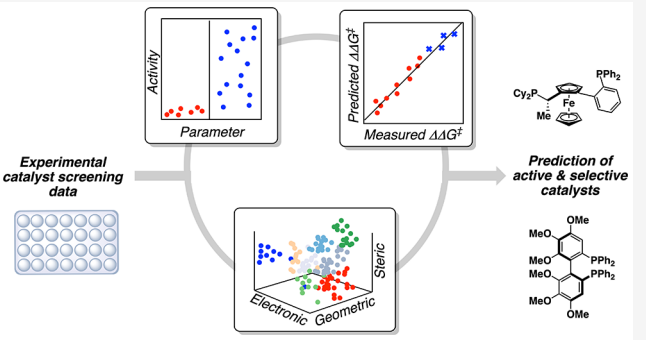
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

Data-Driven Multi-Objective Optimization Tactics for Catalytic Asymmetric Reactions Using Bisphosphine Ligands

Jordan J. Dotson,[‖] Lucy van Dijk,[‖] Jacob C. Timmerman, Samantha Grosslight, Richard C. Walroth, Francis Gosselin, Kurt Püntener, Kyle A. Mack,* and Matthew S. Sigman*

Cite This: J. Am. Chem. Soc. 2023, 145, 110−121

ABSTRACT: Optimization of the catalyst structure to simultaneously improve multiple reaction objectives (e.g., yield, enantioselectivity, and regioselectivity) remains a formidable challenge. Herein, we describe a machine learning workflow for the multi-objective optimization of catalytic reactions that employ chiral bisphosphine ligands. This was demonstrated through the optimization of two sequential reactions required in the asymmetric synthesis of an active pharmaceutical ingredient. To accomplish this, a density functional theory-derived database of >550 bisphosphine ligands was constructed, and a designer chemical space mapping technique was established. The protocol used classification methods to identify active catalysts, followed by linear regression to model reaction selectivity. This led to the prediction and validation of significantly improved ligands for all reaction outputs, suggesting a general strategy that can be readily implemented for reaction optimizations where performance is controlled by bisphosphine ligands.



INTRODUCTION

Transition-metal-catalyzed organic reactions are essential to the synthesis of pharmaceutical, agrochemical, and fine chemical ingredients. For many such reactions, performance is controlled by the structural features of the auxiliary ligand attached to the transition metal. The relationship between the chemical structure of the ligand and the behavior of the catalyst is typically difficult to decipher intuitively.[1−4] Consequently, reaction optimization is a resource-intensive, trial-and-error-based campaign. Data science tools have recently emerged to streamline this effort through the development of methods to relate the chemical structure to function.[5−7] In particular, the calculation of descriptors has evolved to incorporate diverse, refined properties,[8,9] which can be scaled to produce databases.[10−12] The resulting descriptors can then be fed into a range of data science workflows to optimize a particular objective such as the yield, reaction rate, regioselectivity, or stereoselectivity. These workflows include machine learning (ML) algorithms such as multivariate linear regression (MLR) to regress the structure to function,[13] classification tools to explore reactivity cliffs,[14] and dimensionality reduction techniques to map the chemical space to visualize reactivity patterns.[9,15−17] More sophisticated algorithms can also be employed to explore the multi-dimensionality of reaction optimization including the exploitation of Bayesian optimization tools.[18−20] However, a limited effort has thus far been reported for the simultaneous optimization of multiple objectives in homogeneous catalysis.[21−29] This is an underlying challenge in organic chemistry as high performance in one objective (e.g., yield) does not necessarily correlate with the desired performance in another (e.g., stereoselectivity).

The challenge of multidimensional optimization is exemplified in catalytic asymmetric reactions, and the most impactful ligand class in this reaction category is undoubtedly bidentate organophosphorus ligands (Figure 1A). This is highlighted by the >50 year evolution of chiral bisphosphines as privileged ligands in asymmetric catalysis. This effort has led to an exceptional range of ligand scaffolds that have had a tremendous impact on the synthesis of molecules for the treatment of human disease. Following seminal examples such as DIPAMP[30] and DIOP,[31] diverse structural motifs have been introduced that include axial chirality and ferrocenyl backbones as well as phospholane rings as seen in BINAP,[32] Josiphos,[33] and DuPhos,[34] respectively. Additionally, various C_2-symmetric biaryl ligands such as MeO-BIPHEP[35] and Segphos[36] have seen widespread application. The evolution of ever more structurally complex motifs has paralleled the vast utilization of

Received: August 10, 2022
Published: December 27, 2022

https://doi.org/10.1021/jacs.2c08513
J. Am. Chem. Soc. 2023, 145, 110−121

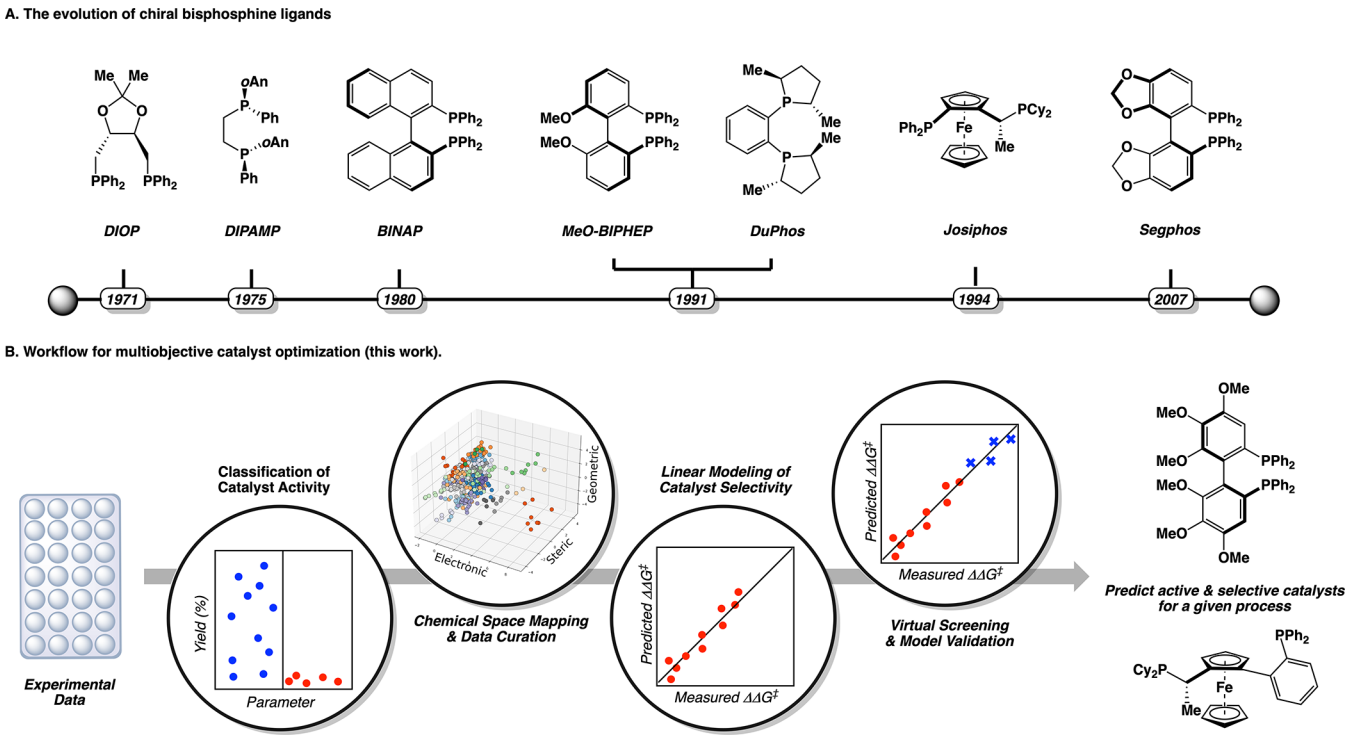


Figure 1. (A) The evolution of chiral bisphosphines has led to the development of numerous high-performance ligands with diverse structural motifs. (B) We report a multi-objective workflow for catalyst optimization. This consists of initial experimental data collection followed by the combined use of classification and linear regression supervised ML models, chemical space-based data curation, and virtual screening to identify ligands with superior performance.

these ligands in diverse reactions operating with unique metal–ligand coordination structures. However, the myriad of accessible structures is also a challenge for optimization campaigns where one may find the requisite reactivity but not high stereoselectivity (or vice versa).

How one navigates this traditionally difficult multi-optimization problem has inspired us to develop a workflow to synchronously improve the yield and selectivity by sequentially using ML tools to analyze experimental data (Figure 1B). The workflow consists of three key stages: (1) use of classification algorithms to assess reactivity (i.e., conversion or yield), (2) MLR to evaluate selectivity (regio- and/or stereoselectivity), and (3) virtual ligand screening to predict both reactivity and enhanced selectivity. Ultimately, this process de-risks subsequent experimental testing of extrapolations by modeling multiple objectives[37] and through the use of a chemical space analysis technique to understand the domain of applicability for a given model. This study encompasses the culmination of multiple data science tools and molecular featurization techniques that have been developed by our group and others.

We showcase this workflow in the optimization of two consecutive steps on the route to the synthesis of an active pharmaceutical ingredient (API) for the treatment of asthma:[38] a Pd-catalyzed Hayashi–Heck reaction[39,40] and a Rh-catalyzed alkene hydroformylation reaction. The successful implementation of this strategy was enabled by the construction of a density functional theory (DFT)-derived descriptor database of >550 bisphosphine ligands and the development of an interpretable chemical space mapping technique. Stringent purity requirements for APIs and the high cost of chromatographic purification on a multi-kilogram scale demand that catalysts used in process chemistry routes impart exquisite chemo-, stereo-, and regioselectivity coupled to a high-yielding reaction. This endeavor succeeded in identifying ligands that furnished a simultaneous and significant improvement of all objectives—yield, stereoselectivity, and regioselectivity—to improve the scalability of the current route to the API. More generally, we showcase how the multi-objective workflow can be successfully applied to reactions using different transition metals that require distinct bisphosphine ligands to be effective. This large virtual library and the requisite data processing scripts developed for this study have been made publicly available, significantly enabling practitioners in the academia and industry to shorten timelines for catalyst optimization campaigns where performance is controlled by bisphosphine ligands.

RESULTS AND DISCUSSION

Bisphosphine Virtual Library Construction and Visualization. As our catalyst optimization would rely on calculated ligand features, a comprehensive descriptor library was constructed incorporating bisphosphine ligands that are commercially available or prevalent in the literature, along with synthetically accessible derivatives (>550 ligands including >200 that are commercially available). This database was inspired by seminal contributions from Fey and co-workers.[41,42] We set out to construct a bisphosphine descriptor library that would be applicable to reactions with varying ligand coordination environments. The resulting parameters needed to be comparable across bisphosphine ligand scaffolds with varying symmetries (C_1, C_2, etc.) so that disparate bisphosphine ligands could be meaningfully compared. Additionally, ligand enantiomers would require specific parameters to allow prediction of their performance on chiral substrates.

https://doi.org/10.1021/jacs.2c08513
J. Am. Chem. Soc. 2023, 145, 110–121

With these challenges in mind, quantum mechanical methods were used to calculate the geometries and descriptors for a wide range of bidentate organophosphorus ligands using a square planar [ligand]$PdCl_2$ complex as the model system (Figure 2A).[43] We have previously used a similar strategy with pyridine–oxazoline ligands and found that parameters could be effective in models for reactions that use other transition metals.[44] The computational workflow included a molecular mechanics-based method to generate conformations of the model complex, which were subsequently optimized using DFT. Steric, electronic, and geometric parameters such as those shown in Figure 2B were collected from the DFT-optimized structures of both the Pd complex and the ligand without $PdCl_2$.[45,46]

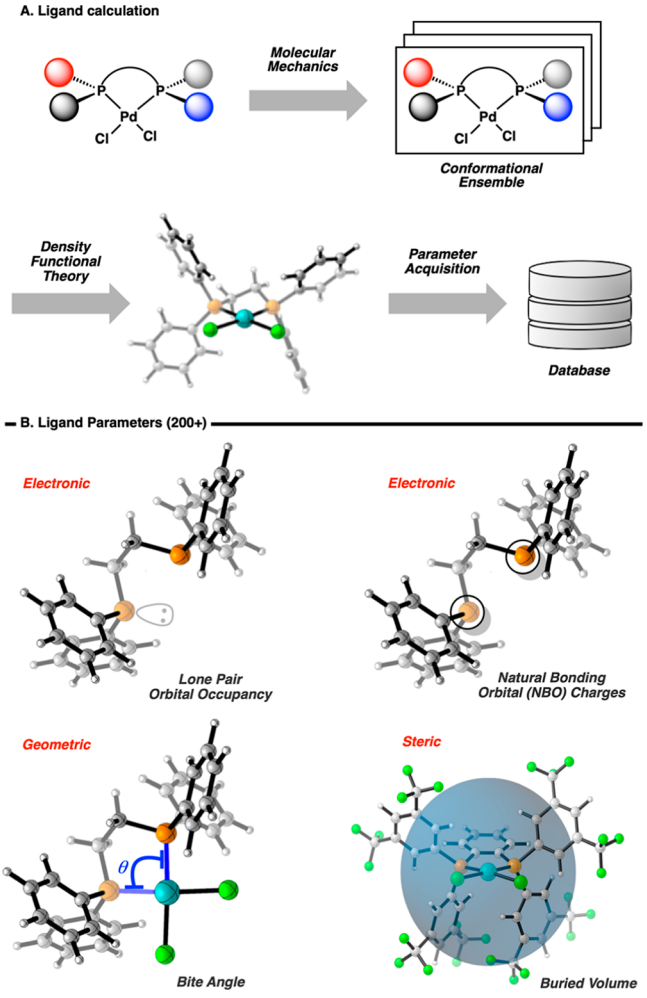


Figure 2. (A) Workflow for the calculation of ligand parameters. Conformational ensembles for the [ligand]$PdCl_2$ complexes were generated using molecular mechanics. Selected conformers were then optimized using DFT, and parameters were obtained from the resultant structures. (B) Selection of calculated ligand parameters. Parameters can be categorized as being electronic, geometric, and steric in nature.

Of particular importance, parameters were required to reflect quadrant-specific descriptors for ligands containing different symmetry elements. To illustrate this, one may consider the percent buried volume (V_{bur}) of the northwestern quadrant[47−49] for a C_{2v} symmetric ligand such as Xantphos (Figure 3). For these symmetric ligands, all four quadrants are symmetry equivalent. However, the optimized ground-state geometries are often significantly distorted out of C_{2v} symmetry, and as a result,

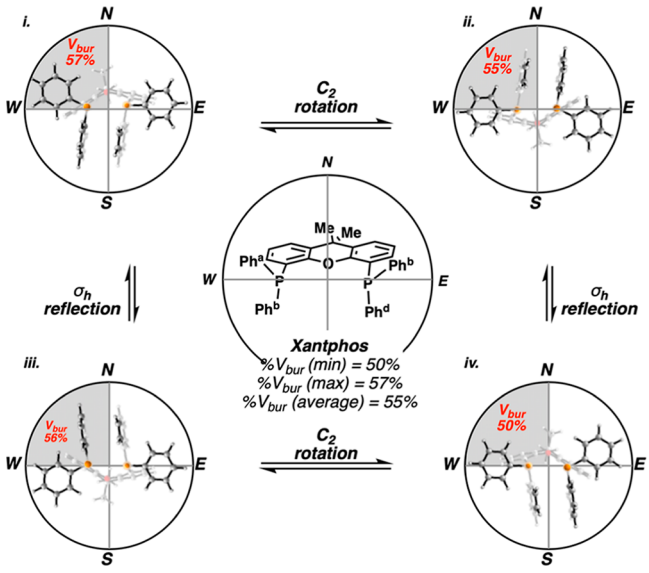


Figure 3. Symmetry-adapted parameters were required to effectively compare disparate ligand backbones. This process is illustrated for the quadrant-specific percent V_{bur} of a C_{2v} symmetric ligand.

the quadrant buried volume parameters are not equivalent. This apparent contradiction of symmetry-equivalent quadrants with markedly different parameter values can be understood by considering the ligand's conformational equilibrium—applying the two symmetry operations of the C_{2v} point group (a C_2 rotation and a σ_h reflection) produces degenerate conformers of Xantphos (Figure 3).[50] As each quadrant can access four different degenerate conformations, we report the minimum, maximum, and average values of the symmetry-equivalent quadrant-specific parameters. Analogous symmetry considerations were applied to ligands of the C_s and C_2 point groups (see the Supporting Information for further discussion). This parameterization process enabled the effective comparison of disparate ligand backbones.[51]

While the parameter library was computed using [ligand]-$PdCl_2$ complexes, its applicability to non-Pd-catalyzed reactions warrants further reflection. Given the significant computational cost of re-calculating this library for each transition metal commonly used in homogeneous catalysis, this $PdCl_2$-based library is intended to be generally applicable. We posit that this will be the case for a given reaction if the following two criteria are met. First, any errors in the parameter values (i.e., differences between the $PdCl_2$ complex and the catalytically active species) are systematic. Second, if a parameter is not reflective of the catalytically active species (i.e., has frequent non-systematic errors), it will not be correlated with the observed reaction outcome.[52]

After descriptor library construction, we turned our attention to the visualization of the ligand parameter space (Figure 4). For many virtual molecular libraries, chemical space representations are generated using dimensionality reduction techniques such as principal component analysis (PCA).[9,10,40,41,53] This results in a graphical depiction of the property space in which the proximity of ligands reflects their similarity. A potential drawback of such dimensionality-reduced maps, especially those encompassing numerous descriptors, is that their axes can be inherently challenging to interpret, as is the case for our bisphosphine descriptor library. Therefore, we set out to develop a chemical space representation that would include more readily interpret-

https://doi.org/10.1021/jacs.2c08513
J. Am. Chem. Soc. 2023, 145, 110−121

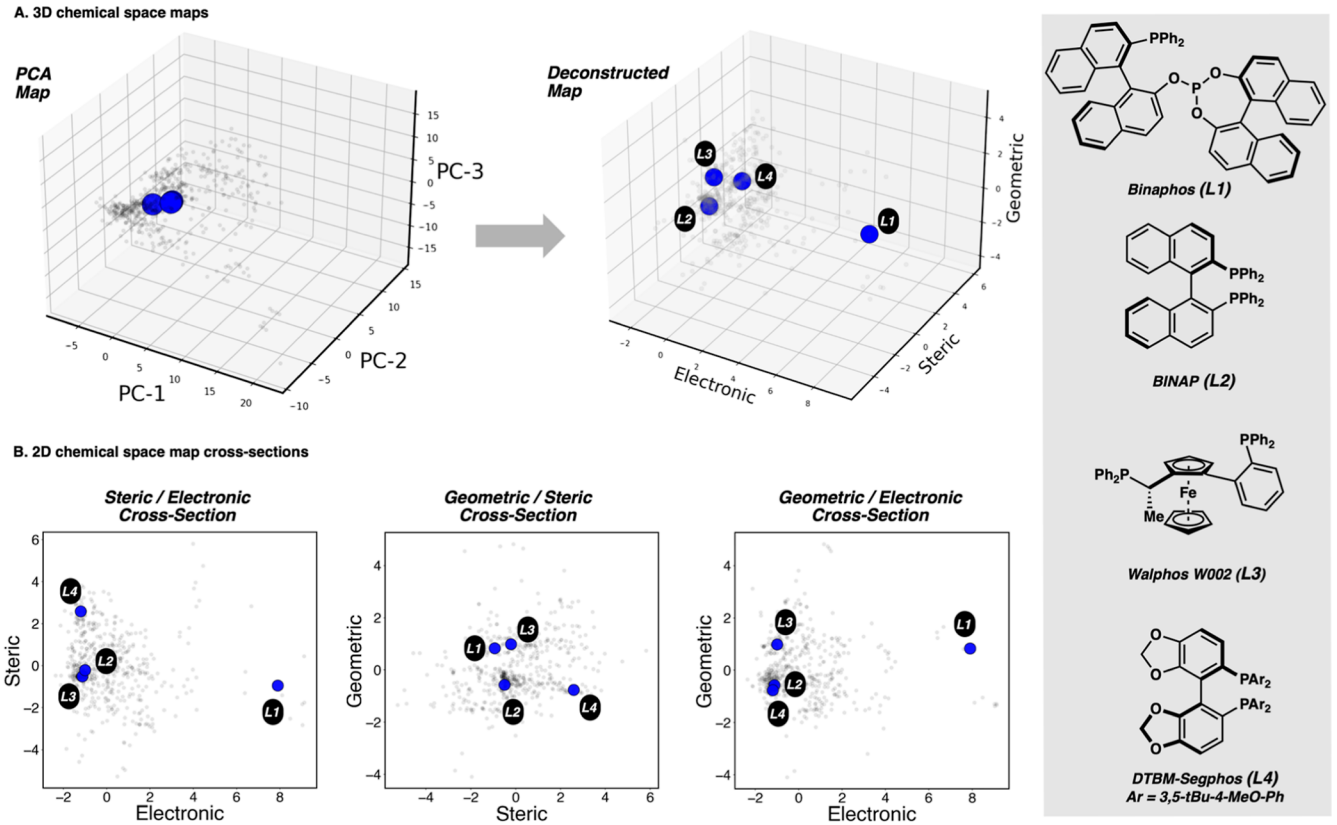


Figure 4. (A) A chemical space plot was generated with three axes that separately reflect the ligand sterics, electronics, and geometry. (B) Three cross-sections (steric/electronic, geometric/steric, and geometric/electronic) are used to compare ligands **L1**−**L4**.

able axes akin to a classical Tolman map,[54] while also incorporating multi-dimensional information of modern PCA plots. To accomplish this, parameters were selected from recognizable bins: steric, electronic, and geometric. Examples from each of these categories are depicted in Figure 2B. Individual descriptor categories were submitted to PCA, and the resulting first principal component of each of these parameter subsets was plotted to generate the 3D map shown in Figure 4A (for more details of the parameter selection, see the Supporting Information). It should be noted that there was no correlation between the steric, electronic, or geometric axes.

To illustrate the map's interpretability, we evaluated four representative bisphosphine ligands (**L1**−**L4**) in the designer map as well as in a conventional PCA map, which incorporates all of the descriptors from the data library (Figure 4). From the ligand location in the designer PCA map, one can glean more detailed information when compared to that in a conventional PCA map. This is best illustrated when one plots 2D cross-sections of the 3D space (Figure 4B, steric/electronic, geometric/steric, and geometric/electronic representations). Specifically, differences in the ligand structure can be readily assigned to physically meaningful ligand attributes. For instance, phosphite **L1** is similar to **L3** in terms of the steric dimension but has a significantly different electronic profile. Likewise, DTBM-Segphos **L4** is similar to **L2** in terms of electronics and geometric features, but the key differentiating factor is sterics. As showcased below, the designer maps can be applied to various downstream data science steps in multi-objective optimization campaigns. In addition to the reaction optimization applications presented (vide infra), it can be envisioned that this calculated parameter library and chemical space map could be used for many other potential applications such as training set design,[55] novel ligand generation,[56] and mechanistic understanding.[57]

Multi-Objective Optimization.[58] The challenge of multi-objective optimization of catalytic reactions is magnified in the field of process chemistry. Due to purity requirements for APIs and the cost of purification on a multi-kilogram scale, catalysts must impart exquisite chemo-, stereo-, and regioselectivity along with providing a high-yielding reaction. As such, the first two steps of the synthesis of TRPA1 inhibitor **1**,[37] an API, is an excellent case study (Figure 5). In the first step, the absolute stereochemistry of the API is set through an enantioselective Pd-catalyzed Hayashi−Heck reaction of 2,3-dihydrofuran (**3**) with aryl triflate **2** to produce enantioenriched enol ether (***R***)-**4**.[38,39] Besides the demand for high yield and enantioselectivity, this process delivers the undesired isomeric alkene **5** in low enantioselectivity. This impurity cannot be readily removed by distillation, the practical method of purification on a manufacturing scale for low-molecular-weight oils. In the second step, a Rh-catalyzed hydroformylation of (***R***)-**4** sets the second stereocenter of API **1**.[59,60] In this case, regioisomer **7** and C_4 epimer (***R,R***)-**6** are generated in addition to the desired aldehyde (***R,S***)-**6**. Downstream intermediates stemming from **7** could not be easily purged, leading to unacceptable amounts of the regioisomer and related impurities. Thus, both steps require an exquisite control of the reaction outcomes. Regioselectivity was identified as the most important metric targeted for improvement while maintaining the high stereoselectivity and overall yield of the desired products.

In parallel to the descriptor library construction, a conventional high-throughput experimental campaign was undertaken to examine the performance of a selection of ligands for both

https://doi.org/10.1021/jacs.2c08513
J. Am. Chem. Soc. 2023, 145, 110−121

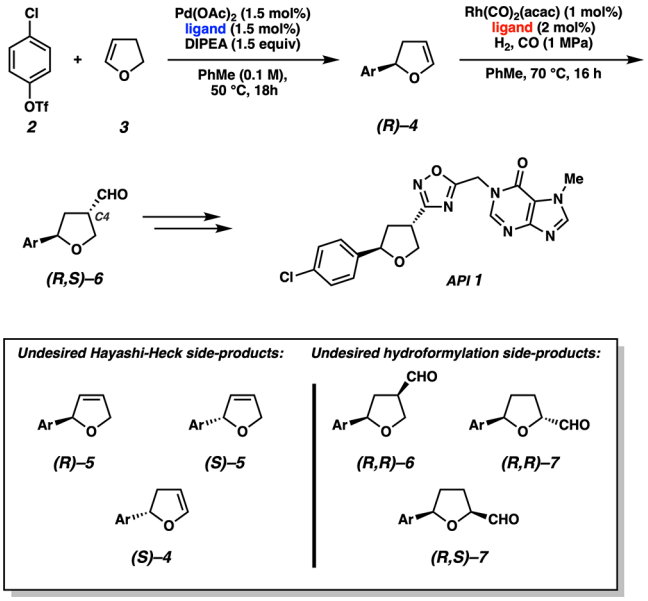


Figure 5. Process chemistry route to the API **1** and undesired side products in the Hayashi–Heck and hydroformylation reactions.

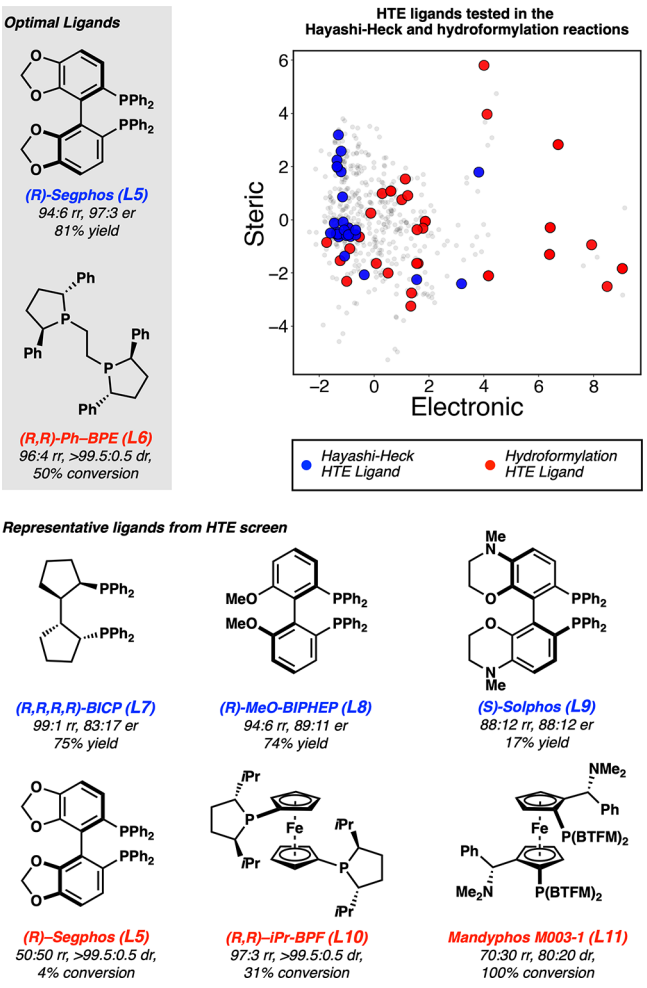


Figure 6. HTE results for representative ligands for the catalytic synthesis of (*R*)-**4** and (*R,S*)-**6**. The ligands used for the Hayashi–Heck and hydroformylation reactions are plotted on the chemical space graph. BTFM = bis(trifluoromethyl)phenyl.

reactions (Figure 6). For the Hayashi–Heck reaction, most ligands tested in the first step had a C_2 biaryl backbone as ligands of this type have previously proven effective in similar transformations.[61] This is reflected by the relatively narrow region of electronic but broad steric diversity for most of the bisphosphines tested (see the steric/electronic chemical space map in Figure 6). The initial results revealed that the use of Segphos **L5** furnished (*R*)-**4** in a regioisomeric ratio of 94:6, an enantiomeric ratio of 97:3, and 81% yield. As a wide range of bidentate organophosphorus ligands are known to be effective in alkene hydroformylation, a diverse selection of ligand scaffolds were evaluated including bisphosphines, mixed phosphine–phosphonites, and bisphosphonites. This is illustrated in the chemical space representation by the relatively wide distribution of tested ligands (Figure 6). Bisphosphine (*R,R*)-Ph-BPE **L6** was identified as a top performer for the hydroformylation reaction, providing 50% conversion of (*R*)-**4** to (*R,S*)-**6** with a regioisomeric ratio of 96:4 and a diastereomeric ratio of >99.5:0.5.[62]

Although the results of the HTE campaign are seemingly outstanding by academic and medicinal chemistry standards, significant improvements, especially in regioselectivity (>98:2), were desired for more efficient manufacturing. The lack of intuitive trends relating the ligand structure to reaction performance made it difficult to suggest improved ligands. For example, when Segphos **L5** was replaced with **L7**, increased regioselectivity is observed but with a concomitant reduction in enantioselectivity (Figure 6). Similarly, in the Rh-catalyzed hydroformylation, replacing **L6** with **L11** improved the reaction conversion but with an associated erosion of regioselectivity. This highlights the complexity of the multi-objective optimization problem and why a data-driven workflow was pursued. Additionally, we recognized the opportunity to evaluate how the workflow manages the contrasting nature of the two catalytic steps. Not only do the reactions use different transition metals (i.e., Pd vs Rh), but the coordination geometries of putative reaction intermediates are also distinct. While the Heck reaction primarily involves square-planar Pd species,[59] the hydroformylation reaction is thought to proceed through trigonal bipyramidal Rh intermediates.[57]

Our optimization commenced with an analysis of reactivity trends (Figure 7A, step 1) by applying our recently reported method to classify active and inactive ligands using a single-node decision tree.[13] Essentially, this algorithm partitions the ligands as a function of a ligand descriptor at a user-defined reactivity threshold. This step allows for a rigorous approach to data curation in subsequent analyses, while also providing a means to de-risk future ligand selection to those that would produce active catalysts. Analysis of the Hayashi–Heck data set using a threshold of 5% yield resulted in an excellent reactivity classification defined by the phosphorus lone pair occupancy (Figure 7A, graph 1 and Figure 7B).[63] It should be noted that this parameter is related to ligand electron richness, with highly electron rich ligands falling above the threshold value and providing a low yield. Ligand electronic effects have been previously identified as key factors in Hayashi–Heck reactions.[59] However, it should be noted that the correlations observed in this and all models discussed below are not necessarily causal, and therefore, mechanistic conclusions cannot be obtained on their basis alone. However, correlations such as these can be useful in the generation of mechanistic hypothesis, as highlighted below. Validation of the threshold was

https://doi.org/10.1021/jacs.2c08513
J. Am. Chem. Soc. 2023, 145, 110–121

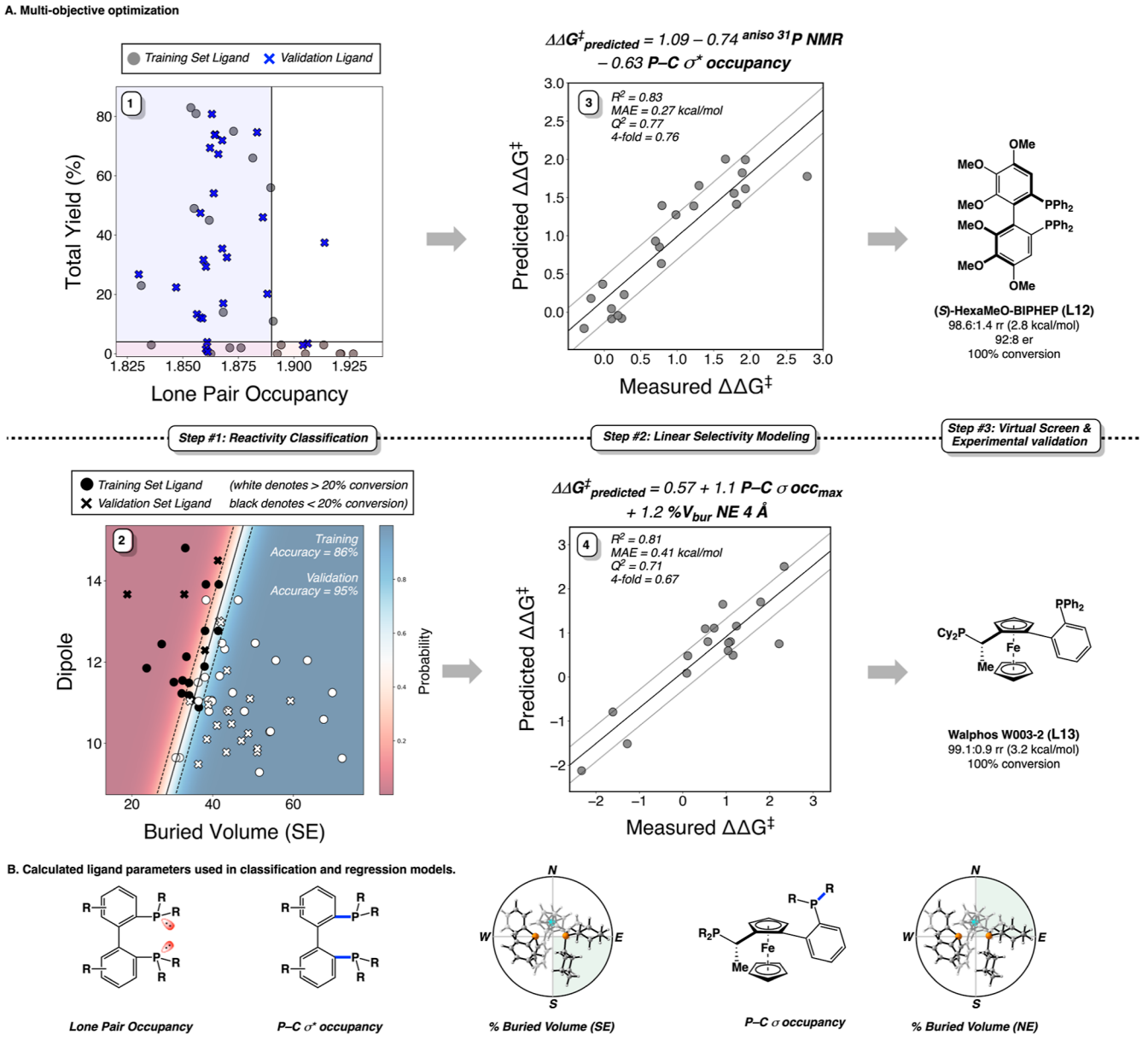


Figure 7. (A) Multi-objective optimization workflow applied to the optimization of catalysts for both the Hayashi−Heck reaction (top) and hydroformylation (bottom). Step 1. A classification model is used to relate the binary reactivity (reactive or unreactive) to ligand parameters. A single-parameter threshold analysis was used in the Hayashi−Heck reaction, while a two-parameter logistic regression was deployed for the hydroformylation reaction. Step 2. MLR was used to correlate the experimental regioselectivity ($\Delta\Delta G^{\ddagger}$) to computed parameters. Step 3. Virtual screening using the defined classifications and correlations led to the identification of optimal ligands **L12** and **L13**. (B) Schematic representations depicting the ligand parameters used in these models.

accomplished by exploring ligands predicted to be active (Figure 7A, graph 1).

In the case of the hydroformylation reaction, our attempts to classify ligands as active or inactive using a single-node decision tree were unsuccessful.[64] Therefore, we sought to use a logistic regression classification algorithm. Logistic regression converts a continuous variable (percent conversion) to a binary one (1 = active or 0 = inactive) with a user-defined threshold value. The data distribution in the training set was relatively bimodal with the conversion falling either below 13% or above 30%. Therefore, a threshold value between these two populations (20% conversion) was chosen. This classifier produces the probability of a catalyst being active as the output. While logistic regression is an established ML technique, it is underutilized in organic chemistry.[65−68] The algorithm identified a bivariate classification using the buried volume and total ligand dipole (Figure 7A, graph 2 and Figure 7B). In the plot, dark blue denotes a high probability that the ligand is active, while red represents a low probability. The solid black line reflects 50% probability of a ligand being active, while the dashed lines define the 25 and 75% probability boundaries. To validate this model, an array of additional ligands were experimentally evaluated, resulting in a 95% prediction accuracy (Figure 7B, graph 2). Interestingly, this led to many ligands with C_2 biaryl backbones being grouped in the inactive region of the logistic regression plot. We hope that this successful implementation of logistic regression will encourage the catalysis community to more widely adopt this powerful algorithm.

https://doi.org/10.1021/jacs.2c08513
J. Am. Chem. Soc. 2023, 145, 110−121

After successfully applying a reactivity classification step, we proceeded with the interrogation of the regioselectivity for both the Hayashi−Heck and hydroformylation reactions (Figure 7A, step 2) by regressing ligand parameters to the experimentally measured $\Delta\Delta G^{\ddagger}$ using MLR.[12] For the Hayashi−Heck reaction, regioselectivity (i.e., formation of olefin **4** vs **5**) was correlated to two parameters, ^{aniso31}P NMR and P−C σ* occupancy, the computed anisotropic phosphorus NMR shielding and the occupancy of the σ* orbitals of the P−C bonds, respectively (Figure 7B).[69] While the latter term has been postulated to correlate with the metal-to-ligand backbonding interactions,[70] the former was more difficult to interpret. However, more detailed analysis suggests that ^{aniso31}P NMR is likely a hybrid parameter reflecting both distal sterics and phosphorus electron richness (see the Supporting Information for further discussion). It should be noted that acceptable models could only be found when ligands were first curated using the yield threshold.[71] MLR was also used successfully to examine the enantioselectivity for this transformation (see Supporting Information Figure S12).[72]

Statistical models were pursued for regioselectivity of the hydroformylation reaction (e.g., formation of aldehyde **6** vs **7**) by evaluating the data with greater than 20% conversion (the same reactivity threshold as was used in the logistic regression model). An additional curation step was found to be necessary to produce acceptable linear models based on the hypothesis that disparate ligand types may impart selectivity through unique mechanisms.[73] As the training set was diverse, interrogation of the chemical space revealed a group of ligands that occupied a position in the chemical space remote from the remainder of the ligands evaluated (see Supporting Information Figure S15).[74] These incompatible ligands were removed from the training set, resulting in a robust two-term model for regioselectivity using an electronic parameter P−C σ occupancy and a steric parameter % V_{bur} NE (Figure 7B). Here, it is the symmetry-adapted quadrant-specific % V_{bur} term that enables the model to differentiate between the performance of catalyst enantiomers on this enantiomerically enriched substrate (***R***)-**4**. A larger buried volume in this quadrant seems to favor the formation of the desired regioisomer, perhaps by hindering the formation of an intermediate where the Rh is coordinated to the substrate at the C_5 position. However, this bivariate model indicates that electronics also plays a vital role in determining the regioselectivity, and mechanistic studies beyond the scope of this work would be required to validate the causation underpinning these correlations. The use of the chemical space map is noteworthy as it provided a method to visualize and curate the data set.

As the ultimate step in the workflow, virtual screening of the entire bisphosphine database was performed to identify ligands that were first predicted to be "active" and then predicted to be selective from the classification and linear regression models, respectively. Numerous accessible ligands were selected and tested experimentally with both extrapolations and interpolations included to gauge model robustness (vide infra). For the Hayashi−Heck reaction, HexaMeO-BIPHEP[34] **L12** was found to result in superb reaction performance and provided the desired product **4** in a regioisomeric ratio of 98.6:1.4 consistent with the predicted outcome (98:2). This represents an *∼1 kcal/mol extrapolation in regioselectivity* relative to that of the previous optimal ligand, Segphos **L5**. The reaction also resulted in an enantiomeric ratio of 92:8 (predicted 95:5) and 100% conversion. The multi-objective workflow was also successful in identifying a significantly improved ligand for the hydroformylation with Walphos **L13** furnishing (***R,S***)-**6** with an exceptional regioisomeric ratio of 99.1:0.9 (predicted 98:2) and 100% conversion.[75] **L13** also provides an *∼1 kcal/mol improvement in regioselectivity* relative to (*R,R*)-Ph-BPE **L6**, the previous best ligand.

These significant improvements could be realized on gram scale with the first step providing (***R***)-**4** in 77% isolated yield with 98:2 rr and 93.5:6.5 er and the second providing (***R,S***)-**6** with 99:1 rr and 97:3 dr (Figure 8). Owing to its lability, (***R,S***)-**6**

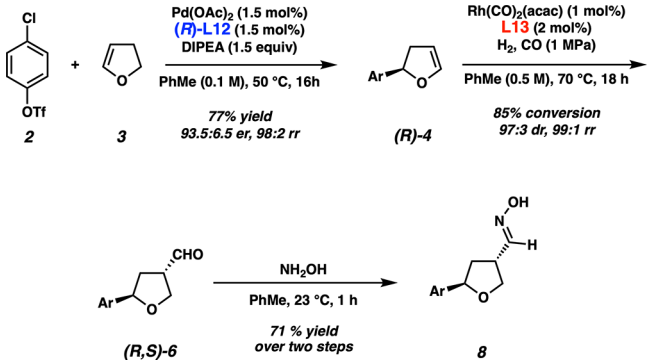


Figure 8. Validation of optimized ligands **L12** and **L13** on gram scale.

was directly converted to the corresponding stable oxime (***R,S***)-**8** in 71% isolated yield over two steps. In addition to producing the desired product in high yield, these superior catalysts provided regio- and stereochemical purity significant enough to obviate extensive purification, thus streamlining the route toward API **1**.

The identification of optimal ligands HexaMeO-BIPHEP **L12** and Walphos **L13** highlights several notable strengths of this approach. First, the optimization of the Hayashi−Heck reaction demonstrates the ability of the statistical models to predict performance improvements resulting from subtle changes in the ligand structure (Figure 9A). Comparing HexaMeO-BIPHEP **L12** to ligands **L5**, **L8**, **L14**, and **L15**, it would be difficult to intuitively predict that one ligand would perform significantly better than the others as all share similar C_2-biaryl backbones decorated with electron-donating substituents. Nevertheless, the statistical model correctly identified **L12** as a high-performing ligand for the Hayashi−Heck reaction. While **L12** has been known since 1991,[34] to the best of our knowledge, there have only been two reports of its use in the subsequent peer reviewed literature.[76,77]

Therefore, this strategy led to the identification of a seldom-used ligand that would not typically be included in a conventional optimization. Additionally, the modeling workflow enables "scaffold hopping"—the prediction of optimized ligands with distinct backbones (Figure 9B).[78,79] This was critical to the successful improvement of the hydroformylation reaction as all readily available derivatives of (*R,R*)-Ph-BPE **L6** had been previously tested and did not demonstrate improved performance. Three Walphos derivatives were included in the training set, but each resulted in poor performance (either in terms of conversion or regioselectivity), giving no indication that this ligand class would intuitively be a promising avenue to pursue. However, the workflow correctly identified Walphos-**L13** as an optimal ligand, despite it being a ligand with different symmetry (i.e., C_1 vs C_2) and element of chirality (i.e., a chiral backbone instead of a phospholane ring) from that of the previously

https://doi.org/10.1021/jacs.2c08513
J. Am. Chem. Soc. 2023, 145, 110−121

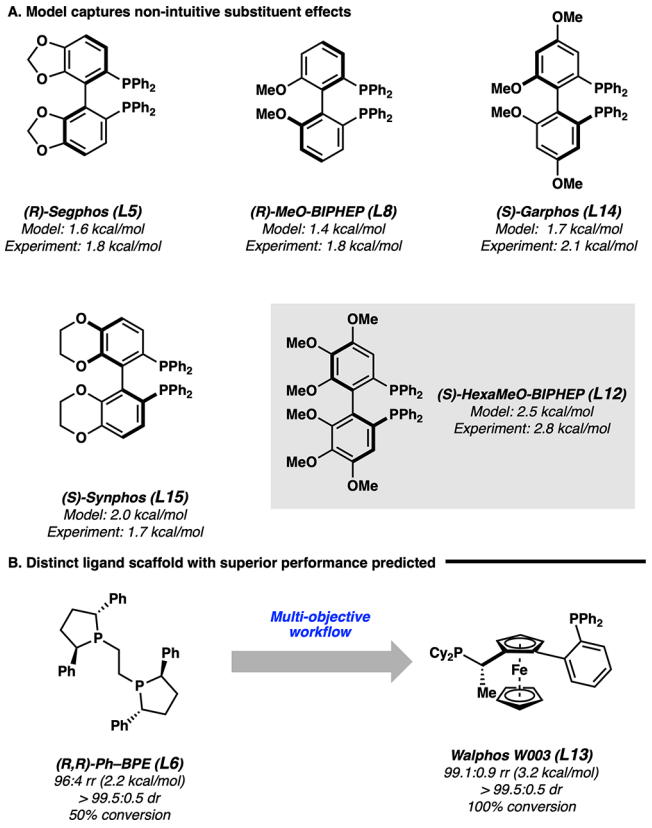


Figure 9. (A) Models capture the dramatic effects that subtle changes to the catalyst structure can have in the Hayashi–Heck reaction. (B) Models allow for ligand scaffold hopping to identify a significantly different superior ligand for the alkene hydroformylation.

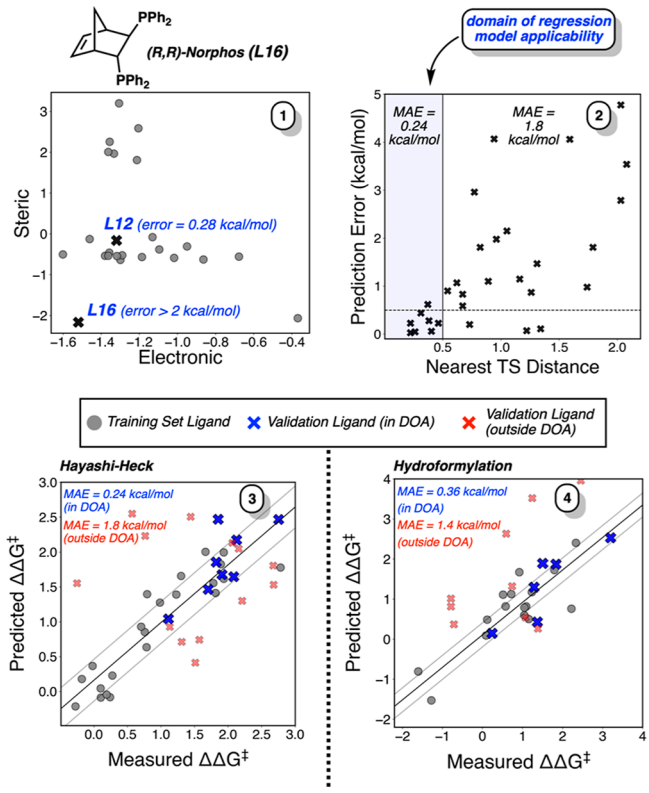


Figure 10. Domain of applicability, as defined by the distance in the chemical space, for the linear regression models describing regioselectivity in both the Hayashi–Heck and hydroformylation reactions. TS = training set and DOA = domain of applicability.

optimized ligand **L6**. Finally, although each ligand in the virtual library was calculated as the [ligand]$PdCl_2$ complex, the resultant parameters can be used for reactions that employ different closed-shell transition metal complexes. This is evidenced by the successful logistic and linear modeling of the Rh-catalyzed olefin hydroformylation reaction using these parameters.

While the workflow ultimately led to significant improvements in the reaction performance, we were interested in evaluating what factors could lead to erroneous predictions as understanding the limitations of the models could inform future implementations of this workflow (Figure 10). We observed that ligands with significant prediction errors tended to be farther in the chemical space from the training set ligands. For example, in the Hayashi–Heck reaction, Norphos **L16** was poorly predicted by the selectivity model and was relatively remote from the training set ligands (Figure 10, graph 1). In contrast, the optimal ligand HexaMeO-BIPHEP (**L12**) was comparatively close in the chemical space to several training set ligands, and its performance was accurately predicted. In fact, a plot of the distance of each ligand to the nearest training set neighbor demonstrates a domain of model applicability—an allowable distance from the training set ligands within which the regression model is highly accurate. The concept of a domain of applicability is well-established in quantitative structure–activity relationship (QSAR) models[80] and has previously been reported in reaction informatics.[81] Generally, the uncertainty in a particular prediction is correlated to the similarity between that molecule and the molecules used to construct the model.

Using Euclidean distance in the chemical space to quantify the domain of applicability (see the Supporting Information for details on distance calculation), ligands tested that were 0.5 normalized parameter space units or less from a training set ligand had a prediction mean absolute error (MAE) of 0.24 kcal/mol for the Hayashi–Heck reaction (Figure 10, graph 2 and graph 3).[82] In contrast, ligands with >0.5 normalized parameter space units from a training set ligand were poorly predicted (MAE = 1.8 kcal/mol). An extended domain of applicability was observed for the hydroformylation regioselectivity model (1.6 distance units to the nearest training set neighbor, see Figure 10 graph 4 and the Supporting Information), which is likely due to the training set covering a much broader area of the chemical space. The chemical space coverage in the training set is likely a key influencing factor on a model's domain of applicability, along with ligand-dependent changes to the reaction mechanism. With the growing use of large databases of calculated properties in conjunction with data science tools, the quantification of a particular model's domain of applicability is likely to become essential.

CONCLUSIONS

We computed a virtual library of >550 bisphosphine ligands and then leveraged it in conjunction with data science tools to optimize the first two steps in the synthesis of an API. This led to 1 kcal/mol improvements in regioselectivity for both synthetic steps. Key to the success of this endeavor was the combined use of linear regression and classification algorithms to model selectivity and reactivity respectively as well as chemical space analysis to understand outliers. The effective multi-objective

https://doi.org/10.1021/jacs.2c08513
J. Am. Chem. Soc. 2023, 145, 110–121

optimization of these two consecutive processes with different transition metal catalysts, distinct mechanisms, and training sets occupying different regions of the chemical space highlights the utility of this workflow. We anticipate that this strategy combined with the intuitive chemical space map of bidentate organophosphorus ligands has the potential to enable the development and optimization of a wide array of transition-metal-catalyzed reactions that employ bisphosphine ligands. In order to extend these concepts to encompass all bidentate ligand classes, virtual libraries of P,N and N,N ligands are currently under construction and will be reported in due course.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.2c08513.

The full parameter library and all Python scripts used for data analysis, which have also been made available at https://github.com/SigmanGroup/Multiobjective_Optimization (ZIP)

Multi-objective parameters and symmetry adapted parameters (XLSX)

Detailed experimental procedures, ligand parameters, and compound characterization data (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Kyle A. Mack − *Department of Small Molecule Process Chemistry, Genentech, Inc., South San Francisco, California 94080, United States;* orcid.org/0000-0002-4986-4040; Email: mack.kyle@gene.com

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

Authors

Jordan J. Dotson − *Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States*

Lucy van Dijk − *Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States;* orcid.org/0000-0002-1899-087X

Jacob C. Timmerman − *Department of Small Molecule Process Chemistry, Genentech, Inc., South San Francisco, California 94080, United States*

Samantha Grosslight − *Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States*

Richard C. Walroth − *Department of Small Molecule Process Chemistry, Genentech, Inc., South San Francisco, California 94080, United States*

Francis Gosselin − *Department of Small Molecule Process Chemistry, Genentech, Inc., South San Francisco, California 94080, United States;* orcid.org/0000-0001-9812-4180

Kurt Püntener − *Synthetic Molecules Technical Development, Process Chemistry & Catalysis, F. Hoffmann-La Roche Limited, CH-4070 Basel, Switzerland*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.2c08513

Author Contributions

‖J.J.D. and L.V.D. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Colin Masui (Genentech) and Adam Childs (Genentech) for performing high-throughput experimentation, Isabelle Duffour (Roche) and Steven Chin (Genentech) for analytical support, Anna-Lena Glass (Roche) for experimental support, Professor Clark Landis (University of Wisconsin-Madison) for helpful discussions regarding olefin hydroformylation, Tobias Gensch (TU Berlin) for his insights into parameter selection, and Ellyn Peters and Hanna Clements (University of Utah) for discussions regarding data analysis. M.S.S. and L.v.D acknowledges financial support from the NSF under the CCI Center for Computer Assisted Synthesis (CHE-1925607 and CHE-2202693). J.J.D. acknowledges financial support from the NIH (F32-GM140529).

■ REFERENCES

(1) Efforts to deconvolute structure−function relationships in homogeneous catalysis have led to important ligand descriptors and stereochemical models. These include the use of bite angle (ref 2) and cone angle (ref 3) as parameters and the quadrant stereochemical model for C_2-symmetric ligands (ref 4).

(2) Birkholz, M.-N.; Freixa, Z.; van Leeuwen, P. W. N. M. Bite angle effects of diphosphines in C-C and C-X bond forming cross coupling reactions. *Chem. Soc. Rev.* **2009**, *38*, 1099−1118.

(3) Tolman, C. A. Phosphorus ligand exchange equilibriums on zerovalent nickel. Dominant role for steric effects. *J. Am. Chem. Soc.* **1970**, *92*, 2956−2965.

(4) Knowles, W. S. Asymmetric Hydrogenation. *Acc. Chem. Res.* **1983**, *16*, 106−112.

(5) Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V. The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS Cent. Sci.* **2021**, *7*, 1622−1637.

(6) Zahrt, A. F.; Athavale, S. V.; Denmark, S. E. Quantitative Structure-Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem. Rev.* **2020**, *120*, 1620−1689.

(7) Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning. *Chem. Rev.* **2021**, *121*, 9927−10000.

(8) Gallegos, L. C.; Luchini, G.; St John, P. C.; Kim, S.; Paton, R. S. Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.* **2021**, *54*, 827−836.

(9) The rational tuning of ligand descriptors to optimize catalyst performance has been reported in the literature. For representative examples, see the following review and the references reported therein: Gillespie, J. A.; Dodds, D. L.; Kamer, P. C. J. Rational design of diphosphorus ligands - a route to superior catalysts. *Dalton Trans.* **2010**, *39*, 2751−2764.

(10) Durand, D. J.; Fey, N. Building a Toolbox for the Analysis and Prediction of Ligand and Catalyst Effects in Organometallic Catalysis. *Acc. Chem. Res.* **2021**, *54*, 837−848.

(11) 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.

(12) Virtual libraries have gained widespread use in the field of medicinal chemistry. For further discussion see the following review: Walters, W. P. Virtual Chemical Libraries. *J. Med. Chem.* **2019**, *62*, 1116−1124.

(13) 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.

(14) Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Peters, E.; Gensch, T.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G. Univariate Classification of Phosphine Ligation State and Reactivity in Cross-Coupling Catalysis. *Science* **2021**, *374*, 301−308.
(15) See, X. Y.; Wen, X.; Wheeler, T. A.; Klein, C. K.; Goodpaster, J. D.; Reiner, B. R.; Tonks, I. A. Iterative Supervised Principal Component Analysis Driven Ligand Design for Regioselective Ti-Catalyzed Pyrrole Synthesis. *ACS Catal.* **2020**, *10*, 13504−13517.
(16) Kutchukian, P. S.; Dropinski, J. F.; Dykstra, K. D.; Li, B.; DiRocco, D. A.; Streckfuss, E. C.; Campeau, L.-C.; Cernak, T.; Vachal, P.; Davies, I. W.; Krska, S. W.; Dreher, S. D. Chemistry Informer Libraries: A Chemoinformatics Enabled Approach to Evaluate and Advance Synthetic Methods. *Chem. Sci.* **2016**, *7*, 2604−2613.
(17) Fey, N.; Garland, M.; Hopewell, J. P.; McMullin, C. L.; Mastroianni, S.; Orpen, A. G.; Pringle, P. G. Stable Fluorophosphines: Predicted and Realized Ligands for Catalysis. *Angew. Chem., Int. Ed.* **2012**, *51*, 118−122.
(18) 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*, 89−96.
(19) Nambiar, A. M. K.; Breen, C. P.; Hart, T.; Kulesza, T.; Jamison, T. F.; Jensen, K. F. Bayesian Optimization of Computer-Proposed Multistep Synthetic Routes on an Automated Robotic Flow Platform. *ACS Cent. Sci.* **2022**, *8*, 825−836.
(20) Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018**, *4*, 1134−1145.
(21) Chen, H.; Yamaguchi, S.; Morita, Y.; Nakao, H.; Zhai, X.; Shimizu, Y.; Mitsunuma, H.; Kanai, M. Data-Driven Catalyst Optimization for Stereodivergent Asymmetric Synthesis by Iridium/Boron Hybrid Catalysis. *Cell Rep. Phys. Sci.* **2021**, *2*, 100679.
(22) Blázquez-Barbadillo, C.; Aranzamendi, E.; Coya, E.; Lete, E.; Sotomayor, N.; González-Díaz, H. Perturbation Theory Model of Reactivity and Enantioselectivity of Palladium-Catalyzed Heck-Heck Cascade Reactions. *RSC Adv.* **2016**, *6*, 38602−38610.
(23) Aguado-Ullate, S.; Guasch, L.; Urbano-Cuadrado, M.; Bo, C.; Carbó, J. J. 3D-QSPR Models for Predicting the Enantioselectivity and the Activity for Asymmetric Hydroformylation of Styrene Catalyzed by Rh-Diphosphane. *Catal. Sci. Technol.* **2012**, *2*, 1694−1704.
(24) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chem. Sci.* **2018**, *9*, 7642−7655.
(25) Sagmeister, P.; Ort, F. F.; Jusner, C. E.; Hebrault, D.; Tampone, T.; Buono, F. G.; Williams, J. D.; Kappe, C. O. Autonomous Multi-Step and Multi-Objective Optimization Facilitated by Real-Time Process Analytics. *Adv. Sci.* **2022**, *9*, 2105547.
(26) Schweidtmann, A. M.; Clayton, A. D.; Holmes, N.; Bradford, E.; Bourne, R. A.; Lapkin, A. A. Machine Learning Meets Continuous Flow Chemistry: Automated Optimization Towards the Pareto Front of Multiple Objectives. *Chem. Eng. J.* **2018**, *352*, 277−282.
(27) In parallel to the field of small-molecule organic synthesis, the field of materials science has benefited from data science tools for multi-objective optimizations. For representative examples see refs 28 and 29.
(28) Wu, S.; Kondo, Y.; Kakimoto, M.-a.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; Yoshida, R. Machine-Learning-Assisted Discovery of Polymers with High Thermal Conductivity Using a Molecular Design Algorithm. *npj Comput. Mater.* **2019**, *5*, 66.
(29) Lv, H.; Chen, X. Intelligent control of nanoparticle synthesis through machine learning. *Nanoscale* **2022**, *14*, 6688−6708.
(30) Knowles, W. S.; Sabacky, M. J.; Vineyard, B. D.; Weinkauff, D. J. Asymmetric Hydrogenation with a Complex of Rhodium and a Chiral Bisphosphine. *J. Am. Chem. Soc.* **1975**, *97*, 2567−2568.
(31) Dang, T. P.; Kagan, H. B. The Asymmetric Synthesis of Hydratropic Acid and Amino-Acids by Homogeneous Catalytic Hydrogenation. *J. Chem. Soc. D.* **1971**, 481.
(32) Miyashita, A.; Yasuda, A.; Takaya, H.; Toriumi, K.; Ito, T.; Souchi, T.; Noyori, R. Synthesis of 2,2′-bis(diphenylphosphino)-1,1′-binaphthyl (BINAP), an atropisomeric chiral bis(triaryl)phosphine, and its use in the rhodium(I)-catalyzed asymmetric hydrogenation of alpha-(acylamino)acrylic acids. *J. Am. Chem. Soc.* **1980**, *102*, 7932−7934.
(33) Togni, A.; Breutel, C.; Schnyder, A.; Spindler, F.; Landert, H.; Tijani, A. A Novel Easily Accessible Chiral Ferrocenyldiphosphine for Highly Enantioselective Hydrogenation, Allylic Alkylation, and Hydroboration Reactions. *J. Am. Chem. Soc.* **1994**, *116*, 4062−4066.
(34) Burk, M. J. C2-symmetric bis(phospholanes) and their use in highly enantioselective hydrogenation reactions. *J. Am. Chem. Soc.* **1991**, *113*, 8518−8519.
(35) Schmid, R.; Foricher, J.; Cereghetti, M.; Schönholzer, P. Axially Dissymmetric Diphosphines in the Biphenyl Series: Synthesis of (6,6′-Dimethoxybiphenyl-2,2′-diyl)bis(diphenylphosphine)("MeO-BIPHEP") and Analoguesviaanortho-Lithiation/IodinationUllmann-Reaction Approach. *Helv. Chim. Acta* **1991**, *74*, 370−389.
(36) Saito, T.; Yokozawa, T.; Ishizaki, T.; Moroi, T.; Sayo, N.; Miura, T.; Kumobayashi, H. New Chiral Diphosphine Ligands Designed to Have a Narrow Dihedral Angle in the Biaryl Backbone. *Adv. Synth. Catal.* **2001**, *343*, 264−267.
(37) An important distinction between multi-objective and multi-parameter optimization should be noted by the reader. Multi-objective optimization refers to the simultaneous optimization of multiple reaction outcomes (i.e., yield, regioselectivity, and enantioselectivity) while multi-parameter optimization points to the simultaneous optimization of multiple reaction parameters (e.g., solvent, temperature, ligand) for a given reaction. The present study is focused exclusively on multi-objective optimization.
(38) Terrett, J. A.; Chen, H.; Shore, D. G.; Villemure, E.; Larouche-Gauthier, R.; Déry, M.; Beaumier, F.; Constantineau-Forget, L.; Grand-Maître, C.; Lépissier, L.; et al. Tetrahydrofuran-Based Transient Receptor Potential Ankyrin 1 (TRPA1) Antagonists: Ligand-Based Discovery, Activity in a Rodent Asthma Model, and Mechanism-of-Action via Cryogenic Electron Microscopy. *J. Med. Chem.* **2021**, *64*, 3843−3869.
(39) Ozawa, F.; Kubo, A.; Hayashi, T. Catalytic Asymmetric Arylation of 2,3-Dihydrofuran with Aryl Triflates. *J. Am. Chem. Soc.* **1991**, *113*, 1417−1419.
(40) Ozawa, F.; Kubo, A.; Matsumoto, Y.; Hayashi, T.; Nishioka, E.; Yanagi, K.; Moriguchi, K. Palladium-Catalyzed Asymmetric Arylation of 2,3-Dihydrofuran with Phenyl Triflate. A Novel Asymmetric Catalysis Involving a Kinetic Resolution Process. *Organometallics* **1993**, *12*, 4188−4196.
(41) Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Murray, P.; Orpen, A. G.; Osborne, R.; Purdie, M. Computational Descriptors for Chelating P,P- and P,N-Donor Ligands1. *Organometallics* **2008**, *27*, 1372−1383.
(42) 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 Chelating P,P-Donor Ligands (LKB-PP). *Organometallics* **2012**, *31*, 5302−5306.
(43) It should be noted that ligands with bite angles ≥120° were not considered in this study. This is because all of the parameters were based on the cis-chelated complexes. Since ligands with wide bite angles are known to chelate in a trans geometry, the cis complexes would not be physically meaningful.
(44) Guo, J.-Y.; Minko, Y.; Santiago, C. B.; Sigman, M. S. Developing Comprehensive Computational Parameter Sets To Describe the Performance of Pyridine-Oxazoline and Related Ligands. *ACS Catal.* **2017**, *7*, 4144−4151.
(45) Parameters were acquired from the corresponding DFT output files using an Python3 script (see Supporting Information for details). Our script implemented functions from the Morfeus Python library (https://github.com/kjelljorner/morfeus, version 0.5.3).
(46) It should be noted that the bite angles reported in this virtual library are distinct from but highly correlated with the natural bite angles reported by van Leeuwen (see refs 19 and 20 in the Supporting Information). For further discussion, see Figure S36 in the Supporting Information.

https://doi.org/10.1021/jacs.2c08513
J. Am. Chem. Soc. 2023, 145, 110−121

(47) Note that the buried volume measurements depicted in Figure 3 were measured using a sphere with a 4 Å radius, except for the logistic regression model where a sphere with a 7 Å radius is used. For further discussion of quadrant buried volume, see refs 48 and 49.
(48) 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 Topographic Steric Maps. *Organometallics* **2016**, *35*, 2286−2293.
(49) Poater, A.; Cosenza, B.; Correa, A.; Giudice, S.; Ragone, F.; Scarano, V.; Cavallo, L. S.V. Samb V ca: A Web Application for the Calculation of the Buried Volume of N-Heterocyclic Carbene Ligands. *Eur. J. Inorg. Chem.* **2009**, 1759−1766.
(50) The effect of molecular conformation on quadrant-based parameters has been discussed in the following reference: Zahrt, A. F.; Rinehart, N. I.; Denmark, S. E. A. A Conformer-Dependent, Quantitative Quadrant Model. *Eur. J. Org. Chem.* **2021**, 2343−2354.
(51) The generation of minimum, maximum, and average parameters based on the symmetry of the ligand is done automatically by the parameter acquisition script that we have made available on GitHub (https://github.com/SigmanGroup/Multiobjective_Optimization).
(52) A consequence of the latter assumption is that the absence of a correlation cannot be used to make a conclusion about a reaction mechanism since it could either be due to a mechanistic feature or non-systematic parameter errors.
(53) Fey, N.; Koumi, A.; Malkov, A. V.; Moseley, J. D.; Nguyen, B. N.; Tyler, S. N. G.; Willans, C. E. Mapping the Properties of Bidentate Ligands with Calculated Descriptors (LKB-bid). *Dalton Trans.* **2020**, *49*, 8169−8178.
(54) Tolman, C. A. Steric Effects of Phosphorus Ligands in Organometallic Chemistry and Homogeneous Catalysis. *Chem. Rev.* **1977**, *77*, 313−348.
(55) Gensch, T.; Smith, S. R.; Colacot, T. J.; Timsina, Y. N.; Xu, G.; Glasspoole, B. W.; Sigman, M. S. Design and Application of a Screening Set for Monophosphine Ligands in Cross-Coupling. *ACS Catal.* **2022**, *12*, 7773−7780.
(56) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(57) Lustosa, D. M.; Milo, A. Mechanistic Inference from Statistical Models at Different Data-Size Regimes. *ACS Catal.* **2022**, *12*, 7886−7906.
(58) All ML modeling was conducted by compiling relevant data (yield, regioselectivity, etc.) and ligand IDs corresponding to the virtual library into an excel spreadsheet. Our Python3 modeling scripts then read the data and perform the statistical modeling. For more details on the setup of the excel files and the execution of the scripts see our GitHub repository.
(59) Chikkali, S. H.; Bellini, R.; de Bruin, B.; van der Vlugt, J. I.; Reek, J. N. H. Highly Selective Asymmetric Rh-Catalyzed Hydroformylation of Heterocyclic Olefins. *J. Am. Chem. Soc.* **2012**, *134*, 6607−6616.
(60) Adint, T. T.; Wong, G. W.; Landis, C. R. Libraries of Bisdiazaphospholanes and Optimization of Rhodium-Catalyzed Enantioselective Hydroformylation. *J. Org. Chem.* **2013**, *78*, 4231−4238.
(61) Mc Cartney, D.; Guiry, P. J. The Asymmetric Heck and Related Reactions. *Chem. Soc. Rev.* **2011**, *40*, 5122−5150.
(62) The ligands plotted in Figure 6 were selected without the use of data science tools. The data resulting from these ligands was then used to train the models that, in turn, informed the next ligands selected (i.e., **L12**, **L13**, and the ligands plotted in Figure 10).
(63) It should be noted that the choice of threshold value for the Hayashi-Heck reaction was made based on the data distribution. For further discussion, see Figure S7B in the Supporting Information.
(64) It should be noted that logistic regression and decision tree analysis were both reasonable candidates to classify reactivity in the Hayashi−Heck and hydroformylation reactions. In each instance, the best technique was selected based on trial-and-error.
(65) Moskal, M.; Beker, W.; Szymkuć, S.; Grzybowski, B. A. Scaffold-Directed Face Selectivity Machine-Learned from Vectors of Non-covalent Interactions. *Angew. Chem., Int. Ed.* **2021**, *60*, 15230−15235.
(66) Banerjee, S.; Sreenithya, A.; Sunoj, R. B. Machine Learning for Predicting Product Distributions in Catalytic Regioselective Reactions. *Phys. Chem. Chem. Phys.* **2018**, *20*, 18311−18318.
(67) Yada, A.; Nagata, K.; Ando, Y.; Matsumura, T.; Ichinoseki, S.; Sato, K. Machine Learning Approach for Prediction of Reaction Yield with Simulated Catalyst Parameters. *Chem. Lett.* **2018**, *47*, 284−287.
(68) Boni, Y. T.; Cammarota, R. C.; Liao, K.; Sigman, M. S.; Davies, H. M. L. Leveraging Regio- and Stereoselective C(sp3)-H Functionalization of Silyl Ethers to Train a Logistic Regression Classification Model for Predicting Site-Selectivity Bias. *J. Am. Chem. Soc.* **2022**, *144*, 15549−15561.
(69) It should be noted that the validation statistics Q^2 and 4-fold were both modest (0.77 and 0.76 respectively). The lower values relative to the R^2 (0.8) are likely due to the training set size and structure. In addition, there are no training set ligands with similar performance to BICP (2.78 kcal/mol). Therefore, the model changes significantly when this is left out.
(70) De Jesus Silva, J.; Bartalucci, N.; Jelier, B.; Grosslight, S.; Gensch, T.; Schünemann, C.; Müller, B.; Kamer, P. C. J.; Copéret, C.; Sigman, M. S.; Togni, A. Development and Molecular Understanding of a Pd-Catalyzed Cyanation of Aryl Boronic Acids Enabled by High-Throughput Experimentation and Data Analysis. *Helv. Chim. Acta* **2021**, *104*, No. e2100200.
(71) The ligand HOBIPHEP (ID-576) was excluded since it provided < 1% combined yield of the product, therefore, regioisomeric ratios could not accurately be quantified. The ligand TMBTP (ID-544) was also excluded as it was found to be an outlier that led to poor models (see Figure S8 for ligand structures).
(72) Ligands featuring 3,5-tBu-Ph moieties (ligand IDs 189, 196, 707, 708, 709, and 717) on the phosphorus could not be modeled with rest of the ligand data, possibly due to a distinct mechanism for enantioinduction (see Figure S9 for ligand structures).
(73) Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Acc. Chem. Res.* **2021**, *54*, 3136−3148.
(74) It should be noted that the validation statistics Q^2 and 4-fold were both modest (0.71 and 0.67 respectively). This is likely because the training set was quite small (only 17 ligands) and relatively sparse. In particular, only three ligands showed selectivity below 0 kcal/mol. As such the model changes significantly when these are left out of the training set, thus leading to diminished validation statistics.
(75) Specific studies comparing the turnover number (TON) or turnover frequency (TOF) between ligands were outside of the scope of in the present work. However, for the hydroformylation reaction, it was observed that the rate was increased by ca. 4× for W002-2 relative to (*R*,*R*)-Ph-BPE (see Supporting Information for ligand structures). This may imply that TOF was significantly improved in this optimization.
(76) Cederbaum, F.; Lamberth, C.; Malan, C.; Naud, F.; Spindler, F.; Studer, M.; Blaser, H.-U. Synthesis of Substituted Mandelic Acid Derivativesvia Enantioselective Hydrogenation: Homogeneous versus Heterogeneous Catalysis. *Adv. Synth. Catal.* **2004**, *346*, 842−848.
(77) Wang, L.-P.; Feng, W.-H.; Jiao, X.-Z.; Xie, P.; Liang, X.-T. Improvement on the Synthesis of Hexa-MeO-BIPHEP Diphosphine Ligand. *Chin. J. Synth. Chem.* **2007**, *15*, 108−110.
(78) The concept of scaffold hopping is also important in the field of medicinal chemistry. For further discussion see the following reference: Zhao, H. Scaffold Selection and Scaffold Hopping in Lead Generation: A Medicinal Chemistry Perspective. *Drug Discovery Today* **2007**, *12*, 149−155.
(79) The importance of scaffold-hopping is evidenced by numerous recent literature reports. For further discussion, see the following review: Jurczyk, J.; Woo, J.; Kim, S. F.; Dherange, B. D.; Sarpong, R.; Levin, M. D. Single-Atom Logic for Heterocycle Editing. *Nat. Synth.* **2022**, *1*, 352−364.

https://doi.org/10.1021/jacs.2c08513
J. Am. Chem. Soc. 2023, 145, 110−121

(80) Weaver, S.; Gleeson, M. P. The Importance of the Domain of Applicability in QSAR Modeling. *J. Mol. Graph. Model.* **2008**, *26*, 1315−1326.
(81) Haywood, A. L.; Redshaw, J.; Hanson-Heine, M. W. D.; Taylor, A.; Brown, A.; Mason, A. M.; Gärtner, T.; Hirst, J. D. Kernel Methods for Predicting Yields of Chemical Reactions. *J. Chem. Inf. Model.* **2022**, *62*, 2077−2092.
(82) Note that the MLR graph is only plotted from 0 to 3 kcal/mol. Four of the ligands that were not in the domain of applicability fell out of this range. These can be seen in graph 2 but not graph 3.

https://doi.org/10.1021/jacs.2c08513
J. Am. Chem. Soc. 2023, 145, 110−121