKline, Collegeville, Pennsylvania 19426, United States;* orcid.org/0000-0002-0664-7556; Email: janelle.e.steves@gsk.com

Authors

Jennifer M. Crawford − *Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States;* Present Address: Chemical Development, GlaxoSmithKline, 1250 S. Collegeville Road, Collegeville, Pennsylvania 19460, United States.; orcid.org/0000-0003-2936-4038

Tobias Gensch − *Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States;* Present Address: Department of Chemistry, TU Berlin, Straße des 17. Juni 135, Sekr. C2, 10623 Berlin, Germany; orcid.org/0000-0002-1937-0285

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.oprd.1c00357

Author Contributions

The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare the following competing financial interest(s): J.M.C., J.M.E., and J.E.S. were employees of GlaxoSmithKline while conducting the work described in this manuscript. GlaxoSmithKline participated in the review and approval of this manuscript.

■ ACKNOWLEDGMENTS

The authors thank GlaxoSmithKline for support during the preparation of this manuscript. The authors thank Rebecca Colandrea for computing cone angles with the program, Solid-G.[25] The authors also thank Dr. Jing-Yao Guo for her detailed tutorial on conducting linear regression modeling on our data set. The authors are grateful to Dr. Doo-Hyun Kwon for reviewing the manuscript and to Dr. Damian Hruszkewycz and Greg Stockdale for helpful discussions. J.M.C. acknowledges early support from the NSF-GRFP. M.S.S. would like to thank the NSF under the CCI Center for Computer Assisted Synthesis (CHE-1925607) for additional support. T.G. thanks the Leopoldina Fellowship Programme of the German National Academy of Sciences Leopoldina (LPDS 2017-18). The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged.

■ REFERENCES

(1) For selected perspectives on HTE in pharmaceutical process chemistry, see: (a) Krska, S. W.; DiRocco, D. A.; Dreher, S. D.; Shevlin, M. The evolution of chemical high-throughput experimentation to address challenging problems in pharmaceutical synthesis. *Acc. Chem. Res.* **2017**, *50*, 2976−2985. (b) Selekman, J. A.; Qiu, J.; Tran, K.; Stevens, J.; Rosso, V.; Simmons, E.; Xiao, Y.; Janey, J. High-Throughput Automation in Chemical Process Development. *Annu. Rev. Chem. Biomol. Eng.* **2017**, *8*, 525−547. (c) Mennen, S. M.; Alhambra, C.; Allen, C. L.; Barberis, M.; Berritt, S.; Brandt, T. A.; Campbell, A. D.; Castañón, J.; Cherney, A. H.; Christensen, M.; Damon, D. B.; Eugenio de Diego, J.; García-Cerrada, S.; García-Losada, P.; Haro, R.; Janey, J.; Leitch, D. C.; Li, L.; Liu, F.; Lobben, P. C.; MacMillan, D. W. C.; Magano, J.; McInturff, E.; Monfette, S.; Post, R. J.; Schultz, D.; Sitter, B. J.; Stevens, J. M.; Strambeanu, I. I.; Twilton, J.; Wang, K.; Zajac, M. A. The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. *Org. Process Res. Dev.* **2019**, *23*, 1213−1242.

(2) (a) Fey, N.; Tsipis, A. C.; Harris, S. E.; Harvey, J. N.; Orpen, A. G.; Mansson, R. A. Development of a ligand knowledge base, part 1: computational descriptors for phosphorus donor ligands. *Chem. - Eur. J.* **2006**, *12*, 291−302. (b) 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*, 6245−6258. (c) Niemeyer, Z. L.; Milo, A.; Hickey, D. P.; Sigman, M. S. Parameterization of phosphine ligands reveals mechanistic pathways and predicts reaction outcomes. *Nat. Chem.* **2016**, *8*, 610−617. (d) Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R. Enantiodivergent Pd-catalyzed C-C bond formation enabled through ligand parameterization. *Science* **2018**, *362*, 670−674. (e) Durand, D. J.; Fey, N. Computational ligand descriptors for catalyst design. *Chem. Rev.* **2019**, *119*, 6561−6594.

(3) (a) 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. *J. Am. Chem. Soc.* **2022**,

144, 1205−1217. (b) Gensch, T.; Smith, S.; Colacot, T.; Timsina, Y.; Xu, G.; Sigman, M. S. Design and Application of a Training Set for Monophosphine Ligands in Metal Catalysis *ChemRxiv* DOI: 10.33774/chemrxiv-2021-fgm7v. (c) 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*, No. 112.

(4) For references describing use of RMSD as a metric to assess molecular geometry differences, see: (a) Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. *J. Chem. Inf. Model.* **2017**, *57*, 2719. (b) Lim, V. T.; Hahn, D. F.; Tresadern, G.; Bayly, C. I.; Mobley, D. Benchmark Assessment of Molecular Geometries and Energies from Small Molecule Force Fields. *ChemRxiv* **2020**, *9*, No. 1390. (c) Ehrman, J.; Lim, V. T.; Bannan, C. C.; Thi, N.; Kyu, D.; Mobley, D. Improving Small Molecule Force Fields By Identifying and Characterizing Small Molecules with Inconsistent Parameters. *J. Comput.-Aided Mol. Des.* **2021**, *35*, 271−284.

(5) (a) *JMP*, Versions 14.2.0 and 15.2.0; SAS Institute Inc.: Cary, NC, 1989−2021. (b) *Spotfire*, Version 10.10.2.53; TIBCO Software Inc.: Palo Alto, CA.

(6) See Supporting Information for details on software and methods used to create dendrogram plots.

(7) For the initial development of percent buried volume, see: Hillier, A. C.; Sommer, W. J.; Yong, B. S.; Peterson, J. L.; Cavallo, L.; Nolan, S. P. A Combined Experimental and Theoretical Study Examining the Binding of N-Heterocyclic Carbenes (NHC) to the Cp*RuCl (Cp* = η5-C5Me5) Moiety: Insight into Stereoelectronic Differences between Unsaturated and Saturated NHC Ligands. *Organometallics* **2003**, *22*, 4322−4326.

(8) For a description of the software used to compute percent buried volume, see: Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L. SambVca 2. A Web Tool for Analyzing Catalytic Pockets with Topographical Steric Maps. *Organometallics* **2016**, *35*, 2286−2293.

(9) See ref 2d for a description of how to compute V_{min} and the following reference for a description of how V_{min} is used to quantify electronics of phosphines: Suresh, C. H.; Koga, N. Quantifying the Electronic Effect of Substituted Phosphine Ligands via Molecular Electrostatic Potential. *Inorg. Chem.* **2002**, *41*, 1573−1578.

(10) Newman-Stonebraker, S.; Smith, S.; Borowski, J.; Peters, E.; Gensch, T.; Johnson, H.; Sigman, M.; Doyle, A. Linking Mechanistic Analysis of Catalytic Reactivity Cliffs to Ligand Classification. *Science* **2021**, *374*, 301−308.

(11) We were unable to locate a crystal structure of JackiePhos as a Pd complex, and thus removed it from this analysis.

(12) (a) Barder, T. E.; Buchwald, S. L. Rationale Behind the Resistance of Dialkylbiaryl Phosphines toward Oxidation by Molecular Oxygen. *J. Am. Chem. Soc.* **2007**, *129*, 5096−5101. (b) Ikawa, T.; Barder, T. E.; Biscoe, M. R.; Buchwald, S. L.Pd-Catalzyed Amidations of Aryl Chlorides Using Monodentate Biaryl Phosphine Ligands: A Kinetic, Computational, and Synthetic Investigation. *J. Am. Chem. Soc.* **2007**, *129*, 13001−13007.

(13) Gioria, E.; del Pozo, J.; Martínez-Ilarduya, J. M.; Espinet, P. Promoting Difficult Carbon-Carbon Couplings: Which Ligand Does Best? *Angew. Chem., Int. Ed.* **2016**, *55*, 13276−13280.

(14) See Supporting Information for details on software and methods used to conduct *k*-means clustering.

(15) Arrington, K.; Barcan, G. A.; Calandra, N. A.; Erickson, G. A.; Li, L.; Liu, L.; Nilson, M. G.; Strambeanu, I. I.; VanGelder, K. F.; Woodard, J. L.; Xie, S.; Allen, C. L.; Kowalski, J. A.; Leitch, D. C. Convergent synthesis of the NS5B inhibitor GSK8175 enabled by transition metal catalysis. *J. Org. Chem.* **2019**, *84*, 4680−4694.

(16) For the initial report of Pd-catalyzed borylation and general mechanism, see: Ishiyama, T.; Murata, M.; Miyaura, N. Palladium(0)-Catalyzed Cross-Coupling reaction of Alkoxydiboron with Haloarenes: A Direct Procedure for Arylboronic Esters. *J. Org. Chem.* **1995**, *60*, 7508−7510.

(17) For references describing the multivariate linear regression process, see: (a) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **2018**, *9*, 2398−2412. (b) Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A. The Development of Multidimensional Analysis tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **2016**, *49*, 1292−1301. (c) Reid, J. P.; Sigman, M. S. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nat. Rev. Chem.* **2018**, *2*, 290−305.

(18) For a reference describing application of Fukui functions to phosphines, see: (a) Morales-Bayuelo, A.; Caballero, J. New Insights into Steric and Electronic Effects in a Series of Phosphine Ligands from the Perspective of Local Quantum Similarity using the Fukui Function. *J. Mol. Model.* **2015**, *21*, No. 45.

(19) For computation of CM5 charges, see: Marenich, A. V.; Jerome, S. V.; Cramer, C. J.; Truhlar, D. G. Charge Model 5: An Extension of Hirshfeld Population analysis for the Accurate Description of Molecular Interactions in Gaseous and Condensed Phases. *J. Chem. Theory Comput.* **2012**, *8*, 527−541.

(20) Buried volume range is represented by the difference between the maximum and minimum buried volume quadrants divided by the total buried volume.

(21) Autschbach, J.; Zheng, S.; Schurko, R. W. Analysis of electric field gradient tesnors at quadrupolar nuclei in common structural motifs. *Concepts Magn. Reson., Part A* **2010**, *36A*, 84−126.

(22) (a) Gordon, C. P.; Shirase, S.; Yamamoto, K.; Andersen, R. A.; Eisenstein, O.; Copéret, C. NMR chemical shift analysis decodes olefin oligo-and polymerization activity of d0 group 4 metal complexes. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E5867−E5876. (b) Engl, P. S.; Santiago, C. B.; Gordon, C. P.; Liao, W. C.; Fedorov, A.; Copéret, C.; Sigman, M. S.; Togni, A. Exploiting and understanding the selectivity of Ru-N-heterocyclic carbene metathesis catalysts for the ethenolysis of cyclic olefins to α, ω-dienes. *J. Am. Chem. Soc.* **2017**, *139*, 13117−13125.

(23) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **2019**, *9*, 2313−2323.

(24) For a recent example of industrial implementation of the phosphine features described in ref 3a, see: Zell, D.; Kingston, C.; Jermaks, J.; Smith, S. R.; Seeger, N.; Wassmer, J.; Sirois, L. E.; Han, C.; Zhang, H.; Sigman, M. S.; Gosselin, F. Stereoconvergent and -divergent Synthesis of Tetrasubstituted Alkenes by Nickel-Catalyzed Cross-Couplings. *J. Am. Chem. Soc.* **2021**, *143*, 19078−19090.

(25) Guzei, I. A.; Wendt, M. *Dalton Trans.* **2006**, *128*, 3991−3999.

https://doi.org/10.1021/acs.oprd.1c00357
Org. Process Res. Dev. XXXX, XXX, XXX−XXX