Decoding Regioselectivity in Cu(I)-Catalyzed Borylation of Alkynes: Insights from Machine Learning and Artificial Intelligence

Guillermo Marcos-Ayuso[1,2,*], David Quesada[2], Sara Fernández-Moyano[1], Carlos Lendínez[1], Pablo Mauleón[1,*],Ramón Gómez Arrayás[1,*]

[1] Department of Organic Chemistry, Institute for Advanced Research in Chemical Sciences (IAdChem) UAM, 28049 Madrid, Spain.

[2] AItenea Biotech, Parque Científico de Madrid, Ciudad Universitaria de Cantoblanco, Calle Faraday, 7, 28049 Madrid, Spain.

KEYWORDS (Word Style "BG_Keywords"). If you are submitting your paper to a journal that requires keywords, provide significant keywords to aid the reader in literature retrieval.

ABSTRACT: Vinyl boronates are highly valuable intermediates in chemical synthesis, extensively used in C–C bond-forming reactions such as catalytic cross-coupling. Transition metal-catalyzed hydroboration of alkynes has emerged as a key method for synthesizing these building blocks. While classical approaches rely on noble metals like rhodium and iridium, copper-catalyzed hydroboration offers a sustainable and cost-effective alternative. This strategy utilizes bench-stable reagents under mild conditions, delivering highly stereoselective *trans*-vinylboronates. However, predicting regioselectivity remains a challenge due to the complex interplay of ligand structures, alkyne substitution patterns, and reaction conditions. To address this, we employed a combination of experimental data, high-throughput computational calculations, and machine learning (ML) to develop predictive models for regioselectivity. Ligand and catalyst descriptors were derived from DFT calculations and literature databases, forming a robust dataset used to train ML algorithms. Further optimization proved effective in guiding experimental efforts by identifying promising ligands and improving hydroboration yields. This workflow integrates experimental and computational tools to achieve a stereocontrolled synthesis of substituted alkenyl boronates from alkynes. As a case study, we demonstrate the successful application of ML-guided optimization, reducing copper catalyst loading while improving yields and regioselectivity.

INTRODUCTION

Vinylboronates are widely employed in a range of C–C bond-forming reactions, including catalytic cross-coupling processes, and are considered highly valuable entities in chemical synthesis.[1,2] One of the most important methods for the synthesis of these building blocks is by transition metal catalyzed hydroboration.[3] Traditionally, metals such as rhodium or iridium and boranes have been employed for these transformation, but the emergence of copper-catalyzed hydroboration has provided a powerful alternative to classical approaches. Copper-based catalysis enables the hydroboration of alkynes under mild conditions, often using bench-stable reagents like $B_2(pin)_2$ (Bis(pinacolato)diboron) or HBpin (Pinacolborane)[4,5]. This strategy, which involves nucleophilic boron species, offers excellent stereoselectivity, typically delivering *trans*-vinylboronates with high precision. Additionally, Cu catalysts are cost-effective and earth-abundant, representing an environmentally sustainable alternative to more expensive noble metal systems.[6] Compared with the traditional cross-coupling reactions, this strategy, in which the vinyl copper species is formed by migratory insertion of the Cu–B across a C–C unsaturation, has the following advantages: (a) abundant and stable boranes are used instead of air- and moisture-sensitive organometallic reagents; (b) the vinyl-Cu intermediate can be further transformed by subsequent reaction with suitable electrophiles; and (c) two different products can be obtained from the same starting materials if a regiodivergent hydroboration is achieved.[7]

However, predicting regioselectivity remains challenging due to the complex interplay of factors such as ligand structure, substrate substitution patterns, and reaction conditions.[8] The regioselectivity of hydroboration reactions is not solely governed by electronic and steric factors of the substrate but is also heavily influenced by the ligand environment around the Cu center and the reaction conditions employed.[9] This complexity is further compounded by the vast diversity of ligands, including phosphines, N-heterocyclic carbenes (NHCs), and chiral bidentate ligands like bis-oxazolines, each capable of modulating the reactivity and selectivity of the copper catalyst in distinct ways.[10] The factors providing a greater stabilization in the transition states are the steric hindrance and the electronic properties of the ligand, along with the electronic characteristics and substitution pattern of the alkyne, which govern the boron insertion on the insaturation.[11] Thus, remarkable differences have been observed between ligands: for example, NHC type ligands promote one selectivity while phosphine ligands the other. Furthermore, the different substitution on the alkyne seems to guide the insertion process, as depicted in Figure 1.

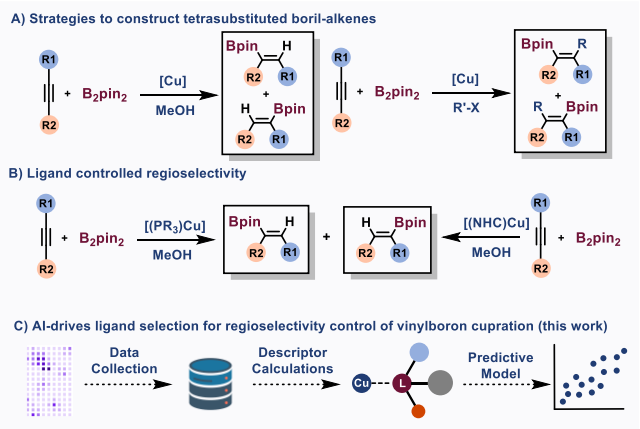


Figure 1. Ligand and reactant controlled regioselective hydroboration.

Given this multifaceted problem, a traditional empirical approach to predict regioselectivity by testing individual ligands and substrates in an experimental setting is both time-consuming and resource-intensive. As a result, alternative methodologies capable of systematically analyzing and predicting regioselectivity are of great interest. Among these, machine learning (ML) has emerged as a powerful tool to address complex problems in chemical synthesis. By utilizing ML algorithms, one can develop predictive models that correlate structural and experimental parameters with reaction outcomes, thus streamlining the design and optimization of regioselective hydroboration reactions.[12,13,8,14,15,16] In particular, the calculation of variables has evolved to incorporate diverse,[17,18,19,20] refined properties,[21] which can be scaled to produce databases.[22] The resulting descriptors can then be fed into a range of data science workflows to optimize a particular objective such as yield, stereoselectivity,[23] or regioselectivity.[24] For simplicity, molecular descriptors will be referred to as "descriptors" throughout this work, and they will be treated as variables in the context of data analysis. This terminology does not specifically refer to descriptors used in machine learning models but is adopted here for linguistic simplicity.

The workflow typically employed in projects of this nature is both well-defined and streamlined. The first step involves data collection, which can be achieved through various approaches.[25] Experimental reactions may be conducted in the laboratory using high-throughput experimentation.[26] Alternatively, data can be sourced from the literature or generated through high-throughput computational calculations. The second step requires comprehensive descriptor collection to characterize ligands, catalysts, and reactants. This involves computational calculations at various levels to analyze the chemical properties and functionalities of the selected systems, including catalysts, reactants, solvents, or additives.[27] A notable example of this methodology is the *Kraken Monophosphines Database*, as reported by Sigman et al.[28] The third step focuses on identifying relationships between the collected descriptors and the target variables by training machine learning models. These models are then employed to predict or classify ligands, catalysts, reactants, or additives based on the specified objective variable. Finally, validation is essential to assess the performance of the predictive models or classification functions. This is achieved by employing error metrics, such as root mean square error (RMSE), to evaluate accuracy, and using cross-validation techniques to ensure robust and unbiased assessments of the dataset.

Lastly, optimizing a chemical reaction involves more than just regioselectivity, requiring the evaluation of multiple parameters such as substrates, catalysts, reagents, additives, solvents, temperature, and reactor type. However, practical constraints in laboratories often limit the exploration of conditions, even with advances in high-throughput experimentation (HTE) that allow the collection of thousands of data points. The vast number of possible configurations makes it challenging to identify optimal conditions. In this context, artificial intelligence, particularly Bayesian optimization, has emerged as a powerful tool to streamline reaction optimization.[29] By efficiently guiding experimental efforts toward the most promising conditions, these techniques enhance efficiency and reduce the need for exhaustive experimentation. Herein, we report *a method for a predictive stereocontrolled synthesis of substituted alkenyl boronic esters from simple alkynes, that combines our experimental results and DFT calculations, along with data obtained from the literature. Further optimization led to increased yields of the hydroboration product, reducing the catalytic amount of Cu while achieving excellent yields*.[30]

RESULTS AND DISCUSSION

1. Experimental data collection.

1. 1. Regioselective synthesis of the α-isomer.

As a case study, we tackled the challenge of predicting the regioselectivity of a reaction using only the SMILES representation of the ligand. Our initial focus was to understand how ligands influence the regioselectivity in copper-catalyzed hydroboration/carboboration reactions of internal alkynes. To achieve this, we conducted a detailed investigation of ligand effects in a model hydroboration reaction. The reaction was carried out at room temperature in toluene, employing CuCl (10 mol%), ligand (12 mol%), and NaO^{t}Bu (15 mol%) as the base (further details are provided in the ESI). For this study, we selected (3-methoxyprop-1-yn-1-yl)benzene **1** as the model substrate due to its distinct advantage of showcasing regio-divergence when tested with various ligand families. This divergence, reflected in the α/β ratio expressed as ΔΔG‡, arises from the two distinct functional groups flanking the alkyne (OMe and Ph), each capable of directing the borylcupration step in opposite directions, and is essential for exploring the full spectrum of regioselective outcomes. The variable ΔΔG‡ (difference in activation free energies) is calculated to compare the energy barriers between two competing pathways in a chemical reaction. It quantifies how much one pathway is favored over the other and is directly related to the ratio of the product distribution (α/β). The relationship can be expressed mathematically as:

$$\Delta\Delta G^{\ddagger} = -RTln(\frac{\alpha}{\beta})$$

In this specific case, the α/β ratio represents the relative regioselectivity of the borylcupration step, which is influenced by the directing effects of the two distinct functional groups (OMe and Ph) flanking the alkyne. A positive or negative ΔΔG‡ indicates which pathway is more favorable, with smaller magnitudes correlating to less pronounced selectivity. It is possible to calculate ΔΔG‡ using activation energies obtained through Density Functional Theory (DFT) calculations or by experimental methods such as differences in integrals observed in ^{1}H NMR spectra. DFT calculations provide activation free energies (ΔG‡) for each pathway, allowing direct comparison of their relative energy barriers. Alternatively, ^{1}H NMR can measure the product distribution (α\β) by integrating the signals corresponding to the regioisomeric products. Since the ratio α/β is linked to ΔΔG‡ through the equation ΔΔG‡=−RTln(α/β)), the experimental data enables calculation of the free energy difference without requiring detailed computational modeling. Both approaches provide complementary insights into the factors governing regioselectivity.

We examined ligand families including phosphines, phosphites, and NHC carbenes, while excluding N-donor ligands due to the poor yields observed under these conditions. The results, detailing both yield and regioselectivity, are summarized in Figure 2. As observed, the reaction yields exceed 90% in most cases, with regioselectivity ratios ranging from 99/1 for PCy_3 to 40/60 for the IPr ([1,3-Bis(2,6-diisopropylphenyl)-imidazol-2-ylidene]) carbene. These results align closely with the ratios previously reported by Hoveyda et al.[11] and our group.[7] Notably, steric hindrance emerges as a critical factor influencing regioselectivity, as demonstrated by the data for IPr and Imes (1,3-Bis(2,4,6-trimethylphenyl)-1,3-dihydro-2H-imidazol-2-ylidene) carbenes. Additionally, electronic effects play a signifi-

cant role, as evidenced by the differences in regioselectivity between triphenylphosphine and its *para*-substituted analogs. This information provides a valuable foundation for training predictive models.

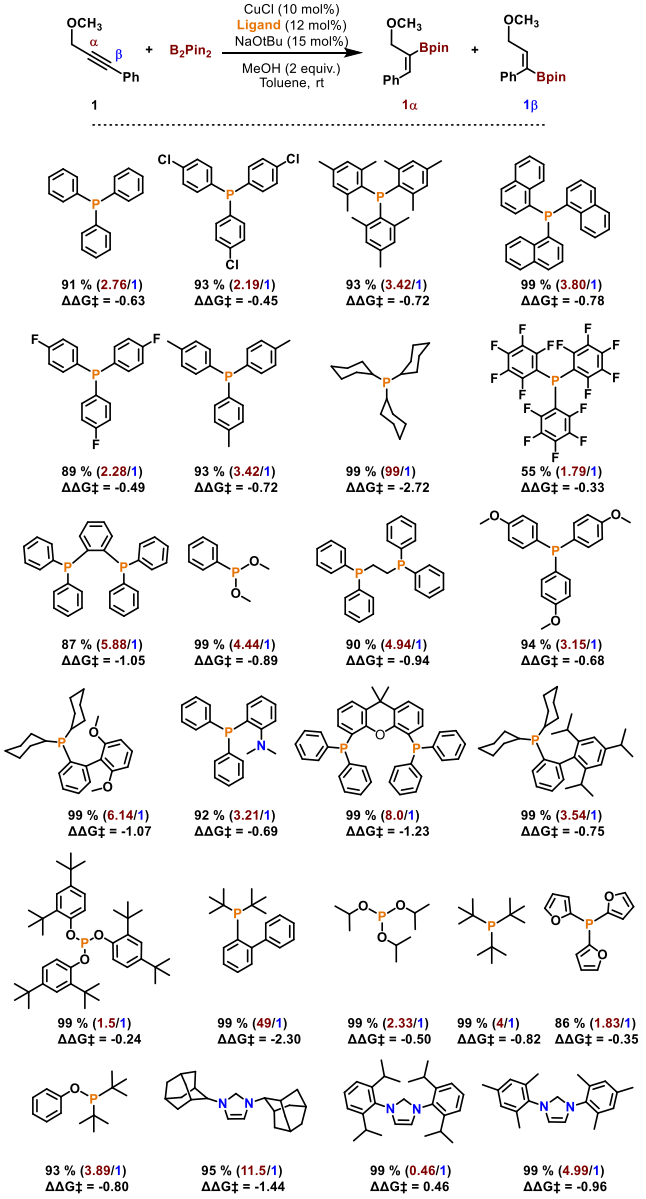


Figure 2 Regioselectivity using different ligands (experimental).

2. DFT data collection and initial ML models.

To further enhance our understanding of regioselectivity, we aimed to expand the ligand dataset and perform DFT calculations. Accordingly, we computed DFT-level geometries for all Ligand-Cu-Bpin adducts and generated complete energetic profiles for 17 selected ligands.

Based on the provided data for PPh_3 case, which is one of the most correlated values between experimental and computation, the α insertion pathway is more favorable than the β pathway due to key differences in the structural changes required to reach the transition state. In the α pathway, the distance between the boron atom and the α-carbon of the alkyne decreases from 3.092 Å in the intermediate to 2.219 Å in the transition state, a change of 0.873 Å. In contrast, the β pathway involves a slightly larger change, from 3.194 Å to 2.191 Å, corresponding to a difference of 1.003 Å. This indicates that the α pathway requires less structural reorganization, suggesting a lower energy barrier for the reaction. Additionally, the transition state angle for the α pathway is slightly larger (68.89°) compared to the β pathway (67.10°). This difference, while subtle, may reflect improved orbital alignment in the α pathway, facilitating stronger interactions between the alkyne and the coordinating species. Combined with the smaller structural changes required along the reaction coordinate, the α pathway benefits from a more efficient and energetically favorable transition to the product. These factors explain why the α pathway is preferred over the β pathway under the studied conditions in this specific case.

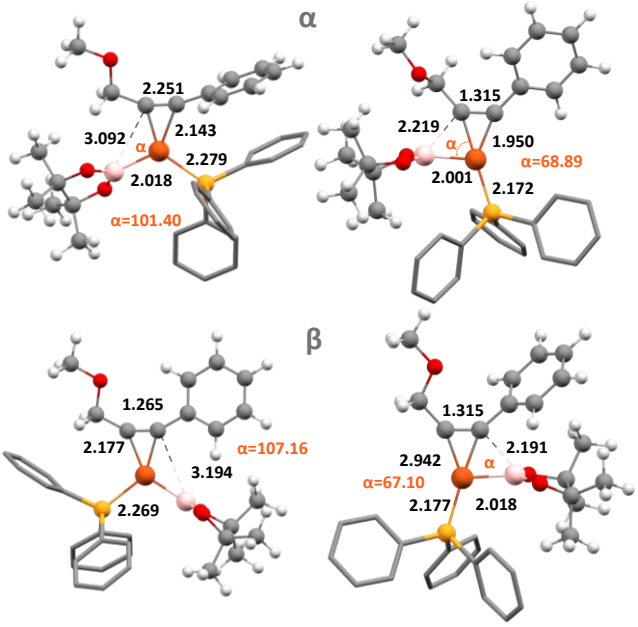


Figure 3. Molecular structure of the transition states TSα and TSβ and the previous intermediates for PPh_3. Representative distances (Å) and angles (°) are indicated.

These calculations contribute critical descriptors for capturing the nuanced interplay of steric and electronic effects in the regioselectivity process. DFT calculations offer valuable insights for mechanistic investigations, providing access to information on reaction intermediates and transition state energies. In our study, we compared the calculated energies of the experimental ligand scope with our energetic profile and observed that, despite the intrinsic error associated with DFT, the predicted major regioisomer generally aligns with the experimentally observed outcomes. This consistency enables us to integrate these computational data into the experimental dataset, enhancing its robustness. The primary results obtained in this regard are summarized in Figure 4. As illustrated in Figures 3 and 4, the α/β ratios generated by DFT calculations closely match the experimental results. For instance, the $P(p\text{-tol})_3$ ligand produced an experimental value of -0.72, compared to -0.718 from DFT calculations. Similarly, the dppe bisphosphine ligand yielded a ΔΔG‡ of -0.94, consistent between ^1H NMR measurements and DFT predictions. However, some discrepancies were observed, such as with $P(C_6F_5)_3$, which provided an experimental value of -0.33 versus -0.55 from DFT. Overall, the observed trends are highly consistent across both methods, validating the incorporation of the entire dataset for training a machine learning model capable of predicting regioselectivity based on the ligand structure.

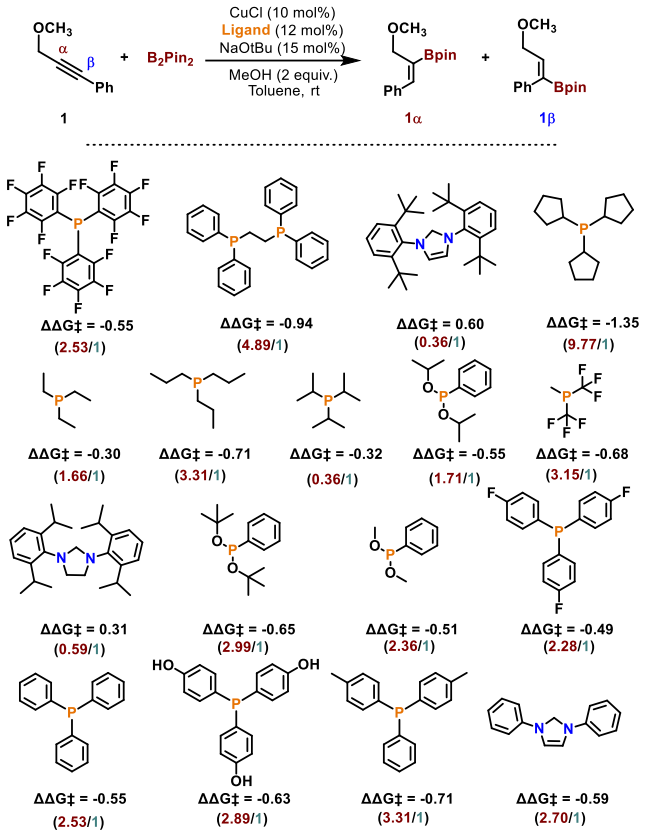


Figure 4. Stereodivergence using different ligands (DFT level, (B3LYP/6-31G+(d,p) (C,H,B,O,P,N,F), SDD (Cu), 298K, 1atm).

2.1. Workflow for the calculation of ligand/complex parameters.

To optimize the catalyst, we relied on calculated ligand features and developed a comprehensive descriptor library. This library includes commercially available or widely studied ligands such as monophosphines, bisphosphines, phosphites, and carbenes, along with synthetically accessible derivatives. Quantum mechanical methods were employed to calculate geometries and descriptors for a diverse range of ligands, using a linear Ligand-Cu-Bpin complex as model system. [31] The computational workflow (depicted in Figure 5) began with a molecular mechanics-based method to generate conformations of the model complex. These initial conformations were subsequently optimized using DFT to obtain accurate structural representations. From these DFT-optimized structures, steric, electronic, and geometric parameters were extracted. Additionally, the software Morfeus,[31] adapted to operate at the GFN2-xTB level of theory, was utilized to enhance descriptor collection, ensuring a robust dataset for subsequent analyses.

2.2. Machine Learning models.

Generating sufficient data is a significant challenge for chemists aiming to leverage Artificial Intelligence (AI) to enhance their systems. While modern High-Throughput Experimentation (HTE) and High-Throughput Calculations (HTC) enable the rapid acquisition of substantial datasets, these approaches remain uncommon outside the pharmaceutical industry. As a result, many studies operate under low-data regimes with fewer than 100 data points, making the development of accurate and robust predictive models particularly challenging. In our case, the application of experimental methods and computational calculations allowed us to compile a dataset containing descriptors for 34 ligands (Figure 6). This dataset was subsequently used to train various predictive models to estimate regioselectivity in our system. The models tested included ridge regression, support vector regression, random forests, and Gaussian processes. Among these, the Gaussian process model delivered the most accurate predictions, with the lowest error metrics. By contrast, while random forest and ridge regression models are widely used in similar projects,[29] they produced predictions that deviated further from the experimental results. These findings underscore the importance of model selection when working in low-data environments.

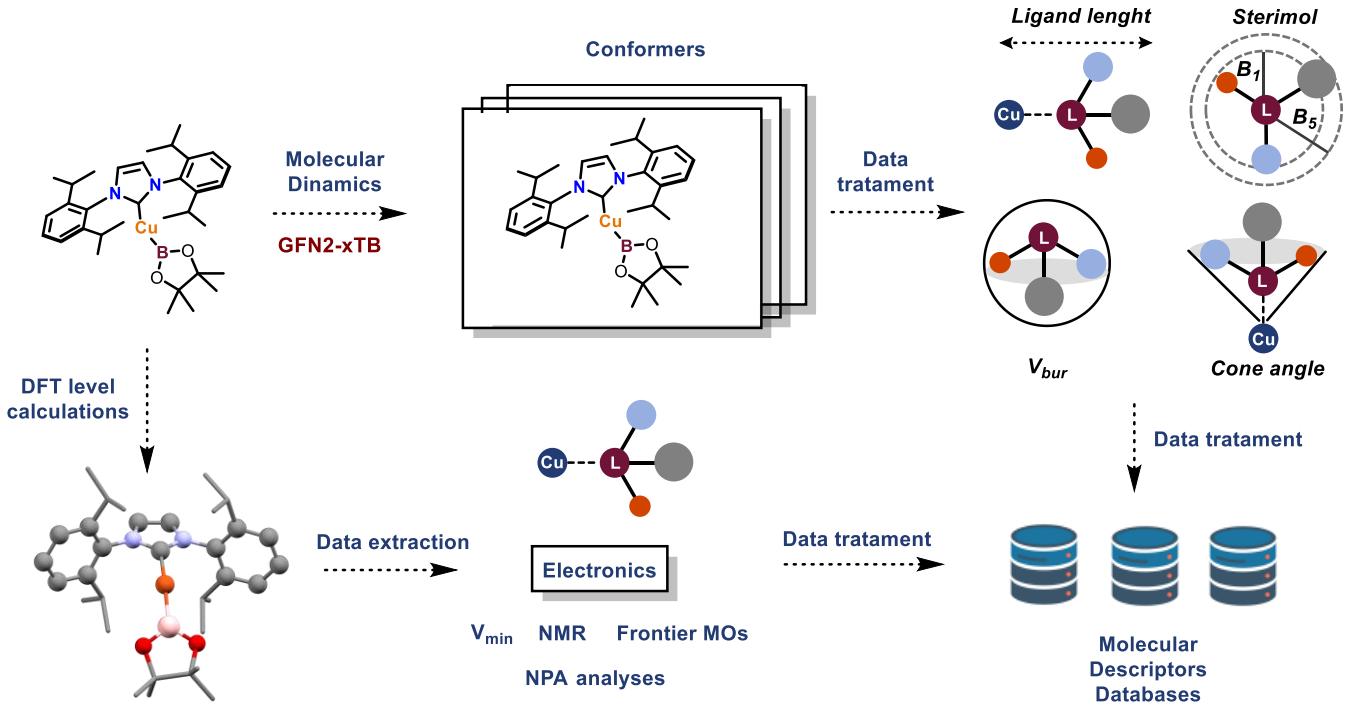


Figure 5. Computational workflow for the descriptors collection.

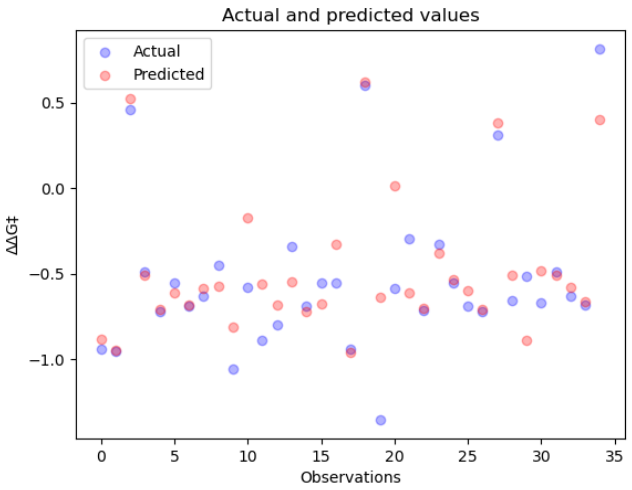


Figure 6. Actual *vs* predicted values of the obtained data from experimentation and calculations using Gaussian process.

Given the limited dataset of 34 data points, we utilized a Leave-One-Out Cross-Validation (LOOCV) approach to validate the predictive models. This approach involves training as many models as there are data points, excluding one instance from the training set during each iteration and treating it as unseen data for prediction. This method closely simulates real-world scenarios and ensures that every data point is used for both training and testing. Once all evaluations are completed, the results are averaged to produce a final RMSE score for the model. Although LOOCV can be computationally intensive for larger datasets due to the high number of models it requires, it is particularly advantageous in low-data regimes, where splitting the dataset into distinct training and testing subsets could result in the loss of valuable instances. Using this approach, the Gaussian process model demonstrated the best predictive performance, achieving a Root Mean Square Error (RMSE) of 0.14. Our results, showcasing the relationship between actual and predicted values for each ligand across all cross-validation iterations, are compiled in Figure 7.

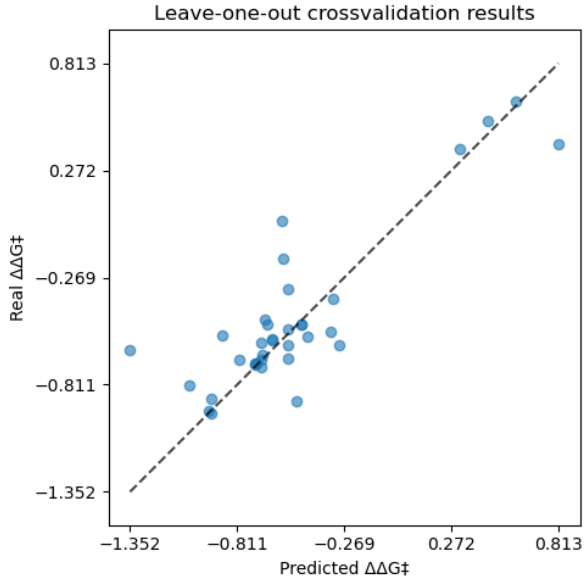


Figure 7. Actual *vs* predicted values of the obtained data from experimentation and calculations using Gaussian process with LOOCV. The ratio displayed on the Y-axis represents the value of ΔΔG‡

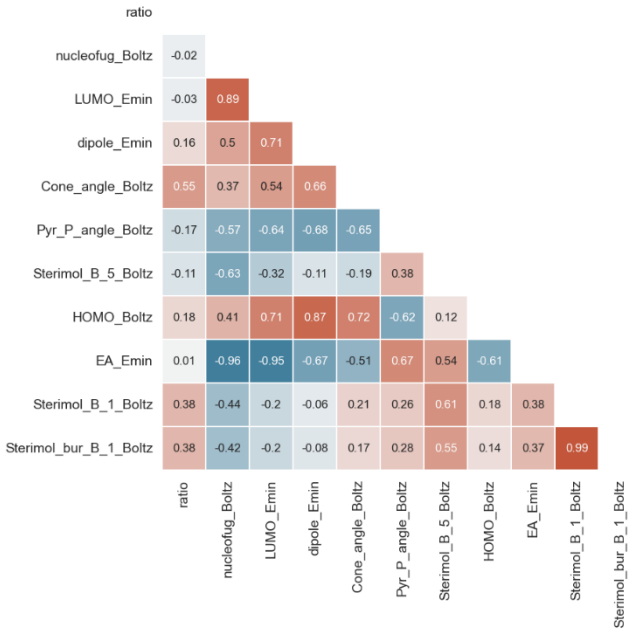


Figure 8. Heatmap of the selected descriptors of the ligand employed to estimate the α/β ratio.

The heatmap illustrates variables with high correlations, which is attributed to the fact that certain variables are derived from the same molecular property calculated under different conditions. For instance, the cone angle is represented in its minimum energy structure, across a set of conformers, and as the maximum achievable cone angle based on conformational sampling. This overlap in variable definitions naturally results in high correlations. The inclusion of these descriptors, despite their interrelation, is justified as they emerged as the most significant after a thorough variable selection and cleaning process. These descriptors were retained because they capture essential structural or electronic characteristics of the system and exhibit strong predictive potential in the context of the analysis. Their redundancy can later be addressed using dimensionality reduction techniques or feature selection in subsequent modeling stages, ensuring that their importance to the model is preserved while minimizing collinearity.

Additionally, we analyzed the descriptors utilized in the Gaussian process predictions to gain deeper insights into the factors influencing regioselectivity. By visualizing the data used by the predictive model in a heatmap, it becomes evident that *the electronic properties and steric hindrance of the ligands significantly impact the regioselectivity of the insertion step*. As illustrated in Figure 8, key parameters such as nucleofugality, cone angle, LUMO energy, and the Sterimol B5 descriptor are the most influential contributors. These findings align with prior studies on the insertion process in copper-catalyzed hydroboration and carboboration of alkynes, reaffirming the critical role of steric and electronic factors. However, this investigation provides a more detailed understanding of ligand contributions, offering valuable insights into how specific descriptors modulate the reaction outcome. Such knowledge underscores the utility of combining computational models with experimental data to refine our understanding of catalytic processes.

2.3. Application of our model to the regioselective synthesis of the β-isomer.

Achieving a predominantly β-regioselective product poses a significant challenge due to the low levels of regioselectivity

observed in prior experiments with our model substrate. However, the integration of AI with advanced mechanistic insights should enable the identification of ligands capable of favoring the formation of this regioisomer with high selectivity. To address this, we utilized AI-driven tools to guide the search for promising ligands. To that end, we evaluated over 100 commercially available ligands, including carbenes and phosphines, to identify those predicted to optimize regioselectivity. Using predictions generated by the Gaussian process model, we identified the top-performing ligands based on $\Delta\Delta G^{\ddagger}$ values (Figure 9 and 10). Consistent with expectations, the ligands predicted to enhance β-regioselectivity were all carbene-type ligands characterized by high steric hindrance, highlighting their potential in directing the regioselectivity of the insertion step. Through this analysis, we identified several ligands based on the IPr carbene framework that hold potential for improving the regioselectivity achieved to date Among these, as illustrated in Figure 9, ligand **L16** exhibited a predicted ΔΔG value of 0.6. This corresponds to an approximate regioselectivity ratio of 30:70, representing a slight improvement over the results obtained with the IPr ligand, which yielded a ratio of approximately 35:65.

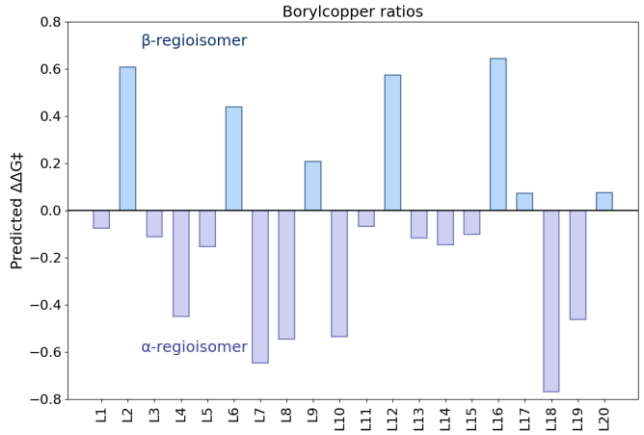


Figure 9. Prediction on commercially available ligands and recommended ligands (**L2**, **L6**, **L12** and **L16**) based on the output.

These results are in good agreement with both computational predictions and experimental trends, as well as with the key descriptors highlighted in the heat map in Figure 8. Particularly noteworthy are ligands **L2**, **L6**, **L12**, and **L16**, which strongly favor the formation of the β-regioisomer. Conversely, ligands not previously encountered by the model, such as Buchwald-type phosphines, display significant values favoring the α-regioisomer, although close to a statistical mixture. This behavior is likely due to the model's lack of prior data on this class of ligands, leading it to make conservative predictions. Lastly, it is striking that ligands **L10** and **L19**, despite having steric profiles significantly larger than that of the IPr carbene (characteristics that would seemingly make them ideal for β-regioisomer formation) exhibit the opposite behavior, with predictions favoring the α-regioisomer as the major product.

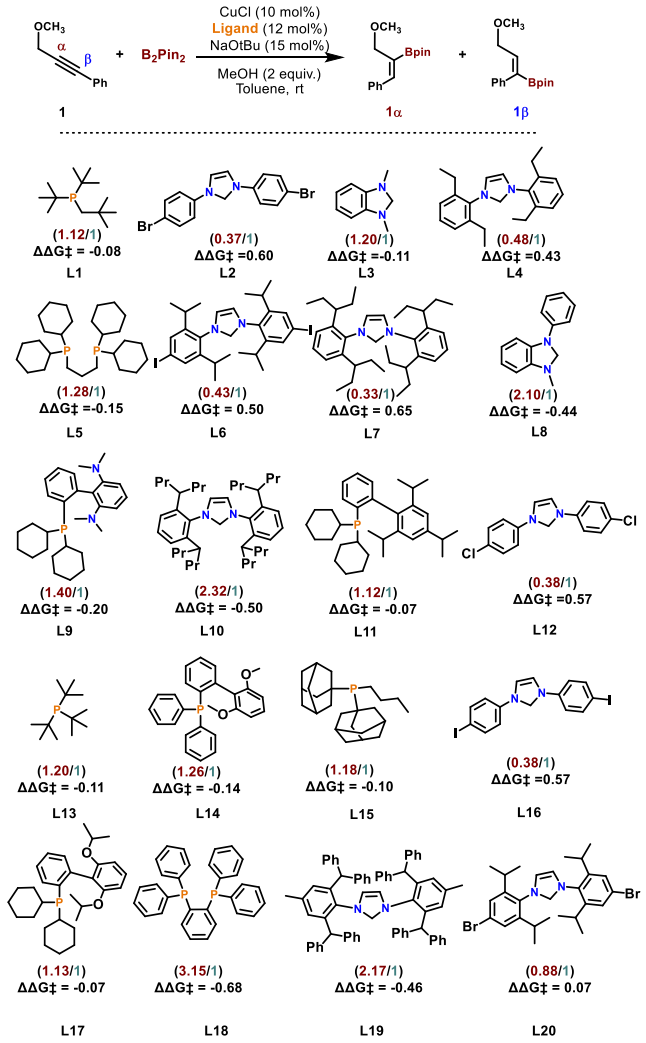


Figure 10. Regioselectivity using different ligands (predicted).

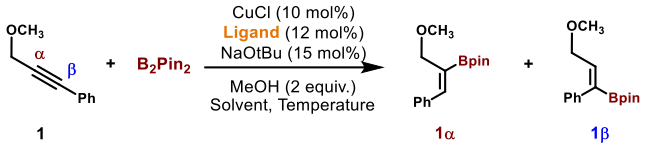
2.4. Optimization: final adjustments

Achieving high regioselectivity while reducing catalytic loading is a key challenge in Cu-catalyzed hydroborations. Traditional loadings of ~10% are necessary for acceptable rates and yields but lead to increased costs, environmental impact, and by-product formation. Oxidative decomposition of copper catalysts under reaction conditions likely contributes to this limitation. To address these issues, we integrated experimental data, AI-assisted analysis, and a trained machine learning model. This approach reoptimized reaction conditions to enhance β-regioselectivity and minimize catalytic loading without sacrificing efficiency or selectivity, promoting more sustainable catalytic systems for hydroboration reactions.

Results compiled in Table 1 show that regioselectivity is influenced primarily by the ligand, though solvent and temperature also play roles. Using the IPr ligand, tetrahydrofuran (THF) yielded the best combination of conversion and regioselectivity (30:70), outperforming toluene (35:65). Coordinating solvents like DMF showed similar regioselectivity but reduced yields, while chlorinated solvents like CH_2Cl_2 showed negligible conversion, likely due to poor solubility of catalytic intermediates. Lowering temperature enhanced β-regioisomer formation (20:80 at 0°C in THF) but reduced overall conversion, dropping below 10% at -10°C. Further optimization allowed for a reduction in catalyst loading to 5% by extending the reaction time to

24 hours at 0°C in THF, maintaining the 20:80 regioisomer ratio. Furthermore, the catalytic load can be reduced to 0.5 mol% by extending the reaction time to 24 hours at room temperature. Remarkably, this adjustment does not compromise the regioselectivity of the reaction, maintaining the same high level of precision in product distribution. This highlights the efficiency and robustness of the catalytic system, even under reduced catalyst concentrations, offering a more sustainable and cost-effective approach to the reaction.

Table 1. Optimization process for hydroboration reaction using different conditions.



1 1α 1β

Entry	Ligand	Solvent	T	Time	Yield	Ratio
1	IPr	Toluene	25	3 h	99%	35/65
2	IPr	CH_2Cl_2	25	3 h	>1%	-
3	IPr	DMF	25	3 h	70%	34/66
4	IPr	THF	25	3 h	85%	30/70
5	IPr	THF	10	6 h	70%	25/75
6	IPr	THF	0	6 h	72%	20/80
7	IPr	THF	-10	12 h	31%	26/74
8	Iodo-IPr	THF	25	3 h	93%	30/70
9	Iodo-IPr	THF	0	4 h	70%	21/79
10	IPr	THF	25	3 h[a]	63%	31/69

0.5 mmol scale in substrate. The yield was determined by ^{1}H NMR from the crude mixture, using 1,3,5-Trimethoxybenzene as internal standard. [a] [Cu] = 0.5 mol%

These results represent significant progress toward more sustainable and efficient catalytic hydroboration processes. In the absence of a ligand, the reaction demonstrates significantly low conversion rates and yields, achieving only approximately 17% in THF at 25°C after 3 hours. Moreover, the regioselectivity observed under these conditions shows an alpha/beta ratio of around 70/30. These suboptimal results can likely be attributed to the poor solubility of the salt in the reaction medium, which limits its availability for catalytic activity, as well as the inherent instability of the catalytic species under these conditions. These factors collectively hinder the efficiency and selectivity of the reaction, emphasizing the critical role of a ligand in stabilizing the catalytic species and improving the overall process. The introduction of the novel Iodo-IPr ligand, proposed by the machine learning model, enhanced regioselectivity compared to the standard IPr ligand. While the IPr ligand achieved a regioselectivity of 35:65 under standard conditions, the Iodo-IPr ligand improved this to 30:70, representing a slight shift toward the desired β-regioisomer. Furthermore, a modest improvement in regioselectivity was observed by lowering the reaction temperature, with a more pronounced β-selectivity (20:80) at 0°C, albeit at the cost of reduced conversion. The choice of solvent also proved critical, with THF consistently delivering the best balance of conversion and regioselectivity compared to other solvents such as toluene or DMF. Notably, it was demonstrated that the catalyst loading could be reduced tweentyfold (to 0.5 mol%) by extending the reaction time to 4 hours at room temperature, maintaining the high level of regioselectivity achieved with higher catalyst concentrations. These findings highlight the robustness and efficiency of the catalytic system, underscoring the potential of ML-designed ligands, such as Iodo-IPr, to drive advancements in sustainable and efficient hydroboration processes.

3. Data extracted from scientific literature.

3.1. Nature of the alkyne.

Although the results obtained using different ligands with the same reactant are capable of correctly explaining the outcomes, it is well-known that the nature of the alkyne is critical in determining the regioselectivity of such processes. Terminal and internal alkynes exhibit markedly different behaviors, further compounded by the complexity observed with different ligands. For this reason, and considering the information gathered over the past decades in the literature, we undertook the task of compiling regioselectivity data from the most relevant studies in this field.[29,30,31] In this endeavor, we collected more than 100 examples involving diverse alkynes, aiming to cover the widest chemical space possible, ranging from terminal and internal alkynes to those with different functional groups, including directing groups. This provided a comprehensive perspective on the borylcupration process. Among these varied examples, different ligands, such as IPr, IMes, P(p-Tol)$_3$ or P(Cy)$_3$, were employed. Generally, the dataset includes cases exhibiting regiodivergent behaviors; however, to provide a complete view, we also incorporated some results demonstrating complete regiocontrol over the reaction. It is worth noting that not all results showing complete regioselectivity could be included in the model, as the number of structurally diverse reactions exhibiting absolute regioselectivity towards the substrate is far greater than those showing regiodivergence. Including all these cases would result in an unrepresentative data distribution for the main objective of this study. Consequently, we propose a general distribution, as depicted in Figure 11, where we show a histogram of the number of instances in our dataset based on their ΔΔG‡ and grouped by terminal alkynes, internal alkynes with a methyl group and internal alkynes, categorized as aryl-aryl, aryl-alkyl, or alkyl-alkyl with varying degrees of substitution and functional groups.

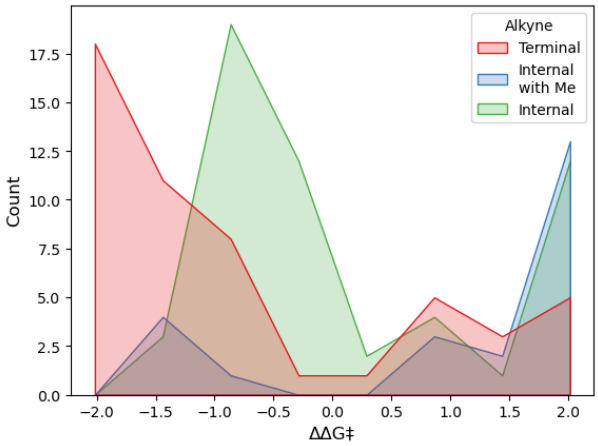


Figure 11. Diagram of the different nature of alkynes in our dataset.

With all this information extracted from the literature, we repeated the workflow described above to generate molecular descriptors, this time focusing on the alkynes. Using molecular dynamics simulations and DFT calculations, we obtained various steric and electronic descriptors for the different alkynes. This process allowed us to construct a comprehensive database

of alkynes and catalysts to explain the behavior of the reaction ratio.

3.2. Machine Learning models.

Once this phase was completed, we tested several machine learning models to explain the regioselectivity of each alkyne based on the ligand used in the insertion process. After extracting the descriptors of each ligand and reagent, our final dataset consisted of 128 instances and 113 possible variables that can be used to predict the regioselectivity of the process. As such, given the high number of potential ligand and reactive descriptors that can be used, each tested model underwent two subsequent feature subset selection methods: a Greedy Forward Selection and a Genetic Algorithm Selection. The greedy forward algorithm starts from an empty set of features and adds variables one by one based on the accuracies obtained with each one, until no variable can be found that improves previous results. This first selection serves as a baseline of some of the most important descriptors for each model. Afterwards, a genetic algorithm is applied to try to improve the subset of descriptors found by the greedy forward algorithm. Once a promising subset of descriptors is found, a hyperparameter tuning based on differential evolution[31] is performed for each model that allows it. Among all the models evaluated, the Random Forest (RF) algorithm provided the best results, achieving a RMSE of 0.4 and an R^2 of 0.89 as shown in Table 2 and in Fig. 12. The instances distributed in vertical lines both at -2.3 and 2.3 ΔΔG‡ values are different alkynes from the literature with reported 1 to 99 or 99 to 1 regioselectivity. This translates into the same ratio value for many alkynes with different descriptors, which in turn makes models have inaccuracies with that specific shape in those extremes.

	Ridge	*GP*	*RF*	*SVR*
RMSE	0.62	0.54	0.40	0.53
R^2	0.73	0.81	0.89	0.82

Table 2. Prediction results for different models fitted to our dataset.

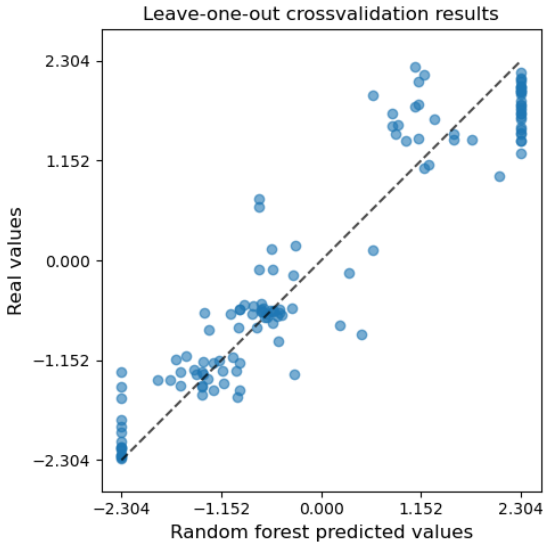


Figure 12. Observed *vs.* predicted results for the Random Forest model using leave-one-out cross-validation.

The Random Forest algorithm is particularly well-suited for modeling and explaining behaviors in metal-catalyzed chemical reactions using electronic and steric molecular descriptors of both reactants and catalysts. RF is an ensemble learning method that builds multiple decision trees during training and combines their outputs to improve predictive accuracy and robustness. Its ability to handle high-dimensional, non-linear, and complex datasets makes it an excellent choice for studying chemical systems, where multiple factors often interplay in determining the outcome of a reaction. It is important to note that, in low data regimes like this one, we need to constraint the hyperparameters of the model so that it can be able to properly generalize its results to unseen instances and it does not end up overfitting. Parameters like the depth of the internal decision trees and the number of trees in the forest should be kept relatively low, while the minimum number of instances per leaf node and the minimum number of instances to perform splits should be reasonable in comparison with our total number of instances.

In the context of metal-catalyzed reactions, electronic and steric descriptors provide critical information about the nature of the reactants, intermediates, and catalysts, as well as their interactions. These descriptors often exhibit non-linear relationships with the reaction outcome, such as yield, regioselectivity, or stereoselectivity. These non-linear relationships between descriptors and objective variables go beyond the traditional understanding of reaction mechanisms due to the complexity of their associations. For this reason, the implementation of such studies is of vital importance to further advance and delve into the complexity of reaction mechanisms. With the inclusion of the literature data, the shape of our dataset changed drastically. When instances share the same ligand or reactive, they share the same descriptors too, which in turn means that now there are blocks of instances where values from a continuous space remain constant. In a continuous space like that of our descriptors, we explore a kind of discrete space mapped from each one of the studied alkynes. An ensemble model like RF is particularly well suited to this kind of scenario, and can capture the non-linearities effectively due to its tree-based architecture. Its underlying decision trees perform similar cuts in the space to those already present in our data due to its unique nature, and this helps in obtaining the best predictive results of all the tested models. Another advantage of RF is its robustness to overfitting, especially when the number of samples is relatively low compared to the number of descriptors, a common scenario in reaction datasets. This robustness arises from the bootstrap aggregation technique, which reduces variance by averaging predictions from multiple trees.

3.3. Mechanistic lessons extracted from the more relevant descriptors.

Furthermore, RF inherently evaluates feature importance, allowing researchers to identify which descriptors—such as electronic charge distributions, steric hindrance parameters, or coordination geometries—have the most significant influence on the reaction. This allowed us to uncover certain explanations for the reaction behavior by considering the nature of the predictive model employed. For instance, some of the most utilized descriptors in our RF model's decision trees to explain the regioselectivity of the hydroboration are: i) The ligand's electronegativity, directly related to its electron-donating or electron-accepting character, highlighting significant differences between phosphines and carbenes. ii) Steric factors of the ligand derived from Sterimol parameters. iii) The dipole moment of the alkyne, which is directly related to the degree of

substitution of the alkyne and the presence or absence of functional groups. iv) The HOMO energy of the alkyne, associated with its electronic density. In the following paragraphs, we represent the ΔΔG‡ relative to these descriptors.

3.3.1. ΔΔG‡ vs. electronegativity of the ligand.

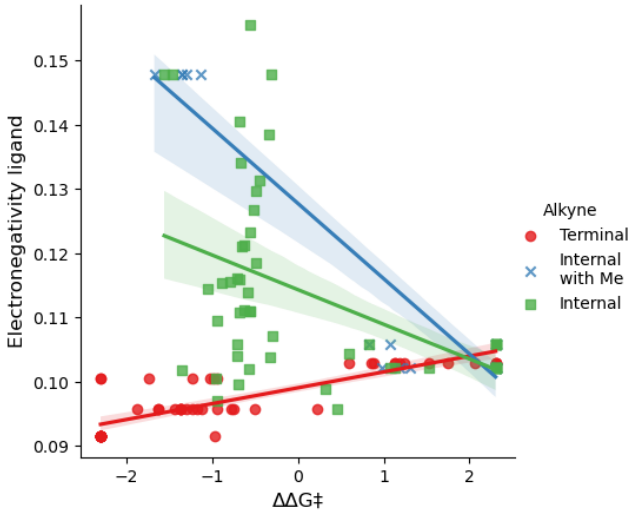


Figure 13. Ratio *vs.* ligand electronegativity grouped by alkyne.

Firstly, considering the ligand's electronegativity in Fig. 13, we observe that terminal alkynes exhibit an almost linear relationship with this descriptor. This shows how an apparently non-linear space can give rise to simpler subspaces that can be fitted by the RF algorithm, and it indicates that *for electronegativity values above 0.10, the results -regardless of the nature of the terminal alkyne- show a clear dominance of the α-regioisomer over the β-regioisomer. In cases of higher electronegativity, the α-regioisomer is exclusively formed. Conversely, for lower electronegativity ranges, the opposite behavior is observed.*

Additionally, this behavior is exclusive to this descriptor. As shown in Fig. 14, it cannot be simplified to the nature of the ligand by merely dividing them into carbenes and phosphines, as no clear relationship is observed.

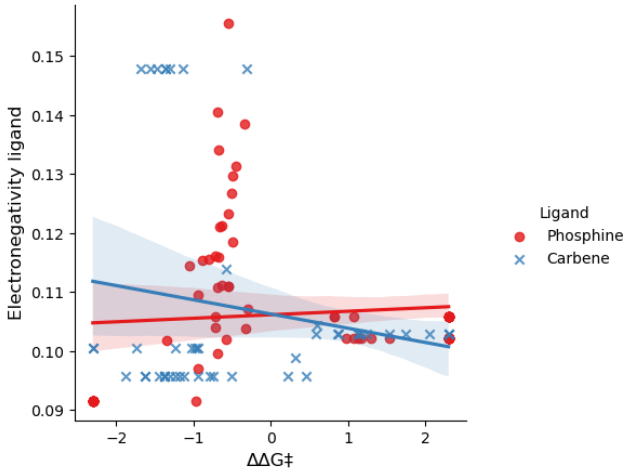


Figure 14. Ratio *vs.* ligand electronegativity grouped by ligand type.

3.3.2. ΔΔG‡ vs. alkyne structure.

On the other hand, certain trends can be observed in the case of internal alkynes substituted with a Me group or internal alkynes with varying degrees of substitution when exclusive reactant descriptors, such as dipole moment or HOMO energy, are used. Again, the data demonstrates a *clear linear correlation between terminal alkynes and the ligand's electronegativity*, as shown in Figure 13 by the red points in the graph. This strong correlation likely arises from the electronic asymmetry of terminal alkynes, where one carbon in the triple bond exhibits a partial negative charge. This inherent charge distribution makes the regioselectivity of terminal alkynes particularly sensitive to the electronic properties of the coordinating ligand, leading to straightforward predictive trends.

For internal alkynes, however, no such correlation is observed with ligand electronegativity. Instead, trends emerge when molecular descriptors such as *dipole moment* and *HOMO energy* (Figures 15 and 16, respectively) are employed. The variability of these descriptors in internal alkynes is heavily influenced by the nature of their substituents, highlighting the interplay between electronic and steric effects. Methyl-substituted internal alkynes generally introduce minor steric hindrance and modest electronic effects, resulting in relatively subtle modulations of dipole moment and HOMO energy. In contrast, bulkier substituents or different functional groups with strong electron-donating or electron-withdrawing capabilities (e.g., ethers, amines, amides, or esters) significantly alter these descriptors. Functional groups like ethers and esters tend to increase the dipole moment due to their polar nature, while amines and amides can act as *electron-donating groups* via inductive or resonance effects, shifting the electronic environment around the triple bond. Larger, sterically demanding groups can also influence the spatial accessibility of the reactive site, adding further complexity to the regioselectivity prediction.

These effects are particularly challenging to capture through simple linear correlations because the relationships between substituent properties and regioselectivity are highly non-linear. This is where the strength of the RF model becomes evident. RF models are ensemble learning techniques that operate by constructing multiple decision trees, each trained on random subsets of the data, and then aggregating their predictions. Unlike linear regression models, RF does not assume any specific functional form of the relationships between variables. Instead, it relies on decision boundaries determined by splitting the data iteratively based on the most significant features at each step. This structure allows RF to account for complex interactions between multiple descriptors, such as steric and electronic effects, without requiring explicit linear dependencies. As a result, while the trends in dipole moment or HOMO energy provide valuable insights, the RF model integrates these descriptors alongside other features to uncover nuanced patterns that may not be immediately apparent from the individual graphs. This also explains why the predictions for internal alkynes reflect subtle shifts influenced by substituent effects, even in the absence of clear linear relationships. Conversely, the simpler electronic behavior of terminal alkynes leads to more straightforward trends that align closely with ligand-based descriptors like electronegativity, making their behavior easier to interpret directly.

In sharp contrast with these observations, no relationships are observed *for terminal alkynes: the impact of substituents on dipole moment or HOMO energy is minimal due to the absence of adjacent substituents on the triple bond*. The regioselectivity in terminal alkynes is thus governed primarily by the electronic

interactions between the ligand and the alkyne, rather than by internal structural modifications. This distinction underscores why ligand-based descriptors like electronegativity yield robust correlations for terminal alkynes, while internal alkynes require a more detailed analysis of substituent effects to predict regioselectivity effectively.

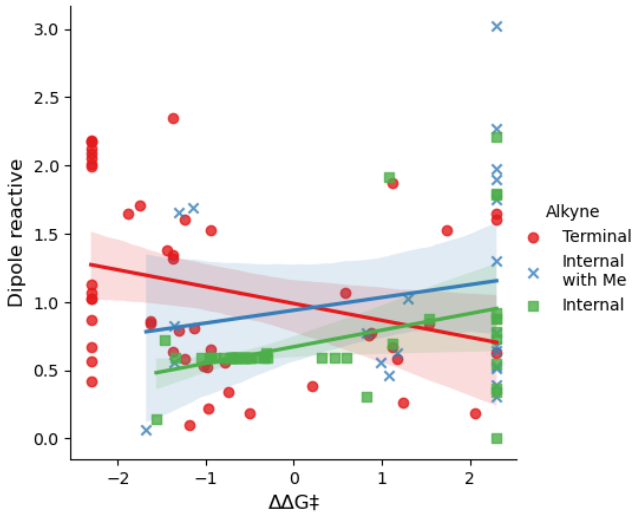


Figure 15. Ratio *vs.* dipole moment of the alkyne.

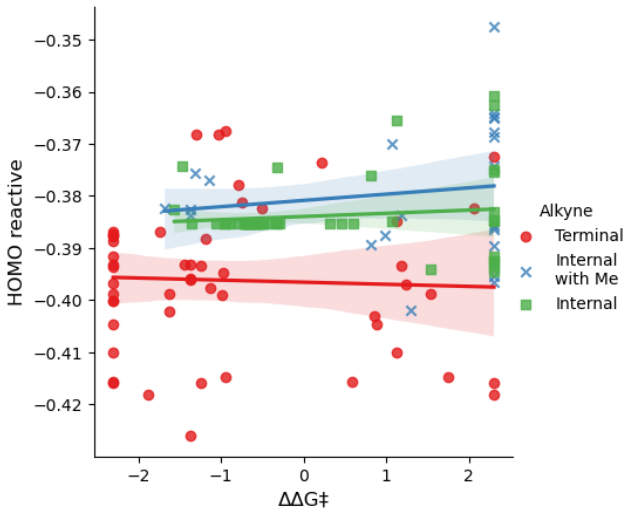


Figure 16. Ratio *vs.* HOMO energy of the alkyne.

3.3.3. Real vs. predicted ΔΔG‡ of phosphine and carbene ligands.

The results above not only highlight the complexity of descriptors and the linear and non-linear relationships that exist when attempting to explain the behavior of a chemical reaction but also demonstrate the promising potential of AI-based predictive tools for elucidating complex processes. Furthermore, this opens the door to mechanistic explanations that are much more intricate than those traditionally used in such processes, providing richer information and offering a more nuanced and comprehensive perspective on chemical processes. In this way, it is shown that with just over 40 distinct ligands and 100 different alkynes, a better understanding of a complex process such as the borylcupration of alkynes can be achieved. The final results obtained show that the fitted RF model is able to predict the regioselectivity of the reaction accurately for both phosphine and carbene ligands, whose performance can be seen separately in Fig. 17.

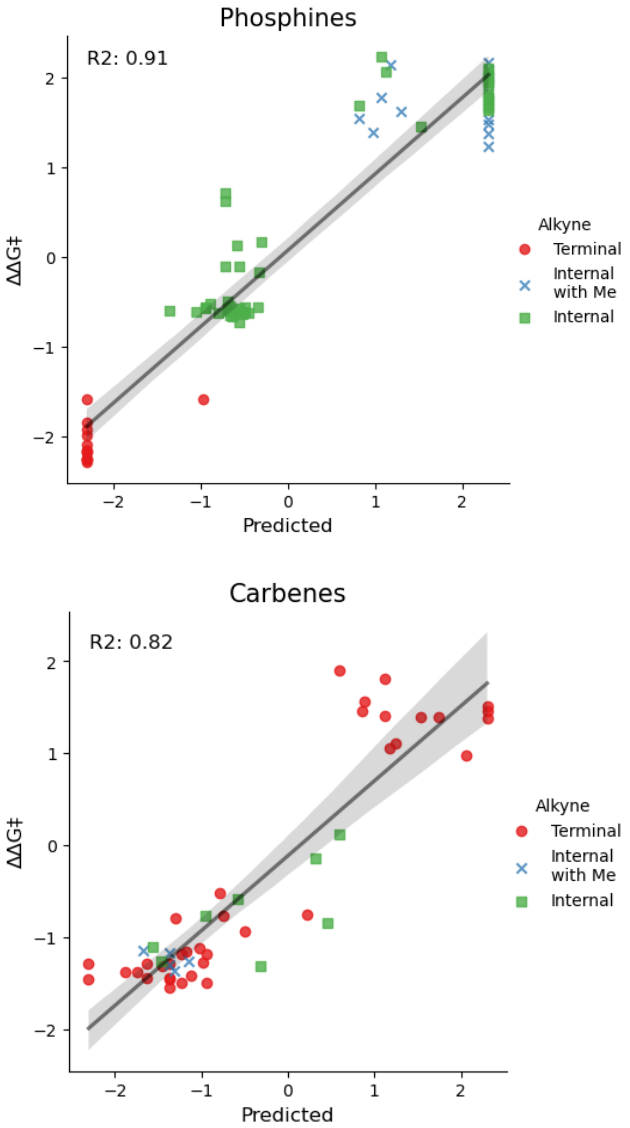


Figure 17. Real *vs.* predicted ratios of phosphine and carbene ligands.

CONCLUSION.

This study demonstrates the feasibility of predicting the regioselectivity of organometallic transformations, i.e. the Cu-catalyzed hydroboration of alkynes, using Artificial Intelligence tools. A Random Forest machine learning model was developed and trained with experimental regioselectivity data for a diverse set of ligands, enriched with information from the existing literature. The results reveal that regioselectivity can be effectively explained and predicted using simple molecular descriptors, such as electronic and steric parameters, easily obtained through established computational protocols like molecular dynamics and DFT-based approaches. The key findings of this study on the insertion mechanism and behaviour are as follows:

1. The electronegativity of the ligand emerges as a crucial parameter, particularly for terminal alkynes, where its influence is paramount.

2. In contrast, the behaviour of internal alkynes is better explained using descriptors such as the dipole moment and the energy of the HOMO orbital, reflecting the impact of substituents or functional groups on both sides of the triple bond.

3. Furthermore, the nature of the ligand proves to be decisive; descriptors related to electron density and steric hindrance, such as the cone angle or ligand size, play a pivotal role in determining the regioselectivity of the process.

This highlights a clear and direct relationship between simple molecular descriptors and complex catalytic phenomena. Without the advanced tools of AI, uncovering these intricate relationships would have been impossible, as they require processing and analysing vast amounts of data far beyond traditional methods.

This model not only provides a deeper and more comprehensive mechanistic understanding compared to traditional methods but also highlights the transformative potential of AI-driven methodologies in chemistry. By enabling synthetic chemists to access predictive tools, it becomes possible to systematically select the optimal ligand for a specific substrate or predict the outcome for any ligand-substrate combination, significantly reducing trial-and-error experimentation. This approach accelerates the discovery and optimization of catalytic systems while fostering a deeper understanding of reaction mechanisms, ultimately contributing to the development of more efficient and sustainable chemical processes.

ASSOCIATED CONTENT

.

AUTHOR INFORMATION

Corresponding Author

Guillermo Marcos-Ayuso - Department of Organic Chemistry, Faculty of Science; Institute for Advanced Research in Chemical Sciences (IAdChem); and Centro de Innovación en Química Avanzada (ORFEO–CINQA), Universidad Autónoma de Madrid (UAM), 28049 Madrid, Spain;

AItenea Biotech, Parque Científico de Madrid, Ciudad Universitaria de Cantoblanco, Calle Faraday, 7, 28049 Madrid, Spain

orcid.org/0000-0002-9443-578X Email: Guillermo.marcos@aitenea.es

Pablo Mauleón – Department of Organic Chemistry, Faculty of Science; Institute for Advanced Research in Chemical Sciences (IAdChem); and Centro de Innovación en Química Avanzada (ORFEO–CINQA), Universidad Autónoma de Madrid (UAM), 28049 Madrid, Spain; orcid.org/0000-0002-3116-2534; Email: pablo.mauleon@uam.es

Ramón Gómez Arrayás – Department of Organic Chemistry, Faculty of Science; Institute for Advanced Research in Chemical Sciences (IAdChem); and Centro de Innovación en Química Avanzada (ORFEO–CINQA), Universidad Autónoma de Madrid (UAM), 28049 Madrid, Spain; orcid.org/0000-0002-5665-0905; Email: ramon.gomez@uam.es

Author Contributions

The manuscript was written through contributions of all authors. / All authors have given approval to the final version of the manuscript. / ‡These authors contributed equally. (match statement to author names with a symbol)

Funding Sources

Ministerio de Ciencia e Innovación (MICINN) and for financial support (Agencia Estatal de Investigación/Projects PID2021-127655NB-I00, and TED2021-129970B-C22).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

We thank the Ministerio de Ciencia e Innovación (MICINN) and for financial support (Agencia Estatal de Investigación/Projects PID2021-127655NB-I00, and TED2021-129970B-C22). G.M.A thanks the Ministerio de Ciencia, Innovación y Universidades and Agencia Estatal de Investigación (MCIU/AEI), for Torres-Quevedo postdoctoral fellowship (PTQ2023-012950) and the Centro de Computación Científica at the Universidad Autónoma de Madrid for their generous allocation of computer time.

ABBREVIATIONS

CCR2, CC chemokine receptor 2; CCL2, CC chemokine ligand 2; CCR5, CC chemokine receptor 5; TLC, thin layer chromatography.

REFERENCES

1. Bose, S. K.; Mao, L.; Kuehn, L.; Radius, U.; Nekvinda, J.; Santos, W. L.; Westcott, S. A.; Steel, P. G.; Marder, T. B. First-Row d-Block Element-Catalyzed Carbon–Boron Bond Formation and Related Processes. *Chem. Rev.* **2021**, *121*, 13238–13341.
2. Kim-Lee, S.-H.; Mauleón, P.; Gómez Arrayás, R.; Carretero, J. C. Dynamic Multiligand Catalysis: A Polar to Radical Crossover Strategy Expands Alkyne Carboboration to Unactivated Secondary Alkyl Halides. *Chem* **2021**, *7*, 2212–2226.
3. Rej, S.; Das, A.; Panda, T. K. Overview of Regioselective and Stereoselective Catalytic Hydroboration of Alkynes. *Adv. Synth. Catal.* **2021**, *363*, 4818–4840.
4. Rej, S.; Das, A.; Panda, T. K. Overview of Regioselective and Stereoselective Catalytic Hydroboration of Alkynes. *Adv. Synth. Catal.* **2021**, *363*, 4818–4840.
5. Yang, G.; Wang, Z.-Q.; Engle, K. M. Synthesis of Stereodefined 1,1-Diborylalkenes via Copper-Catalyzed Diboration of Terminal Alkynes. *Org. Lett.* **2020**, *22*, 5235–5239.
6. Ur Rasool, J.; Ali, A.; Ahmad, Q. N. Recent Advances in Cu-Catalyzed Transformations of Internal Alkynes to Alkenes and Heterocycles. *Org. Biomol. Chem.* **2021**, *19*, 10259–10287.
7. Moure, A. L.; Gómez Arrayás, R.; Cárdenas, D. J.; Alonso, I.; Carretero, J. C. Regiocontrolled CuI-Catalyzed Borylation of Propargylic-Functionalized Internal Alkynes. *J. Am. Chem. Soc.* **2012**, *134*, 7219–7222.
8. Romer, N. P.; Min, D. S.; Wang, J. Y.; Walroth, R. C.; Mack, K. A.; Sirois, L. E.; Gosselin, F.; Zell, D.; Doyle, A. G.; Sigman, M. S. Data Science Guided Multiobjective Optimization of a Stereoconvergent Nickel-Catalyzed Reduction of Enol Tosylates to Access Trisubstituted Alkenes. *ACS Catal.* **2024**, *14*, 4699–4708.
9. Jang, H.; Zhugralin, A. R.; Lee, Y.; Hoveyda, A. H. Highly Selective Methods for Synthesis of Internal (α-) Vinylboronates through Efficient NHC–Cu-Catalyzed Hydroboration of Terminal Alkynes. Utility in Chemical Synthesis and Mechanistic Basis for Selectivity. *J. Am. Chem. Soc.* **2011**, *133*, 7859–7871.

10. Su, W.; Gong, T.-J.; Zhang, Q.; Zhang, Q.; Xiao, B.; Fu, Y. Ligand-Controlled Regiodivergent Copper-Catalyzed Alkylboration of Unactivated Terminal Alkynes. *ACS Catal.* **2016**, *6*, 6417–6421.
11. Jang, H.; Zhugralin, A. R.; Lee, Y.; Hoveyda, A. H. Highly Selective Methods for Synthesis of Internal (α-) Vinylboronates through Efficient NHC–Cu-Catalyzed Hydroboration of Terminal Alkynes. Utility in Chemical Synthesis and Mechanistic Basis for Selectivity. *J. Am. Chem. Soc.* **2011**, *133*, 7859–7871.
12. Yano, J.; Gaffney, K. J.; Gregoire, J.; Hung, L.; Ourmazd, A.; Schrier, J.; Sethian, J. A.; Toma, F. M. The Case for Data Science in Experimental Chemistry: Examples and Recommendations. *Nat. Rev. Chem.* **2022**, *6*, 357–370.
13. For Chemists, the AI Revolution Has yet to Happen. *Nature* **2023**, *617*, 438–438.
14. Samha, M. H.; Karas, L. J.; Vogt, D. B.; Odogwu, E. C.; Elward, J.; Crawford, J. M.; Steves, J. E.; Sigman, M. S. Predicting Success in Cu-Catalyzed C–N Coupling Reactions Using Data Science. *Sci. Adv.* **2024**, *10*, eadn3478.
15. Dantas, J. A.; Hardy, M. A.; Costa, M. O.; De Oliveira, C. P.; Cabral, T. L. G.; Tormena, C. F.; Sigman, M. S.; Ferreira, M. A. B. Synthesis of Dihydropyrazoles via Palladium-Catalyzed Cascade Heterocyclization/Carbonylation/Arylation of β,γ-Unsaturated N-Tosyl Hydrazones. *Adv. Synth. Catal.* **2024**, *366*, 1120–1127.
16. Haas, B. C.; Lim, N.-K.; Jermaks, J.; Gaster, E.; Guo, M. C.; Malig, T. C.; Werth, J.; Zhang, H.; Toste, F. D.; Gosselin, F.; Miller, S. J.; Sigman, M. S. Enantioselective Sulfonimidamide Acylation via a Cinchona Alkaloid-Catalyzed Desymmetrization: Scope, Data Science, and Mechanistic Investigation. *J. Am. Chem. Soc.* **2024**, *146*, 8536–8546.
17. 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.
18. 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.
19. 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.
20. 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.
21. Hueffel, J. A.; Sperger, T.; Funes-Ardoiz, I.; Ward, J. S.; Rissanen, K.; Schoenebeck, F. Accelerated Dinuclear Palladium Catalyst Identification through Unsupervised Machine Learning. *Science* **2021**, *374*, 1134–1140.
22. Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A. K.; 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.
23. Dotson, J. J.; van Dijk, L.; Timmerman, J. C.; Grosslight, S.; Walroth, R. C.; Gosselin, F.; Püentener, K.; Mack, K. A.; Sigman, M. S. Data-Driven Multi-Objective Optimization Tactics for Catalytic Asymmetric Reactions Using Bisphosphine Ligands. *J. Am. Chem. Soc.* **2023**, *145*, 110–121.
24. Romer, N. P.; Min, D. S.; Wang, J. Y.; Walroth, R. C.; Mack, K. A.; Sirois, L. E.; Gosselin, F.; Zell, D.; Doyle, A. G.; Sigman, M. S. Data Science Guided Multiobjective Optimization of a Stereoconvergent Nickel-Catalyzed Reduction of Enol Tosylates to Access Trisubstituted Alkenes. *ACS Catal.* **2024**, *14*, 4699–4708.
25. 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.
26. Nie, W.; Wan, Q.; Sun, J.; Chen, M.; Gao, M.; Chen, S. Ultra-High-Throughput Mapping of the Chemical Space of Asymmetric Catalysis Enables Accelerated Reaction Discovery. *Nat. Commun.* **2023**, *14*, 6671–6682.
27. Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. AutoQChem: An Automated Workflow for the Generation and Storage of DFT Calculations for Organic Molecules. *React. Chem. Eng.* **2022**, *7*, 1276–1284.
28. Gensch, T.; Gomes, G. d. P.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A. K.; 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.
29. Xu, L.; Zhu, J.; Shen, X.; Chai, J.; Shi, L.; Wu, B.; Li, W.; Ma, D. 6-Hydroxy Picolinohydrazides Promoted Cu(I)-Catalyzed Hydroxylation Reaction in Water: Machine-Learning Accelerated Ligands Design and Reaction Optimization. *Angew. Chem. Int. Ed.* **2024**, e202412552.
30. 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.
31. 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.

Insert Table of Contents artwork here