Electronic Supplementary Material (ESI) for Organic & Biomolecular Chemistry.

A Predictive and Mechanistic Statistical Modelling Workflow for Improving Decision Making in Organic Synthesis and Catalysis

*Isaiah O. Betinol and Jolene P. Reid**

Correspondence to: jreid@chem.ubc.ca

Department of Chemistry, University of British Columbia, 2036 Main Mall, Vancouver, British Columbia, V6T 1Z1, Canada

Table of Contents

Regression method comparison and model development

In organic chemistry applications, multivariate linear regression (MLR) models relate an output (e.g., $\Delta\Delta G^{\ddagger}$) to a molecular structure described by input features (e.g., Sterimol values, NBO charges, IR vibrations, etc.). A successful model fits the entire scope of empirical data (good and bad values). While model fitting over a data range is essential for 1. determining mechanistically relevant parameters and 2. drawing mechanistic conclusions, this approach can lessen the importance of the borderline cases. These are defined as near or on the experimentalist's threshold to perform a follow-up experiment and are often more scrutinized in reaction and catalyst design campaigns. In such endeavours, accurately predicting a $\Delta\Delta G^{\ddagger}$ value at the extremes is less important than determining if a reaction will be successful or not, and the associated probability with each outcome. Thus, the scenario begins to mimic a classification task that may be handled better by another generalized linear regression method – logistic regression.

To perform multivariate logistic regression (MLoR) analysis, outputs must be classified as successes or failures depending on a user defined threshold. Near equal sized categories and realistic values are important considerations when deciding this value (see below).[1] In contrast to MLR which attempts to predict a value given the parameters, MLoR only attempts to determine how the log(odds) of success change with the linear equation. Consequently, a probability of a given item being a success or failure can then be obtained by the equation:

$$P = \frac{1}{1 + \exp(-\beta)}$$

where β represents the general linear equation. An example of the resulting logistic function can be visualized in Figure S1B.

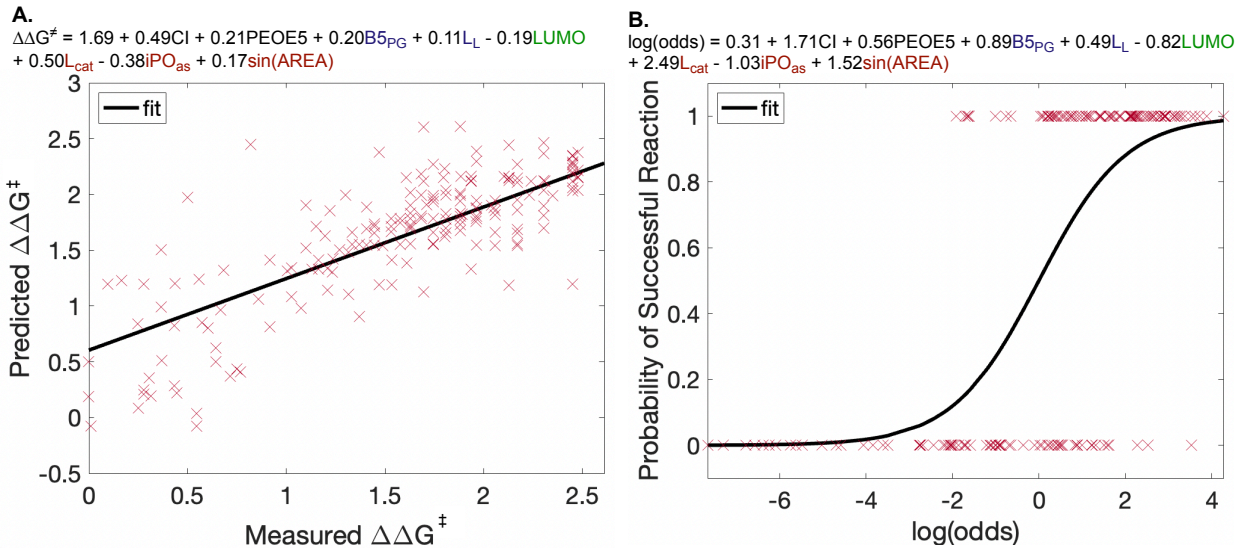

Figure S1. Comparison of MLR (left) and MLoR (right) models describing the chiral phosphoric acid catalyzed nucleophilic addition to *E* – imines.

In the context of this study, the main advantage of logistic regression is that it treats all successful and non-successful reactions the same. As such, the focus of the input-output relationship is positioned on the region that separates successful reactions from non-successful. This is in significant contrast to MLR which emphasises the high-low data range. Thus, this workflow is better suited to find reactivity cliffs or commonalities that are present in all successful reactions.

The workflow is designed to be a simple but useful extension of the well-established MLR modeling process with MLoR at the final readout step. Thus, our statistical model building procedure utilizes some scripts and functions previously developed.[2] Following the construction of a MLR model as described in detail previously,[2–4] reactions are coded as 0's and 1's (binary response value) to label them as successes or failures. Using this data along with the parameters deemed significant by MLR analysis, MLoR models can be created. These were generated in MATLAB_R2021a with built-in functions (mnrfit) for the logistic regression.[5]

Generated models were evaluated using Brier score,[6] a scoring rule that measures the accuracy of the assigned probabilities via the formula:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$$

where f_i is the assigned probability and o_i is the classified outcome (i.e., 0 or 1 for successful or non-successful results). This scoring rule is analogous to the commonly used mean square error.

MLR assumes that the output value (e.g. $\Delta\Delta G^{\ddagger}$) is a linear function of the parameters whereas MLoR assumes the response is Bernoulli distributed given the parameters. Consequently, the absolute parameter coefficients cannot be directly compared as they have fundamentally different meanings (how each parameter changes the log(odds) of success or the predicted $\Delta\Delta G^{\ddagger}$). Thus, to allow for easy comparison between relative parameter weights (i.e. which parameters are most/least important in each model), mean normalization was performed on the absolute values of the coefficients according to the formula:

$$x' = \frac{x - mean(x)}{range(x)}$$

Data curation

Relevant data were manually extracted from the following sources:

(1) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, 571 (7765), 343–348. https://doi.org/10.1038/s41586-019-1384-z.

(2) Reid, J. P.; Proctor, R. S. J.; Sigman, M. S.; Phipps, R. J. Predictive Multivariate Linear Regression Analysis Guides Successful Catalytic Enantioselective Minisci Reactions of Diazines. *J. Am. Chem. Soc.* **2019**, 141 (48), 19178–19185. https://doi.org/10.1021/jacs.9b11658.

(3) Li, J.; Grosslight, S.; Miller, S. J.; Sigman, M. S.; Toste, F. D. Site-Selective Acylation of Natural Products with BINOL-Derived Phosphoric Acids. *ACS Catal.* **2019**, 9 (11), 9794–9799. https://doi.org/10.1021/acscatal.9b03535.

(4) Connecting and Analyzing Enantioselective Bifunctional Hydrogen Bond Donor Catalysis Using Data Science Tools. *J. Am. Chem. Soc.* **2020**, 142 (38), 16382–16391. https://doi.org/10.1021/jacs.0c06905

(5) Ravasco, J. M. J. M.; Coelho, J. A. S. Predictive Multivariate Models for Bioorthogonal Inverse-Electron Demand Diels–Alder Reactions. *J. Am. Chem. Soc.* **2020**, 142 (9), 4235–4241. https://doi.org/10.1021/jacs.9b11948.

(6) Reid, J. P.; Hu, M.; Ito, S.; Huang, B.; Hong, C. M.; Xiang, H.; Sigman, M. S.; Toste, F. D. Strategies for Remote Enantiocontrol in Chiral Gold(iii) Complexes Applied to Catalytic Enantioselective γ,δ-Diels–Alder Reactions. *Chem. Sci.* **2020**, 11 (25), 6450–6456. https://doi.org/10.1039/D0SC00497A.

(7) Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S. A Data-Intensive Approach to Mechanistic Elucidation Applied to Chiral Anion Catalysis. *Science* **2015**, 347 (6223), 737–743. https://doi.org/10.1126/science.1261043.

(8) Orlandi, M.; Escudero-Casao, M.; Licini, G. Nucleophilicity Prediction via Multivariate Linear Regression Analysis. *J. Org. Chem.* **2021**, 86 (4), 3555–3564. https://doi.org/10.1021/acs.joc.0c02952.

All parameters and numerical prediction results can be found in the supplementary excel sheet.

Extra Case Studies Explained

Bifunctional Hydrogen Bond Donor Catalysis

An 8-parameter MLR model was reported describing bifunctional hydrogen bond donor catalysis (ref 4 from the above list). To first validate the MLoR model, 150 reactions were pseudorandomly split into 50:50 training set:validation set, the model trained and evaluated. The low validation set Brier score (BS = 0.04) indicates the probabilities assigned are close to the actual values. To use the MLoR model as a mechanistic probe, the model was then retrained on the full 150 reaction training set and the coefficients compared to the MLR model.

$$\Delta\Delta G^{\neq} = 1.40 - 0.06PEOE1 + 0.11Pol + 0.57NBO_X + 0.23B5_{avg} + 0.31i_{NH} + 0.11NBO_N - 0.35NBO_H - 0.25B1$$

$$\log(odds) = -0.43 + 0.03PEOE1 + 0.53Pol + 3.64NBO_X + 2.31B5_{avg} + 1.42i_{NH} - 0.63NBO_N - 2.69NBO_H - 1.08B1$$

Though direct comparisons between absolute values of coefficients cannot be made, relative weights can be compared. To this end, two main points can be drawn when comparing both mathematical equations. First, both equations emphasize the NBO_X term as the most important parameter, owing to the importance of the nucleophile heteroatom in determining enantioselectivity. Second, the MLoR model weights the nucleophile B5 term much more heavily than the respective MLR model. Interestingly, the authors do find that there is a minimum steric size of nucleophile required for high enantioselectivity, though this observation was obtained through graphical analysis in the original report. Thus, the MLoR model presents a quantitative way of finding such reactivity bins. In practice, this would give a hint as to what parameters to vary first in reaction design campaigns before fine tuning other parameters.

Inverse Electron Demand Diels-Alder Reactions

The same protocol described above was applied to a comprehensive MLR model describing the rate of inverse-electron demand Diels-Alder reactions (ref 5 from the above list). After partitioning the data into a 50:50 ts:vs split, the low validation set brier score (BS = 0.08) indicated that the model had predicting skill. Built on the full training set, the following equation was obtained and compared to the MLR model:

$$\Delta G^{\neq} = 16.98 - 1.15\varepsilon + 5.90\alpha_2 - 4.90d_{cc} + 0.66B1_1 - 2.47\Lambda + 2.18d_{cc} + 0.84d_{CN} - 0.98L_L$$

$$\log(odds) = -4.78 - 2.24\varepsilon + 22.2\alpha_2 - 19.5d_{cc} + 1.44B1_1 - 6.49\Lambda + 6.27d_{cc} + 0.31d_{CN} - 0.99L_L$$

For this case study, the same terms were emphasized in both MLR and MLoR models. Taken together, both models conclude that the most direct path to increase reaction rate is through increasing ring strain by either tightening the angle between dienophile trans-substituents or weakening the dienophile C-C bond. As both models reached the same conclusion, this case study ultimately demonstrates that the logistic model can be used to build further confidence in results obtained from MLR. Further, this case study shows that the MLoR application is not limited to enantioselectivity cases and can be generalized to other scenarios in which MLR is applicable.

Catalytic Enantioselective γ,δ-Diels–Alder Reactions

Following the same protocol on a MLR model describing catalytic enantioselective γ,δ-Diels–Alder reactions (ref 6 from the above list), a MLoR model was built and validated (validation set BS = 0.1) before generating the following equation with all reactions in the training set:

$$\Delta\Delta G^{\neq} = 0.79 + 0.13E_{int} + 0.42L_{C2} - 0.26B5_{C2}$$

$$\log(odds) = -4.94 + 3.97E_{int} + 8.45L_{C2} - 2.28B5_{C2}$$

From high-level DFT calculations in the original report, the authors find that stacking interactions between the substrate and catalyst are important in stabilizing the major transition state for the optimal catalyst system. Though both the MLR and logistic regression model place the most emphasis on the catalyst L term, the MLoR model more heavily weights the interaction energy term, E_{int}, when compared to the MLR model. Thus, this case study again supports that MLoR can be used as a mechanistic probe to find important interactions in well performing reactions.

Predicting Mayr Nucleophilicities from Structural Parameters

The change in parameter weighting can perhaps be visualized most intuitively when interrogating a system that is much more well defined than NCIs in catalysis. To this end, Orlandi and coworkers reported a MLR model that can predict a given molecules nucleophilicity from structural parameters (ref 8 from the above list). Both the MLR and logistic regression models identify the protonation energy as the most important parameter for predicting a given molecules nucleophilicity.

$$N = -2.84E_{PA} + 1.93S_{Nu} - 1.63S_{int} + 0.20S_{H} - 0.07\varepsilon + 0.38e_{HOMO} + 0.98q + 0.18B1 - 0.17\%V$$

$$\log(odds) = -3.32 - 11.5E_{PA} + 7.75S_{Nu} - 8.04S_{int} - 0.70S_{H} - 1.12\varepsilon + 5.17e_{HOMO} + 2.85q + 1.03B1 - 0.43\%V$$

The interpretation of such terms varies depending on the model. For instance, the MLR model relationship essentially states increasing the protonation energy is associated with higher nucleophilicities. In contrast, the logistic regression model would be stating that for a nucleophile to be a good nucleophile (N>17 in this case), it should have a high protonation energy (i.e., be anionic). This simple yet important distinction is key when trying to interpret complex MLR questions where the parameter meanings may be more obscure. Specifically, situations in reaction and catalyst development wherein modulating a parameter that fine tunes an outcome is less important than first ensuring that key design features are satisfied.

Finding Appropriate Thresholds for MLoR Analysis

A key aspect of utilizing MLoR for chemical systems with continuous outputs (e.g. enantioselectivity) is finding an appropriate threshold between 'successes' and 'failures'. As logistic regression is formally a tool for classification, one must choose a threshold that leads to a relatively balanced training set due to the noted limitations of classification algorithms in classifying events sparsely represented in the training set.[1] On the other hand, applications would be minimal if the set threshold is far below what would be of interest when describing a chemical system. To illustrate how threshold choice affects the output, we performed MLoR analysis of the CPA catalyzed Minisci reaction (ref 2 from the above list) at three different threshold values. The training/validation sets were held constant for all three analyses.

Setting the threshold value at 50% ee for successes leads to the most balanced training set (47% of training set classified as 'successes'), and the classification performance of the MLR model reflects this (training set BS = 0.08, validation set BS = 0.02) (Figure S2).

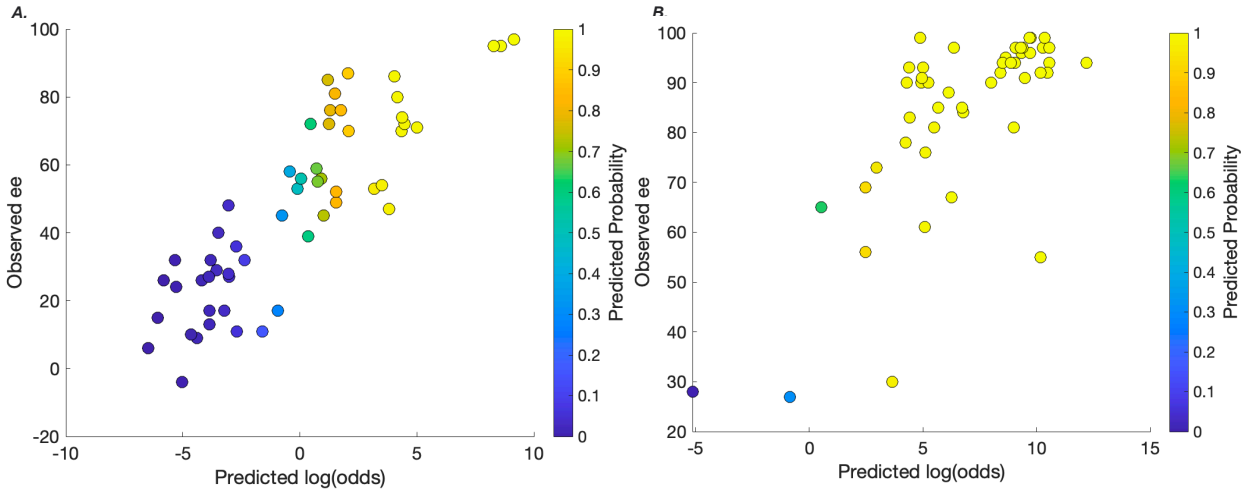

Figure S2. Training set (A) and prediction set (B) obtained for the CPA catalyzed Minisci reaction with a success threshold of 50% ee. The equation for this model is:

$$\log(\text{odds}) = -0.22 - 2.52\ iPO_{as} + 0.55\ NBO_{C2} + 1.58\ NBO_{RAE} + 1.39\ NBO_{Nhet} - 2.00\ L_{RAE} + 1.85\ B1_{Nhet}$$

Though this threshold leads to great predictions, the significance of a 50% ee reaction in this scenario is minimal. In other words, finding reactions above 50% ee is not an overly interesting endeavour when developing and expanding this reaction, and consequently the insights gained via such models would be of little value, despite how accurate they are.

In contrast, setting the threshold to 80% ee would be of interest to practitioners, however this threshold also leads to quite an imbalanced dataset (15% of training set classified as 'successes'). Evidently, despite a perfect training set classification, the validation set classification performance

of the MLR model suggests that the model is not adequately describing the system and is overfit to predict the few successful training set reactions (training set BS = 0, validation set BS = 0.61). Visual inspection of the validation set classification more clearly shows the lack of prediction power with this threshold value as the model incorrectly assigns absolute probabilities to various reactions (Figure S3).

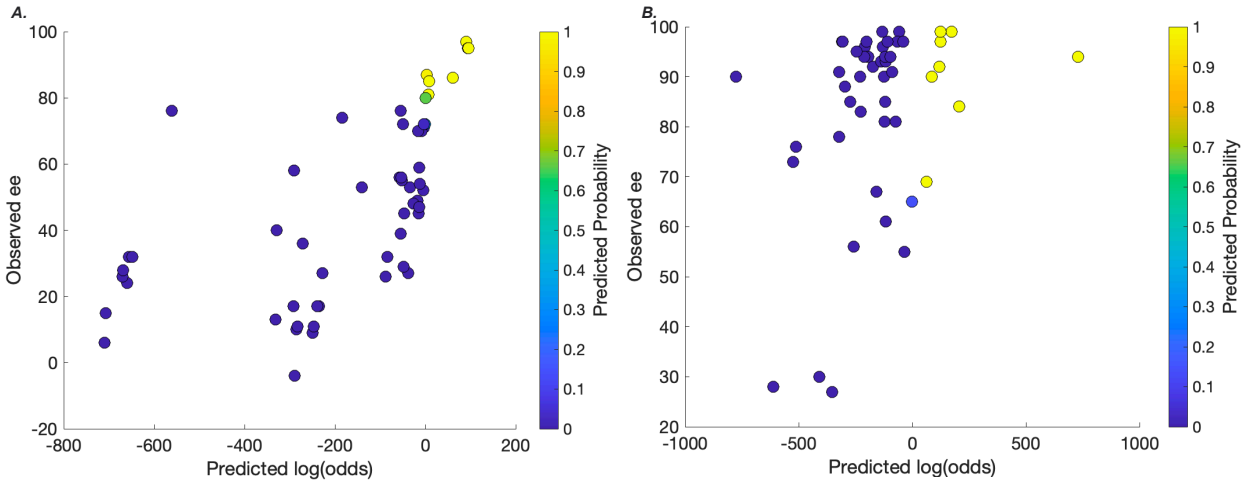

Figure S3. Training set (A) and prediction set (B) obtained for the CPA catalyzed Minisci reaction with a success threshold of 90% ee. The equation for this model is:

$$\log(\text{odds}) = -175 - 31\ iPO_{as} + 27.6\ NBO_{C2} - 7.27\ NBO_{RAE} + 72.8\ NBO_{Nhet} - 241\ L_{RAE} + 77.2\ B1_{Nhet}$$

In general, we found that having ~30% of the training set as 'successes' led to the best balance of classification and chemical significance. For this system, setting the threshold to 70% ee leads to both adequate classification (training set BS = 0.08, validation set BS = 0.07) as seen in Figure S4 and chemically relevant insights.

The thresholds set in this study were limited to clean integer values on significant borders where applicable (e.g. thresholds were set at 85% ee or 90% ee and not intermediate values such as 87% ee).

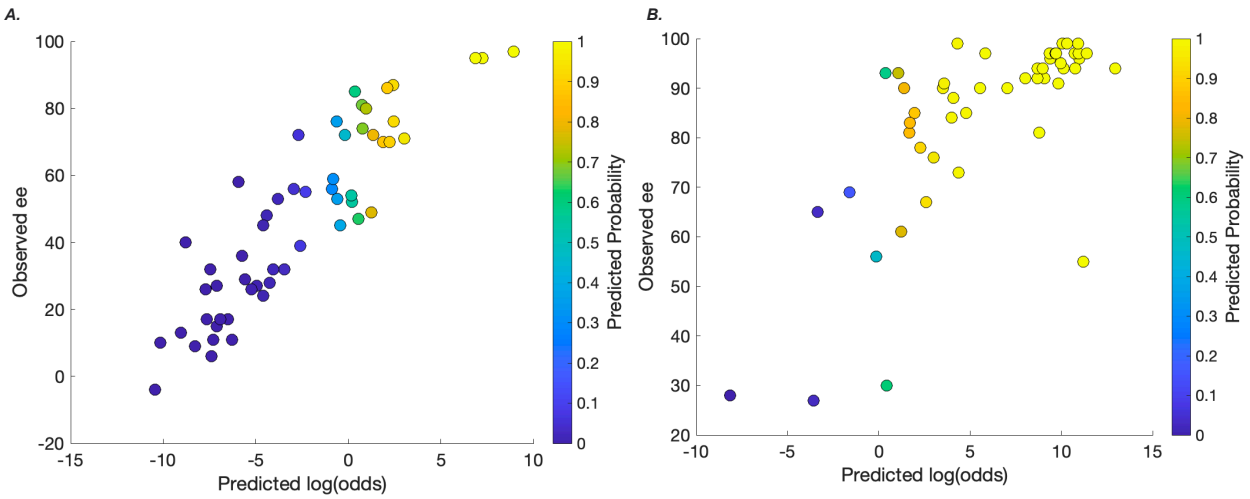

Figure S4. Training set (A) and prediction set (B) obtained for the CPA catalyzed Minisci reaction with a success threshold of 70% ee. The equation for this model is:

$$\log(\text{odds}) = -2.64 - 2.33\ iPO_{as} + 1.52\ NBO_{C2} + 1.96\ NBO_{RAE} + 1.90\ NBO_{Nhet} - 1.61\ L_{RAE} + 3.14\ B1_{Nhet}$$

Radar Plots

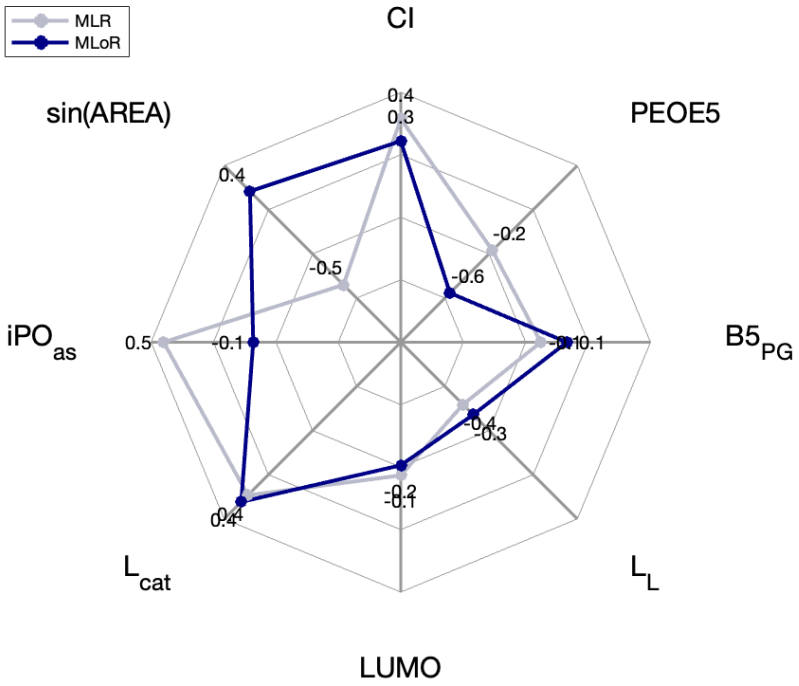

Figure S5. Radar plot comparing MLR and MLoR coefficients for the chiral phosphoric acid catalyzed nucleophilic addition to imines.

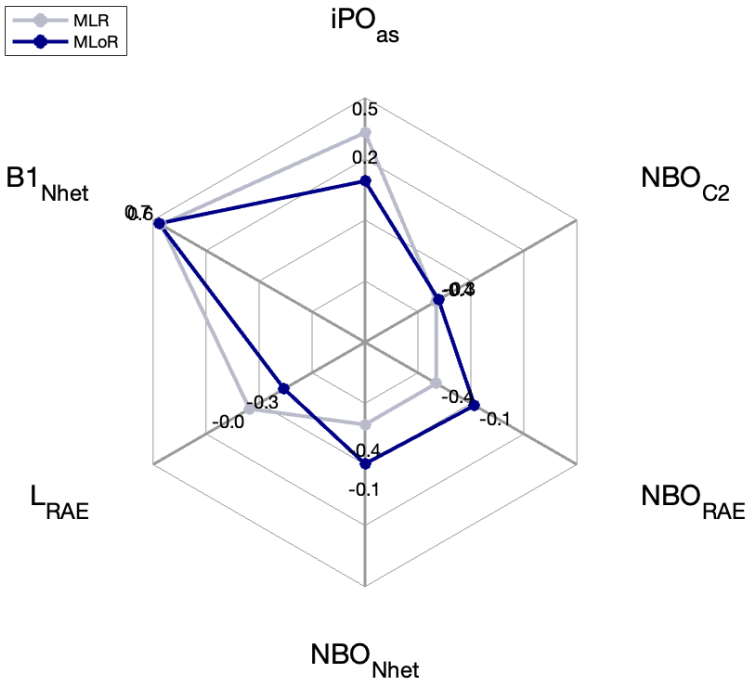

Figure S6. Radar plot comparing MLR and MLoR coefficients for the chiral phosphoric acid catalyzed Minisci reaction.

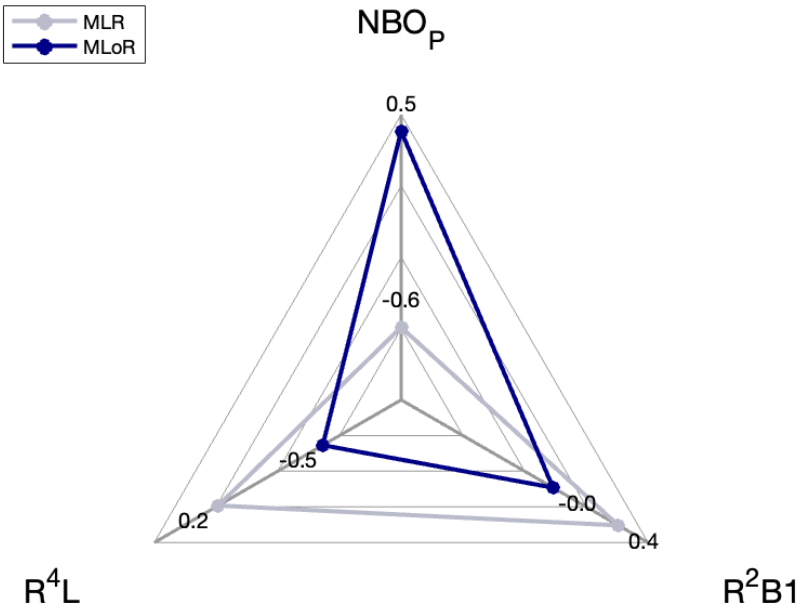

Figure S7. Radar plot comparing MLR and MLoR coefficients for the chiral phosphoric acid catalyzed site-selective acylation.

References

(1) King, G.; Zeng, L. Logistic Regression in Rare Events Data. *Polit. Anal.* **2001**, *9* (2), 137–163. https://doi.org/doi:10.1093/oxfordjournals.pan.a004868.

(2) Guo, J.-Y.; Minko, Y.; Santiago, C. B.; Sigman, M. S. Developing Comprehensive Computational Parameter Sets To Describe the Performance of Pyridine-Oxazoline and Related Ligands. *ACS Catal.* **2017**, *7* (6), 4144–4151. https://doi.org/10.1021/acscatal.7b00739.

(3) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571* (7765), 343–348. https://doi.org/10.1038/s41586-019-1384-z.

(4) Reid, J. P.; Proctor, R. S. J.; Sigman, M. S.; Phipps, R. J. Predictive Multivariate Linear Regression Analysis Guides Successful Catalytic Enantioselective Minisci Reactions of Diazines. *J. Am. Chem. Soc.* **2019**, *141* (48), 19178–19185. https://doi.org/10.1021/jacs.9b11658.

(5) *MATLAB and Statistics Toolbox Release R2021a*; The MathWorks, Inc.: Natick, Massachusetts, United States.

(6) Brier, G. W. Verification of Forecasts Expressed in Terms of Probability. *Mon. Weather. Rev.* **1950**, *78*, 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2.