Machine Learning Enables a Top-Down Approach to Mechanistic Elucidation

Isaiah O. Betinol[1], Yutao Kuang[1], Junshan Lai[1], Christopher Yousofi[1], and Jolene P. Reid[1]*

[1]Department of Chemistry, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada

*Corresponding author. Email: jreid@chem.ubc.ca

Abstract: General reaction behavior is rarely reported in asymmetric catalysis, not simply because it is difficult to achieve, but also due to the methods used for its identification and study. Traditional approaches involve compartmentalization, where the impact of individual components is first analyzed, followed by assimilation using simple response and structure matching techniques. However, extending this method to accommodate complex conditions and diverse reactions proves challenging. Here, we present a data-driven method that relies on clusterwise linear regression to derive and predictively apply general mechanistic models of enantioinduction, with minimal human intervention. When applied to the palladium-catalyzed decarboxylative asymmetric allylic alkylation (DAAA) reaction, unexpected interactions governing enantioselectivity are revealed, supported by high-level computations and additional experiments. Our results demonstrate this workflow as a powerful new tool for automating mechanistic elucidation and effectively identifying general reaction performance.

Chemical reactions are traditionally analyzed first individually to determine how the molecular features of the reaction components contribute to the mechanistic aspects of the transformation (*1*–*4*). Subsequently, the insight gathered from this process is matched, either qualitatively (*5*–*7*) or quantitatively (*8*–*10*), to a wide array of experimental observations to derive general mechanistic principles applicable to a broader range of reactions. While this bottom-up approach to mechanistic elucidation has been invaluable for guiding the rational design of catalytic systems (*11*–*13*), its limitations lie in its potential oversimplification, dependence on complete datasets (*14*, *15*), and challenges in extrapolating meaningful mechanistic insights to transformations that utilize diverse structures and complicated reaction conditions (*16*). These issues are particularly common in asymmetric catalysis, an area that has evolved to impart high-levels of enantioinduction through the application of complex catalysts that engage in various interactions with reactants, often noncovalent in nature and energetically weak (*17*, *18*). Such transformations, although seemingly mechanistically related, are typically adjusted to perform optimally with distinct conditions and catalyst structures. Accordingly, applying simple response and structural matching techniques to establish broad mechanistic principles remains a challenge (*19*, *20*). In order to address such systems, we envisioned a top-down strategy for mechanistic study involving the application of data-driven techniques. This approach relies on the use of clusterwise linear regression (*21*) to autonomously discover subsets of reaction space that operate generally (Fig. 1A). Here, we demonstrate that this technique enables in-depth mechanistic analysis of the features that govern enantioselectivity, affording nonintuitive insight into the origin of general asymmetric induction and guiding rational experimental design.

Our platform for probing the presence of linear relationships as an indicator of mechanistic continuity was inspired by Hammett analysis. Simply put, this type of physical organic experiment demonstrates that a transformation's mechanistic features are embodied in its unique response to changes in reaction component structure and conditions (*22*). Traditional assessments constrain the chemical space for evaluation by probing the impact of one or few changes to a single structure, ultimately allowing for trends of best fit to be established with a single parameter (*23*). The resultant correlation, or pair of correlations in the case of non-linearity (*24*), can provide insight into the molecular feature impacting one or a set of key transition state structures (*25*–*27*). If more than one molecular property influences the transition state structure, the approach can be scaled to incorporate these multiple factors (*28*). Expanding this idea to complex datasets containing multiple multidimensional linear relationships is highly appealing and an unmet challenge. Such an approach would lead to the identification of reactions that operate similarly, producing testable hypotheses regarding the structural origins of general reaction behavior (Fig. 1B).

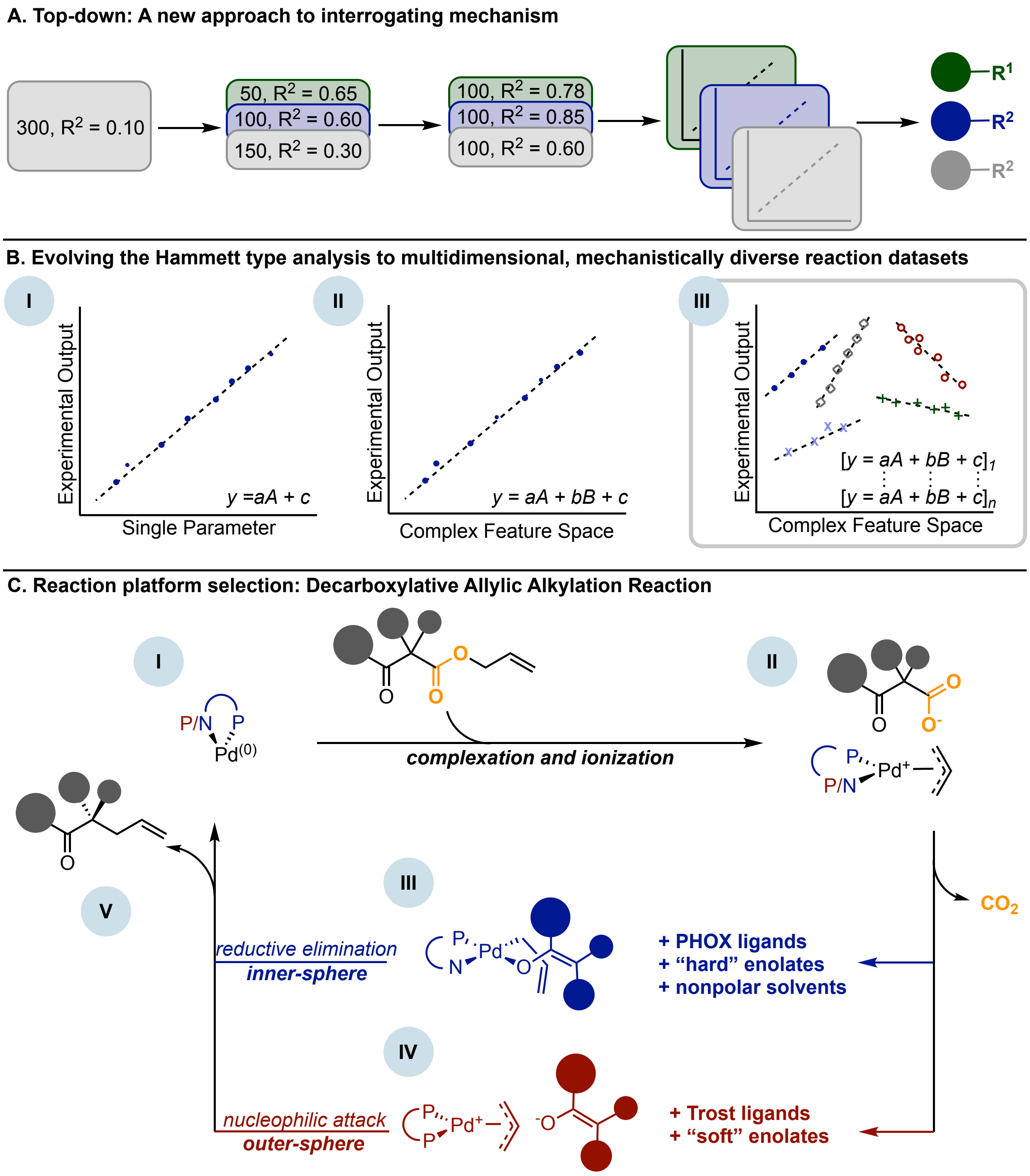

Fig. 1: Studying general reaction behavior through structure-function relationships. (A) Overview of the clusterwise linear regression algorithm for automating the construction of general mechanistic models and predictive application to experimental design. **(B)** I–III demonstrates the

evolution of the Hammett type analysis from uniparameter to multivariate correlations. This work focuses on algorithm development for the automated identification of multiple multivariate correlations and mechanistic analysis of the resulting statistical models. (**C**) General mechanism (I-V) and recognized empirical trends of the palladium-catalyzed DAAA reactions, the case study to be studied in depth.

Workflow Design and Implementation

Perhaps the greatest impediment to performing mechanistic analysis in this manner is selecting data subsets that can be correlated linearly from a much larger, potentially unstructured dataset. Current training set selection methods aim to minimize within cluster variance based on input features only (e.g. constraining reaction space or *k*-means clustering); however, applying this approach will lead to ill-defined groupings in cases where similar transformations proceed through distinct pathways. Including the response variable to be correlated in the clustering step is essential; thus, by design, the mechanistic features that linearly connect a set of reactions are identified in the process. To leverage regression for clustering problems, we implemented an optimization framework in Python specifically aimed at identifying groups of reactions that operate generally. By utilizing principal component regression with a single principal component for clustering, we minimize the effect of overfitting on the final results and our implementation is lightweight and computationally efficient. Despite the iterative nature of the clustering algorithm, we obtain high-quality, reproducible cluster configurations. However, in questioning the proper deployment of the algorithm, we were initially met with several challenges. Of these, perhaps the most important concerned the determination of the optimal cluster number, a critical hyperparameter. Reasoning that in ideal clustering the average of within-cluster R^2 should be high, close to 1, we decided to implement a modified elbow method. This feature allows for determining a point where adding more clusters does not significantly improve the cluster quality. Because in some cases structurally related starting materials perform comparably under similar conditions, we purposely deploy *k*-means to initialize cluster labels and allow the algorithm to adjust the assignment for each data point as necessary. Given the sequential nature of the process, we anticipate that the final cluster configurations will be particularly altered by the order of operations performed during optimization, and this will vary widely. To address this limitation in simulating optimization runs iteratively, we deploy a customized ensembling technique to ensure that the final cluster labels correspond to the most common set of groupings. These data classifications can then be analyzed further using a broad suite of mechanistic interrogation techniques including multivariate linear regression (MLR) and DFT transition state analysis.

Reaction Platform Selection

After evaluating our algorithm for the identification of linear relationships (see SI), we next sought to perform mechanistic analysis on a system that has proved challenging to generalize using traditional techniques. Accordingly, the field of transition-metal catalysis was appealing because despite the seemingly straightforward proposed factors that determine enantioselectivity, general models of stereoselectivity – even of a qualitative nature – have been rare. Recent reports of machine learning applied to this arena demonstrate that there are correlations to be found (*29*, *30*), but that either (a) the reactions are difficult to study computationally meaning broad mechanistic insight must be assembled mostly through opaque chemical intuition; or perhaps more notably (b) several reaction paths contribute to enantioselectivity and their distinct sensitivities to reaction component structure are difficult to deconvolute. In this context, the palladium-catalyzed decarboxylative asymmetric allylic alkylation (DAAA) reaction was identified as a prototypical example (*31*).

An intriguing mechanistic feature of this reaction class is highlighted by the observation that they often proceed with high levels of enantioselectivity with two disparate ligand classes, namely PHOX and Trost. Although these data allude to selectivity determination via a set of interactions common to different ligand designs and mechanistic pathways (i.e., inner- and outer-sphere), such general determinants would be difficult to further characterize by traditional techniques. Indeed, major efforts by several groups using traditional bottom-up approaches have revealed the fundamental mechanistic features of the system and have found particular use in explaining reaction outcomes that align with one of two broad trends (Fig. 1C) (*32*–*38*). Whilst these simplistic trends describe a large amount of DAAA chemistry, they are necessarily incomplete (*39*, *40*). This highlights the limitation in our current mechanistic understanding of the transformation, emphasizing the need to avoid relying solely on simple structure and response matching techniques, which may overlook valuable experimental insights and impose constraints on our capacity to conduct rational experimental design. Moreover, despite the potential for modularity the structural variation within the ligand classes is small, and although there is some catalyst-substrate matching, few catalysts have seen extensive use. In significant contrast to many asymmetric reaction classes, including those facilitated by organocatalysts, enantioselectivity improvements are often made through reaction condition selection, which could introduce effects that have not yet been studied using data-driven approaches. Consequently, our goal was to apply clusterwise linear regression as a comprehensive mechanistic platform allowing for reaction generalization and the evaluation of how subtle structural features affect selectivity, using this reaction as an important case study.

Data Set Modelling

To initiate this workflow toward a more complete mechanistic assessment, an expanded inventory of 329 reactions was curated from the available literature. To achieve a significant spread of enantioselectivity values and to ensure the inclusion of complex mechanistic features, this comprehensive data search was designed to purposefully sample the reaction component and conditions that profoundly impacted the experimental outcomes. As a result, we obtained a dataset with a significant spread of absolute enantioselectivity, comprising a $\Delta\Delta G^{\ddagger}$ window of 3.6 kcal/mol. This dataset includes combinations of 6 ligands including both PHOX and Trost types, 178 enolates, 5 allyl groups, and 14 unique solvents.

Because our goal is to create statistical models characterized by simplicity and robust interpretability, we carefully considered the molecular features to be provided to the algorithm for model building (Fig. 2A). Parameter selection is typically accomplished using candidate structures that best represent the species relevant to the catalytic cycle. In this case, we implemented a truncation strategy which treated the allylic and enolate components of the starting material separately. We viewed this as a crucial but simple means to interrogate the impact that these individual components have on the overall process. Initially, catalyst parameterization efforts focused on computing the free ligand only; however, we expanded the process to include the base catalyst structure as a means to fix the ligand in the most relevant conformation. To define the parameter library, DFT optimizations were performed on these structures at the M06/def2-TZVP level of theory, and various descriptors were collected to probe structural effects of the reaction partners and ligands. This included NBO charges (*41*), molecular orbital energies, bond distances and bond angles. Sterimol descriptors (*42*) were collected to measure the size of the substituents at the different positions of the intermediate structures. To describe the asymmetry between the two ligand types and their surrounding environments, the catalyst structures were projected on a quadrant diagram and aligned. This approach allowed for the collection of Sterimol measurements

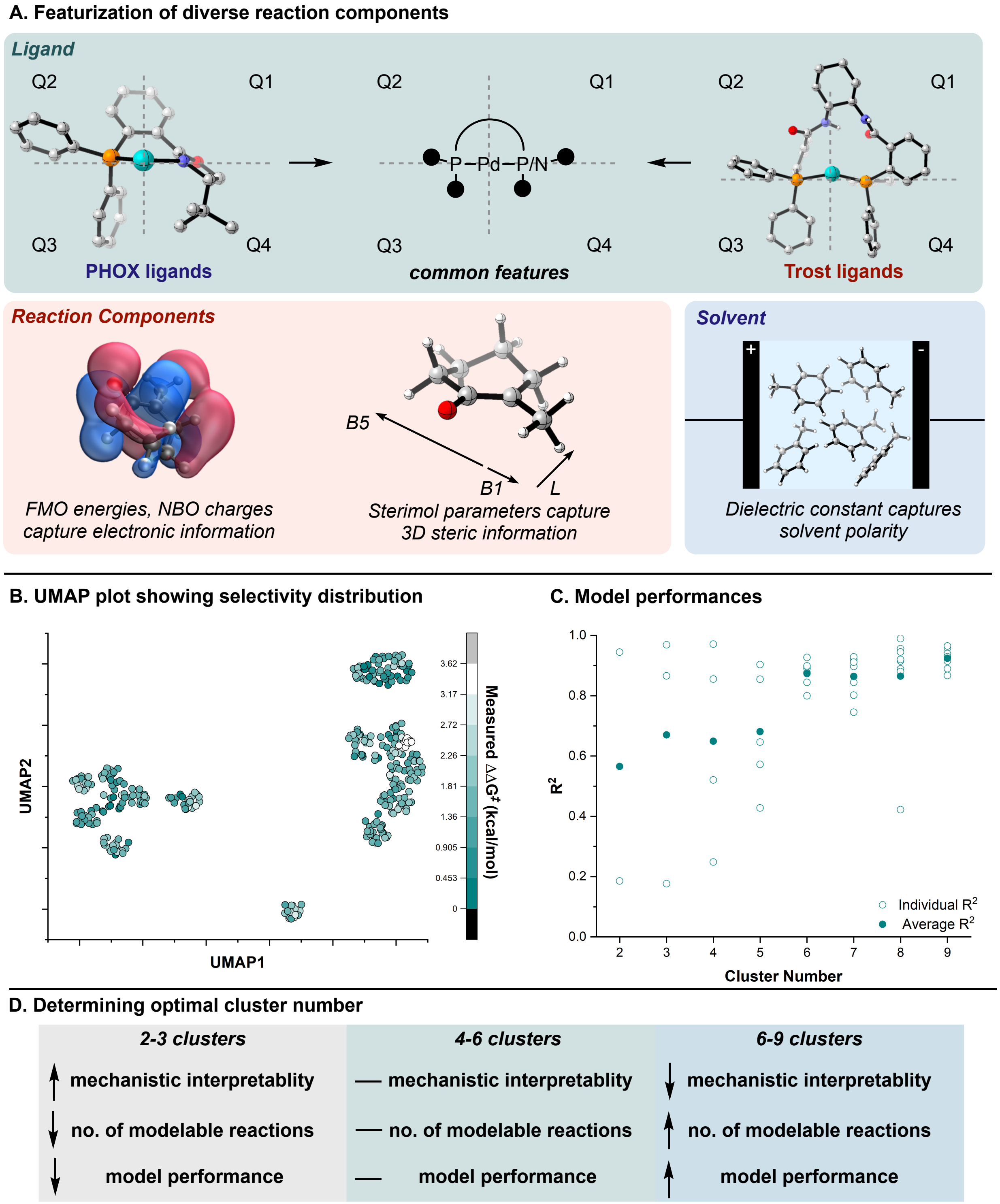

Fig. 2: Investigation of clusterwise linear regression for identifying general reaction behavior. (A) An overview of the molecular features collected for analysis and parameterization strategy for diverse ligand structures. Frontier molecular orbital (FMO) energies are a traditionally used descriptor for assessing reactivity. Charges from Natural bond orbital (NBO) analysis are deployed to capture electron density. Sterimol descriptors B1 and B5 describe the lowest and

highest width of a substituent perpendicular to a set axis, and Sterimol L describes the substituents length along that axis. **(B)** UMAP plot showing both high and low performing reactions distributed throughout chemical space. **(C)** Visual demonstrates the point where adding more clusters does not significantly improve the cluster quality. **(D)** Factors considered in determining the optimal cluster number.

that represent the size of the ligand portions occupying each quadrant. Given the clear mechanistic importance of the solvent in determining the enantioselectivity we also sought to describe its impact through physical meaningful continuous variables like dielectric constant (see SI for full details).

Our algorithm was then applied to the entire dataset shown in Figure 2B to identify correlations between the molecular structure of every reaction variable and the experimentally determined enantioselectivity, $\Delta\Delta G^{\ddagger}$. To determine the optimal set of groupings and probe the algorithm behavior as the cluster number increases, a communicative visualization of the results is crucial. Thus, we elected to present the average R^2 across clusters alongside each individual cluster R^2 (Fig. 2C). Analysis of this data reveals that increasing the cluster number generally improves model quality but reduces interpretability. Certainly, low cluster numbers (fewer than 4) lead to more comprehensible separations, requiring fewer model comparisons (Fig. 2D); however, this can result in non-optimal clusters, where individual cluster R^2 is much less than the average R^2. This demonstrates that fewer cluster numbers result in groupings that include some reactions which do not behave analogously. A change in cluster number from 5 to 6 results in a sharp uptick in average R^2, reflecting significantly improved model quality across all clusters and essentially no non-optimal clusters were generated. This factor combined with the minimal statistical gain observed for cluster numbers higher than 6, are the reasons for why we selected these cluster configurations to be optimal for downstream mechanistic analysis.

To quantify the improvement in model quality, provided by our approach, we established several baselines for comparison. This involved benchmarking our model performances against those obtained from qualitative rule-based systems and those including similar structures. We envisioned this would provide a general assessment of the features that impact a key mechanistic step or related reactions, capturing the essential aspects a model should have to perform similarly to human experts in data-driven mechanistic analysis. Consequently, we segmented the data into subsets, categorized by ligand type (PHOX or Trost), as these are hypothesized to lead to structurally distinct interactions with other components. In other words, this organizational scheme was viewed as a traditional means to facilitate the identification of the molecular features that affect particular mechanistic pathways identified by an expert. The second baseline strategy involved generating individual regressions from data groupings determined by *k*-means. Compared to the baselines, our models achieved improved accuracies and outperforms statistical models constructed by "experts" in both training set fit and prediction accuracy (see SI).

We find that the clustering algorithm yielded unexpected groupings, as evidenced by situations in which seemingly similar substrates were placed into distinct clusters based on their enantiomeric excess results (Fig. 3A) (*39*). Although the individual clusters shown in Fig. 2B establishes the capacity of our algorithm to readout general aspects of this system, the ultimate goal of our workflow is to discern subtle underlying mechanistic phenomena. To truly interrogate the precise molecular features responsible for enantioselectivity, linear regression algorithms were then applied to the individual data groupings (see SI). Subsequently, analysis and refinement of the resultant models were used to produce explicit mechanistic hypothesis. The models depicted in Figure 3C were identified for each cluster and in each case a good correlation was determined ($R^2 = 0.77 – 0.89$) using a small set of molecular descriptors. Combining the prediction statistics from all 6 models yields an R^2 of 0.84 and test mean average error of 0.33 kcal/mol (Fig. 3B).

The simplicity of the linear regression models encouraged us to further explore the impact of the reaction components on the enantioselectivity outcome. Although the relatively small number of overlapping features between the models serves as a validation of our approach, it does create a challenge for model comparison. To address this issue, we sought to consolidate some of the mechanistic insights from the statistical models into a simple chart, illustrating the absolute importance of features grouped by reaction component structure and subclassified either as steric or electronic. Analysis of this data arrangement clearly highlights the significance of the enolate structure in determining the enantioselectivity result. By juxtaposing the enolate terms included in models I-II, III-IV, and V-VI, it becomes apparent that certain outcomes are more sensitive to the enolate steric profile than to the electronic features, and vice versa. Regarding the evaluation of ligand effects, several overlapping terms suggest that bulky PHOX or Trost ligands are compatible with the reaction if the enolate molecular requirements are met. This observation is reinforced by the construction of model I, where only the enolate structure influences the enantioselectivity result. The most compelling difference between the models is that for two of them, a solvent descriptor is not included as a weighted term in the equation. This suggests that for some reactions, solvent properties do not have much effect on the stereoselectivity outcome, despite the large structural variance. This result, especially in the case of PHOX ligands, is non-obvious. Indeed, using non-polar solvents with PHOX ligands is an established practice, likely well-known among specialized groups, to favor selective inner-sphere mechanistic pathways (*35*, *37*). Thus, an interpretation of the categorization would suggest that the enolate structure primarily governs the solvent effects, regardless of the ligand structure.

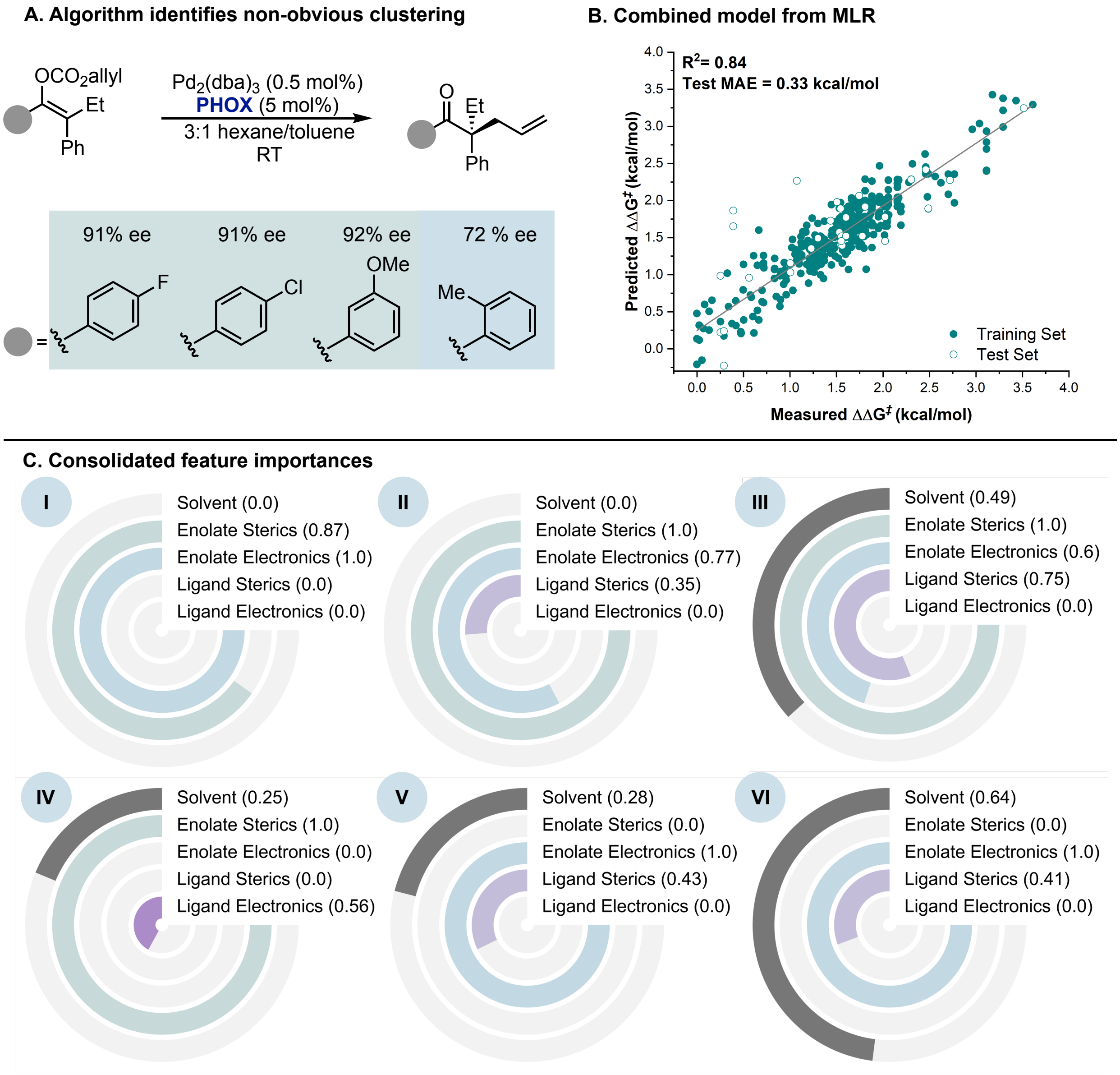

Fig. 3: Mechanistic analysis by multidimensional clustering and regression modeling. (A) Clustering algorithm identifies non-intuitive groupings that match subtle changes in structure with variations in enantioselectivity. **(B)** Statistical model representing the combination of 6 individual MLR models created for each cluster. R^2 represents goodness of fit, MAE = mean average error

(C) Illustration and casual interpretation of the model terms. Across all the reaction components and models, steric descriptors include Sterimol L, B1 and B5 parameters, and ligand bite angle. Likewise, NBO charges, FMO energies, and FMO derived features like hardness and softness are recognized as electronic features.

To interrogate this hypothesis, we attempted to link an enolate structural feature to the solvent dependent enantioselectivity outcomes (*43*) (Fig. 4A). Whereas the enolate descriptors did not provide any clear cut-offs for PHOX reactions involving non-polar solvents (i.e., dielectric constant < 3), the NBO charge on oxygen (NBO O) allowed for classification of enantioselectivity performance into selective and unselective bins with polar solvents (i.e., dielectric constant > 3), though several outliers were present with the former (Fig. 4B). The NBO O descriptor revealed a sharp cutoff in selectivity at a value of around -0.74, revealing the electronic requirement for good enantioselectivity in polar solvents with PHOX ligands. The observation that this distinctive charge profile at the oxygen was essential for good enantioselectivities in polar solvents raised questions about the relative energy differences of the competing pathways (Fig. 4A). Overall, these results suggest that a Pd-O interaction is established between the catalyst and the enolate intermediate during the enantioselectivity determining step, the strength of which is modulated by local electron density and solvent properties. Essentially, if a less basic enolate is paired with a polar solvent, the interaction weakens. This results in structures featuring Pd-O bonds being disfavored and raised closer in energy to those that do not rely on the Pd-O bond (i.e., outer-sphere pathways). Based on these hypotheses, and to specifically probe this putative interaction, we set out to calculate some of the key structures involved in selective and unselective reactions incorporating THF or Et_2O as the polar solvent. The lower energy structures located along the inner-sphere pathway corresponded with a stronger Pd-O bond, as indicated through a shorter bond, and a reflection of the increased electron density at the oxygen atom with this enolate (Fig. 4C, teal and purple). This favorable interaction stabilizing the key isomerization step along the inner-sphere pathway is compromised in the example involving the less basic enolate (Fig. 4C, grey), leading to small energy differences between the competing pathways and the low levels of enantioselectivity found experimentally. Collectively, these studies demonstrate the physical significance of the NBO O descriptor, and the -0.74 cutoff value observed experimentally.

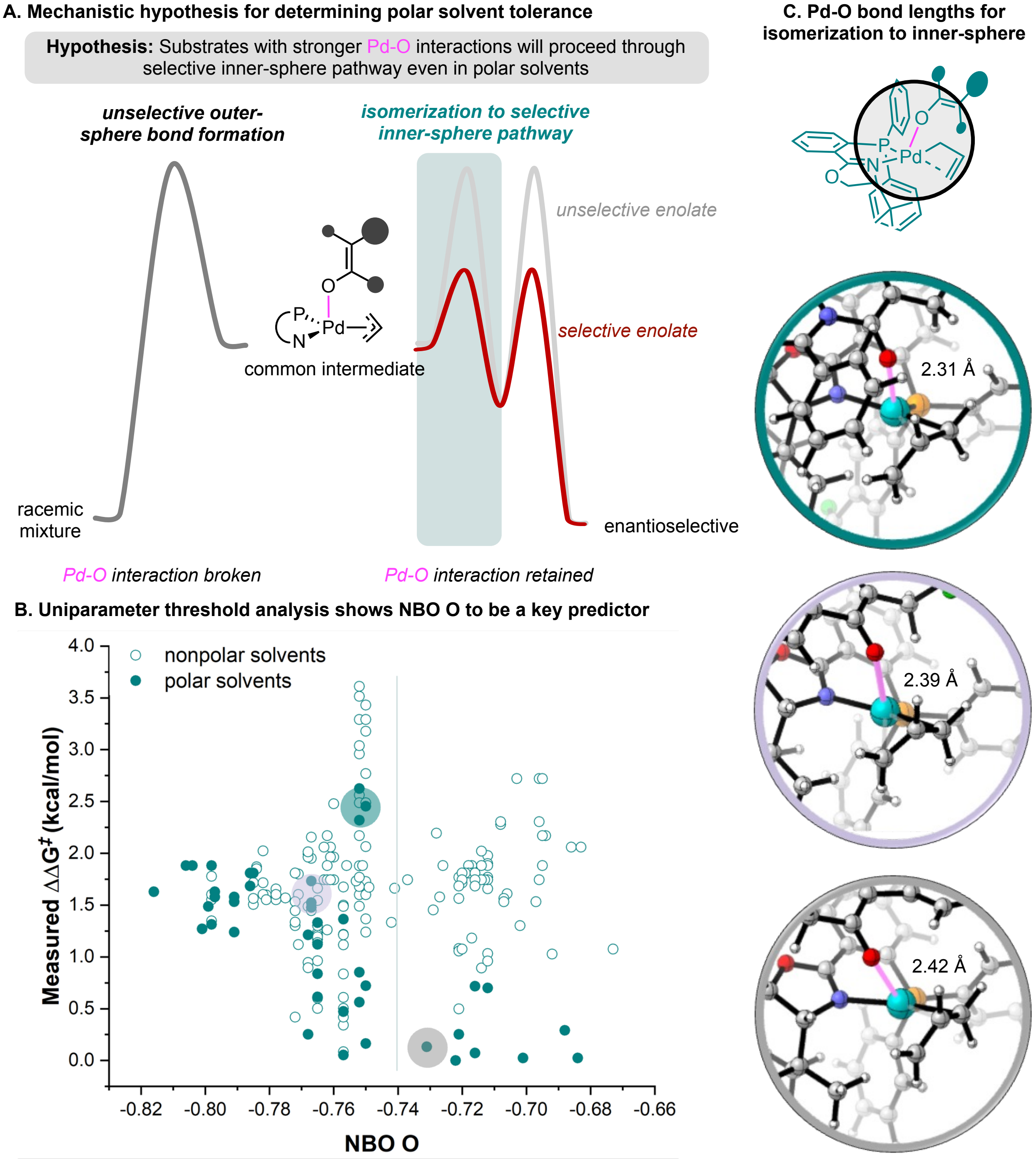

Fig. 4: Studies deployed to probe the nonintuitive solvent dependent selectivity with PHOX ligands. **(A)** Mechanistic hypothesis for the identification of selective or unselective enolates in polar solvents. **(B)** Analysis of enantioselectivities against NBO O demonstrates the dependence of charge density at the enolate oxygen on solvent-dependent enantioselectivity outcomes. Light green points represent reactions evaluated in nonpolar solvents, green points represent reactions performed in polar solvents. **(C)** Visualizing a portion of the transition state structures located for

isomerization to the inner-sphere pathway. These structures are computed at the SMD-M06/def2-TZVP//BP86-D3(BJ)/def2-TZVP(Pd)-def2-SVP level of theory. Palladium-oxygen bond is highlighted in pink with the distance labelled in Å.

Based on these hypotheses, we evaluated our capacity to strategically modify the enolate and solvent components to achieve the anticipated reaction outcomes (Fig. 5). From the enolates surveyed, we identified a set of structures that had the potential to achieve good selectivity in polar solvents having calculated NBO O values below the observed threshold of -0.74 (Fig. 5A). These structures were previously evaluated experimentally with nonpolar solvents only (*44*); however, upon reassessment of several reactions with a polar solvent, we found moderate to good enantioselectivities, consistent with the classification and NBO O analysis (i.e., provide average $\Delta\Delta G^{\ddagger}$ values greater than those found with NBO O > -0.74). Next, we identified a seemingly similar set of enolates that contains the same types of substituents but differs only by a switch in the positions of the heteroaromatic and aromatic groups (*45*). The more positive NBO O characterized (i.e., NBO O > -0.74) these substrates similar to those that provide poor enantioselectivities in polar solvents but experimentally we found these led to no reaction (Fig. 5B). We recognized this result as an extreme circumstance where the enolate properties have a profound impact on the reaction outcome, leading to virtually no levels of catalytic activity in polar solvents but high reaction yields when operated in a nonpolar medium. This further highlights the dependence of relative barrier heights on enolate basicity and solvent properties.

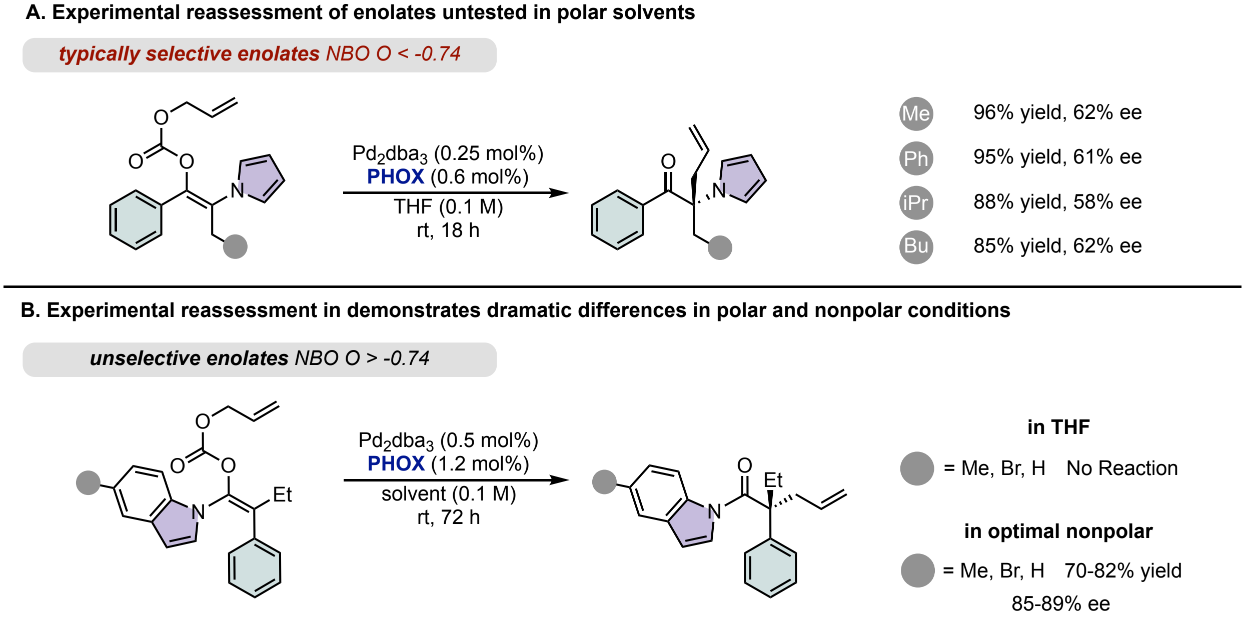

Fig. 5: Experimental reassessment of enolates characterized as (A) typically selective and (B) unselective in polar solvents.

Summary and Outlook

We have developed a workflow that utilizes a clusterwise linear regression algorithm for the automated identification of distinct mechanistic profiles from large, unstructured datasets. The applicability of this data-driven approach for mechanistic analysis was assessed, demonstrating its usefulness in predicting and studying complex asymmetric catalysis outcomes. Through the development of mechanism-specific correlations, this method reveals reaction similarities and reaction-specific mechanistic principles. For example, in the case of the palladium-catalyzed decarboxylative asymmetric allylic alkylation (DAAA) reaction, targeted mechanistic analysis provided new insights into the solvent-dependent enantioselectivity outcomes. Overall, this approach holds potential for simplifying the challenging process of mechanistic generalization and analysis, making it less subjective and more accurate.

References and Notes

1. K. N. Houk, P. H.-Y. Cheong, Computational prediction of small-molecule catalysts. *Nature* **455**, 309–313 (2008).

2. Q. Peng, F. Duarte, R. S. Paton, Computing organic stereoselectivity – from concepts to quantitative calculations and predictions. *Chem. Soc. Rev.* **45**, 6093–6107 (2016).

3. Y. Lam, M. N. Grayson, M. C. Holland, A. Simon, K. N. Houk, Theory and Modeling of Asymmetric Catalytic Reactions. *Acc. Chem. Res.* **49**, 750–762 (2016).

4. R. B. Sunoj, Transition State Models for Understanding the Origin of Chiral Induction in Asymmetric Catalysis. *Acc. Chem. Res.* **49**, 1019–1028 (2016).

5. S. Bahmanyar, K. N. Houk, H. J. Martin, B. List, Quantum Mechanical Predictions of the Stereoselectivities of Proline-Catalyzed Asymmetric Intermolecular Aldol Reactions. *J. Am. Chem. Soc.* **125**, 2475–2479 (2003).

6. B. M. Trost, M. R. Machacek, A. Aponick, Predicting the Stereochemistry of Diphenylphosphino Benzoic Acid (DPPBA)-Based Palladium-Catalyzed Asymmetric Allylic Alkylation Reactions: A Working Model. *Acc. Chem. Res.* **39**, 747–760 (2006).

7. J. P. Reid, L. Simón, J. M. Goodman, A Practical Guide for Predicting the Stereochemistry of Bifunctional Phosphoric Acid Catalyzed Reactions of Imines. *Acc. Chem. Res.* **49**, 1029–1041 (2016).

8. J. P. Reid, M. S. Sigman, Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343–348 (2019).

9. A. Shoja, J. Zhai, J. P. Reid, Comprehensive Stereochemical Models for Selectivity Prediction in Diverse Chiral Phosphate-Catalyzed Reaction Space. *ACS Catal.* **11**, 11897–11905 (2021).

10. J. P. Reid, I. O. Betinol, Y. Kuang, Mechanism to model: a physical organic chemistry approach to reaction prediction. *Chem. Commun.* **59**, 10711–10721 (2023).

11. S. Mitsumori, H. Zhang, P. Ha-Yeon Cheong, K. N. Houk, F. Tanaka, C. F. Barbas, Direct Asymmetric *a nti* -Mannich-Type Reactions Catalyzed by a Designed Amino Acid. *J. Am. Chem. Soc.* **128**, 1040–1041 (2006).

12. L. C. Burrows, L. T. Jesikiewicz, G. Lu, S. J. Geib, P. Liu, K. M. Brummond, Computationally Guided Catalyst Design in the Type I Dynamic Kinetic Asymmetric Pauson–Khand Reaction of Allenyl Acetates. *J. Am. Chem. Soc.* **139**, 15022–15032 (2017).

13. S.-S. Meng, P. Yu, Y.-Z. Yu, Y. Liang, K. N. Houk, W.-H. Zheng, Computational Design of Enhanced Enantioselectivity in Chiral Phosphoric Acid-Catalyzed Oxidative Desymmetrization of 1,3-Diol Acetals. *J. Am. Chem. Soc.* **142**, 8506–8513 (2020).

14. P. Raghavan, B. C. Haas, M. E. Ruos, J. Schleinitz, A. G. Doyle, S. E. Reisman, M. S. Sigman, C. W. Coley, Dataset Design for Building Models of Chemical Reactivity. *ACS Cent. Sci.* **9**, 2196–2204 (2023).

15. D. M. Lustosa, A. Milo, Mechanistic Inference from Statistical Models at Different Data-Size Regimes. *ACS Catal.* **12**, 7886–7906 (2022).

16. I. O. Betinol, Y. Kuang, J. P. Reid, Guiding Target Synthesis with Statistical Modeling Tools: A Case Study in Organocatalysis. *Org. Lett.* **24**, 1429–1433 (2022).

17. R. R. Knowles, E. N. Jacobsen, Attractive noncovalent interactions in asymmetric catalysis: Links between enzymes and small molecule catalysts. *Proc. Natl. Acad. Sci.* **107**, 20678–20685 (2010).

18. A. J. Neel, M. J. Hilton, M. S. Sigman, F. D. Toste, Exploiting non-covalent π interactions for catalyst design. *Nature* **543**, 637–646 (2017).

19. A. J. Neel, A. Milo, M. S. Sigman, F. D. Toste, Enantiodivergent Fluorination of Allylic Alcohols: Data Set Design Reveals Structural Interplay between Achiral Directing Group and Chiral Anion. *J. Am. Chem. Soc.* **138**, 3863–3875 (2016).

20. A. Milo, A. J. Neel, F. D. Toste, M. S. Sigman, A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science* **347**, 737–743 (2015).

21. H. Späth, Algorithm 39 Clusterwise linear regression. *Computing* **22**, 367–373 (1979).

22. L. P. Hammett, The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **59**, 96–103 (1937).

23. P. R. Wells, Linear Free Energy Relationships. *Chem. Rev.* **63**, 171–219 (1963).

24. J. O. Schreck, Nonlinear Hammett relationships. *J. Chem. Educ.* **48**, 103 (1971).

25. E. N. Jacobsen, W. Zhang, M. L. Guler, Electronic tuning of asymmetric catalysts. *J. Am. Chem. Soc.* **113**, 6703–6704 (1991).

26. M. Palucki, N. S. Finney, P. J. Pospisil, M. L. Güler, T. Ishida, E. N. Jacobsen, The Mechanistic Basis for Electronic Effects on Enantioselectivity in the (salen)Mn(III)-Catalyzed Epoxidation Reaction. *J. Am. Chem. Soc.* **120**, 948–954 (1998).

27. M. S. Sigman, J. J. Miller, Examination of the Role of Taft-Type Steric Parameters in Asymmetric Catalysis. *J. Org. Chem.* **74**, 7633–7643 (2009).

28. M. S. Sigman, K. C. Harper, E. N. Bess, A. Milo, The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **49**, 1292–1301 (2016).

29. J. Wahlers, J. Margalef, E. Hansen, A. Bayesteh, P. Helquist, M. Diéguez, O. Pàmies, O. Wiest, P.-O. Norrby, Proofreading experimentally assigned stereochemistry through Q2MM predictions in Pd-catalyzed allylic aminations. *Nat. Commun.* **12**, 6719 (2021).

30. J. Lu, S. Donnecke, I. Paci, D. C. Leitch, A reactivity model for oxidative addition to palladium enables quantitative predictions for catalytic cross-coupling reactions. *Chem. Sci.* **13**, 3477–3488 (2022).

31. J. James, M. Jackson, P. J. Guiry, Palladium-Catalyzed Decarboxylative Asymmetric Allylic Alkylation: Development, Mechanistic Understanding and Recent Advances. *Adv. Synth. Catal.* **361**, 3016–3049 (2019).

32. J. A. Keith, D. C. Behenna, J. T. Mohr, S. Ma, S. C. Marinescu, J. Oxgaard, B. M. Stoltz, W. A. Goddard, III, The Inner-Sphere Process in the Enantioselective Tsuji Allylation Reaction with (*S*)- *t* -Bu-phosphinooxazoline Ligands. *J. Am. Chem. Soc.* **129**, 11876–11877 (2007).

33. B. M. Trost, J. Xu, T. Schmidt, Palladium-Catalyzed Decarboxylative Asymmetric Allylic Alkylation of Enol Carbonates. *J. Am. Chem. Soc.* **131**, 18343–18357 (2009).

34. D. C. Behenna, J. T. Mohr, N. H. Sherden, S. C. Marinescu, A. M. Harned, K. Tani, M. Seto, S. Ma, Z. Novák, M. R. Krout, R. M. McFadden, J. L. Roizen, J. A. Enquist, D. E. White, S. R. Levine, K. V. Petrova, A. Iwashita, S. C. Virgil, B. M. Stoltz, Enantioselective Decarboxylative Alkylation Reactions: Catalyst Development, Substrate Scope, and Mechanistic Studies. *Chem. – Eur. J.* **17**, 14199–14223 (2011).

35. J. A. Keith, D. C. Behenna, N. Sherden, J. T. Mohr, S. Ma, S. C. Marinescu, R. J. Nielsen, J. Oxgaard, B. M. Stoltz, W. A. Goddard, III, The Reaction Mechanism of the Enantioselective Tsuji Allylation: Inner-Sphere and Outer-Sphere Pathways, Internal Rearrangements, and Asymmetric C–C Bond Formation. *J. Am. Chem. Soc.* **134**, 19050–19060 (2012).

36. K. E. McPherson, M. P. Croatt, A. T. Morehead, A. L. Sargent, DFT Mechanistic Investigation of an Enantioselective Tsuji–Trost Allylation Reaction. *Organometallics* **37**, 3791–3802 (2018).

37. A. Q. Cusumano, B. M. Stoltz, W. A. Goddard, III, Reaction Mechanism, Origins of Enantioselectivity, and Reactivity Trends in Asymmetric Allylic Alkylation: A Comprehensive Quantum Mechanics Investigation of a C(sp3)–C(sp3) Cross-Coupling. *J. Am. Chem. Soc.* **142**, 13917–13933 (2020).

38. C. P. Butts, E. Filali, G. C. Lloyd-Jones, P. O. Norrby, D. A. Sale, Y. Schramm, Structure-Based Rationale for Selectivity in the Asymmetric Allylic Alkylation of Cycloalkenyl Esters Employing the Trost 'Standard Ligand' (TSL): Isolation, Analysis and Alkylation of the Monomeric form of the Cationic η^3-Cyclohexenyl Complex [(η^3- *c* -C_6H_9)Pd(TSL)]$^+$. *J. Am. Chem. Soc.* **131**, 9945–9957 (2009).

39. E. J. Alexy, H. Zhang, B. M. Stoltz, Catalytic Enantioselective Synthesis of Acyclic Quaternary Centers: Palladium-Catalyzed Decarboxylative Allylic Alkylation of Fully Substituted Acyclic Enol Carbonates. *J. Am. Chem. Soc.* **140**, 10109–10112 (2018).

40. E. J. Alexy, S. C. Virgil, M. D. Bartberger, B. M. Stoltz, Enantioselective Pd-Catalyzed Decarboxylative Allylic Alkylation of Thiopyranones. Access to Acyclic, Stereogenic α-Quaternary Ketones. *Org. Lett.* **19**, 5007–5009 (2017).

41. E. D. Glendening, J. K. Badenhoop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, P. Karafiloglou, C. R. Landis, F. Weinhold, NBO 7.0., Theoretical Chemistry Institute, University of Wisconsin, Madison (2018).

42. A. Verloop, W. Hoogenstraaten, J. Tipker, "Development and Application of New Steric Substituent Parameters in Drug Design" in *Drug Design* (Elsevier, 1976; https://linkinghub.elsevier.com/retrieve/pii/B9780120603077500109), pp. 165–207.

43. S. H. Newman-Stonebraker, S. R. Smith, J. E. Borowski, E. Peters, T. Gensch, H. C. Johnson, M. S. Sigman, A. G. Doyle, Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. *Science* **374**, 301–308 (2021).

44. R. Lavernhe, E. J. Alexy, H. Zhang, B. M. Stoltz, Palladium-Catalyzed Enantioselective Decarboxylative Allylic Alkylation of Acyclic α- *N* -Pyrrolyl/Indolyl Ketones. *Org. Lett.* **22**, 4272–4275 (2020).

45. E. J. Alexy, T. J. Fulton, H. Zhang, B. M. Stoltz, Palladium-catalyzed enantioselective decarboxylative allylic alkylation of fully substituted *N* -acyl indole-derived enol carbonates. *Chem. Sci.* **10**, 5996–6000 (2019).

Acknowledgments:

Funding:

Financial support to J.P.R. was provided by the University of British Columbia, the Natural Sciences and Engineering Research Council of Canada (NSERC) and the CFI John R. Evans Leaders Fund. Computational resources were provided from the Digital Research Alliance of Canada and the Advanced Research Computing (ARC) center at the University of British Columbia. I.O.B. acknowledges NSERC for a PGSD Research Fellowship. Y.K. thanks the Chinese Scholarship Council (CSC) for a graduate research fellowship.

Author contributions:

I.O.B. and J.P.R. conceptualized the project. I.O.B. implemented the clusterwise linear regression algorithm and code. Y.K. and C.Y. constructed the database for statistical modeling. I.O.B., Y.K., and C.Y. computed the structures. J. L. performed the experiments. I.O.B., Y.K., and J.P.R. analyzed the data. J.P.R directed the project. J.P.R wrote the manuscript with contributions from I.O.B. and Y.K.

Competing interests: Authors declare that they have no competing interests.